

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Maiza Ferreira Martins Rocha

**Proposta de um experimento para avaliar os
mecanismos usados por spammers para obter
contas de e-mail**

Uberlândia, Brasil

2017

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Maiza Ferreira Martins Rocha

**Proposta de um experimento para avaliar os mecanismos
usados por spammers para obter contas de e-mail**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, Minas Gerais, como
requisito exigido parcial à obtenção do grau
de Bacharel em Sistemas de Informação.

Orientador: Rodrigo Sanches Miani

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2017

Maiza Ferreira Martins Rocha

Proposta de um experimento para avaliar os mecanismos usados por spammers para obter contas de e-mail

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 22 de dezembro de 2017:

Rodrigo Sanches Miani
Orientador

Professor

Professor

Uberlândia, Brasil
2017

Resumo

O recebimento de mensagens eletrônicas não solicitadas (*spam*) é um problema que cresceu com a expansão da Internet e a popularização do correio eletrônico, gerando um grande custo para empresas, organizações, provedores e usuários. Antes de disseminar tais mensagens indesejadas, os *spammers* utilizam diversos mecanismos para obter endereços eletrônicos, que vão desde a compra de banco de dados até a criação de suas próprias listas, geradas a partir de técnicas como: *harvesting* (ou colheita), geração de contas usando técnicas de força bruta e dicionário e uso de programas maliciosos. Nesta monografia, analisa-se tais mecanismos a partir da proposta de um experimento utilizando contas de e-mail reais. O objetivo do experimento é expor endereços de e-mail em alguns meios como por exemplo, sites públicos da Web, redes sociais ou fornecer o endereço de e-mail em cadastros realizados em lojas e verificar a influência desses grupos propostos no recebimento de spam. Além disso, procura-se verificar se o provedor de serviços o qual a conta foi cadastrada tem relação com o recebimento de mensagens não solicitadas. Resultados preliminares indicam que os meios nos quais as contas foram divulgadas têm influência no recebimento de spam, assim como a identidade do provedor de serviços utilizado.

Palavras-chave: Segurança da informação, E-mail, Spam, Experimento.

Lista de ilustrações

Figura 1 – Volume de Spam global em percentagem do tráfego de e-mail total. Retirado de (STATISTA, 2017).	8
Figura 2 – Protocolo SMTP	14
Figura 3 – Endereço de e-mail como figura.	16
Figura 4 – Código HTML utilizado nas páginas para exposição das contas de e-mail.	21
Figura 5 – Forma como as contas foram expostas no sítio do <i>stackoverflow</i>	22
Figura 6 – Exemplo de cadastro realizado na Empresa 1.	23
Figura 7 – Grupos de exposição criados.	24
Figura 8 – Exposição das contas no grupo ANUNCIE BRASIL do WhatsApp.	25
Figura 9 – Fluxograma da metodologia proposta.	26
Figura 10 – Caixa de entrada da conta c6.	30
Figura 11 – Conteúdo de um dos e-mails recebidos pela conta c6.	31
Figura 12 – E-mail recebido na conta c1.	31
Figura 13 – E-mail recebido na conta c2	32
Figura 14 – E-mail recebido na conta c3	33
Figura 15 – E-mails recebidos na conta c4.	34
Figura 16 – Conteúdo de um dos e-mails recebidos na conta c4.	34
Figura 17 – Distribuição de spams por grupo de exposição.	35

Lista de tabelas

Tabela 1 – Exemplo de nomes e palavras no dicionário e e-mails a serem testados.	17
Tabela 2 – Distribuição das contas de acordo com o provedor e o grupo de exposição.	26
Tabela 3 – Quantidade de spam recebida em 8 semanas	29
Tabela 4 – Quantidade de spams por provedor.	33
Tabela 5 – Quantidade de spams por grupo de exposição.	35

Lista de abreviaturas e siglas

CERT.br	Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil
TIC	Tecnologia de Informação e Comunicação
SMTP	<i>Simple Mail Transfer Protocol</i>
TCP	<i>Transmission Control Protocol</i>
IP	<i>Internet Protocol</i>
CGI.br	Comitê Gestor da Internet no Brasil

Sumário

1	INTRODUÇÃO	8
2	FUNDAMENTAÇÃO TEÓRICA	11
2.1	Segurança da Informação	11
2.2	O E-mail, a Internet e o Spam	12
2.3	Trabalhos Correlatos	17
3	DESENVOLVIMENTO	19
3.1	Metodologia	19
3.2	Detalhamento da proposta	20
4	ANÁLISE DOS RESULTADOS PRELIMINARES	28
5	CONCLUSÃO	36
	REFERÊNCIAS	37

1 Introdução

O correio eletrônico (e-mail) é um dos meios de comunicação mais utilizados em todo o mundo (OLIVO; SANTIN; OLIVEIRA, 2015). Seu sucesso advém de várias vantagens, entre elas o fato de ser um mecanismo eficaz, econômico e ecológico. Por outro lado, o e-mail também possui suas desvantagens, dentre elas, as mensagens eletrônicas não solicitadas, mais conhecidas como spams.

A popularização do e-mail e seu volume sempre crescente resultou no aumento de spams. Entre os anos de 2002 e 2010 a média de spam enviados por dia passou de 2,4 bilhões para 300 bilhões (ALMEIDA, 2010). Contudo, esse cenário vem sendo modificado. Segundo o portal de estatística Statista houve uma queda no volume de spam em nível global do ano de 2014 à 2017. A Figura 1 mostra esse declínio. Apesar de tal fato, spams classificados como maliciosos continuam representando uma grave ameaça aos negócios de Tecnologia de Informação e Comunicação (TIC) (OLIVEIRA, 2016).

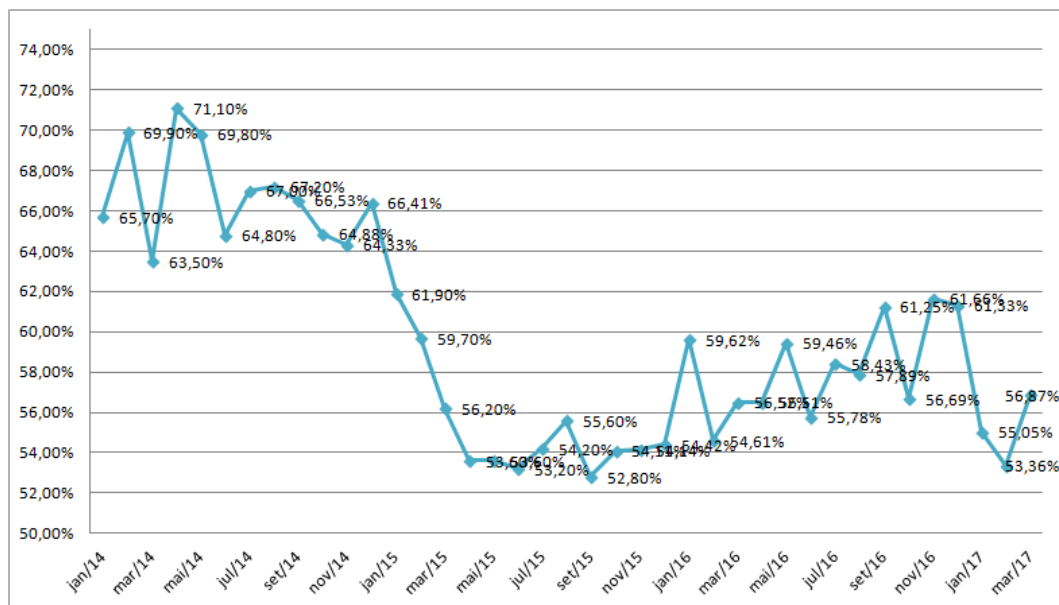


Figura 1 – Volume de Spam global em porcentagem do tráfego de e-mail total. Retirado de (STATISTA, 2017).

Chamadas telefônicas, correspondências e folhetos promocionais são exemplos de mensagens não eletrônicas que podem ser comparadas ao spam. Mas diferente desses meios que possuem certo custo de recursos para quem envia, o spam eletrônico não custa praticamente nada para quem envia, e ainda consegue alcançar um número massivo de pessoas sem maiores dificuldades. Por isso o e-mail se tornou o principal alvo de ataque entre os *spammers*: quem enviam spam.

Existem várias estratégias utilizadas pelos *spammers* para obter endereços de e-mails. De acordo com o Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (CERT.br), algumas dessas estratégias são: compra de banco de dados com listas de e-mails de empresas/organizações; programas maliciosos que se infiltram no computador do usuário atrás de informações, como o endereço de e-mail; colheita de e-mails feita por programas automatizados que varrem páginas web em busca de tais informações; ataques a dicionários, onde o *spammer* através de nomes, palavras de dicionário, entre outros, tenta descobrir endereços de e-mail válidos através de combinações de palavra+provedor.

Cada uma dessas técnicas utilizadas para coletar endereços de e-mail serão explicadas na seção 2.2, no próximo capítulo deste trabalho.

No Brasil não existem leis para fiscalizar e punir os autores de spam (ALMEIDA, 2010). Contudo, existem vários recursos empregados para evitar o problema do spam. Um mecanismo comumente usado são os filtros de spam. Os filtros classificam e-mails em spam usando conhecimento sobre o próprio remetente e também métodos para identificação de conteúdo malicioso no próprio e-mail. Tais filtros devem ser constantemente modificados para que não percam a eficiência frente as novas metodologias que os *spammers* utilizam para burlar as ferramentas de proteção.

Conhecer melhor a forma como atuam os *spammers* é um progresso para ajudar os formuladores de políticas, provedores de serviços de internet, fornecedores de software e estudiosos a acabar com esse problema do correio eletrônico. Alguns estudos já foram feitos, mas a pesquisa científica sobre o setor de spam não é muito investigada (HANN et al., 2006).

Hann et al. (2006) realizaram um estudo para confirmar se o envio de spams seguem algum tipo de orientação ou se são distribuídos aleatoriamente. Se o spam não for transmitido aleatoriamente, desejava-se saber quais os fatores que determinam a taxa de spam. Para esse experimento eles criaram várias contas de e-mail em alguns provedores. Concluíram que o spam não é um evento aleatório, mas especificamente estão mais presentes em certos provedores de e-mail e pessoas com interesses declarados em determinados produtos ou serviços. Concluíram também que alguns fatores como interesses declarados, idade e nacionalidade foram determinantes na taxa de recebimento de e-mails.

