Uma metodologia para identificação adaptativa e caracterização de phishing

Pedro Henrique B. Las-Casas[1], Osvaldo Fonseca[1], Elverton Fazzion[1]
Cristine Hoepers[2], Klaus Steding-Jessen[2], Marcelo H. P. Chaves[2],
Ítalo Cunha[1], Wagner Meira Jr.[1], Dorgival Guedes[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

[2]CERT.br - Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança
NIC.br - Núcleo de Informação e Coordenação do Ponto BR

{pedro.lascasas,osvaldo.morais,elverton}@dcc.ufmg.br

{cristine,klaus,mhp}@cert.br

{cunha, meira, dorgival}@dcc.ufmg.br

Abstract. *Phishing remains one of the most significant Internet security problems, causing financial damage to organizations and users. This type of attack combines social engineering and other sophisticated techniques with which the attacker attempts to deceive the victim to steal personal information. Usually, the fight against phishing is done with some of the same techniques used to combat spam, such as those that focus on sets of words and characteristic terms. However, there are few studies that specifically address phishing and there are doubts about the stability and dynamics of message characteristics over time and how they can be leveraged to improve the system's defenses. In this work, we present an adaptive method to identify phishing messages and apply it on a large data set to identify and characterize hundreds of phishing campaigns.*

Resumo. *Phishing continua sendo um dos mais significativos problemas de segurança da Internet, causando prejuízos financeiros a organizações e usuários. Esse tipo de ataque combina engenharia social e outras técnicas sofisticadas com as quais o atacante tenta ludibriar a vítima para roubar informações pessoais desta. Usualmente, o combate ao phishing é feito com algumas das mesmas técnicas usadas no combate ao spam, como aquelas que focam em conjuntos de palavras e termos característicos. Entretanto há poucos trabalhos que tratam especificamente de phishing e há dúvidas sobre a estabilidade ou evolução de características das mensagens ao longo do tempo e sobre como elas podem ser exploradas para melhorar as defesas do sistema. Neste trabalho apresentamos um método adaptativo para identificação de mensagens de phishing e aplicamos este método em um grande conjunto de dados para identificar e caracterizar centenas de campanhas de phishing.*

1. Introdução

Atualmente, phishing é uma das atividades criminosas mais lucrativas na Internet. Um estudo de 2007 mostra que ataques de phishing nos Estados Unidos causaram prejuízos

de mais de 3,2 bilhões de dólares, afetando cerca de 3,6 milhões de usuários. De acordo com um relatório da Kaspersky Lab, de junho de 2012 à junho de 2013, 37,3 milhões de pessoas reportaram terem sido vítimas de tais ataques, representando um aumento de 87% em relação ao ano anterior. Phishing combina engenharia social e técnicas de ataque sofisticadas para enganar usuários, levando-os a acreditar que as entidades em questão são legítimas, de forma a roubar dados pessoais destas vítimas. Em geral, atacantes tentam ludibriar os usuários, mascarando emails e páginas web com marcas confiáveis para roubar informações sensíveis das vítimas, como senhas, números de identificação pessoal e dados do cartão de crédito.

Apesar de haver um grande esforço por parte da comunidade mundial para combater esse tipo de fraude [Fette et al. 2007, Kumaraguru et al. 2007, Bergholz et al. 2010], o fato notório é que o *phishing* ainda persiste, o que nos leva a crer que o *phisher*[1] evolui suas técnicas para ludibriar os diversos métodos de mitigação criados. Essa motivação deixa visível que tão importante quanto construir ferramentas para combater o *phishing*, é tentar entendê-lo em sua essência para permitir a evolução e o aprimoramento de mecanismos que o combatam. Essa atividade, aparentemente simples, demanda uma grande complexidade para obtenção de um conjunto de dados representativos o suficiente para tornar o trabalho relevante para a comunidade, bem como para criação ou adaptação de técnicas de aprendizado de máquina.

Neste trabalho, buscamos entender as principais características de *phishing* que norteiam sua disseminação atual utilizando 13 *honeypots* de baixa interatividade distribuídos em diferentes pontos do mundo. Esses *honeypots*, configurados para parecerem servidores de emails vulneráveis, coletaram, durante um período de 91 dias, 1,13 bilhões de mensagens de emails mescladas entre duas categorias: *spam* e *phishing*. Para separar os dois grupos, utilizamos termos característicos de *phishing* como "banco", "clique" e "crédito" e comparamos com os termos representativos de cada email, obtidos com TF-IDF (*Term Frequency–Inverse Document Frequency*). Para cada mensagem, assinalamos um escore de compatibilidade entre as palavras do email e de *phishing* e comparamos a um limiar mínimo global para classificar o email como pertencente ao conjunto de *phishing*.

A primeira contribuição do trabalho é mostrar uma lista atualizada de palavras relacionadas a *phishing* encontradas em nossa base de dados, de forma a elucidar os principais tipos de técnicas utilizadas pelos *phishers*. Estando de acordo com nosso argumento inicial, onde caracterizamos o *phisher* como um personagem dinâmico, utilizamos as técnicas encontradas em [Mikolov et al. 2013] para expandir nossa lista de palavras, obtida na literatura, com outros termos que estejam relacionados aos termos iniciais. Com essa simples expansão, pouco mais de um milhão de mensagens que continham apenas uma palavra do conjunto inicial e que teriam, portanto, um escore baixo, agora tem seus escores maiores devido à adição das novas palavras e serão classificadas como *phishing* com maior precisão. Acreditamos que essa aboragem seja essencial para melhor diferenciar as mensagens de *phishing* das demais.

Fomos ainda capazes de apresentar características interessantes do phishing, como referente ao seu envio, majoritariamente através do protocolo SMTP. Além disso, mostramos que poucos endereços IP e poucas campanhas são responsáveis pela maior parte dos

[1] Aquele que envia o *phishing*.

emails relacionados ao *phishing*, mostrando que, ao combater estes poucos casos, torna-se possível reduzir consideravelmente os prejuízos causados por estes ataques.

