Caracterização e análise do comportamento de destinatários de SPAMs

Thaína Amélia de Oliveira Alves[1], **Humberto T. Marques-Neto**[1]

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

thaina.alves@sga.pucminas.br, humberto@pucminas.br

Abstract. *The e-mail exchange among internet users has a structure like a complex network, which can be studied by the analysis of its characteristics, such as the popularity of their nodes. This paper characterizes users recipients of eletronic messages registered in two datasets. The first dataset was generated by a real filter spam running in a corporative e-mail provider and the second dataset has data about electronic messages exchange with the users of the mentioned provider. Analyzing the popularity of the recipients, we observed that there was a concentration of spamming and e-mails for a restrict group of users. This group was analyzed in order to understand their characteristics and their behavior to improve the management of the e-mail accounts of the provider.*

Resumo. *A troca de mensagens eletrônicas entre usuários tem estrutura semelhante a uma rede complexa, a qual pode ser estudada a partir da análise de suas características, como por exemplo a popularidade de seus nós. Este artigo apresenta a caracterização de usuários destinatários de mensagens eletrônicas presentes em dois datasets. O primeiro foi gerado pelo filtro de spam de um provedor de Internet corporativo brasileiro e o segundo dataset gerado a partir de e-mails legítimos deste mesmo provedor. Através da análise da popularidade dos destinatários, verificou-se uma concentração de envio de spams e de e-mails para um conjunto restrito de usuários do provedor. Este conjunto foi analisado com o propósito de entender as suas características e comportamento, afim de aprimorar o gerenciamento de contas de e-mail pelo provedor de serviços.*

1. Introdução

O desenvolvimento e a popularização da Internet vem crescendo muito nos últimos anos. Neste contexto, percebe-se que o spam se tornou um grande problema de abuso da infraestrutura utilizada para comunicação dos seus usuários, inclusive em provedores de serviço brasileiros. Relatórios mostram que o Brasil está entre os três países que mais disseminam spams na rede [IronPort Email and Web Security 2011]. Mesmo com a existência de mecanismos dedicados ao bloqueio de spams, ainda nota-se a chegada deste tipo de mensagem na caixa de entrada dos destinatários. O relatório do [Message Labs 2011] mostra que em Julho de 2011 a taxa de spams por e-mail trafegado aumentou 4,9%, se comparado ao mês anterior. Enfim, em alguns casos, os spams representam cerca de 80% de todos e-mails de um provedor de serviços [Symantec 2011].

O aumento do tráfego de spams é prejudicial em vários aspectos. Por exemplo, o relatório da [Nucleus Research 2007] evidencia que o spam filtrado pelo próprio

usuário, acarreta uma queda de produtividade em cerca de 1,2% por empregado ao ano, causando, portanto, prejuízos para as empresas. Além disso, essa grande quantidade de spams que circulam na Internet pode afetar também o desempenho dos provedores de e-mail e, ainda, causar desperdício de recursos para o tratamento do tráfego de spam na rede [Dan Twining 2004]. Prejuízos dessa natureza podem ser minimizados com a melhoria dos mecanimos *antispam* e também com um melhor gerenciamento das contas de e-mails por parte dos provedores. Ou seja, é importante analisar as redes de e-mails em duas abordagens: a partir da análise do comportamento dos spammers e a partir da análise do comportamento dos destinatários de spams. Esta última abordagem é proposta neste artigo, o qual apresenta análises até o nível do usuário final.

A interação entre os usuários de e-mails tem estrutura semelhante a uma rede complexa, a qual é definida em [Easley and Jon 2009] como uma rede que possui uma topologia com características específicas, tais como, alto coeficiente de agrupamento e reciprocidade. Além disso, várias características dos nós das redes complexas podem ser modeladas com distribuições de probabilidade de causa pesada. Este trabalho sugere que os remetentes e destinatários sejam estudados como uma rede complexa, com o propósito de se compreender algumas dimensões do comportamento dos usuários de e-mail, principalmente, as implicações da popularidade dos nós dessas redes. De acordo com [Easley and Jon 2009], a popularidade é um fenômeno caracterizado por desequilíbrios extremos. Espera-se, portanto, que poucos destinatários que recebem muitos spams, tenham um comportamento, relacionado ao envio/recepção de e-mails, diferente da maioria dos destinatários que recebem uma menor quantidade de spams. Desta forma, a partir da análise da popularidade, pretende-se entender o comportamento dos destinatários "foco" de spams, diferentemente de outros estudos que buscam avaliar apenas o comportamento dos spammers, ou seja, os remetentes desse tipo de mensagem eletrônica.

Este trabalho caracteriza e analisa o comportamento de destinatários de spams e de e-mails legítimos de um provedor de correio eletrônico. As análises são realizadas com base, principalmente, na popularidade dos nós de duas redes de troca de informações entre remetentes e destinatários de mensagens eletrônicas, reconstituídas a partir de dois *datasets*. O primeiro conjunto de dados foi gerado por um filtro *antispam* de um provedor corporativo de e-mails (rede de spams) e o segundo contém informações dos e-mails legítimos dos usuários deste mesmo provedor (rede de e-mails).