Outro estudo realizado na área foi de Oliveira (2016) onde ela replicou o estudo feito por (HANN et al., 2006). Nesse estudo a autora verificou quais fatores influenciam no recebimento de spams. Para isso foram criados alguns grupos de exposição com base em algumas características de navegação e comportamento de um usuário da Web como: fazer compras online, frequentar redes sociais, participar de fóruns de discussão e utilizar ferramentas Web para compartilhamento de arquivos. Os resultados obtidos mostram que as taxas de spam seriam maiores nos grupos de compras online e rede social. Sendo

liderada por redes sociais, especificamente pelo Facebook.

Com base nos trabalhos anteriores, o objetivo deste estudo também é melhorar a compreensão sobre o spam. No entanto, diferente dos estudos conduzidos por [Hann et al. \(2006\)](#) e [Oliveira \(2016\)](#), não serão investigadas características dos usuários que recebem spam e sim as estratégias utilizadas por *spammers* para obter endereços de e-mail. Ou seja, deseja-se melhorar a compreensão sobre as origens do spam, buscando entender quais são as formas mais adotadas pelos *spammers* para conseguir endereços de e-mail.

Para isso, foi proposto um experimento usando novas contas de e-mail criadas em três provedores distintos: *Hotmail*, *Gmail* e *Yahoo*. Foram criados também novos grupos de exposição, como por exemplo, incluir o endereço de e-mail em sites ou fornecer o endereço de e-mail em cadastros realizados em lojas. Redes sociais também foram consideradas como uma estratégia usada por *spammers* pra a obtenção de contas de e-mail.

Ao final deste trabalho, foi realizada uma comparação entre as taxas de spams dos grupos de exposição e também foi verificado qual desses grupos de exposição é o maior alvo de ataque de *spammers*. Com esses resultados espera-se contribuir para melhorar a compreensão sobre o comportamento dos *spammers*.

O trabalho apresentado estrutura-se em quatro seções, além desta introdução. No Capítulo 2 é apresentado o referencial teórico, fundamentando o tema e assuntos relacionados, fortalecendo assim a pesquisa; no Capítulo 3, encontram-se os aspectos metodológicos; o Capítulo 4 mostra a discussão dos resultados, finalizando, na última seção, com as considerações finais do trabalho.

2 Fundamentação teórica

Este capítulo apresenta um levantamento bibliográfico sobre a segurança da informação, o e-mail, a internet e o spam, conceitos que fundamentam o trabalho.

2.1 Segurança da Informação

Segurança da informação remete a proteção da informação e que pessoas não autorizadas estejam fora do alcance de acessá-las (DIAS, 2000). Pode-se considerar a informação como um ativo cada vez mais importante e valorizado para uma organização. Esse importante bem, por outro lado, é também uma fonte de problemas a partir de diversos tipos de falhas de segurança.

Segundo Spanceski (2004), os usuários de sistemas de informação esperam que suas informações cumpram os princípios da segurança digital, ou seja, sejam confiáveis, corretas, mantidas fora do alcance de pessoas não autorizadas, e que estejam disponíveis no momento e local que determinar. Essas expectativas do usuário podem ser traduzidas nos seguintes princípios ou pilares de segurança: autenticidade, confidencialidade, integridade e disponibilidade, que são definidos por Spanceski (2004) da seguinte forma:

- Autenticidade: está relacionada com a identificação de um usuário ou computador. A autenticidade é provida por meio da autenticação e garante que a entidade que está tentando se comunicar é quem afirma ser.
- Confidencialidade: é a garantia de que a informação será acessível apenas por pessoas autorizadas.
- Integridade: garante que as informações não sofreram modificação não autorizada, e sejam recebidas conforme foram enviadas.
- Disponibilidade: princípio que garante que a informação esteja sempre disponível e utilizável por aqueles usuários autorizados pelo proprietário da informação.

Stallings (2008) propõe um quinto pilar da segurança da informação, chamado irretratabilidade (ou não-repúdio):

- Irretratabilidade (não-repúdio): impede que o emissor ou o receptor negue a autoria da mensagem transmitida. Ou seja, quando uma mensagem é enviada, o receptor pode provar que o emissor realmente enviou a mensagem. Da mesma forma, quando

uma mensagem é recebida, o emissor pode provar que o receptor alegado realmente recebeu a mensagem.

Fragilidades de segurança podem comprometer os princípios da segurança. [Teotônio \(2013\)](#) cita algumas das ameaças que vão contra os princípios de segurança, são elas:

- Integridade: ameaças de ambiente (fogo, enchente, tempestade), erros humanos, fraudes e erro de processamento;
- Divulgação da informação: divulgação de informações premeditada e divulgação de informações acidentalmente. Determinados tipos de spam podem estar diretamente relacionados a tal problema, quando, por exemplo, banco de dados contendo dados pessoais são alvos de vazamento;
- Indisponibilidade: falhas em sistemas ou nos diversos ambientes computacionais podem tornar as informações indisponíveis;
- Alterações não autorizadas: alteração premeditada e alteração acidental.

2.2 O E-mail, a Internet e o Spam

O e-mail ou correio eletrônico é um serviço disponível na Internet, que através do protocolo SMTP (*Simple Mail Transfer Protocol* ou Protocolo Simples de Transporte de Mensagens), realiza a troca de mensagens entre seus usuários. Ao contrário do que se possa imaginar, o e-mail surgiu antes da Internet, e foi uma ferramenta crucial para a criação da rede internacional de computadores, a Internet ([ALMEIDA, 2010](#)).

A Internet surgiu a princípio voltada para atender interesses específicos, como ensino e pesquisa ([TEIXEIRA, 2001](#)). Segundo dados da *World Internet Usage Statistics*, no final do ano de 1996, ano em que a Internet já era de uso comum ([TEIXEIRA, 2001](#)), a Internet possuía aproximadamente 36 mil usuários. Ao fim de 2016, esse número superou a marca de 3 bilhões e 600 milhões usuários, representando um crescimento superior a 999,99% e correspondendo a aproximadamente 49.5% de toda a população mundial.

Seu indiscutível crescimento impactou e revolucionou a vida das pessoas proporcionando diversos recursos de informações e serviços, como e-mail, bate-papo em tempo real, transferência e compartilhamento de arquivos, jogos *on-line*, inter-conectividade de documentos hipertexto, além de outros recursos ([ALMEIDA, 2010](#)). Com certeza esse panorama tem inúmeras vantagens. Mas nem tudo é vantagem na Internet. Muitos fazem o mal uso desses recursos e acabam por utilizar a Internet para comportamentos desagradáveis, como: se aproveitar da exposição das pessoas para prática de golpes, roubos,

fraudes, entre outros crimes. Muitas vezes as pessoas se escondem atrás de anonimato para cometer tais crimes. Há muita ilegalidade, pequenas fraudes são cometidas a todo momento e ações como essas mostram um lado ruim da internet. Prejudicando aqueles que fazem o bom uso dessa ferramenta.

O envio indiscriminado de mensagens eletrônicas não solicitadas, conhecido como spam, é um dos comportamentos que infringe os bons costumes na Internet. Segundo Szendrodi (2005) a prática do spam teve origens na “flexibilidade” com que o protocolo SMTP foi criado. Desenvolvido na época militar em meados de 69, o protocolo tinha de ser simples, como o próprio nome sugere, pois naquela época a comunicação tinha de ser rápida, ou seja, um protocolo complexo dificultaria a troca de mensagens entre os soldados. O protocolo SMTP como já mencionado, possibilita a troca de mensagens entre seus usuários, sendo responsável por descrever o formato da mensagem e a forma de comunicação dos servidores para se realizar a troca de e-mails entre eles. Algumas das características do protocolo SMTP, são:

- Possui basicamente três entidades: Agente Usuário, Servidor SMTP do emissor e Servidor SMTP do receptor.
- É orientado a conexão, sendo transmitido sobre TCP (*Transmission Control Protocol* ou Protocolo de controle de transmissão), que é completado pelo IP (*Internet Protocol* ou Protocolo da Internet), garantindo assim que os dados sejam entregues aos destinatários em ordem e completos.
- A comunicação é resumidamente feita da seguinte forma: 1) o Cliente (Agente usuário) envia a mensagem, 2) o servidor do emissor (Emissor-SMTP) recebe a mensagem 3) o Emissor-SMTP envia a mensagem para o servidor do receptor (Receptor-SMTP). Conforme mostra a Figura 2.
- Os comandos SMTP necessários para enviar uma mensagem são basicamente: HELO (o cliente identifica-se), MAIL from (endereço origem), RCPT to (endereço destino), DATA (mensagem) e QUIT (fim). Aqui nenhuma verificação de autenticidade da mensagem é feita. Ou seja, não há verificação de validade dos campos FROM ou TO. Permitindo assim que o endereço de quem enviou a mensagem seja falsificado, por exemplo.

SMTP.PNG

Com o surgimento e a popularização da Internet e, conseqüentemente, do uso do e-mail, esse veículo tornou-se o favorito entre os *spammers* (OLIVO; SANTIN; OLIVEIRA, 2015), nome designado aos autores de *spamming* ou envio de spam. Já que, uma das características dos *spammers* é, justamente, espalhar mensagens indesejadas para um número cada vez maior de destinatários.

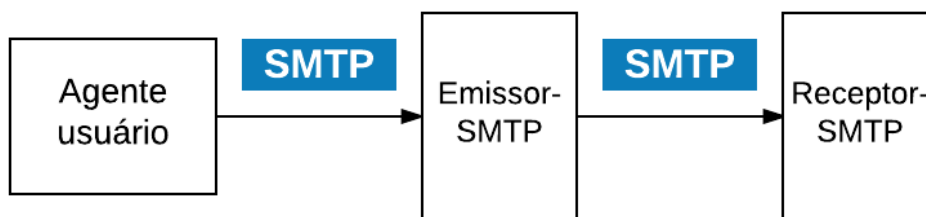


Figura 2 – Protocolo SMTP

De acordo com (TEIXEIRA, 2001) os tipos mais comuns de spam, são:

- Boatos: são textos que contam histórias falsas que chamam a atenção e instigam o leitor a continuar sua divulgação. Há ainda casos de boatos que difamam empresas ou produtos.
- Correntes: semelhantes aos boatos, correntes são textos que estimulam o leitor a enviar várias cópias a outras pessoas, gerando um processo contínuo de propagação. Normalmente esses textos prometem sorte e riqueza aos que derem continuidade a propagação, e azar aos que quebrarem a corrente.
- Propagandas: empresas com o intuito de divulgar seus produtos, mandam diversos e-mails de propagandas de seus produtos ou serviços aos usuários sem a autorização dos mesmos.
- Outros: alguns spams são enviados com o intuito de fazer ameaças, brincadeiras de mau gosto ou apenas por diversão. Essas práticas também são consideradas spam.