Este trabalho está estruturado da forma que se segue. Na Seção 2 apresentamos trabalhos relacionados aos ataques de *phishing*. Na Seção 3 descrevemos, a um nível mais refinado, os coletores de mensagens e o processo para separar as mensagens de *spam* e *phishing*. Na Seção 4, focamos nas principais características do *phishing* e as principais campanhas encontradas em nossa base de dados. Concluimos o trabalho na Seção 5, mostrando os trabalhos futuros que podem ser feitos utilizando os resultados deste trabalho como inspiração.

2. Trabalhos Relacionados

Phishers utilizam diversas técnicas para ludibriar suas potenciais vítimas. Alguns exemplos das técnicas utilizadas são *DNS cache poisoning*, sequestro de servidores web, imitação de serviços Web ou engenharia social; enganando o usuário e roubando suas informações. Neste trabalho, focamos no último caso, analisando mensagens que buscam aparentar legítimas. Uma forma comum de proteção contra ataques de phishing é o uso de *blacklists*. Entretanto, *blacklists* não provêem proteção durante o ínicio dos ataques de *phishing*, pois a URL ou site deve ser inserido na *blacklist* para que essa forma de proteção comece a funcionar. Além disso, como as URLs utilizadas pelos *phishers* possuem tempo de vida curto, a utilização de *blacklists* acaba sendo ineficiente. De acordo com [Sheng et al. 2009], 63% das campanhas de phishing analisadas pelos autores duraram menos de duas horas.

Com relação às técnicas baseadas no conteúdo para detecção de phishing, Marchal et al. analisa as características das URL's presentes na mensagem para classificá-la como phishing [Marchal et al. 2014]. Entretanto, o método proposto não é capaz de analisar URL's encurtadas, que se mostram cada vez mais presentes. Como analisamos todo o conteúdo da mensagem, nosso método é capaz de identificar mensagens que apresentam tais URL's. Zhang et al. propôs CANTINA [Zhang et al. 2007, Xiang et al. 2011], um método utiliza TF-IDF para extrair os principais termos de um documento, checa-os em máquinas de busca e então, caso o domínio do documento em questão apareça na busca, a página é considerada legítima. Em nosso trabalho, utilizamos o algoritmo TF-IDF para identificar os e-mails de phishing, como mostrado na Seção 3. Em [Aggarwal et al. 2014], os autores utilizam processamento de linguagem natural para identificar e-mails contendo *phishing*. Em sua abordagem, os autores identificam características chave deste tipo de mensagem, como a presença de referências a dinheiro, e utilizam estas características para classificar as mensagens como *phishing*. As características propostas pelos autores foram utilizadas em nosso método de detecção dos e-mails contendo *phishing*.

3. Metodologia

Nesta seção explicamos como realizamos a captura das mensagens de e-mail, o processo de filtragem e as técnicas utilizadas para identificar mensagens de *phishing* e separá-las dos demais spams.

3.1. Coleta das mensagens

As mensagens foram capturadas por treze honeypots de baixa interatividade [Steding-jessen et al. 2007] instalados em diferentes países, sendo dois no Brasil, dois

nos Estados Unidos, e um coletor em cada um dos seguintes países: Argentina, Áustria, Austrália, Chile, Equador, Hong Kong, Holanda, Noruega, Taiwan e Uruguai. Os honeypots são configurados para simular servidores vulneráveis, como *proxies* HTTP e SOCKS e *relays* SMTP abertos. Quando um spammer se conecta ao servidor SMTP de um honeypot, ele é levado a crer que está interagindo com um servidor SMTP operando como um *relay* aberto. Quando uma máquina se conecta a um honeypot através dos protocolos HTTP ou SOCKS, é levada a crer que é capaz de estabelecer conexões com outros servidores SMTP na rede. Com a finalidade de esconder a identidade dos *spammers* e evitar que suas máquinas sejam colocadas em *blacklists*, *proxies* e *relays* são frequentemente utilizados para o envio de spam. Além disso, como esses protocolos são orientados a conexão, é improvável a ocorrência de IP *spoofing*, que só seria possível se feito ao longo da rota de retorno dos pacotes e durante toda a duração da conexão.

Como os *honeypots* não prestam serviço para nenhuma rede e não são anunciados publicamente, assumimos que todas as mensagens recebidas por eles provém de *spammers*. Toda interação com os *honeypots* é registrada e as mensagens são armazenadas localmente, e, diariamente, copiadas para os servidores centrais do projeto. Nossos *honeypots* nunca encaminham as mensagens recebidas, com exceção daquelas mensagens cujos conteúdos indicam, de acordo com regras pré-definidas[2], que são mensagens de teste utilizadas pelos *spammers* para verificar se *proxies* e *relays* abertos estão funcionando. Neste trabalho analisamos mensagens coletadas entre 01/08/2015 e 31/10/2015.

3.2. Pré-processamento da base de spam

Como possuímos pontos de coleta em diferentes pontos do mundo, nossa base de spam é composta por mensagens dos mais variados idiomas. Porém, como neste trabalho objetivamos analisar o conteúdo das mensagens para identificar aquelas que caracterizam phishing, optamos por selecionar apenas as mensagens em inglês. Cabe ressaltar que nossa metodologia é extensível aos demais idiomas.