Os autores deste artigo realizaram em [Alves and Marques-Neto 2011] a análise de uma carga de trabalho inicial, contendo apenas o *dataset* de spam de um período de 2 meses. Como principal contribuição adicional ao artigo [Alves and Marques-Neto 2011], este apresenta a análise da evolução temporal dos destinatários de spams. Além disso, a nova metodologia utilizada contempla a análise de uma segunda carga de trabalho, distinta da primeira, a qual possui os usuários de e-mails legítimos do mesmo provedor. Com isso, analisou-se o comportamento dos destinatários de spams em conjunto com o comportamento dos usuários de mensagens legítimas. Desta forma, foi possível verificar a relação entre os usuários dessas duas cargas de trabalho e ainda classificá-los em dois grupos: *Corporativo* e *Pessoal*. Enfim, o uso da metodologia proposta neste artigo busca identificar possíveis destinatários "foco" de recepção de spam. Na visão do provedor, a constatação de usuários que recebem muito spams e e-mails e a identificação do comportamento desses usuários pode auxiliar a detecção, por exemplo, de contas desativadas.

Assim, os resultados deste trabalho pode contribuir com um melhor gerenciamento das contas de usuários de correio eletrônico de um provedor de serviços.

O restante deste trabalho está organizado em mais 5 seções. Os trabalhos relacionados são discutidos na seção 2. Na seção 3 é descrita a metodologia de caracterização. A seção 4 apresenta e analisa os resultados relevantes para este trabalho. Por fim, na seção 5 apresenta-se a conclusão e os possíveis trabalhos futuros.

2. Trabalhos Relacionados

Devido ao fato do spam ter se tornado um dos grandes problemas a ser enfrentado pelos usuários de e-mail, diversas técnicas foram desenvolvidas com a intenção de contê-los. O estudo de [Cook et al. 2006], apresenta técnicas de classificação dos usuários quanto a sua confiabilidade, adicionando em *blacklists* os usuários não confiáveis e em *whitelists* os usuários ou *hosts* considerados confiáveis. Porém, essas técnicas podem ser facilmente contornadas por e-mails com remetentes forjados. Para conter tal problema, alguns trabalhos sugerem estratégias para identificar o emissor, o que pode ser visto em [Ramachandran and Feamster 2006], onde é estudado como os spammers exploram a infraestrutura da rede para enviar suas mensagens, tentando determinar faixas de endereços IP que são mais usadas para se enviar spam. [Guerra et al. 2010] também caracteriza os remetentes de spams, contudo, a partir da análise das listas de destinatários dos spammers. Essas listas são consideradas relevantes, pois, por mais que o spammer altere sua origem na rede, o destinatário das mensagens não pode ser ofuscado.Os resultados mostram que grande parte dos destinatários são alvo de spammers que possuem endereços IPs geograficamente próximos na rede.

Além das estratégias citadas acima, que buscam sempre dificultar o envio de spam, técnicas que propiciam o entendimento do comportamento dos spammers afim de definir estratégias para contê-los a partir de suas características observáveis, também devem ser utilizadas para mitigar a quantidade de spams que circulam na rede. Na busca por diferenciar comportamentos maliciosos daqueles classificados como legítimos, o trabalho de [Gomes et al. 2007] caracteriza uma carga de trabalho de e-mails com base em critérios, tais como, tamanho das mensagens, processo de chegada e localidade temporal de endereços remetentes. Os resultados mostram que o processo de chegada de e-mails pode ser representado por uma distribuição de Pareto e que o tamanho dos e-mails é mais aproximado a uma distribuição Log Normal. Além disso, ao caracterizar também o comportamento dos spammers, identificou-se que o comportamento malicioso e o legítimo se diferenciam em muitos dos aspectos.

Dando continuidade a um trabalho anterior, [Gomes et al. 2009] estende a discussão sobre a caracterização da extensa carga de trabalho de spams apresentada. Neste trabalho, foram derivados modelos matemáticos para representar a taxa de chegada de spams e o tamanho das mensagens, apresentando comparações com cargas de trabalhos legítimas. Foi visto, por exemplo, que enquanto o envio de mensagens legítimas exibe padrões temporais diários e semanais característicos, o envio de spam não exibe nenhuma diferença significativa ao longo do período analisado.

Em [Alves and Marques-Neto 2011] foi caracterizada uma carga de trabalho contendo um *dataset* de spam por um período de 2 meses. Os remetentes e destinatários de spams foram analisados como uma rede complexa, para identificar, principalmente, as

implicações da popularidade dos nós dessas redes no tráfego de e-mails de um provedor de serviços de correio eletrônico. De acordo com [Easley and Jon 2009], a popularidade de uma rede complexa é representa por uma lei de potência [M. E. J. Newman 2006] e é modelada utilizando a distribuição *Zipf-like* ($Prob(acessarumobjetoi) = C/i^{\alpha}$) onde $\alpha > 0$ e C é a constante de normalização [Kluckhohn 1950]. Quando uma função decresce a medida que o valor de α cresce, esta função segue uma lei de potência.

Em extensão ao trabalho de [Alves and Marques-Neto 2011], o artigo aqui apresentado mostra a evolução temporal dos destinatários de spams e ainda contempla a análise de uma segunda carga de trabalho, a qual possui mensagens de usuários de e-mails legítimos. Portanto, este novo trabalho utiliza as teorias citadas acima, juntamente com uma nova metodologia de caracterização do comportamento de usuários de e-mail para verificar possíveis usuários destinatários "foco" de spam e de e-mail. Na visão do administrador de redes, a identificação de tais usuários, é uma informação que pode aprimorar o gerenciamento de contas de usuários de e-mails que estejam prejudicando o provedor.