Os danos causados aos usuários vítimas de spam podem ir muito além de um simples aborrecimento. O site Antispam.br (2017), que é mantido pelo Comitê Gestor da Internet no Brasil (CGI.br), constitui uma fonte de referência sobre o spam, cita exemplos de como o spam pode afetar os usuários ameaçando a produtividade e segurança:

- Não recebimento de e-mails: Boa parte dos provedores de Internet limita o tamanho da caixa postal do usuário no seu servidor. Caso o número de spams recebidos seja grande, ele corre o risco de ter sua caixa postal lotada com mensagens não solicitadas. Se isto ocorrer, passará a não receber e-mails e, até que possa liberar espaço em sua caixa postal, todas as mensagens recebidas serão devolvidas ao remetente. Outro problema é quando o usuário deixa de receber e-mails nos casos em que regras anti spam ineficientes são utilizadas, por exemplo, classificando como spam mensagens legítimas.

- Gasto desnecessário de tempo: Para cada spam recebido, o usuário necessita gastar um determinado tempo para ler, identificar o e-mail como spam e removê-lo da caixa postal.
- Perda de produtividade: Para quem usa o e-mail como ferramenta de trabalho, o recebimento de spams aumenta o tempo dedicado à tarefa de leitura de e-mails, além de existir a chance de mensagens importantes não serem lidas, serem apagadas por engano ou lidas com atraso.
- Conteúdo impróprio ou ofensivo: Como a maior parte dos spams é enviada para conjuntos aleatórios de endereços de e-mail, é bem provável que o usuário receba mensagens com conteúdo que julgue impróprio ou ofensivo.
- Prejuízos financeiros causados por fraude: O spam tem sido amplamente utilizado como veículo para disseminar esquemas fraudulentos, que tentam induzir o usuário a acessar páginas clonadas de instituições financeiras ou a instalar programas maliciosos, projetados para furtar dados pessoais e financeiros. Esse tipo de spam é conhecido como *phishing/scam*. O usuário pode sofrer grandes prejuízos financeiros, caso forneça as informações ou execute as instruções solicitadas nesse tipo de mensagem fraudulenta.

É possível dividir o funcionamento do spam em duas etapas: a primeira é a obtenção de dados, que será usada para encontrar destinatários para o spam. A segunda é o envio dos e-mails, que inclui os mecanismos para enviar uma grande quantidade de e-mail, dentre esses mecanismos o mais utilizado segundo (MOREIRA, 2014) é através de computadores “zumbis”. Uma rede “zumbi” é composta por milhares de computadores que são invadidos por spammers e infectados por programas que permitem que estes computadores sejam os emissores de spams.

Como mencionado no Capítulo 1, os *spammers* utilizam diversos mecanismos para obtenção de endereços de e-mail. Segundo (Marluce Peron, 2015) as principais estratégias utilizadas, são:

- Programas maliciosos: após o computador alvo ser atingido por um ataque, o programa malicioso se infiltra no mesmo atrás de informações, como o endereço de e-mail, por exemplo, na lista de endereços do usuário. Os dados são recolhidos e repassados para *spammers*.

Por isso é importante atenção ao receber e-mails sobre brindes, promoções ou descontos. Verificar a procedência dessas informações clicar em links suspeitos. Esse tipo de ataque é comumente conhecido como *phishing*. Ter um filtro anti-spam instalado, ou ainda, usar os recursos anti-spam oferecidos pelo seu provedor de acesso. São algumas práticas que podem ser tomadas para prevenção contra o spam.

- *Harvesting* (colheita) ou máquina de busca: esse mecanismo consiste em coletar dados através de um rastreador da rede (em inglês, *web crawler*). O *crawler* é um software desenvolvido para realizar uma varredura na Internet de maneira sistemática através de informação vista como relevante a sua função. Esse software é então programado para realizar rotinas de busca por palavras chaves que possam direcionar a possíveis endereços de e-mail (como *tags malito* em HTML ou simplesmente @). Tais rotinas são executadas em diversos tipos de páginas como páginas pessoais, listas de discussões, listas de cadastro (como em sítios de universidade), redes sociais, entre outros.

Uma forma de mitigar este tipo de ataque é expor os endereços de e-mail em figuras, como mostra a Figura 3. Outra forma é, por exemplo, a substituição da @ por *at*. Ao invés de fulano@dominio.org se coloca fulanoatdominio.org. Contudo, *crawlers* mais sofisticados podem ser programados para lidar com tais situações.

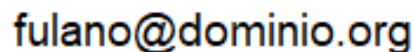


Figura 3 – Endereço de e-mail como figura.

- Ataques a dicionários: o *spammer* forma endereços de e-mail a partir de listas de nomes de pessoas, de palavras presentes em dicionários e/ou da combinação de caracteres alfanuméricos, gerando possíveis endereços. A Tabela 1 mostra um exemplo de como isso pode ser feito. Como os e-mails são formados aleatoriamente, os spammers usam técnicas a fim de identificar os endereços válidos. (MOREIRA, 2014) cita algumas dessas técnicas:
 - Enviando uma mensagem: é enviado um e-mail simples para o endereço de e-mail a ser validado. Se o servidor não retornar erro, o spam tem então a confirmação da validade daquele endereço.
 - Utilização de imagem: inserir no e-mail spam imagens do tipo *Web bug*, projetada para monitorar o acesso a uma página Web ou e-mail. Quando o usuário abre o spam, o *Web bug* é acessado e o spammer recebe a confirmação que aquele endereço de e-mail é válido.
 - Links externos: usuários pouco experientes são vulneráveis e, facilmente, clicam em links com mensagens do tipo “veja nossas fotos”, “se não quiser receber mais mensagem, clique aqui” ou de uma falsa notícia. Ao clicar o usuário confirma a validade do e-mail.

Para evitar este tipo de técnica, o usuário pode configurar seu servidor de e-mail ou seu programa de e-mail para que só receba mensagens em formato txt ou formatado

sem figuras. Isso faria com que os links automáticos ou aparecessem apenas como um endereço web ou que as figuras não fossem exibidas, evitando o *Web bug* (MOREIRA, 2014). É interessante também evitar utilizar endereços de e-mails simples, como aqueles formados apenas pelo primeiro nome.

Tabela 1 – Exemplo de nomes e palavras no dicionário e e-mails a serem testados.

Dado	E-mail a ser testado
Pedro	pedro@dominio.ogr
João	joao@dominio.org
Maria	maria@dominio.org
Suporte	suporte@dominio.ogr
compras	compras@dominio.ogr
sac	sac@dominio.ogr

- Banco de dados: outra forma de obtenção é através de um banco de dados com listas de e-mails. Bases de dados de empresas, organizações e órgãos do governo que vendem seu banco de dados com e-mails para spammers. Esse tipo de prática pode ser encontrada na própria Internet ou em centros que possuem atividade de pirataria. Além disso, também é que comum que hackers ou *spammers* vendam ou troquem informações com outros *spammers* (MOREIRA, 2014).

Para evitar esse mecanismo, seria interessante ter, sempre que possível, e-mails separados para assuntos pessoais, profissionais, para as compras e cadastros *on-line*. Certos usuários mantêm um e-mail somente para assinatura de listas de discussão e cadastros feitos para empresas/organizações (ANTISPAM.BR, 2017).

2.3 Trabalhos Correlatos

Esta seção tem como objetivo descrever trabalhos relacionados ao tema principal desse trabalho: análise de spam. Os trabalhos propostos por Hann et al. (2006), Oliveira (2016), Nagamalai, Dhinakaran e Lee (2008) e Prince et al. (2003) serão discutidos a seguir.

Hann et al. (2006) analisaram características dos usuários para confirmar se o envio de spam segue algum tipo de orientação ou se são distribuídos aleatoriamente. Concluíram que o spam não são aleatórios, mas orientados para determinados segmentos, como por exemplo, usuários que são mais propensos a fazer compras online, aqueles que declaram interesse em produtos específicos ou serviços, adultos, e os residentes dos Estados Unidos (EUA). Uma descoberta significativa foi em relação à identidade do prestador de serviços de e-mail. Contas da *Hotmail* receberam um maior número de spam.

Já na proposta de [Oliveira \(2016\)](#), ela analisou o comportamento do usuário Web com diferentes características, como frequentar redes sociais, fazer compras online e utilizar ferramentas Web para compartilhamento de arquivos. Concluiu que as taxas de spam seriam maiores nos grupos de compras online e rede social. A descoberta mais interessante foi de que o fator que exerce a maior influência no recebimento de mensagens eletrônicas não solicitadas é a exposição do endereço de e-mail ao grupo de rede social (Facebook) ([OLIVEIRA, 2016](#)).

No estudo feito por [Nagamalai, Dhinakaran e Lee \(2008\)](#) o objetivo foi analisar as características de contas de e-mails vulneráveis a spam. Durante 14 meses de janeiro de 2006 a fevereiro de 2007 com uma armadilha feita no servidor de correio, foram recolhidos diversos e-mails. Os autores observaram que o spam pode ser classificado em duas grandes categorias, spam com anexo e sem anexo. Também foram analisados os tipos de contas de e-mail que atraem mais spams. Contas de e-mail de usuários "pesados" recebem mais spam do que outros. De acordo com os autores, um usuário pesado foi definido como aqueles que possui seu próprio site, blog e aqueles envolvidos ativamente em fóruns e salas de bate-papo. Observaram também que o volume de tráfego de spam não está relacionado ao tráfego de e-mail legítimo.

[Prince et al. \(2003\)](#) propuseram entender como os spammers roubam os endereços de e-mail. Para eles compreender o comportamento dos colhedores é fundamental para controlar o problema de spam. Além disso também pode ajudar a identificar gangues de spam e dar aos agentes da lei uma nova causa de ação para processos. Para esse fim, o Projeto Honey Pot foi criado para entender a principal maneira pela qual os spammers obtêm novos endereços de e-mail. O Projeto Honey Pot é uma rede de honeypot baseada na web, que é um recurso computacional dedicado a ser sondado, atacado ou comprometido. Uma espécie de armadilha para invasores. O Projeto foi anunciado em 2004 e aberto aos voluntários públicos 14 de outubro de 2004. [Prince et al. \(2003\)](#) analisa dados de um *honeypot*, a partir de colaboradores, enquanto o presente estudo projeta um experimento com contas reais.