Para realizarmos a identificação do idioma das mensagens, extraímos o corpo das mensagens, removendo tags HTML e URLs, bem como desconsiderando o conteúdo dos anexos das mensagens. Uma vez que as mensagens estavam tratadas, pudemos identificar o idioma atráves de uma biblioteca[3] baseada no classificador *Naive Bayes*. A biblioteca utilizada retorna todas as linguagens possíveis para determinado documento. Selecionamos todos aqueles que possuiam inglês dentre os possíveis idiomas. De um total inicial de 1.133.874.435 mensagens, reduzimos o corpo da base de dados para 13.213.796 mensagens em inglês, uma vez que o tráfego de spam é predominantemente composto por mensagens asiáticas. Por fim, o último passo do pré-processamento foi remover *stopwords*, colocar as palavras em caixa baixa, desconsiderar pontuação e eliminar caracteres não UTF-8, de forma a melhorar o desempenho das técnicas de processamento de linguagem natural utilizadas para identificação das mensagens de phishing.

3.3. Identificação das mensagens de phishing

Nosso método para identificação de mensagens de phishing em um conjunto de spams é baseado em técnicas de processamento de linguagem natural. O primeiro passo da

[2]Por exemplo, verificando presença de texto específico no assunto ou corpo da mensagem.

[3]https://pypi.python.org/pypi/langdetect

técnica é a identificação de características chave presentes em mensagens classificadas como phishing. A partir das características propostas por [Aggarwal et al. 2014], identificamos outras capazes de diferenciar phishing das demais mensagens. Tais características foram divididas nas seis categorias descritas a seguir:

Tratamento: Termos utilizados pelos *spammers* para se aproximar e ganhar a confiança da vítima.

Menção monetária: Uma forma comum utilizada por atacantes para convencer usuários a responder seus e-mails é a promessa de dinheiro fácil. Uma vez que a vítima acredita que existe a possibilidade de ganhar dinheiro, ela pode responder ao e-mail enviando suas informações.

Senso de urgência: Os atacantes tentam induzir a vítima a responder a mensagem o mais rápido possível. Uma das razões para isso é que, quando uma pessoa está sob pressão, ou em uma situação de urgência, ela usualmente perde parte de sua razão lógica, tendendo a tomar decisões precipitadas. Além disso, outra razão está no fato de que, o quanto antes a mensagem for respondida, menor a probabilidade da mensagem ou suas URLs terem sido inseridas em alguma *blacklist*.

Pedido de resposta: Outra característica identificada nas mensagens é o pedido de resposta. Como o atacante objetiva obter informações sensíveis do usuário, é necessário que o usuário responda a mensagem (ou acesse alguma URL presente nesta). Portanto, o *phisher* tenta convencer a vítima a responder o e-mail utilizando alguns termos característicos como “reply”, “response”, “answer”, entre outras listadas na Tabela 1.

Formulário: Verificamos que um número grande de mensagens pedem aos usuários para preencher e enviar um formulário com as informações, meios muito pouco utilizados por mensagens de *spam* não-*phishing*.

Segurança: Outra forma de atrair o usuário é mencionar invasão e bloqueio de contas. Como são acontecimentos reais e comuns, esse tipo de abordagem é eficiente.

A seguir apresentamos a nossa metodologia de determinação de phishings e suas respectivas campanhas, que pode ser sintetizada nos 6 passos descritos a seguir:

1. Determinação do conjunto inicial de termos Utilizando os resultados da literatura, identificamos um conjunto inicial de termos que é mostrado na Tabela 1, que é representativo, mas incompleto tendo em vista a evolução das estratégias de phishing e novos contextos que podem ser utilizados, como por exemplo as olimpíadas. Para sermos capazes de evoluir conforme novas abordagens vão sendo usadas pelos *phishers*, utilizamos a técnica Word2Vec [Mikolov et al. 2013] para aumentar o conjunto de palavras utilizado na identificação de *phishing* e torná-lo mais dinâmico, como descrito a seguir.

2. Expansão do conjunto de termos Com base nos vários textos (no nosso caso mensagens), Word2vec identifica palavras que são mais semelhantes e que estão mais relacionadas aos termos de entrada, que são os termos do conjunto inicial. Assim, para cada um dos termos do conjunto inicial, avaliamos a sua saída do Word2Vec e selecionamos os termos associados ao envio de *phishing*. Considerando a base utilizada neste artigo, aumentamos o conjunto de palavras usado na identificação de *phishing* em 156,25%. A

Tabela 1 apresenta os termos adicionados para cada categoria.

Tabela 1. Termos associados a phishing

Categoria	Conjunto Inicial	Termos Adicionados
Tratamento	dear, friend, hello, please	congratulate, valuable, entrusted, congrats, sponsored, nontransferable, expires, regards, authentic, apologize, thank, inconvenience
Menção a dinheiro	bank, money, cash, dollar	credit, customer, funding, purchase, $, transfer, payment, millionaire, profits, accountability, dollars, donate
Pedido de resposta	write, contact, reply, response, foward, send	communication, reapproved, reconfirm, confirming
Urgência	now, today, instantly, straightaway, directly, urgently, urgent, desperately, immediately, soon, shortly, quickly	important
Formulário	form, attach, attached, attachment	information, address, occupation, documentations, subscriber, confidential, zipcode
Segurança	security, violated	detected, correct, authorised, unauthorized, sign, reauthenticate, reliance, spamfiltered, recover, impostors, reactivate, suspects, account, verification

3. Escore da mensagem por categoria O próximo passo da nossa metodologia é estimar a pertinência de cada mensagem a cada categoria de phishing. Essa estimativa é quantificada por um escore baseado no princípio TF-IDF [Baeza-Yates e Ribeiro-Neto 1999] A parte TF do escore deve refletir a ocorrência de termos das categorias, ou seja, quanto mais termos de uma categoria ocorrerem, maior a chance da mensagem pertencer a essa categoria. Para cada mensagem msg e categoria cat, definimos $numtermo_{msg,cat}$ como o número de ocorrências dos termos de cat em msg. Calculamos então $TF_{msg,cat}$ como a razão entre $numtermo_{msg,cat}$ e o maior $numtermo$ que ocorre em msg: $TF_{msg,cat} = numtermo_{msg,cat}/Max_{c \in categoria} numtermo_{msg,c}$. A parte IDF do escore reflete a popularidade das categorias, ou seja, ele é a razão entre o logaritmo do total de mensagens na base ($log(nummsg_*)$) e o número de mensagens assinaladas à categoria cat ($nummsg_{cat}$): $IDF_{cat} = log(nummsg_*)/nummsg_{cat}$. Finalmente, o TF-IDF de uma categoria cat para uma mensagem msg é obtido pela multiplicação das partes.