3. Metodologia de Caracterização

Esta seção apresenta a metodologia de caracterização do comportamento dos destinatários de spams e de e-mail presentes em duas cargas de trabalho reais que foram coletadas de um mesmo provedor corporativo de correio eletrônico. O primeiro conjunto de dados utilizado foi gerado na infraestrutura deste provedor pelo filtro *antispam* denominado "*InterScan Messaging Security Suite 7.0 for Windows*" [Micro 2007]. Este filtro foi desenvolvido pela empresa *Trend Micro* [NSS Labs 2010]. Esta carga de trabalho, nomeada aqui por *Dataset de Spam*, contém o tráfego de e-mails que chegaram ao provedor e foram considerados não legítimos e, portanto, foram bloqueados pelo filtro *antispam*. Os dados foram coletados ao longo de um ano (de julho/2010 a junho/2011) e possui informações tais como: remetentes e destinatários das mensagens, IPs e domínios dos usuários, data e hora.

A segunda carga de trabalho foi gerada também na infraestrutura do mesmo provedor de e-mails e contém todos os e-mails diários enviados e recebidos pelos seus usuários. Este *dataset* foi nomeado aqui como *Dataset de E-mail*, o qual possui os dados do mesmo período do *Dataset de Spam* contendo a mesma estrutura de dados. É válido ressaltar que os dados foram anonimizados pelo administrador do serviço e, portanto, não é possível identificar os usuários na sociedade. Além disso, estes dados estão protegidos por um Acordo de Privacidade que não nos permite aprofundar algumas análises. Devido ao grando volume de dados, as duas cargas de trabalho utilizadas foram organizadas por mês afim de viabilizar as análises aqui realizadas.

Na Seção 2 foi visto que os estudos buscam caracterizar o comportamento de spammers afim de se obter um maior conhecimento para aprimorar técnicas de bloqueio de spams. Neste artigo focamos na análise dos destinatários de spams, pois, conhecer o comportamento desses usuários finais representa um benefício para o gerenciamento do provedor de e-mails. Com isso a análise do *Dataset de Spam* foi considerada sob o ponto de vista dos destinatários. Assim, os registros foram agrupados por cada destinatário distinto, contabilizando a quantidade de spams para cada um deles. Para a análise do *Dataset de E-mail*, os registros foram agrupados pelos remetentes e destinatários distintos, contabilizando a quantidade de e-mails que estes enviaram e/ou receberam.

Inicialmente o *Dataset de Spam* foi analisado com o propósito de se entender a evolução temporal do comportamento dos destinatários de spam durante os 12 meses. Em seguida, os destinatários do *Dataset de E-mail* foram analisados afim de se verificar algumas características dos principais destinatários de spam (identificados na caracterização do primeiro *dataset*) relacionadas ao seu envio/recebimento de mensagens eletrônicas.

3.1. Visão Geral do *Dataset de Spam*

De acordo com a contabilização geral do conteúdo do *Dataset de Spam* no período de 12 meses consecutivos, aproximadamente 25 milhões dos e-mails que chegaram ao provedor foram considerados como não-solicitados e/ou maliciosos e, por isso, foram bloqueados pelo filtro *antispam*. Esses e-mails contêm cerca de 2 milhões de remetentes distintos, os quais estão relacionados a aproximadamente 158 mil domínios.

O provedor de e-mail possui cerca de 5.367 contas de usuários distintos, distribuídos em 25 domínios diferentes. Contabilizando a quantidade de spams que tais usuários receberam durante o período de 12 meses da coleta do *Dataset de Spam*, foi possível verificar a existência de usuários "foco", ou seja, usuários que recebem uma grande quantidade de spams em comparação a média geral. Particularmente, existe um único destinatário que recebeu mais de 100 mil spams em apenas 2 meses. Enquanto isso, outros usuários receberam uma ou duas mensagens classificadas como spam. A Tabela 1 apresenta uma visão geral dos remetentes e destinatários de spam durante os 12 meses.

Tabela 1. Remetentes e destinatários distintos de spam - *Dataset de Spam*

Carga de Trabalho 1 - Dataset de Spam					
Mes/Ano	**Remetentes**		**Destinatários**		**# E-mails**
	# Usuários	**# Domínios**	**# Usuários**	**# Domínios**	
Jul/2010	285.028	25.799	4.129	18	2.404.025
Ago/2010	139.771	20.853	4.070	18	3.358.509
Set/2010	273.020	18.290	4.243	18	2.397.600
Out/2010	138.715	16.200	4.053	18	1.853.430
Nov/2010	236.605	15.640	4.060	18	2.193.334
Dez/2010	260.748	14.982	5.181	19	2.332.453
Jan/2011	293.608	13.609	5.060	20	2.259.342
Fev/2011	103.799	13.335	5.077	22	1.578.987
Mar/2011	92.704	14.475	5.300	19	1.644.344
Abr/2011	104.100	14.187	5.188	20	1.707.125
Mai/2011	100.307	13.405	5.228	25	1.867.562
Jun/2011	89.042	14.113	5.367	17	1.703.342
				Total de E-mails	25.300.053

Para contabilizar os dados entre os remetentes e destinatários de mensagens eletrônicas, os remetentes foram identificados através dos IPs *Senders* utilizados para envio dos spams. Os destinatários foram identificados pelos endereços de e-mails anonimizados contidos no campo *Recipient*.

3.2. Visão Geral do *Dataset de E-mail*

Após a caracterização do *Dataset de Spam*, foi realizada uma contabilização geral do conteúdo do *Dataset de E-mail*. Na Tabela 2 é apresentada uma visão geral dos usuários de e-mail durantes os 12 meses analisados.