3 Desenvolvimento

Esse capítulo tem o seguinte objetivo: apresentar o método proposto neste trabalho, que, baseado nos estudos feitos por [Hann et al. \(2006\)](#) e [Oliveira \(2016\)](#), visa melhorar a compreensão sobre as estratégias adotadas pelos *spammers*, utilizando uma nova base de dados de contas de e-mail e diferentes formas de exposição de tais contas.

3.1 Metodologia

Os seguintes passos foram executados para o desenvolvimento do presente trabalho:

1. Definir grupos de exposição - definiu-se estratégias para exposição das contas de e-mail, levando em consideração as estratégias utilizadas por *spammers* para obtenção de endereços de e-mail. Os grupos, chamados aqui de grupos de exposição, foram definidos da seguinte forma:
 - *Grupo 1* - endereços de e-mail coletados à partir de páginas públicas da Internet usando *crawlers*;
 - *Grupo 2* - endereços de e-mail coletados à partir de registros presentes em bases de dados de organizações;
 - *Grupo 3* - endereços de e-mail coletados à partir de redes sociais.
2. Criar contas de e-mail - foram criadas contas fictícias em três provedores de e-mail gratuitos, sendo eles: *Yahoo*, *Hotmail* e *Gmail*. Cada conta foi criada de forma aleatória em relação as características de cada usuário. Após as contas serem criadas, essas ficaram em quarentena por um período de um mês e meio;
3. Exposição das contas - com as contas de e-mail criadas e passado o período de quarentena, o passo seguinte foi expô-las ao seu respectivo grupo de exposição.
4. Coleta dos spams - as contas de e-mail foram monitoradas por um período pré-estabelecido de dois meses. Neste período as contas foram supervisionadas periodicamente para facilitar a análise;
5. Análise dos spams - os spams coletados foram analisados de acordo com o provedor de serviço e grupo de exposição o qual as contas foram associadas.

Na próxima seção será feito o detalhamento dos passos citados acima.

3.2 Detalhamento da proposta

Para o desenvolvimento desse trabalho, o primeiro passo foi pesquisar as estratégias utilizadas por *spammers* para obter contas de e-mail. Buscou-se então responder a seguinte pergunta “*De onde os spammers conseguem nossos endereços de e-mail?*”.

Conforme discutido na seção 2.2, as principais estratégias utilizadas por *spammers* para obter tais endereços de e-mail são: programas maliciosos, *Harvesting*, ataques a dicionários e banco de dados. Baseado nas estratégias citadas por Marluce Peron (2015), o objetivo consistiu em investigar formas de expor contas de e-mail para que elas estejam a disposição de potenciais *spammers*. Ou seja, aqui buscou-se responder a seguinte pergunta “*Onde estas contas poderiam ser expostas, de modo a serem alvos do ataque de spammers e possivelmente capturadas por umas das estratégias utilizadas por eles?*”.

Baseado na estratégia de *Harvesting*, foram expostas algumas contas de e-mail em sites pela web. A ideia foi verificar se essas contas seriam capturadas por *crawlers* utilizados pelos *spammers* para varrer páginas da web em busca de endereços de e-mail, conforme diz a estratégia de *Harvesting*. Foi então definido o primeiro grupo de exposição para esse experimento, o Grupo 1.

O segundo grupo de exposição escolhido foi baseado na estratégia de banco de dados. Aqui, a ideia foi verificar até que ponto as informações pessoais que enviamos para as empresas estão realmente protegidas. Já que a estratégia de banco de dados utilizada para obter contas de e-mail se dá justamente através do acesso ao banco de dados de empresas com listas de e-mails, é necessário elaborar uma forma de enviar para empresas um certo número de contas de e-mail. Para facilitar o processo, foi definido que o banco de dados de tais empresas seria populado usando a ferramenta de cadastro *on-line* das empresas. Esse foi então definido como sendo o Grupo 2 do experimento.

Por fim, como terceiro grupo de exposição, a ideia é expor contas em redes sociais. Sabe-se que a presença de *bots* (programas automatizados) que espalham spam e conteúdo malicioso em redes sociais é comum (CHU; GIANVECCHIO; WANG, 2010). Além disso, de acordo com (HANN et al., 2006), as redes sociais são uma das mais influentes e importantes formas de *marketing* e o objetivo do spam é promover as vendas. Logo, se os *crawlers* buscam por contas de e-mail em páginas Web, provavelmente também possuem como alvo as redes sociais.

A seguir cada uma das estratégias será detalhada.

Grupo 1

Com o intuito de diversificar os endereços de e-mail coletados à partir de páginas públicas da Internet, foi feita uma divisão em quatro subgrupos dentro do Grupo 1, de

acordo com o tipo de página: a) Páginas pessoal/empresarial, b) Fóruns, c) Blogs e d) Entretenimento.

Para o subgrupo Páginas pessoal/empresarial, já que a ideia era disseminar os endereços de e-mail no arquivo HTML dessas páginas, foram selecionados páginas onde o acesso ao código fonte seria possível de modificação. Foram então selecionadas três páginas da web, duas páginas pessoais de professores da Universidade Federal de Uberlândia e uma de uma empresa de atacado distribuidor de produtos de papelaria e escritório. A figura 2 mostra a forma como isso foi feito nessas páginas. A expressão "display: none" utilizada em cada tag `<div></div>` permitiu que as contas fossem expostas no arquivo HTML da página, mas que ficassem invisíveis visualmente na página.

```
<div style="display: none" id="wb_element_instance58" class="wb_element" style=" line-height: normal;">
  <p>
    <a href="mailto:██████████@yahoo.com" data-type="email" data-url="██████████@yahoo.com">
      ██████████@yahoo.com</a>
    </p>
  </div>
<div style="display: none" id="wb_element_instance59" class="wb_element" style=" line-height: normal;">
  <p>
    <a href="mailto:██████████@gmail.com" data-type="email" data-url="██████████@gmail.com">
      ██████████@gmail.com</a>
    </p>
  </div>
<div style="display: none" id="wb_element_instance60" class="wb_element" style=" line-height: normal;">
  <p>
    <a href="mailto:██████████@hotmail.com" data-type="email" data-url="██████████@hotmail.com">
      ██████████@hotmail.com</a>
    </p>
  </div>
```

Figura 4 – Código HTML utilizado nas páginas para exposição das contas de e-mail.

Para o subgrupo fóruns, de acordo com o site Alexa, um serviço Web que fornece listas de sites ordenados por estatísticas de tráfego, os dois fóruns mais populares são *stackoverflow*¹ e *quora*². Por isso ambos foram escolhidos para exposição de endereços de e-mail.

Para blogs, segundo o site *ebizmba*³, um outro serviço Web que fornece listas de sites ordenados por estatísticas de tráfego, com classificação por categorias mais específicas (como blogs), os dois mais populares são *huffpostbrasil*⁴ e *tmz*⁵. Por isso também foram escolhidos para expor endereços de e-mail.

Para entretenimento, também segundo o sítio *ebizmba* os três sítios de entretenimento mais populares são *Facebook*⁶, *YouTube*⁷ e *Reddit*⁸. Cada um deles foi escolhido para exposição de endereços de e-mail.

¹ <https://pt.stackoverflow.com/>

² <https://www.quora.com/>

³ <http://www.ebizmba.com/>

⁴ <http://www.huffpostbrasil.com/>

⁵ <http://www.tnz.com/>

⁶ <https://www.facebook.com/>

⁷ <https://www.youtube.com/>

⁸ <https://www.reddit.com/>

Diferente do subgrupo Página pessoal/empresarial, nestes outros subgrupos citados (Fóruns, Blogs e Entretenimento) a exposição das contas foi feita da seguinte forma: a princípio, foram selecionadas publicações populares, onde a quantidade de visualizações fossem significativamente numerosas. Após isso, foram feitos comentários em tais publicações selecionadas. Nesses comentários foram inseridos alguns dos endereços de e-mail criados para esse estudo.

Por exemplo, através de uma publicação feita por um usuário no fórum *stackoverflow*, onde havia cerca de 15.886 visualizações, foi realizado um comentário e, no mesmo, incluído três contas de e-mail; um de cada provedor utilizado para esse experimento. A Figura 5 mostra como isso foi feito.

No site do *stackoverflow* em especial, talvez por se tratar de um fórum de perguntas e respostas de assuntos específicos, no caso programação de computadores, ao expor as contas em certas publicações, os usuários classificavam a resposta como fora do contexto e as excluía. Como solução para esse empecilho, foram publicadas respostas que estavam dentro do contexto da pergunta e no meio do texto expostas as contas de e-mail. Dessa forma, conseguiu-se evitar que a resposta fosse excluída. Para os outros sites não houve essa dificuldade.

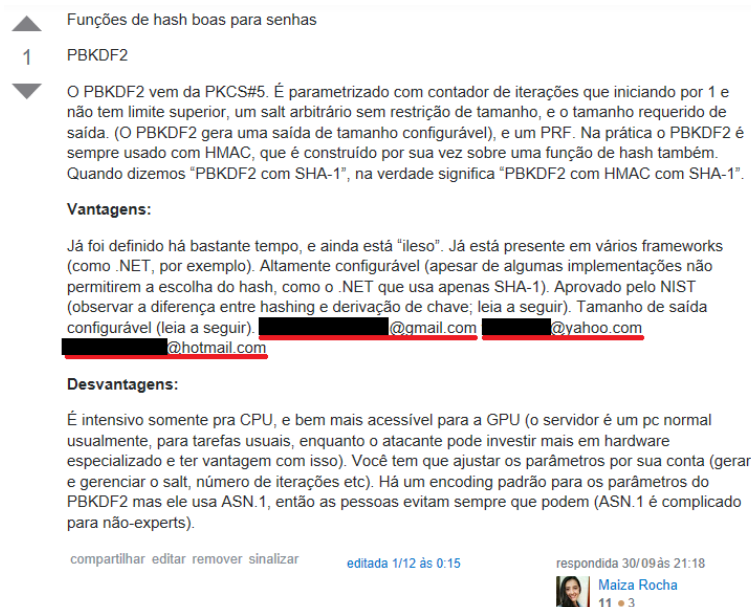


Figura 5 – Forma como as contas foram expostas no sítio do *stackoverflow*.