4. Escore da mensagem O passo seguinte é gerar um escore para cada mensagem que é a soma dos $TF-IDF$ das categorias e a constante α, que quantifica outros aspectos da mensagem que estão comumente associados a phishing. No nosso caso, utilizamos apenas a informação do destinatário para determinar α. Assim, se a mensagem tem apenas um destinatário, α é 1, senão 0, de acordo com a lógica do phishing ser uma mensagem pessoal. Outros critérios como os discutidos na caracterização poderiam ser incorporados ao α. Assim, a equação a seguir determina o escore de cada mensagem msg:

$$ESCORE_{msg} = \alpha + \sum_{cat \in categoria} TFIDF_{cat}$$

5. Classificação de phishing O escore de cada mensagem pode ser utilizado para classificá-la como phishing ou não. Um aspecto importante neste caso é qual o limiar lim que separa os dois grupos com maior acurácia. Para tal, vamos utilizar uma metodologia baseada na curva ROC e na medida AUC [Brown e Davis 2006], que nos permite determinar o valor do escore que maximiza a taxa de acerto numa classificação. Na prática, selecionamos 800 mensagens aleatoriamente da nossa base, considerando um nível de confiança

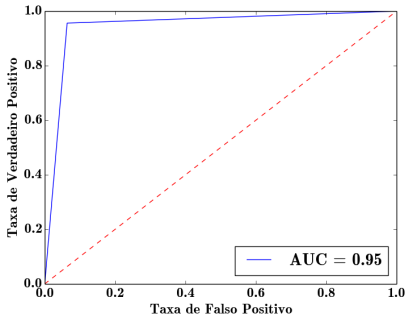


Figura 1. Curva ROC.

de 95% e um erro de $\pm 3,5\%$. A seguir rotulamos manualmente essas mensagens como phishing ou não-phishing. Construímos então a curva ROC (Figura 1) com as mesmas e seus escores e determinamos lim como sendo $1,5$ (ou seja, se $ESCORE_{msg} > lim$, então a mensagem é classificada como *phishing*), o que permite atingir uma AUC de 94,9%.

6. Determinação de Campanhas A partir do conjunto de mensagens classificadas como phishing, agrupamos as mesmas em campanhas (ou seja, grupos de mensagens com a mesma finalidade). O algoritmo utiliza o princípio $TF - IDF$, mas calculado no universo de todos os termos das mensagens. Assim, para cada mensagem msg e termo $termo$, definimos $numtermo_{msg,termo}$ como o número de ocorrências de $termo$ em msg. Calculamos então $TF_{msg,termo}$ como a razão entre $numtermo_{msg,termo}$ e o maior $numtermo$ que ocorre em msg: $TF_{msg,termo} = numtermo_{msg,termo}/Max_{t \in msg} numtermo_{msg,t}$. A parte IDF do escore é o logaritmo da razão entre o total de mensagens na base e o número de mensagens que contem $termo$: $IDF_{termo} = log(nummsg_{*}/nummsg_{termo})$. Finalmente, o $TF - IDF$ de $termo$ para uma mensagem msg é obtido pela multiplicação das partes. Esses valores de $TF - IDF$ compoem um vetor que é comparado posição a posição, com vistas a agrupar mensagens que possuem similaridade maior que 80%. Após desconsiderar as campanhas que possuem menos que 10 emails, encontramos 612 campanhas, que correspondem a mais de 8,5 milhões de mensagens. A Seção 4 apresenta uma caracterização dos resultados da nossa metodologia.

4. Resultados

Nesta seção apresentamos os resultados obtidos ao caracterizar os dados de *phishing* obtidos através da metodologia descrita anteriormente. Inicialmente, mostramos a visão geral dos dados de *phishing*. Em seguida, apresentamos as principais campanhas e as características apresentadas por cada uma delas.

4.1. Visão Geral

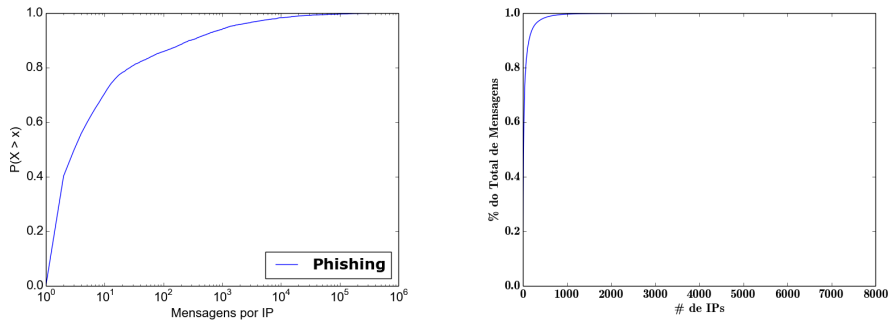
A Tabela 2 oferece uma visão geral das mensagens identificadas como *phishing*. No período de 91 dias, identificamos cerca de 9,7 milhões destas mensagens. Uma característica interessante destes e-mails está na forma como eles são enviados, basicamente através do protocolo SMTP, sendo mais de 99,9% do total das mensagens e mais de 99%

Tabela 2. Visão Geral de *Phishing*

	SMTP(%)	SOCKS(%)	HTTP(%)	Total
Mensagens	9.757.096 (99,94%)	4.550 (0,04%)	807 (0,0%)	9.762.453
Endereços IP	6.651 (99,22%)	52 (0,77%)	4 (0,0%)	6.703
Sistemas Autônomos (AS)	1.701(99,35%)	35 (2,04%)	4 (0,23%)	1.712
***Country Codes* (CC)**	154 (100%)	16 (10,38%)	3 (1,94%)	154

dos endereços IP responsáveis pelo envio das mensagens utilizando o SMTP. Os Sistemas Autônomos relacionados à origem das mensagens apresentam características semelhantes às dos endereços IP. Tal fato advoga em favor da gerencia da porta 25 afinal, desta forma, estas mensagens seriam bloqueadas evitando a propagação de milhões de mensagens de phishing.