Tabela 2. Remetentes e destinatários distintos de E-mail - *Dataset de E-mail*

Carga de Trabalho 2 - Dataset de E-mail					
Mes/Ano	**Remetentes**		**Destinatários**		**# E-mails**
	# Usuários	**# Domínios**	**# Usuários**	**# Domínios**	
Jul/2010	101.159	13.834	58.922	10.488	718.564
Ago/2010	118.000	14.304	83.129	10.705	989.850
Set/2010	65.900	10.305	58.275	6.749	523.107
Out/2010	87.968	11.905	64.089	6.833	756.708
Nov/2010	87.862	11.891	60.346	7.060	545.003
Dez/2010	104.130	11.989	48.687	6.497	497.852
Jan/2011	118.796	12.862	39.718	5.725	637.742
Fev/2011	120.266	13.587	71.018	7.523	958.569
Mar/2011	117.774	14.002	110.227	21.211	1.033.135
Abr/2011	126.273	14.101	73.404	7.942	1.069.526
Mai/2011	144.989	15.578	88.305	11.402	1.299.203
Jun/2011	138.143	15.521	79.571	8.067	1.283.891
				Total de E-mails	9.011.080

Foram enviados e/ou recebidos pelos usuários do provedor aproximadamente 9 milhões de e-mails no período de 1 ano. Esses e-mails contêm cerca de 100 mil usuários remetentes distintos, incluindo os usuários do provedor, os quais estão relacionados a aproximadamente 15 mil domínios diferentes. Quanto aos destinatários dos e-mails deste *dataset*, pôde ser observado a existência de aproximadamente 80 mil usuários destinatários distintos por mês, mostrando que os usuários remetentes desse provedor enviam muitos e-mails para uma grande quantidade de contas diferentes, espalhados em média entre 10 mil domínios distintos. Entretanto, seguindo o objetivo desse trabalho, a análise foi focada somente nos e-mails recebidos pelos usuários do provedor, ou seja, aproximadamente 5 mil usuários distintos.

Após a contabilização geral foi realizada uma análise qualitativa dos dados da rede formada pelos usuários presentes nos *datasets* analisados. A distribuição das probabilidades acumuladas dos usuários foram calculadas para a análise da popularidade dos destinatários da rede de spam, afim de identificar possíveis destinatários "foco" de spam. Além disso, como dito anteriormente, identificou-se no *Dataset de E-mail* os usuários do provedor que mais enviam e recebem e-mails. Em seguida, buscou-se identificar alguma relação entre os usuários que enviam/recebem muitos e-mails e os destinatários "foco" de spam, afim de encontrar alguma característica que justifique essa grande quantidade de spams recebidos no período estudado. Nas próximas seções, serão apresentados os resultados encontrados a partir da aplicação da metodologia proposta neste artigo.

4. Resultados

Esta seção discute os principais resultados encontrados na caracterização dos destinatários de spams e de e-mails. Na próxima seção é apresentada a análise do usuários do *Dataset de Spam*. Em seguida, será apresentada a análise do *Dataset de E-mail* e a relação com os resultados encontrados no *Dataset de Spam*.

4.1. Análise dos Usuários Destinatários de Spam

A análise dos destinatários de spam foi realizada para cada um dos 12 meses disponíveis para análise. Ao contabilizar a quantidade de spams enviados para cada usuário final, encontrou-se foco de destinatários em cada mês. Entre os meses de Setembro e Novembro de 2010, 13% dos destinatários foram alvos de 50% do total de spams, reduzindo para 11% de Dezembro de 2010 a Janeiro de 2011. No período de Fevereiro de 2011 a Junho de 2011, entre 7% e 8% dos usuários receberam 50% dos spams representando uma maior concentração do envio de spams em menor quantidade de destinatários.

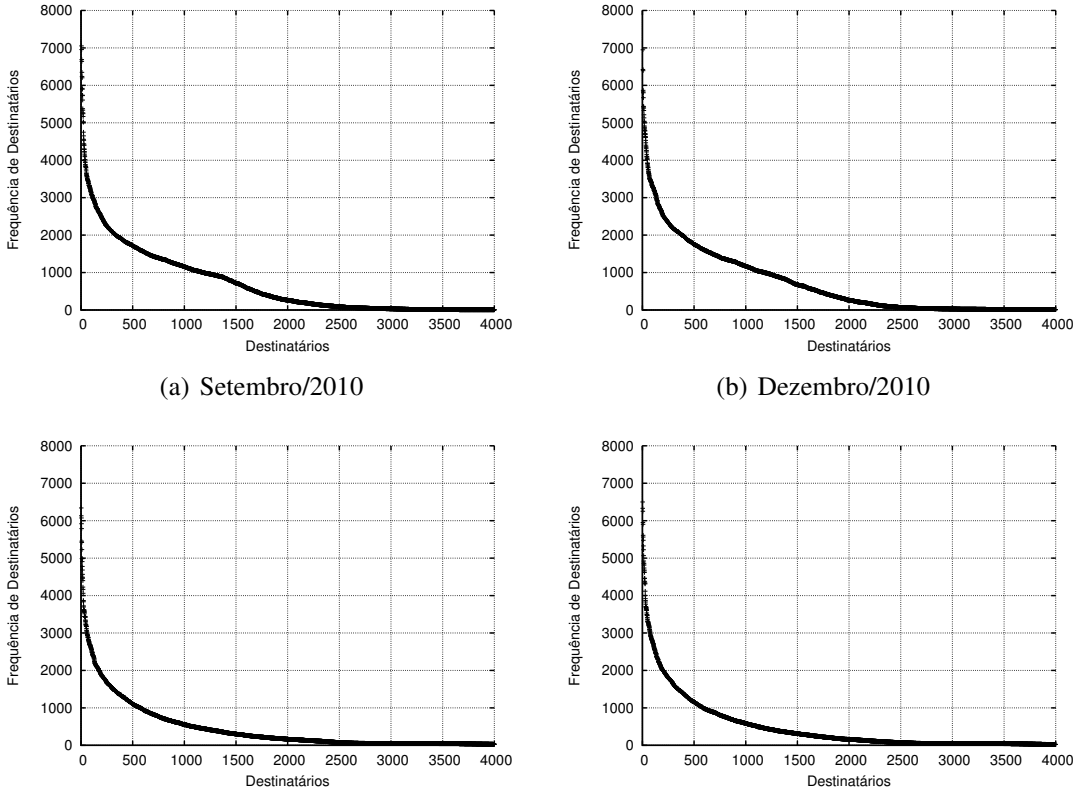
Tais valores podem ser visualizados na Tabela 3 que apresenta uma visão geral dos destinatários de spams durante esse período. Os valores mostram uma queda ocorrendo dos meses de Setembro e Dezembro de 2010 comparados aos meses de Março a Junho de 2011. Porém, tais valores podem ser justificados devido aos meses de 2010 possuir menor quantidade de usuários destinatários que os meses de 2011 e, desta forma, os spams se concentram em um número menor de destinatários. Pode ser observado ainda na Tabela 3 que, apesar de ter uma quantidade de destinatários maior como nos meses de 2011, o percentual dos usuários que recebem pelo menos 50% dos spams apresenta um valor muito baixo, não sendo superior a 13% do total de destinatários.