Grupo 2

Para o grupo 2, foram feitos cadastros online em lojas com algumas das contas de e-mail criadas. O critério usado para escolha das lojas foi a reputação da mesma perante os seus próprios consumidores.

Foram feitas pesquisas no Google sobre "vazamento de dados de lojas", essas pesquisas retornaram notícias e sítios de comentários de consumidores, como o *reclameaqui*⁹. Aquelas lojas que apareceram com maior frequência em relação às pesquisas foram selecionadas. Ou seja, baseado em notícias sobre o vazamento de dados de tais empresas e em comentários dos usuários sobre suas experiências ao efetuarem cadastros nessas lojas, terem seus dados pessoais vazados e/ou por receberem e-mails não solicitados os seguintes sítios de empresas de varejo nacionais foram selecionados três lojas/empresas. Decidiu-se por não revelar a identidade real de tais lojas, por isso, para o presente trabalho as lojas foram identificadas com nomes fictícios da seguinte forma: Empresa 1, 2 e 3.

Ao efetuar o cadastro nessas lojas, havia a opção de receber promoções por e-mail, porém como o objetivo do trabalho é verificar spam, ou seja, e-mails não solicitados, essa opção foi recusada. Garantindo assim, que todos os e-mails da caixa de mensagens são realmente indesejáveis, ou seja, spams.

A figura 6 mostra como isso foi feito.

The screenshot shows a registration form for 'Empresa 1'. At the top, there is a red navigation bar with links for 'Central de atendimento', 'Meus pedidos', and 'Ambiente 100% Seguro'. Below the navigation bar, the page title is 'Identificação'. There are two radio buttons for 'Pessoa física' (selected) and 'Pessoa jurídica'. A 'Já tenho cadastro' button is in the top right. The form is divided into two main sections: 'Dados Pessoais' and 'Dados de acesso ao [redacted]'.
Dados Pessoais:
- **Nome Completo:*** Input field with 'Emily Martins Azevedo' and a green checkmark.
- **CPF:*** Input field with '203.466.063-30' and a green checkmark.
- **Telefone:*** Input field with a dropdown set to 'Celular', '83', and '92674352', all with green checkmarks.
- **Telefone 2:*** Input field with a dropdown set to 'Residencial', '83', and '29538457', all with green checkmarks, and a '+ Adicionar outro' link.
- **Data de Nascimento:*** Input field with '15', '05', and '1983', all with green checkmarks.
- **Sexo:*** Radio buttons for 'Masculino' and 'Feminino' (selected).
Dados de acesso ao [redacted]:
- **E-mail:*** Input field with a redacted email and '@gmail.com', and a green checkmark.
- **Confirmar E-mail:*** Input field with a redacted email and '@gmail.com', and a green checkmark.
- **Senha:*** Input field with '.....' and a green checkmark.
- **Confirmar Senha:*** Input field with '.....' and a green checkmark.
Below the password fields, there are two checkboxes: 'Desejo receber ofertas de produtos e serviços' and 'Desejo receber ofertas por SMS grátis no celular', both unchecked. A red 'Continuar' button is at the bottom right. A note '* Campos Obrigatórios' is at the bottom left of the form area.

Figura 6 – Exemplo de cadastro realizado na Empresa 1.

⁹ <https://www.reclameaqui.com.br/>

Grupo 3

E por fim, para o grupo de exposição 3, o critério utilizado para escolha da rede social foi sua popularidade. Segundo estatísticas feitas pelo sítio Statista (STATISTA, 2017), levando em consideração o número de usuários ativos, as três redes sociais mais populares no mundo são: *Facebook*, *YouTube* e *WhatsApp*, nesta ordem. Como o *Facebook* e o *YouTube* já entraram no Grupo 2 como subgrupo Entretenimento, decidiu por utilizar o *WhatsApp*. A exposição das contas aconteceu em grupos relativamente populares do *WhatsApp*, onde a troca de informações era feita com frequência entre os participantes.

Para isso, foi necessário encontrar listas de grupos públicos de *WhatsApp*. Após uma pesquisa feita por tais grupos no *Google* o seguinte sítio que fornece acesso a grupos públicos de *WhatsApp* foi selecionado o portal *grupowhats*¹⁰. Antes de entrar em determinado grupo o sítio informa a quantidade de participantes de cada grupo disponível.

A partir do sítio *grupowhats*, os seguintes grupos públicos foram selecionados: um grupo de notícias sobre assuntos diversos (204 participantes na época do acesso), outro grupo de compra e venda (187 participantes na época do acesso), e um grupo sobre assuntos relacionados a tecnologia (232 participantes na época do acesso).

A Figura 8 mostra a forma como essas contas foram expostas no grupo de compra e venda, chamado ANUNCIE BRASIL. Foram dez contas expostas por grupo (notícias, compra e venda e tecnologia), ao todo 30 contas divididas igualmente por provedor (*Yahoo*, *Gmail* e *Hotmail*).

A Figura 7 mostra uma árvore dos grupos de exposição utilizados para esse experimento. Para cada estratégia de exposição existem subgrupos que aumentam a diversidade da proposta e permitem fornecer um olhar mais preciso na análise de cada uma das estratégias usadas pelos *spammers* na coleta de endereços de e-mail.

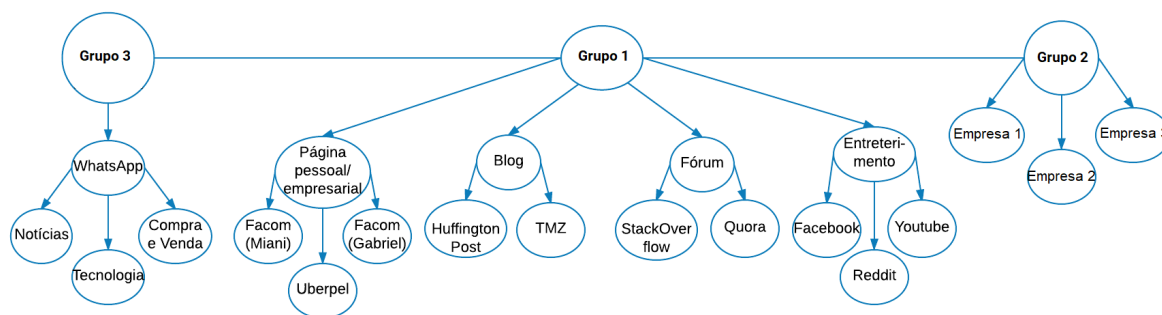


Figura 7 – Grupos de exposição criados.

Seguindo a metodologia, o próximo passo foi a criação das contas de e-mail. Inicialmente, foi empregado um gerador de informações fictícias¹¹, para que dados como

¹⁰ <http://www.gruposwhats.com/>

¹¹ www.4devs.com.br



Figura 8 – Exposição das contas no grupo ANUNCIE BRASIL do WhatsApp.

Nome, Data de nascimento, Telefone, Endereço residencial e Endereço de e-mail fossem gerados e utilizados para o cadastro nos provedores de serviço. Com esses dados foi desenvolvida uma base de dados contendo 105 contas de e-mail distribuídas em três provedores distintos, 35 contas para cada provedor.

Para decidir quais provedores de serviço seriam usados, inicialmente foi realizada uma pesquisa sobre quais são os mais populares, que de acordo com [TECHTIMES \(2014\)](#) seriam: *Yahoo*, *Gmail* e *Hotmail*. Para cada relação provedor x grupo de exposição foram criadas 10 contas, ou seja, 10 (contas) x 3 (provedores) x 3 (grupos), igual a 90 contas, que foram distribuídas nos três grupos de exposição, sendo 30 contas em cada. As 15 contas restantes foram distribuídas igualmente em cada um dos provedores mas não associadas aos grupos de exposição. O objetivo de tais contas é fornecer um controle do experimento, ou seja, comparar o estado das contas em exposição e sem exposição.

Seguindo a nomenclatura do trabalho proposto por [Hann et al. \(2006\)](#), no presente trabalho também foi dado o nome de *contas de exposição* às contas que possuem grupo de exposição e aquelas não possuem grupo de exposição, foi atribuído o nome de *contas de controle*. Temos assim o seguinte número de contas para nossa proposta: 90 contas de exposição e 15 contas de controle, totalizando as 105 contas criadas.

A Tabela 2 sumariza a criação e divisão das contas de acordo com cada provedor

e seu respectivo grupo de exposição.

Tabela 2 – Distribuição das contas de acordo com o provedor e o grupo de exposição.

Provedor	Grupo de Exposição	Número de contas	Total
Yahoo	Grupo 1	10	35
	Grupo 2	10	
	Grupo 3	10	
	Sem grupo de exposição (contas controle)	5	
Gmail	Grupo 1	10	35
	Grupo 2	10	
	Grupo 3	10	
	Sem grupo de exposição (contas controle)	5	
Hotmail	Grupo 1	10	35
	Grupo 2	10	
	Grupo 3	10	
	Sem grupo de exposição (contas controle)	5	
Total de contas criadas			105

As contas de e-mail foram acompanhadas por um período pré-estabelecido de 10 semanas, que foi iniciado no fim de outubro e finalizado no início de dezembro. Nas duas primeiras semanas todas as contas de e-mail passaram por um período chamado de quarentena, onde não foi feita nenhuma atividade nas contas. O intuito é observar se elas receberiam mensagens não solicitadas. Elas foram supervisionadas mensalmente para facilitar a análise da quantidade de spams recebidos ao final do período de análise. A Figura 9 apresenta cada fase do experimento realizado.

O estudo proposto permite verificar a incidência de spams da seguinte forma: a) em cada grupo de exposição isolado, b) em cada provedor isolado e c) em cada relação grupo de exposição x provedor.