O gráfico 2 mostra a distribuição das mensagens pelos endereços IP. Como é possível notar, a maior parte dos endereços enviaram poucas mensagens. Cerca de 60% enviou 10 mensagens ou menos e além disso, como pontuado anteriormente, utilizou-se o protocolo SMTP, característica de máquinas pertencentes à botnets, como mostrado em [Las-Casas et al. 2013]. Outro ponto interessante está na concentração de mensagens em alguns poucos endereços IP. O gráfico 2(b) mostra que 40% destes e-mails estão concentrados em apenas 10 endereços IP e que 90% dos phishing foram enviados de 100 endereços IP distintos. Portanto, uma forma de mitigar grande parte do tráfego de phishing e reduzir possíveis prejuízos causados por estas mensagens é combater estes principais atacantes, responsáveis pela maior parte dos ataques.



(a) CDF das mensagens por endereço IP (b) CDF da concentração das mensagens por IP

Figura 2. Análise do número de mensagens de phishing enviadas por endereços IP distintos.

Considerando estes poucos endereços IP responsáveis por grande parte das mensagens, apresentamos na Tabela 3 os 5 endereços IP que mais enviaram mensagens de phishing. Estes endereços IP possuem características semelhantes quanto ao baixo número de campanhas enviadas. O endereço IP 23.31.87.109, mapeado nos Estados Unidos, foi responsável pelo envio do maior número de mensagens em toda a base analisada. Este atacante enviou mais de 730 mil phishings durante os 3 meses considerados. Um ponto interessante é que, apesar do alto número de mensagens, não há grande variação no tipo de phishing enviado por este atacante, apresentando apenas 6 campanhas distintas. Estas campanhas, apesar de distintas, possuem o objetivo comum de ludibriar a vítima, oferecendo-a dinheiro para, desta forma, conseguir roubar seus dados pessoais.

Os endereços IP 212.227.94.138 e 212.227.255.64 encontram-se no mesmo AS, localizado na Alemanha. Estes endereços possuem uma campanha de phishing em comum (campanha 2, Seção 4.2), levando-nos a crer que são controlados por um mesmo *phisher*. Das mensagens enviadas por estes endereços, 74% e 71%, respectivamente, fazem parte desta campanha. Como será mostrado mais a frente, esta foi a segunda maior campanha encontrada em nossa base de dados e relaciona-se ao acesso a uma conta bancária.

Tabela 3. Top 5 Endereços IP

IP	# de Mensagens	AS	CC	# de Campanhas
23.31.87.109	732.361	7922	US	6
212.227.94.138	524.056	8560	DE	2
65.29.192.68	500.351	10796	US	1
212.227.255.64	384.054	8560	DE	2
186.83.40.72	368.498	10620	CO	2

Com relação à distribuição geográfica de origem, a maior parte das mensagens de *phishing* (mais de 60%) provêm dos Estados Unidos e da Alemanha. Além de ser o país com maior número de mensagens de *phishing*, os Estados Unidos também são aqueles que apresentam o maior número de campanhas distintas (233). Entretanto, estas campanhas possuem um média relativamente baixa de mensagens (15.476), principalmente quando comparado à Alemanha. Este país possui apenas 33 campanhas distintas, mas apresenta a elevada média de 69.945 mensagens por campanha.

Tabela 4. Top 5 Country Codes

CC	# de Mensagens	# de Endereços IP	# de AS's	# de Campanhas
US	3.605.904 (36,93%)	1.406 (20,97%)	302 (17,64%)	233
DE	2.308.181 (23,64%)	175 (2,61%)	53 (3,09%)	33
BA	480.317 (4,92%)	26 (0,38%)	2 (0,11%)	23
CO	388.093 (3,97%)	37 (0,55%)	9 (0,52%)	14
ZA	270.790 (2,77%)	41 (0,61%)	12 (0,70%)	11

4.2. Principais campanhas de *phishing*

Nesta seção apresentaremos as principais campanhas de *phishing* encontradas em nossa base de dados após a separação destas mensagens. Ao todo, foram encontradas 612 campanhas de *phishing*, entretanto, poucas campanhas abrangem a maior parte das mensagens coletadas. Desta forma, descreveremos cada uma das 3 principais campanhas encontradas em nossa base de dados. Estas campanhas foram responsáveis por quase 24% de todas as mensagens classificadas como *phishing*. A Tabela 5 apresenta mais detalhes destas campanhas.