Tabela 3. Visão geral dos destinatários de spam durante os meses

Meses	**Total *Recipients***	**# *Heavy Recipients* (recebem 50% spams)**	**% *Recipients* recebem 50% dos spams**
Jul/2010	4.129	525	12.7%
Ago/2010	4.070	510	12.5%
Set/2010	4.243	560	13.0%
Out/2010	4.053	530	13.0%
Nov/2010	4.060	554	13.0%
Dez/2010	5.181	560	10.8%
Jan/2011	5.060	545	10.8%
Fev/2011	5.077	450	8.9%
Mar/2011	5.379	422	8.2%
Abr/2011	5.188	449	8.7%
Mai/2011	5.228	410	7.8%
Jun/2011	5.367	424	7,9%

Para a análise da popularidade dos destinatários da rede formada pelo *Dataset de Spam* foram calculadas as PDFs dos usuários separados por meses. Devido a limitação de espaço, a Figura 1 apresenta a disposição das curvas das PDFs dos destinatários de spam dos meses com intervalo de 2 meses entre um e outro.

Pode ser visualizado nas PDFs um comportamento típico de uma lei de potência [M. E. J. Newman 2006]. Ou seja, poucos destinatários recebem muitos spams, enquanto a maioria dos usuários finais recebem poucas mensagens não-solicitadas e/ou maliciosas. Em outras palavras, poucos nós destinatários da rede têm alta popularidade entre o total de usuários alvos. Observa-se que a curva da distribuição está sempre bem acentuada e próxima ao eixo y, mostrando que os destinatários de spams se concentram em poucos usuários, mais precisamente menos que 500 usuários do provedor. Com isso, conclui-se



(a) Setembro/2010 (b) Dezembro/2010

(c) Março/2011 (d) Junho/2011

Figura 1. PDFs da popularidade dos destinatários em alguns meses analisados

que a recepção de spam de fato se concentra em poucos destinatários, mostrando uma maior popularidade de certos usuários alvos do provedor. Estes usuários destinatários "foco" de spam foram denominados aqui como *Heavy Spam Recipients*.

Analisando pela visão do administrador de redes, a descoberta da concentração de envio de spam para poucos destinatários pode ser útil na identificação de usuários que atraem spams. Ao serem identificados e tratados devidamente, poderia facilitar o serviço do provedor de e-mails no bloqueio aos spams que tais usuários iriam atrair para suas caixas de entrada. Como consequência, seria possível alcançar um melhor desempenho nos serviços de e-mails, não congestionando a chegada de mensagens válidas [Dan Twining 2004].

4.1.1. Eliminação dos *Heavy Spam Recipients*

Esta seção analisa a retirada dos *Heavy Spam Recipients* encontrados na seção anterior. Foram eliminados os usuários responsáveis por 50% dos spams recebidos e verificou-se o impacto da ausência desses usuários destinatários para cada um dos 4 meses selecionados anteriormente.

Uma métrica comum utilizada para comparar redes é o expoente α obtido através da regressão linear de uma distribuição de lei de potência [Benevenuto 2011]. Tal métrica foi utilizada para analisar as redes complexas após a retirada dos *Heavy Spam Recipients*.

Aplicou-se a regressão linear para calcular o expoente da distribuição antes e depois da eliminação dos usuários foco de spam. A Tabela 4 apresenta os valores dos expoentes α para as distribuições em cada mês analisado.

Tabela 4. Variação do α após eliminação dos *Heavy Spam Recipients*

Meses	Valor do expoente α	
	Todos Destinatários de Spam	**Após remoção dos *Heavy Recipients***
Set/2010	1.77	1.66
Dez/2010	1.36	1.32
Mar/2011	1,28	0.86
Jun/2011	1.25	0.76

Observando os valores da variação dos expoentes α ao retirar os *Heavy Spam Recipients*, foi possível verificar que os valores do expoente das leis de potência diminuem em todas os meses analisados. Em Março/2011 e Junho/2011 houve uma maior queda no valor do expoente α. Tal fato ocorreu pois esses meses possuem uma quantidade maior de destinatários "foco" de spam. Sendo assim, a retirada desses usuários interfere mais na curva da distribuição e, consequentemente, no valor do α. Como pode ser visto em [Clauset et al. 2009], quanto menor for o valor do expoente α da lei de potência em questão, menos forte é o decaimento da curva e mais grossa será a cauda da distribuição.