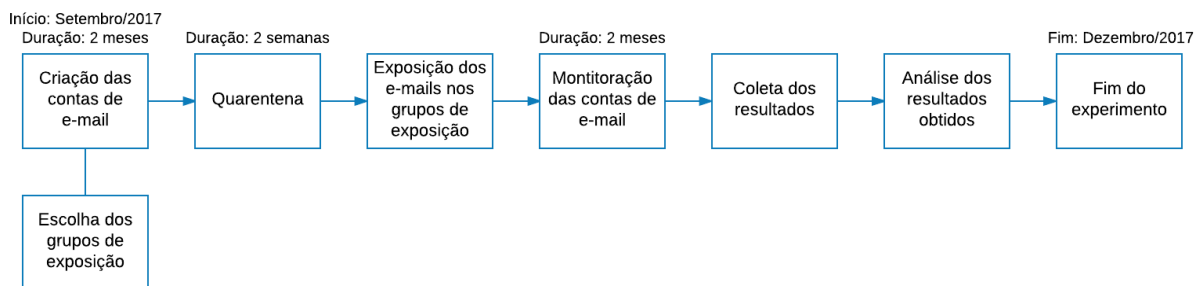


Figura 9 – Fluxograma da metodologia proposta.

O capítulo 4 apresenta uma análise preliminar dos dados obtidos após a exposição das contas por 8 semanas em cada um dos serviços escolhidos. O objetivo é validar a

metodologia proposta e iniciar uma discussão acerca da eficiência das estratégias adotadas pelos spammers na obtenção de contas de e-mail.

4 Análise dos resultados preliminares

No início de Setembro de 2017 começaram a ser criadas manualmente as 105 contas de e-mail utilizadas no experimento. A etapa de criação das contas demandou um certo esforço pelo fato de termos optado por provedores populares (*Yahoo*, *Gmail* e *Hotmail*), os quais exigem medidas de seguranças maiores. Por exemplo, para criar os e-mails nestes provedores foi preciso adicionar um número de telefone e o provedor do serviço enviava um código para esse número como uma forma de confirmação de identidade. Os provedores aceitavam em média 3 números de telefone por conta criada, ou seja, precisou-se em torno de 35 números de telefones diferentes para realizar essa etapa inicial do experimento. Esse controle ajuda a diminuir a influência de atacantes utilizando processos automatizados, o que dificultou a realização da base de dados inicial do experimento, mas não impediu que ela fosse concluída.

Durante o processo de criação das contas nenhuma característica do usuário foi levada em consideração, já que aqui, diferentemente dos estudos propostos por (HANN et al., 2006) e (OLIVEIRA, 2016) não deseja-se verificar característica dos usuários que recebem mais spam. E sim, fazer uma análise das estratégias mais utilizadas por *spammers* para obter contas de e-mail.

Após a criação das contas de e-mail, as mesmas permaneceram sem atividades por um período de duas semanas, o qual recebeu o nome de “quarentena”. Dando prosseguimento ao experimento, as contas de e-mail começaram a ser associadas aos seus respectivos grupos de exposição. A associação das contas com os grupos de exposição foi feita de forma determinada, de modo que uma conta foi associada apenas um grupo de exposição. Isso foi feito para que no momento de coletar os resultados, fosse possível identificar exatamente de onde se originou determinado spam.

As contas ficaram expostas por um período de 8 semanas. No início de dezembro de 2017 o experimento foi concluído. A Tabela 3 reporta detalhes das contas que receberam spam.

É possível notar que foram recebidos um total de 47 spams dentro do período analisado (8 semanas). Esse número é menor do que em outros estudos como por exemplo, os estudos feitos por (HANN et al., 2006) e (OLIVEIRA, 2016), que receberam uma média de 1.580 e 1.032 mensagens de spam, respectivamente. Porém, é importante ressaltar que o tempo de maturação das contas também foi relativamente menor, comparado com os mesmos estudos. (HANN et al., 2006) monitorou suas contas por um período de 7 meses e (OLIVEIRA, 2016) por um período de 5 meses.

Seria necessário um maior tempo de exposição dessas contas para que pudéssemos

Tabela 3 – Quantidade de spam recebida em 8 semanas

ID	Conta	Provedor	Grupo de Exposição	N.º de spams
c1	sofiame***	Gmail	Grupo 2	1
c2	caroli***	Gmail	Grupo 1	1
c3	dso***	Yahoo	Grupo 3	2
c4	kaueg***	Hotmail	Grupo 1	5
c5	luanma***	Hotmail	Grupo 1	6
c6	caiopereira***		Grupo 1	6
c7	diogocr***		Grupo 1	6
c8	luisbarb***		Grupo 2	4
c9	joa***		Grupo 2	5
c10	vitoria***		Grupo 3	6
c11	vitor***		Sem grupo de exposição (conta controle)	6
Total de spams recebidos				48

coletar resultados mais significativos.

Embora o número de spams recebidos tenha sido pequeno, é possível fazer uma análise detalhada das contas que foram alvos de spams e iniciar a discussão sobre as estratégias usadas por spammers na coleta de contas de e-mail.

De acordo com a Tabela 3, 11 contas de e-mail receberam spams (c1 à c11). É possível observar também que os três provedores de serviço (*Yahoo*, *Gmail* e *Hotmail*) utilizados, assim como os três grupos de exposição (Grupo 1, Grupo 2 e Grupo 3) propostos foram afetados. Lembrando que o Grupo 1 é formado por endereços de e-mail coletados à partir de páginas públicas da Internet usando *crawlers*, o Grupo 2 por endereços de e-mail coletados à partir de registros presentes em bases de dados de organizações e o Grupo 3 consiste em endereços de e-mail coletados à partir de redes sociais.

Primeiramente serão analisadas as contas c5, c6, c7, c8, c9, c10 e c11 pois estas receberam o mesmo tipo de spam. A Figura 10 mostra a caixa de entrada com os e-mails recebidos em tais contas e a Figura 11 mostra o conteúdo de um desses e-mails. É possível traçar as seguintes conclusões a respeito dessas contas:

- Os e-mails recebidos foram classificados como spams enviados pelo próprio provedor de serviço utilizado, que neste caso foi o *Hotmail*. Aparentemente são e-mails automáticos enviados pelo provedor. Mas, se esse fosse o caso todos os e-mails cadastrados no *Hotmail* deveriam recebê-los, o que não aconteceu. Apenas 7 das 35 contas cadastradas no provedor do *Hotmail*, receberam esses e-mails. Ou seja, não é uma mensagem automática enviada por padrão a todos os e-mails cadastrados. O

que nos faz pensar: "Qual o critério utilizado pelo provedor em questão para enviar esses e-mails para algumas contas e não para todas?" Inicialmente poder-se-ia pensar em algum interesse declarado do próprio usuário no momento do cadastro. Contudo, isso não faz sentido aqui, já que todas as contas foram cadastradas da mesma forma e sem declaração de interesses.

- Esses spams que chegaram nessas contas específicas não tiveram relação alguma com o grupo de exposição a qual a conta foi exposta. Foi possível concluir isso de duas formas: 1) a conta c11 é uma conta que recebeu os mesmos e-mails e essa é uma conta a qual chamamos de conta de controle, ou seja, que não foi associada a nenhum grupo de exposição. Os e-mails recebidos na caixa de mensagens dessas contas de controle, só poderiam ser oriundos do próprio servidor. 2) Como podemos observar na figura 7, existe um e-mail recebido no dia 07/09/2017, período esse em que as contas estavam em quarentena, ou seja, ainda não estavam expostas em nenhum grupo de exposição.

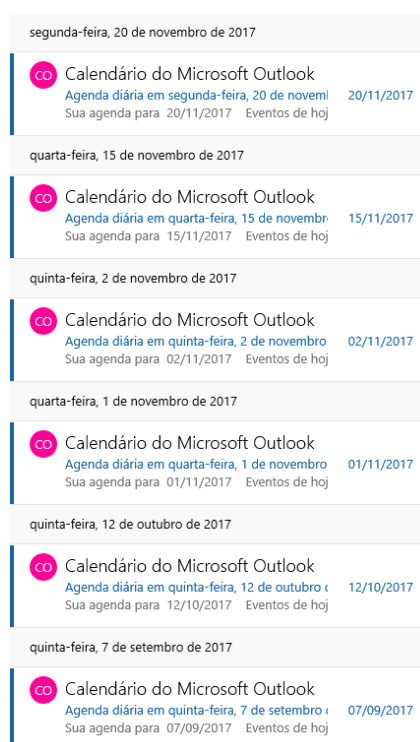


Figura 10 – Caixa de entrada da conta c6.

A próxima conta analisada é a c1. Ela foi cadastrada no provedor de serviços do *Gmail* e exposta no grupo de exposição Lojas online, mais especificamente no site online da Empresa 1. A exposição foi feita por meio de um cadastro onde o e-mail utilizado para efetuar o mesmo foi "sofiameloazevedo@gmail.com". A conta foi exposta do dia 30/09/2017 e no dia 01/12/2017 um e-mail chegou a caixa de entrada, com o seguinte endereço de e-mail remetente: "registration@facebookmail.com". Dentro do período analisado até então,

Agenda diária em segunda-feira, 20 de novembro de 2017



Calendário do Microsoft Outlook
20/11/2017 09:56



Para: Caio Pereira Rocha

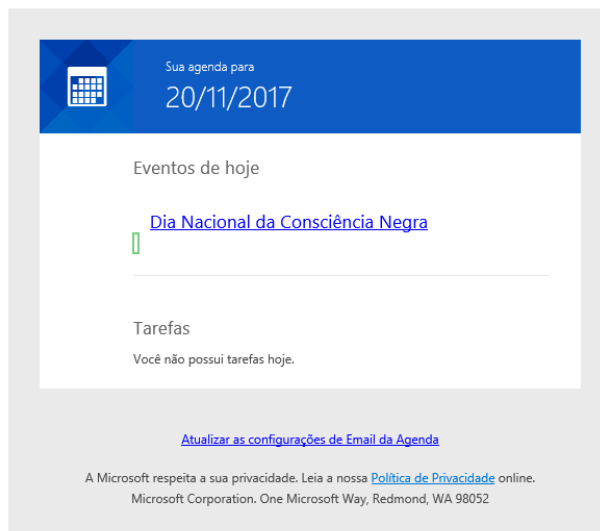


Figura 11 – Conteúdo de um dos e-mails recebidos pela conta c6.

esse foi o único e-mail recebido nesta conta. A Figura 12 mostra o e-mail spam recebido na conta c1.

28417 je vaš potvrdni kôd za Facebook



Facebook
01/12/2017 11:11

Para: Sofia Melo Azevedo



Potrebna je radnja: potvrdite svoj korisnički račun za Facebook

Bok, Sofia Melo,

Nedavno ste se registrirali na Facebook. Za dovršetak registracije, molimo, potvrdite svoj račun.