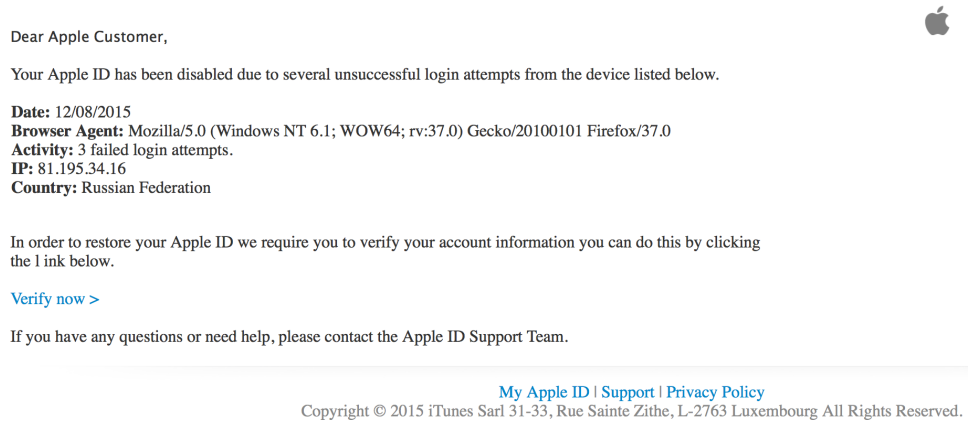
Tabela 5. Top 3 Campanhas

	Características				Categorias					
	Mensagens	**IP**	**AS**	**CC**	**Abordagem**	**Dinheiro**	**Resposta**	**Urgência**	**Formulário**	**Segurança**
C 1	1.124.297	16	5	4	X	X			X	X
C 2	779.359	3	1	1	X			X	X	X
C 3	399.512	30	22	2	X	X	X		X	X

4.2.1. Campanha 1

A maior campanha encontrada em nossa base de dados é composta por 1.124.297 de e-mails, ou seja, 11,52% de todas as mensagens de *phishing* identificadas nos 91 dias de coleta. Como pode ser visto na Figura 3, esta campanha refere-se à um serviço da Apple sendo suspenso e tentando levar a vítima a acessar um link em que será necessário inserir informações a respeito desta conta para reativá-la. Como mostrado na Tabela 5, esta campanha apresenta a catergoria *Abordagem*, uma vez que inicia a mensagem utilizando a saudação *Dear*. As outras categorias presentes são *Dinheiro*, representada pela palavra *Customer*, e *Segurança*, devido as palavras *security* e *account*, além da categoria *Formulário*, pela palavra *information*.

Esta campanha provêm de 16 endereços IP localizados nos seguintes países: Estados Unidos, Alemanha, Colômbia e Grã-Bretanha. Entretanto, apesar dos 16 endereços IP utilizados, 6 destes enviaram baixo número de mensagens, não chagando à 1.000, durante o período analisado. Em contrapartida, o IP 65.29.192.68, listado anteriormente como um dos principais da base, enviou mais de 500 mil *phishing* relacionados a esta campanha.


Dear Apple Customer,

Your Apple ID has been disabled due to several unsuccessful login attempts from the device listed below.

Date: 12/08/2015
Browser Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:37.0) Gecko/20100101 Firefox/37.0
Activity: 3 failed login attempts.
IP: 81.195.34.16
Country: Russian Federation

In order to restore your Apple ID we require you to verify your account information you can do this by clicking the l ink below.

Verify now >

If you have any questions or need help, please contact the Apple ID Support Team.

My Apple ID | Support | Privacy Policy
Copyright © 2015 iTunes Sarl 31-33, Rue Sainte Zithe, L-2763 Luxembourg All Rights Reserved.

Figura 3. Mensagem de exemplo da campanha 1.

4.2.2. Campanha 2

O conteúdo das mensagens da segunda maior campanha, apresentado na Figura 4, refere-se a um banco apresentando um novo processo de autenticação para aumentar a segurança dos seus clientes. Com este pretexto, o atacante induz o usuário a completar informações para utilizar este novo processo, através de um formulário presente acessível por um link presente na mensagem.

Esta campanha também possui alto volume, com aproximadamente 780 mil mensagens, mas que foram enviadas por apenas 3 endereços IP distintos. Estes endereços (212.227.94.138, 212.227.95.8 e 212.227.255.64) se encontram na Alemanha, no AS 8560, e também estão entre os principais endereços IP encontrados em nossa base de dados. Para esta campanha, cada um deles teve uma média de envio de 259 mil *phishings* no período de um mês.

4.2.3. Campanha 3

A terceira campanha de nossa base de dados também refere-se à possível suspensão de uma conta online. No caso do exemplo mostrado, a conta em questão é do banco Wells

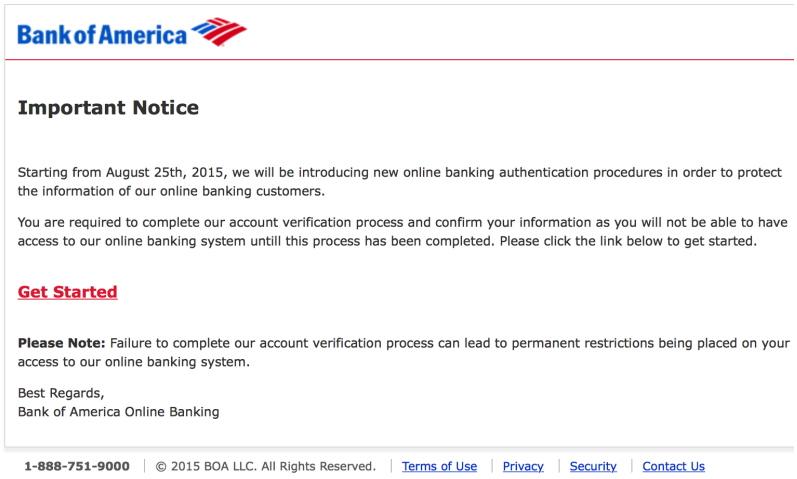


Figura 4. Mensagem de exemplo da campanha 2.

Fargo, entretando, nas demais mensagens desta campanha encontramos referencias a diversos bancos distintos como por exemplo Bank of America e Natwest. Diferente da campanha anterior, que tenta roubar informações do usuário através de um website, acessível por um link presente na mensagem, esta campanha induz o usuário a baixar um documento em anexo, preenchendo-o com informações sensíveis. É possível que o documento em questão represente um *malware*, mas na metodologia do trabalho não cobrimos este tipo de problema. Como trabalho futuro, verificaremos esta questão.