Portanto, ao retirar os usuários que recebem grande quantidade de spams (10% dos usuários) alterou-se a curva da distribuição da lei de potência suavizando seu decaimento. Isso mostra que ao entender o comportamento de poucos usuários destinatários de spam pode-se obter um ganho no gerenciamento de segurança da rede, uma vez que estes usuários "foco" são responsáveis por uma grande quantidade de spams recebidos pelo provedor de correio eletrônico.

4.2. Análise dos Usuários Destinatários de E-mail

A análise dos usuários de e-mail foi realizada com intuito de caracterizar a utilização de mensagens eletrônicas pelos usuários do provedor em estudo. A caracterização foi realizada separada por mês durante o período de Julho/2010 a Junho/2011. Ao contabilizar a quantidade de e-mails recebidos verificou-se uma concentração da recepção de mensagens eletrônicas em poucos destinatários. Como os dados são anonimizados e o objetivo deste trabalho é caracterizar as contas de usuários pertencentes a esse provedor de e-mails, solicitamos ao administrador de rede para identificar quais dos domínios presentes no *Dataset* são gerenciados pelo provedor. Por exemplo, enviamos a lista de domínios anonimizados e dentre esses domínios o administrador de redes informou quais são referentes aos 20 domínios pertencentes ao provedor de e-mails em estudo. Com tal informação, foi possível classificar os usuários como pertencentes ou não ao provedor de e-mails a partir da identificação dos domínios.

Em seguida, foram selecionados os usuários que recebem 50% do total de e-mails. Foi verificado, que o domínio das contas desses usuários que recebem metade do total de e-mails são todos pertencentes ao provedor. Além disso, foi possível visualizar a concentração da recepção de e-mails em poucos usuários, pois dentre as 5.000 contas de e-mail, ao selecionar apenas os usuários responsáveis por receber metade das mensagens eletrô-

nicas, obteve-se um número baixo de usuários em cada um dos meses analisados. Na Figura 2 pode-se visualizar tais valores.

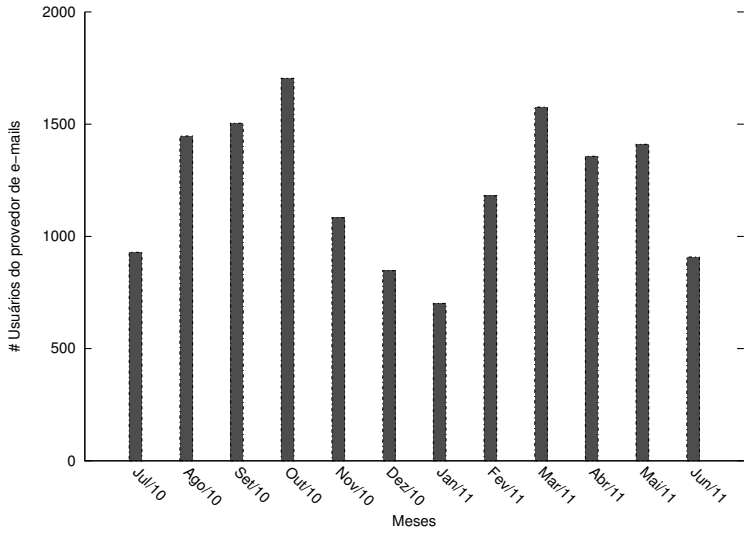


Figura 2. # de usuários do provedor que receberam 50% dos e-mails

Pode ser visto que a quantidade de usuários que recebem metade dos e-mails do provedor sempre se mantém abaixo de 2 mil usuários, mostrando que menos da metade dos destinatários do provedor recebem 50% das mensagens eletrônicas. Para melhor visualização, a Tabela 5 apresenta os percentuais de tais usuários.

Tabela 5. Visão geral dos destinatários dos e-mails durante os 12 meses

Meses	**Total *Destinatários***	**# *Heavy Recipients* (receberam 50% dos e-mails)**	**% Destinatários 50% dos e-mails**
Jul/10	4.420	929	21%
Ago/10	4.235	1.447	34%
Set/10	4.205	1.504	35%
Out/10	4.134	1.704	41%
Nov/10	4.223	1.083	25%
Dez/10	5.203	847	16%
Jan/11	5.195	701	13%
Fev/11	5.413	1.181	22%
Mar/11	5.225	1.576	30%
Abr/11	5.303	1.357	25%
Mai/11	5.386	1.410	26%
Jun/11	5.298	908	17%

Pode ser observado uma variação do percentual de destinatários que recebem 50% das mensagens eletrônicas. Por exemplo, em Outubro de 2010 houve uma quantidade muito maior de usuários comparado ao mês de Dezembro de 2010. Entretanto, durante os 12 meses em média metade dos e-mails são recebidos por 26% dos usuários, tal valor mostra a grande quantidade de e-mails recebidos por um percentual baixo de destinatários. Com tais dados é possível concluir que para o *Dataset de E-mails*, assim como o *Dataset*

de Spam analisado anteriormente, existe um conjunto de usuários "foco" de e-mail. Este conjunto de usuários foram nomeados aqui como *Heavy E-mail Recipients*.

4.2.1. *Heavy Spam Recipients* X *Heavy E-mail Recipients*

As caracterizações realizadas até o momento permitem concluir que existem usuários "foco" de spams, como apresentado nas análises dos *Datasets de Spam* e *E-mail*. Nesta seção busca-se analisar se os usuários destinatários do provedor de e-mails que foram considerados "foco" de spam (*Heavy Spam Recipients*) estão presente entre os usuários classificados como foco de e-mails legítimos (*Heavy E-mail Recipients*).