[Potvrdite svoj korisnički račun](#)

Možda ćete trebati unijeti ovaj potvrdni kôd:

28417

Putem Facebooka možete komunicirati i ostati u vezi sa svim prijateljima. Kad se pridružite, moći ćete dijeliti fotografije, planirati događaje i još mnogo toga.

Ova poruka je poslana na sofiameloazevedo@gmail.com. Ako u buduću ne želite primati ovu e-poštu od Facebooka, [prestanite pratiti](#). Ako niste kreirali račun=8

[Baixar mensagem e imagens \(17.7 KB\)](#)

Figura 12 – E-mail recebido na conta c1.

A conta c2, assim como a conta c1, também foi cadastrada no provedor de servi-

ços do *Gmail*, porém aqui o grupo de exposição a qual ela foi associada foi o Grupo 2, dentro no subgrupo Páginas pessoal/empresarial. A exposição da conta foi feita no arquivo HTML da página. A conta foi exposta no dia 20/09/2017 e no dia 30/11/2017 um e-mail chegou a caixa de entrada, com o seguinte endereço de e-mail remetente: "tim@easymailforgmail.com". Assim como na conta c1, esse também foi o único e-mail recebido até o momento analisado. A Figura 13 mostra o e-mail spam recebido na conta c2.

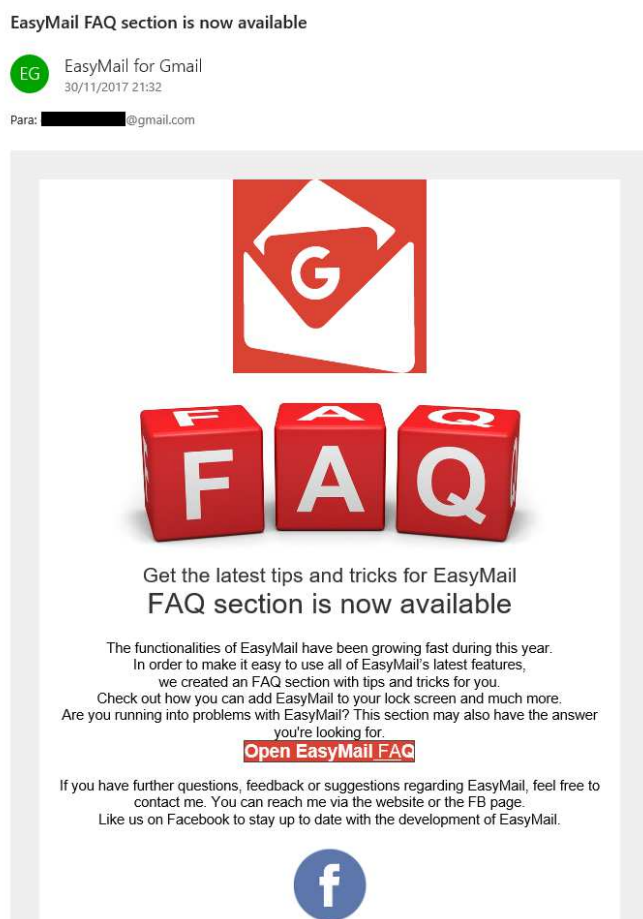


Figura 13 – E-mail recebido na conta c2

A conta c3 foi cadastrada no provedor do *Yahoo* e exposta no Grupo 3. Ela foi exposta em um grupo de compra e venda do WhatsApp, chamado ANUNCIE BRASIL, como mostra a Figura 8 no capítulo 3. A exposição da conta aconteceu no dia 08/07/2017 e no dia 11/12/2017 a conta recebeu em sua caixa de entrada um e-mail com o seguinte remetente: "messages-noreply@linkedin.com".

No dia 14/12/2017 para a mesma conta c3, um outro e-mail foi recebido com o seguinte remetente: "invitations@linkedin.com".

A Figura 14, mostra o conteúdo do primeiro e-mail recebido na conta c3. O segundo e-mail recebido no dia 14/12 apresentou conteúdo semelhante ao e-mail recebido

no dia 11/12: o conteúdo de ambos apresentam uma solicitação de amizade da rede social *LinkedIn*.

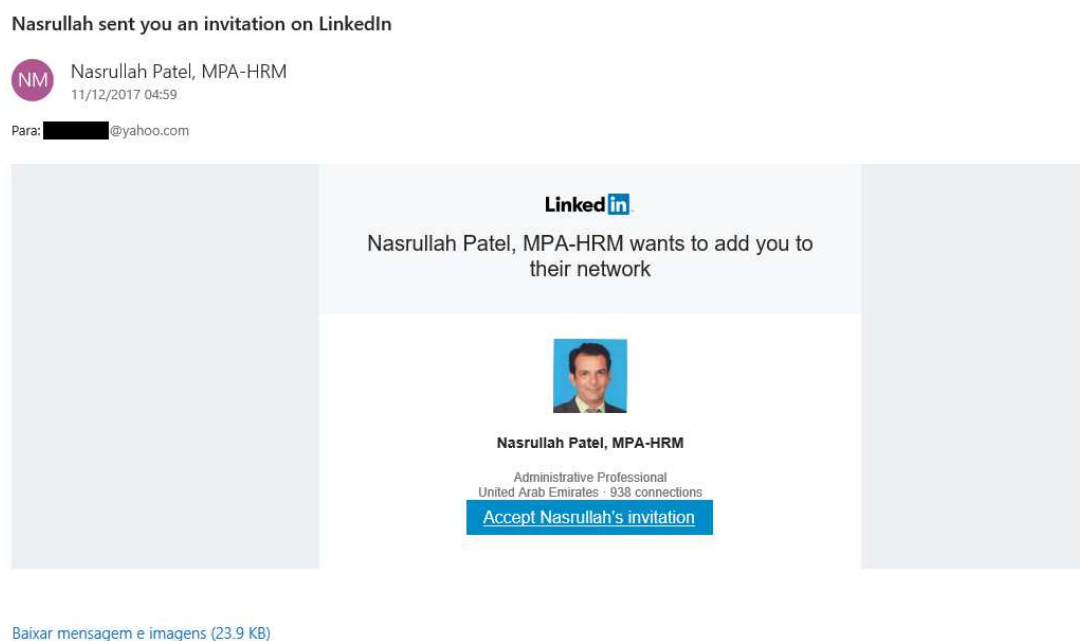


Figura 14 – E-mail recebido na conta c3

A conta c4 foi cadastrada no provedor do *Hotmail* e assim como a conta c2 também foi exposta no Grupo 1, porém dentro do subgrupo Fóruns, no site do StackOverflow. A forma de exposição da conta foi por meio de um comentário feito em uma publicação no dia 30/09/2017, como mostra a figura 3 na sessão 3 deste trabalho. A conta recebeu ao todo 5 e-mails. Diferente das demais contas que receberam seus e-mails na caixa de entrada principal, na conta c4 todos os e-mails recebidos foram marcados como spam e encaminhados para a pasta de lixo eletrônico. A Figura 15 mostra esses e-mails e a data em que cada um foi recebido e a Figura 16 mostra o conteúdo de um desses e-mails recebidos, com o seguinte remetente: "s217676480@nmmu.ac.za".

A Figura 17 ilustra a forma como se deu o recebimento de spams dentro de cada grupo e subgrupo especificadamente. Em vermelho os meios de onde se originou o recebimento de spam dentro de cada grupo e subgrupo.

Tabela 4 – Quantidade de spams por provedor.

Provedor	Número de spams
Yahoo	2
Gmail	2
Hotmail	44

Espera-se que a disseminação de spam para outros grupos ocorra na medida em que o tempo de exposição das contas aumenta. Um resultado semelhante foi encontrado

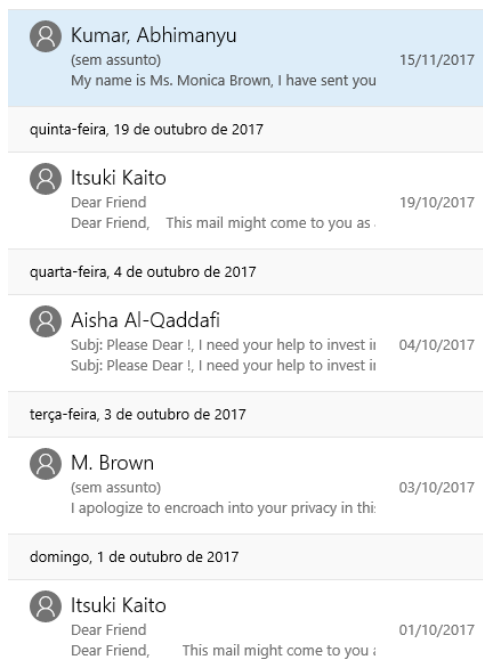


Figura 15 – E-mails recebidos na conta c4.

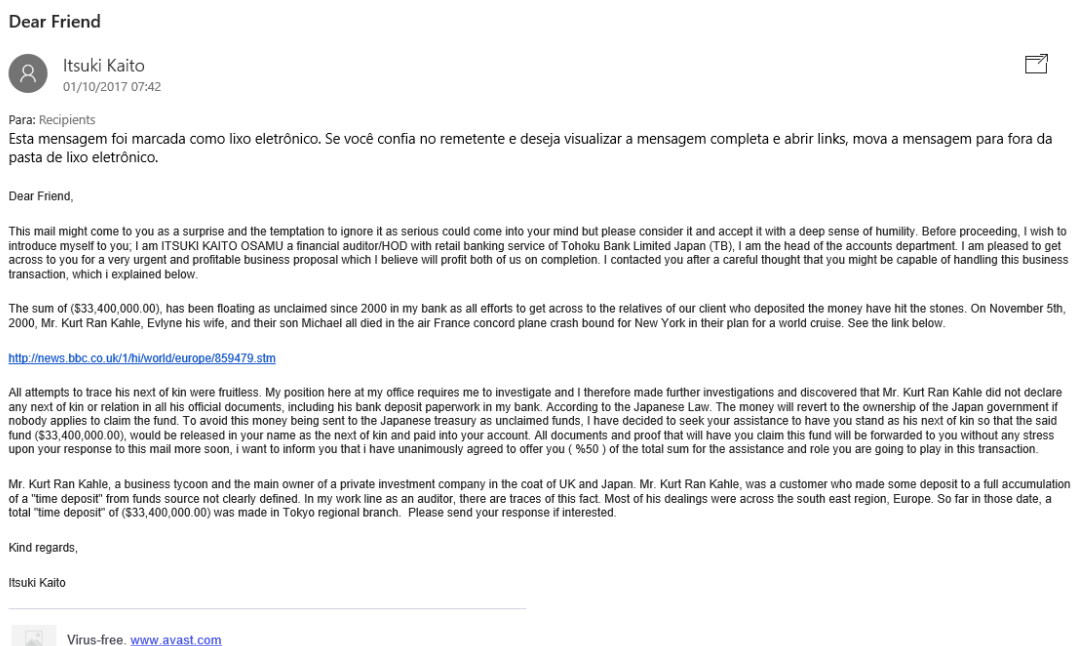


Figura 16 – Conteúdo de um dos e-mails recebidos na conta c4.

por Nagamalai, Dhinakaran e Lee (2008). Os autores discutem que quanto mais antiga a conta de e-mail, maior a probabilidade dela receber spam.