Esta campanha foi enviada por 30 endereços IP distintos presentes nos Estados Unidos e Grã-Bretanha. O fato das mensagens se originarem nestes dois country codes faz sentido, dado que estão tentando abusar de clientes de bancos como Bank of America, localizado nos Estados Unidos e Natwest, originado na Escócia, parte da Grã-Bretanha.

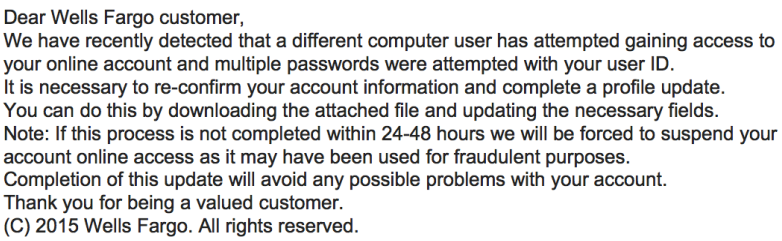


Figura 5. Mensagem de exemplo da campanha 3.

Através destes exemplos, presentes nas principais campanhas identificadas pelo nosso método na base de dados estudada, mostramos a variedade e criatividade dos atacantes na tentativa obter informações sensíveis dos usuários.

4.3. Análise das campanhas de *phishing*

Nesta seção, analisaremos as campanhas de *phishing*. A Tabela 6 apresenta a visão geral das campanhas encontradas. Foram identificadas 612 campanhas distintas. Estas campanhas englobam mais de 8,5 milhões de mensagens, ou seja, mais de 87% de todas as mensagens classificadas como *phishing*. Em média, cada campanha possui 13.984 mensagens, enviadas por apenas 3,16 endereços IP distintos.

Tabela 6. Visão geral das campanhas

Total de Campanhas	612
Média de mensagens	13.984
Média de IPs	3,16
Média de ASes	2,06
Média de CCs	1,82

Como descrito na metodologia, identificamos categorias a serem utilizadas como base para nosso método de detecção. Todas as mensagens identificadas como *phishing* se enquadram em pelo menos uma destas categorias. A Tabela 7 mostra a distribuição das categorias nas campanhas e nas mensagens de *phishing*. A categoria *Abordagem* é aquela aquela que possui maior número de campanhas utilizando-a. Do total de 612 campanhas, 480, com cerca de 7,4 milhões de e-mails, utilizaram alguma forma de abordagem ao usuário. Como dito anteriormente, o objetivo do atacante é ludibriar o usuário e levá-lo a crer que a mensagem é legítima, logo, abordar cordialmente os usuários é um passo fundamental para atingir tal meta. A segunda categoria mais presente nas mensagens da base avaliada é *Dinheiro*. Como descrevemos anteriormente, uma forma fácil de chamar a atenção da vítima é prometendo-a dinheiro fácil. Portanto, os atacantes utilizam muito desta artimanha na tentativa de conseguir informações da vítima. Outra forma utilizada pelos *phishers* para roubar informações dos usuários é fazendo com que ele preencha um formulário com seus dados. Os mais diversos motivos são apresentados pelos atacantes para tentar fazer com que a vítima o faça, como por exemplo a necessidade de atualização de dados cadastrais para que a conta de determinado serviço não seja banida. Com isso, a terceira categoria mais presente foi de *Formulário*. Encontramos 388 campanhas, responsáveis pelo envio de mais de 6 milhões de mensagens apresentando tal comportamento, mostrando que esta é uma categoria importante utilizada nestes tipos de ataque.

Tabela 7. Categoria das campanhas

	Abordagem	**Dinheiro**	**Resposta**	**Urgência**	**Formulário**	**Segurança**	**Total**
Campanhas	480	373	318	236	388	308	612
Mensagens	7.412.701	6.254.321	3.552.518	3.701.086	6.099.531	5.785.346	8.558.237

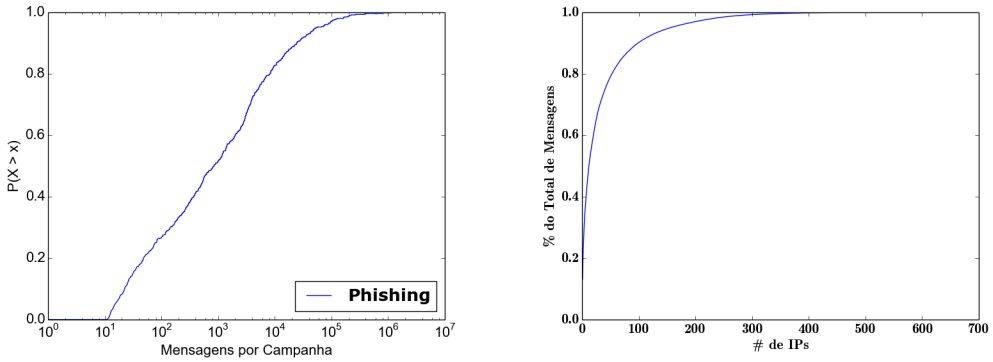
Considerando os conjuntos de categorias às quais as mensagens pertencem, que são 57 de 63 possíveis, podemos afirmar que a nossa metodologia é versátil e capaz de identificar as várias modalidades de phishing. Na Tabela 8, são apresentados os 5 conjuntos que envolveram o maior número de mensagens. O principal conjunto, englobando mais de 1,8 milhão de mensagens, quase 50% do total de *phishing*, envolve *Abordagem*, *Dinheiro*, *Formulário* e *Segurança*. Estas categorias foram encontradas no exemplo da maior campanha encontrada (Figura 3). Além disso, as categorias *Abordagem* e *Formulário* estão presentes em todas estas mensagens, se mostrando características marcantes de mensagens de *phishing*. Uma maneira de evitar estes ataques, como mostrado em [Sheng et al. 2007], é treinar e ensinar aos usuários as diferentes formas de ataque utilizada pelos *phishers*, minimizando a possibilidade de que estes se tornem vítima. Logo, um viés a ser utilizado nas técnicas de ensino aos usuários é apresentar estas características tão importantes na construção das mensagens do *phishing*, como mostrado através dos resultados presentes na Tabela 7.