A análise da relação desses usuários foi realizada para cada mês no período de Julho de 2010 a Junho de 2011. Foram contruídas as listas dos *Heavy Spam Recipients* e dos *Heavy E-mail Recipients* separadas para cada mês. Em seguida, verificou-se a presença de cada usuário na lista de *Heavy Spam Recipients* ou a sua ausência na lista de *Heavy E-mail Recipients*. Na Figura 3 é apresentada a relação dos usuários dessas listas para cada um dos 12 meses analisados.

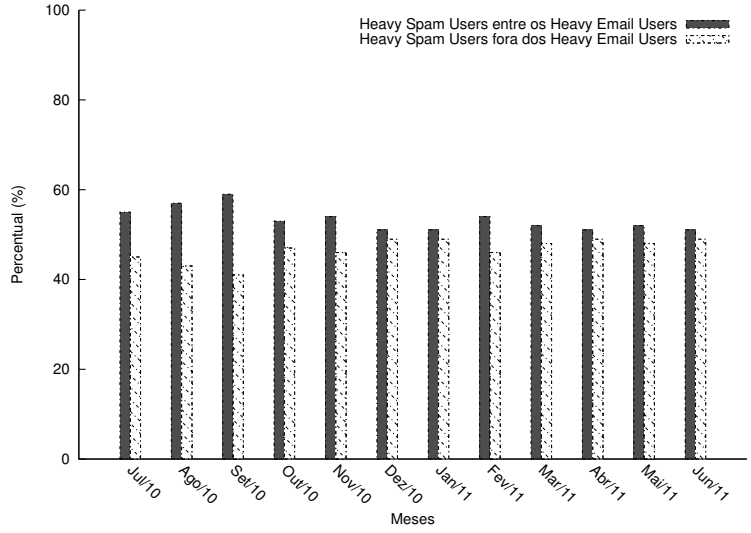


Figura 3. Relação entre *Heavy Spam Recipients* e *Heavy E-mail Recipients*

As barras preenchidas representam os usuários destinatários "foco" de spam que estão presentes na lista dos usuários destinatários "foco" de e-mail. As barras tracejadas representem os usuários destinatários "foco" de spam que estão fora da lista dos *Heavy E-mail Recipients*. Logo, para todos os meses analisados pode ser visto que as barras preenchidas são sempre maiores que as barras tracejadas, representando que a maioria dos *Heavy Spam Recipients* estão entre os *Heavy E-mail Recipients*. Tal informação tem grande importância, uma vez que um usuário que recebe uma grande quantidade de spams e ao mesmo tempo possui grande fluxo de e-mails pode representar um usuário que prejudique os serviços de e-mail a partir do desperdício de recursos para o tratamento de spams e processamento de mensagens eletrônicas que talvez não seja relevante para o usuário final.

4.2.2. Indentificação do Ranking dos *Heavy Spam Recipients*

O objetivo desta subseção é identificar os usuários que sempre se mantiveram no topo das listas mensais *Heavy Spam Recipients* durante o período analisado, ou seja, deseja-se ordenar os usuários que receberam maior quantidade de spams durantes os 12 meses. Para isso, utilizou-se um algoritmo que calcula o peso das posições dos usuários em cada uma das 12 listas. Os pesos são calculados através da soma do inverso das posições dos usuários em cada lista. Por exemplo, um usuário X que ficou em segundo lugar na lista de Julho/2010 e em quarto lugar na lista de Agosto/2010 tem o peso nesses 2 meses de: $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$. Já outro usuário Y que ficou em quinto lugar em Julho/2010 e em vigésimo lugar em Agosto/2010 teria um peso menor ($\frac{1}{5} + \frac{1}{20} = \frac{5}{20} = \frac{1}{4}$) e, portanto, ficaria atrás do usuário X, por este ter permanecido em posições maiores nas listas.

Como resultado do processamento desse algoritmo, possuímos uma única lista dos *Heavy Spam Recipients* para os 12 meses analisados. Lembrando que grande parte destes usuários também são usuários da lista de *Heavy E-mail Recipients*, para o provedor de e-mail a identificação de tais usuários e tratamento deles pode trazer muitos benefícios em questão a desempenho e tempo gasto para processamento de mensagens eletrônicas. Como os dados estão anonimizados, solicitamos ao administrador de rede uma classificação quanto a utilização das contas de e-mails como conta Corporativa, conta Pessoal ou conta inutilizada.

No período dos 12 meses estudados, totalizou-se 702 usuários pertencentes ao grupo dos *Heavy Spam Recipients*, dos quais, foram contabilizados 450 usuários na categoria Pessoal e 252 usuários na categoria Corporativo. Durante a classificação do formulário utilizado, o administrador de rede fez observações sobre a existência de contas inutilizadas, pois, pertencem a pessoas já falecidas ou pessoas que já não trabalham mais na organização que utiliza o provedor de e-mails que gerou os dados desta pesquisa. Para evidenciar que estas contas são inutilizadas, foi analisada no *Dataset de E-mails* como é a utilização dos e-mails legítimos por tais usuários e verificou-se que esses apenas recebem e-mails, não havendo e-mails enviados desses usuários no período estudado. Logo, além dessas contas receberem muito mais spams do que os demais usuários, pois, estão entre os *Heavy Spam Recipients* elas também recebem muito e-mails. Tal fato pode contribuir com o congestionamento das mensagens eletrônicas e piora do desempenho do provedor.