A Tabela 4 mostra o número de spams recebidos por provedor. O provedor de serviços do *Hotmail* recebeu 44 spams no total, uma quantidade alta de spams quando comparado com os provedores do *Yahoo* e *Gmail*, que receberam apenas 2 spams cada. Esse resultado já havia sido confirmado por Hann et al. (2006) em seu estudo, onde os

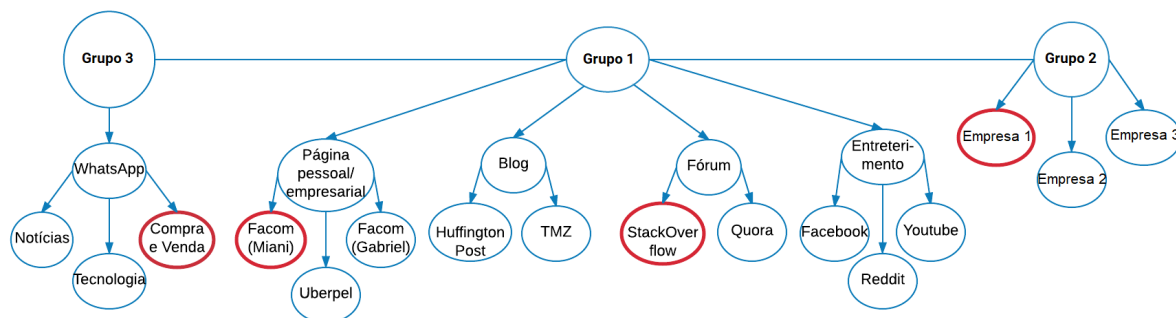


Figura 17 – Distribuição de spams por grupo de exposição.

autores concluem que a identidade do prestador de serviços tem influência no recebimento de spams e que as contas do provedor da *Hotmail* receberam mais spams se comparado os outros provedores de serviço.

Já a Tabela 5 mostra o número de spams recebidos por grupo de exposição. A última linha da tabela "Spams oriundo do provedor" mostra aqueles e-mails que não tiveram relação com o grupo de exposição o qual foi associado, ou seja, os e-mails recebidos foram enviados pelo próprio provedor de serviço a qual a conta foi associada.

Tabela 5 – Quantidade de spams por grupo de exposição.

Grupo de exposição	Número de spams
Grupos 1	6
Grupo 2	1
Grupo 3	2
Spams oriundos do provedor	39

Os resultados preliminares mostram que o Grupo 1, formado pela exposição de contas de e-mail em páginas da Internet, ainda são fontes de dados empregadas pelos *spammers*. Também foi possível validar a escolha pelo Grupo 3 (contas expostas em redes sociais, em particular no aplicativo de mensagem instantânea *WhatsApp*), visto que ao longo do curto período do experimento, duas contas de e-mail já foram alvos de spam usando esse canal de difusão. Por fim, a incidência de spams direcionada ao provedor *Hotmail* mostrou ser maior do que para os outros provedores, comprovando estudos anteriores (HANN et al., 2006) e (OLIVEIRA, 2016). O caso em que uma conta de controle (c11) recebe e-mails indesejados do próprio provedor foi emblemático nesse sentido.

5 Conclusão

Com base nos experimentos conduzidos por [Hann et al. \(2006\)](#) e [Oliveira \(2016\)](#), onde estes investigaram os fatores que influenciam o recebimento de spam baseado em características e comportamentos dos usuários na web, este trabalho propôs também investigar os fatores que influenciam o recebimento de spam, mas baseado nos mecanismos utilizados por *spammers* para obter endereços eletrônicos. O principal objetivo do trabalho é propor um experimento usando dados reais dos principais provedores de e-mail gratuitos do mundo para investigar as preferências e tendências dos *spammers* para conseguir tais endereços de e-mail.

Três formas de exposição de contas de e-mail foram definidas, a saber, páginas da Internet, cadastro em bases de dados de organizações e redes sociais. Após a criação e a associação de cada uma das contas aos grupos propostos o experimento iniciou-se e após oito semanas os resultados preliminares foram investigados.

Apesar do curto tempo de exposição das contas, os três grupos de exposição considerados neste trabalho receberam ao menos um spam, validando o experimento proposto. Acredita-se que a obtenção de resultados mais maduros é apenas uma questão de tempo.

Dos dados obtidos até o fim deste experimento pode-se observar que analogamente ao observado por [Hann et al. \(2006\)](#), a identidade do provedor de serviços tem relação com o recebimento de spam, sendo o *Hotmail* o provedor que recebeu mais spams quando comparado com os outros provedores de serviço considerados no trabalho de [Hann et al. \(2006\)](#) (*Lycos*, *Excite* e *Yahoo*) e também aqui neste trabalho (*Gmail* e *Yahoo*).

Em relação aos grupos de exposição considerados neste estudo (Grupos 1, 2 e 3), é possível observar uma tendência para recebimento de spam em endereços de e-mail coletados à partir de páginas públicas da Internet usando *crawlers*, Grupo 1 considerado neste trabalho.

Como trabalho futuro, espera-se atualizar os resultados obtidos aqui neste estudo. Para isso as contas criadas continuarão por pelo menos mais 5 meses expostas nos seus respectivos grupos de exposição, afim de obter conclusões mais maduras a respeito dos mecanismos utilizados para colheita desses e-mails. Uma importante direção para futuras pesquisas é, por exemplo, envolver as contas utilizadas em cadastro de lojas em transações online e verificar o impacto dessas ações no recebimento de spam.

Referências

- ALMEIDA, T. A. de. *Spam : from the rise to the extinction SPAM : do Surgimento ‘ a Extin , c ~*. Tese (Doutorado), 2010. Citado 3 vezes nas páginas 8, 9 e 12.
- ANTISPAM.BR. *Problemas causados pelo spam*. 2017. Disponível em: <<http://www.antispam.br/problemas/>>. Citado 2 vezes nas páginas 14 e 17.
- CHU, Z.; GIANVECCHIO, S.; WANG, H. Who is Tweeting on Twitter : Human , Bot , or Cyborg ? p. 21–30, 2010. Citado na página 20.
- DIAS, S. Q. *Dicionário do Aurélio Online*. 2000. 42 p. Disponível em: <<http://www.dicionariodoaurelio.com/Pol{í}tic>>. Citado na página 11.
- HANN, I.-h. et al. Who Gets Spammed ? *October*, v. 49, n. 10, p. 83–87, 2006. ISSN 00010782. Citado 10 vezes nas páginas 9, 10, 17, 19, 20, 25, 28, 34, 35 e 36.
- Marluce Peron. *Como os spammers conseguem endereços de e-mail?* 2015. Disponível em: <<https://www.tecmundo.com.br/gmail/2015-como-os-spammers-conseguem-enderecos-de-e-mail-.htm>>. Citado 2 vezes nas páginas 15 e 20.
- MOREIRA, M. S. POLÍTICA DE SEGURANÇA DA INFORMAÇÃO: UMA CONTRIBUIÇÃO PARA O CAMPUS IV. 2014. Citado 3 vezes nas páginas 15, 16 e 17.
- NAGAMALAI, D.; DHINAKARAN, B. C.; LEE, J. K. An In-depth Analysis of Spam and Spammers. v. 2, n. 2, p. 9–22, 2008. Citado 3 vezes nas páginas 17, 18 e 34.
- OLIVEIRA, D. S. Fatores que influenciam o recebimento de spam. 2016. Citado 9 vezes nas páginas 8, 9, 10, 17, 18, 19, 28, 35 e 36.
- OLIVO, C. K.; SANTIN, A. O.; OLIVEIRA, L. E. S. Capítulo 4 Abordagens para Detecção de Spam de E-mail. p. 141–182, 2015. Citado 2 vezes nas páginas 8 e 13.
- PRINCE, M. B. et al. Understanding How Spammers Steal Your E-Mail Address : An Analysis of the First Six Months of Data from Project Honey Pot. n. March, 2003. Citado 2 vezes nas páginas 17 e 18.
- SPANCESKI, F. R. POLÍTICA DE SEGURANÇA DA INFORMAÇÃO – POLÍTICA DE SEGURANÇA DA INFORMAÇÃO –. 2004. Citado na página 11.
- STALLINGS, W. *Criptografia e Segurança de redes*. [S.l.: s.n.], 2008. 512 p. ISBN 85-98569-46-1. Citado na página 11.
- STATISTA. *Most famous social network sites worldwide as of April 2017, ranked by number of active users (in millions)*. 2017. Disponível em: <<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>>. Citado 3 vezes nas páginas 4, 8 e 24.
- SZENDRODI, R. J. C. Universidade Federal do Rio de Janeiro – UFRJ. p. 1–42, 2005. Citado na página 13.

TECHTIMES. What's the most popular email service in the world? 2014. Disponível em: <<http://www.techtimes.com/articles/15802/20140917/most-popular-email-service-in-the-world.htm>>. Citado na página 25.

TEIXEIRA, R. C. O pesadelo do Spam. v. 5, 2001. Disponível em: <<http://sites.usp.br/cetirp/wp-content/uploads/sites/47/2015/02/OPesadeloDoSPAM.pdf>>. Citado 2 vezes nas páginas 12 e 14.

TEOTÔNIO, Í. D. *Entendendo os Fundamentos da Segurança da Informação*. 2013. Disponível em: <<https://www.profissionaisti.com.br/2013/10/entendendo-os-fundamentos-da-seguranca-da-informacao/>>. Citado na página 12.