Quanto ao volume de mensagens das campanhas, podemos observar, através do

Tabela 8. Conjuntos de categorias

Conjuntos de categorias						Características				
Abordagem	**Dinheiro**	**Resposta**	**Urgência**	**Formulário**	**Segurança**	**Mensagens**	**Camp.**	**IP**	**AS**	**CC**
X	X			X	X	1.834.929	33	104	68	31
X	X	X		X	X	839.739	40	94	60	28
X			X	X	X	819.663	8	14	10	14
X	X	X	X	X		555.474	35	85	52	29
X	X	X	X	X	X	511.723	60	172	95	36

gráfico presente na Figura 6(a), que cerca de 50% das campanhas possuem 1.000 mensagens ou menos e 10% das mensagens, ou seja, apenas 61 campanhas tem mais de 24 mil e-mails. Vale ressaltar que, como descrito na metodologia, não consideramos campanhas que possuem 10 mensagens ou menos. Com relação à concentração das mensagens na campanhas, pode ser visto, no gráfico presente na Figura 6(b), que quase 90% das mensagens se encontram nas 100 maiores campanhas. Como dito anteriormente, uma melhoria no combate ao *phishing* é focar os recursos nos principais endereços IP e principais campanhas, mitigando significativamente o tráfego deste tipo de mensagem pela Internet.



(a) CDF do número de mensagens por campanhas.

(b) CDF da concentração de mensagens por campanha

Figura 6. CDF's das campanhas

5. Conclusão e Trabalhos Futuros

Neste trabalho apresentamos um método adaptativo para identificação de mensagens de *phishing* que utiliza categorias bem definidas na literatura e expande-as, através da técnica Word2Vec, para identificar possíveis nuances presentes nas diferentes mensagens de *phishing* ao longo do tempo. Utilizando o método proposto conseguimos identificar, com taxa de acerto de aproximadamente 95%, mais de 9,7 milhões de e-mails de *phishing*, quantidade bastante relevante, principalmente quando comparada à quantidade utilizada por outros trabalhos na literatura.

Além disso, mostramos características do *phishing*, como por exemplo, o fato dele ser enviado quase somente através do protocolo SMTP. Mostramos ainda que alguns poucos endereços de origem, comumente localizados nos Estados Unidos e Alemanha, são responsáveis pela maior parte deste tráfego, e que combatendo-os, tal problema seria significativamente mitigado. Através da geração de campanhas, mostramos que poucas, normalmente ligadas a bancos, respondem por um percentual elevado do total de mensagens

estudados. Como trabalhos futuros, pretendemos aprimorar a técnica de identificação de *phishing*, utilizando outras métricas como a forma de envio. Além disso, pretendemos expandir as análises realizadas, estudando também anexos das mensagens além das URL's que estas possuem.

Agradecimentos

Este trabalho foi parcialmente financiado por NIC.BR Fapemig, CAPES, CNPq, e pelos projetos MCT/CNPq-InWeb (573871/2008-6), FAPEMIG-PRONEX-MASWeb (APQ-01400-14), e H2020-EUB-2015 EUBra-BIGSEA (EU GA 690116, MCT/RNP/CETIC/Brazil 0650/04).

Referências

Aggarwal, S., Kumar, V., e Sudarsan, S. D. (2014). Identification and detection of phishing emails using natural language processing techniques. Em *Proc. of the 7th Int'l Conference on Security of Information and Networks*, New York, USA. ACM.

Baeza-Yates, R. A. e Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Bergholz, A., De Beer, J., Glahn, S., Moens, M.-F., Paaß, G., e Strobel, S. (2010). New filtering approaches for phishing email. *J. Comput. Secur.*

Brown, C. D. e Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80(1):24 – 38.

Fette, I., Sadeh, N., e Tomasic, A. (2007). Learning to detect phishing emails. Em *Proceedings of the 16th International Conference on World Wide Web*, WWW '07.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., e Nunge, E. (2007). Protecting people from phishing: The design and evaluation of an embedded training email system. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Las-Casas, P. H. B., Guedes, D., Jr., W. M., Hoepers, C., Steding-Jessen, K., Chaves, M. H. P., Fonseca, O., Fazzion, E., e Moreira, R. E. A. (2013). Análise do tráfego de spam coletado ao redor do mundo. Em *Anais do simpósio brasileiro de redes de computadores e sistemas distribuídos (SBRC)*. SBC.

Marchal, S., François, J., State, R., e Engel, T. (2014). Phishscore: Hacking phishers' minds. Em *International Conference on Network and Service Management (CNSM)*, p 46–54.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., e Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., e Nunge, E. (2007). Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. Em *Proc. of the 3rd Symp. on Usable Privacy and Security*, SOUPS '07, p 88–99, New York, NY, USA. ACM.

Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., e Zhang, C. (2009). An Empirical Analysis of Phishing Blacklists. Em *Conference on Email and Anti-Spam*.

Steding-jessen, K., Vijaykumar, N. L., e Montes, A. (2007). Using low-interaction honeypots to study the abuse of open proxies to send spam.

Xiang, G., Hong, J., Rose, C. P., e Cranor, L. (2011). Cantina+: A feature-rich machine learning framework for detecting phishing web sites. *ACM Trans. Inf. Syst. Secur.*, 14(2):21:1–21:28.

Zhang, Y., Hong, J. I., e Cranor, L. F. (2007). Cantina: A content-based approach to detecting phishing web sites. Em *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, p 639–648, New York, NY, USA. ACM.