Quanto as contas corporativas e de uso pessoal, verificou-se que apesar do menor número de contas corporativas, estas estão presentes em maior quantidade no topo da lista de *Heavy Spam Recipients*, ou seja, recebem muito mais spams que os demais usuários. As contas classificadas como corporativas geralmente estão vinculadas a um meio de comunicação ou divulgação. Tal fato, justifica esta vulnerabilidade quanto aos spams, pois, são contas mais populares e expostas em meios públicos de fácil acesso por spammers.

Logo, com os resultados apresentados acima, foi possível concluir que existem contas que não são mais relevantes para os usuários que as possui e outras até são totalmente inutilizadas. Através da metodologia apresentada neste trabalho é possível identificar tais usuários que podem causar problemas ao provedor e, portanto, devem ser devidamente tratados.

A Seção 2 mostra que uma das técnicas *antispam* utilizadas são as *Blacklists* de

spams. Seguindo este raciocínio, esses usuários identificados como prejudiciais ao provedor de e-mails poderiam formar uma *Blacklist* de usuários destinatários do provedor. Esta lista poderia ser baseada na classificação dos usuários a partir da sua utilização de e-mails legítimos e recepção de spams como realizado nesta pesquisa.

5. Conclusões e Trabalhos Futuros

Neste artigo foi aplicada uma metodologia de caracterização de duas redes de remetentes e destinatários de um provedor de mensagens eletrônicas, sendo a primeira rede de spams e a segunda de e-mails legítimos. Com a análise da métrica de popularidade, verificou-se a existência de um conjunto restrito de destinatários que recebem muito mais spams do que os demais usuários da rede. Estes usuários foram classificados como *Heavy Spam Recipients*. Quanto aos destinatários de e-mails legítimos, também foram identificados usuários com maior popularidade dentre os demais, os quais foram classificados como *Heavy E-mail Recipients*.

Após esta classificação, foi possível verificar que a maior parte dos usuários presentes no grupo *Heavy Spam Recipients* fazem parte do grupo de *Heavy E-mail Recipients*. Portanto, um destinatário que recebe muitos spams e ao mesmo tempo possui grande fluxo de e-mails pode representar um usuário que prejudique os serviços de e-mail a partir do desperdício de recursos para o tratamento de spams e processamento de e-mails que talvez não seja relevante para o usuário. Para verificar tal suposição, foram identificadas as contas de destinatários de spam que permaneceram sempre no topo da lista dos *Heavy Spam Recipients* e então foi criado um *ranking* dos usuários que receberam mais spams durantes os 12 meses. Esses usuários foram classificados com o auxílio do administrador de redes do provedor de e-mail em estudo. Verificou-se então que as contas classificadas como corporativas, que geralmente estão vinculadas a meio de comunicação ou divulgação, estão presente em maior quantidade no topo da lista de *Heavy Spam Recipients*. Além disso, verificou-se a existência de contas inutilizadas que só recebem spams e e-mails não sendo mais de uso dos usuários do provedor. A aplicação desta metodologia, pode contribur para um melhor gerenciamento de um provedor de e-mails.

Considera-se como principais contribuições deste artigo a proposição da metodologia de identificação das contas de usuários que são prejudicias para o bom funcionamento do provedor, bem como a sugestão de criação de uma *Blacklist* de destinatários do provedor. Até onde se tem conhecimento, a formação dessas listas focando nos usuários destinatário ainda não havia sido sugerido em trabalhos. Como trabalho futuro pretende-se analisar os impacto da implementação da *Blacklist* contendo os usuários destinatários do provedor que são identificados com a metodologia apresentada neste trabalho.

Referências

Alves, T. and Marques-Neto, H. (2011). Caracterização e análise de redes de remetentes e destinatários de spams. In *X Workshop em Desempenho de Sistemas Computacionais e de Comunicação (WPerformance)*. Congresso da Sociedade Brasileira de Computação (CSBC).

Benevenuto, F. (2011). Redes sociais online: Técnicas de coleta, abordagens de medição e desafios futuros.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703.

Cook, D., Hartnett, J., Manderson, K., M, K., and Scanlan, J. (2006). Catching spam before it arrives: Domain specific dynamic blacklists. In *Proceedings of the Australasian Information Security Workshop (Network Security*, pages 193–202.

Dan Twining, Matthew M. Williamson, M. J. F. M. M. R. (2004). Email prioritization: reducing delays on legetimate mail caused by junk mail. In *Distributed Computing Systems 2009 ICDCS '09 IEEE International Conference*.

Easley, D. and Jon, K. (2009). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 1nd edition.

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

Gomes, L. H., Cazita, C., Almeida, J. M., Almeida, V., and Meira, J. W. (2007). Workload models of spam and legitimate e-mails. *Perform. Eval.*, 64(7-8):690–714.

Guerra, P. H. C., Ribeiro, M. T., Guedes, D., Jr., W. M., Hoepers, C., Steding-Jessen, K., and Chaves, M. H. (2010). Identificação e caracterização de spammers a partir de listas de destinatários. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Gramado, RS.

IronPort Email and Web Security (2011). Cisco 2010 annual security report.

Kluckhohn, C. (1950). Human behavior and the principle of least effort. george kingsley zipf. *American Anthropologist*, 52(2):268–270.

M. E. J. Newman (2006). Power laws, Pareto distributions and Zipfs law. page 28. Department of Physics and Center for the Study of Complex Systems.

Message Labs (2011). Message labs intelligence.

Micro, T. (2007). Interscan messaging security suite.

NSS Labs (2010). Consumer anti-malware products group test report.

Nucleus Research (2007). Spam, the repeat offender.

Ramachandran, A. and Feamster, N. (2006). Understanding the network-level behavior of spammers. *SIGCOMM Computing Communitcation Review*, 36:291–302.

Symantec (2011). State of spam e phishing monthly report.