Caracterização do Encadeamento de Conexões para Envio de Spams

Pedro H. Calais Guerra[1], Dorgival Olavo Guedes[1], Wagner Meira Jr.[1]
Cristine Hoepers[2], Klaus Steding-Jessen[2], Marcelo H. P. C. Chaves[2]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte, MG.

[2]CERT.br - Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil
NIC.br - Núcleo de Informação e Coordenação do Ponto br, São Paulo, SP

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

{cristine,jessen,mhp}@cert.br

Abstract. *In this work, we show how spammers exploit open proxies on the Brazilian Internet infrastructure and then chain connections to open relays, bots and other open proxies before delivering spams to the recipients. Our conclusion was based on the analysis of HTTP connections established by spammers to low-interaction honeypots. Although these behaviors are known to security specialists, there are no scientific works that identify and measure such behaviors. Knowing how spammers chain machines in order to send spams may impact the design of reputation-based anti-spam techniques and brings attention to the fact that, although botnets are the most common way to deliver spams nowadays, fighting open proxies is still a need.*

Resumo. *Neste trabalho, mostramos que* spammers *exploram* proxies *abertos na Internet brasileira e, em seguida, encadeiam abusos a* relays *abertos, máquinas de usuários finais que fazem parte de* botnets *e outros* proxies *abertos, antes de entregar as mensagens aos destinatários. Essa conclusão se baseou na análise de conexões HTTP estabelecidas por* spammers *a* honeypots *de baixa interatividade. Embora esses comportamentos sejam conhecidos por especialistas em segurança, não existem trabalhos científicos que identificam e quantificam esses comportamentos. O conhecimento da forma como* spammers *encadeiam máquinas para enviar* spams *pode impactar o projeto de técnicas anti-spam baseadas em reputação de nós e atenta para o fato de que* proxies *abertos ainda precisam ser combatidos, mesmo com a proliferação das* botnets.

1. Introdução

Simultaneamente ao desenvolvimento e popularização da Internet, o *spam* se tornou um dos maiores problemas de abuso da infraestrutura de redes da atualidade [Hayes 2003, Messaging Anti-Abuse Working Group (MAAWG) 2007]. Alguns provedores de serviços de Internet reportam que entre 40% e 80% das mensagens recebidas por seus servidores são *spams* [Whitworth and Whitworth 2004]. Outros estudos [Sipior et al. 2004] avaliam em vários bilhões de dólares o prejuízo que o *spam* acarreta às empresas e à sociedade em geral.

Diversas técnicas para combate ao *spam* têm sido desenvolvidas e aprimoradas, como o uso de *blacklists*, filtros de conteúdo de mensagens [SpamAssassin 2008] e sistemas baseados em reputação de servidores SMTP [Prakash and O'Donnell 2005]. Mesmo com a implementação de tais técnicas, é necessário um esforço contínuo para entender como *spammers* geram, distribuem e disseminam suas mensagens pela Internet, dada a natureza evolutiva do *spam*. Essa evolução acontece tanto na forma como os *spammers* constroem o conteúdo das mensagens [Pu and Webb 2006] quanto no modo como disseminam suas mensagens pela rede, buscando maximizar o volume de mensagens que enviam enquanto mantêm sua identidade oculta.

Para o *spam* ser entregue ao destino, ele deve ser entregue a um servidor SMTP real, que inclua a mensagem no fluxo de correio eletrônico normal. Inicialmente, *spammers* enviavam suas mensagens diretamente ao servidor SMTP responsável pela caixa-postal das vítimas. Essa estratégia foi logo abandonada, pois todo servidor SMTP real registra o endereço IP de origem de cada mensagem. Com base nessa informação, a origem dos abusos poderia ser identificada e bloqueada, ao mesmo tempo que o responsável poderia ser identificado e sofrer penalidades. Por esse motivo, *spammers* passaram, então, a buscar formas de encadear conexões pela Internet antes de alcançar o servidor SMTP de destino, a fim de evitar que sua origem real fosse registrada pelo sistema. Consideramos como encadeamento de conexões o **abuso de duas ou mais máquinas, em sequência, antes que a mensagem seja entregue ao servidor SMTP final da mensagem**. Uma das primeiras formas identificadas para fazê-lo explora uma característica original do protocolo: uma das primeiras formas identificadas para fazê-lo explora uma característica prevista no protocolo, que é a capacidade de fazer o repasse (ou *relay*) de mensagens. Em meados da década de 90, era comum encontrar servidores de correio configurados como *relays* abertos (*open mail relays*), programados para repassar em direção ao destino qualquer mensagem a eles entregues, independente das localizações do emissor e do destinatário. Os *spammers* conseguiam, dessa forma, esconder sua origem real atrás de, pelo menos, mais um servidor SMTP na cadeia.

A resposta dos grupos de combate ao *spam* foi instruir os administradores de rede a reconfigurar os servidores de correio para não agirem como *relays* abertos e também publicar listas de *relays* abertos conhecidos (*Blacklists*). Entretanto, configurações padrão de alguns sistemas ainda trazem essa opção configurada. Mais ainda, vários tipos de *malware* implementam essa funcionalidade quando se instalam em um computador, bem como aqueles que transformam as máquinas invadidas em *bots*, que formam redes de computadores infectados (*botnets*). As *botnets* podem ser usadas em atividades de negação de serviço, esquemas de fraude e envio de *spam*. Por esse motivo, novas máquinas com a funcionalidade de *relays* surgem frequentemente na Internet, criando um sistema distribuído que pode ser usado para disseminar *spam*.

Apesar do *relay* permitir o encadeamento de máquinas, no caso de servidores SMTP reais o endereço de origem da conexão continua sendo registrado nas mensagens que são entregues, podendo levar ao computador de origem do *spam*. Para evitar que isso ocorra, *spammers* passaram a utilizar também diferentes máquinas na Internet que oferecem algum tipo de serviço de *proxy* aberto, via protocolos HTTP e SOCKS. Esses servidores, muitas vezes mal-configurados, aceitam comandos de quem a eles se conecta para que estabeleçam uma conexão a uma outra máquina, repassando então todos os co-

mandos da conexão original para a nova. Dessa forma a identidade da máquina original não é percebida pelo servidor SMTP que recebe a mensagem, que registrará apenas o endereço IP da máquina que executava o serviço que agiu como *proxy*.

Na contínua busca por técnicas de disfarce da origem real das campanhas de *spam*, o nível de sofisticação dos *spammers* vem crescendo, na prática combinando as técnicas descritas: uma conexão inicial a um *proxy* aberto, que pode levar a uma cadeia de *proxies* e possivelmente uma máquina agindo como *relay* SMTP aberto. Apesar desses comportamentos já terem sido registrados informalmente em publicações *online* e listas de discussão, até hoje a caracterização científica desse encadeamento é limitada.

Neste artigo, caracterizamos e quantificamos o comportamento em termos das técnicas utilizadas por *spammers* para entregar suas mensagens pela rede. Para tal, utilizamos *honeypots* de baixa interatividade, máquinas configuradas de modo a simular computadores que atuam como *proxies* e *relays* abertos [Steding-Jessen et al. 2008]. Dessa forma, pudemos observar as origens das conexões aos *honeypots*, os próximos passos tentados no processo de encadeamento e as mensagens enviadas. Nossos resultados indicam que *spammers* abusam *proxies* e *relays* abertos na Internet brasileira e a partir deles encaminham as mensagens de quatro formas distintas:

1. entrega através de *proxies* a servidores de correio final, aqueles responsáveis pelas caixas de correio de um certo domínio de *e-mail*, alvo do *spam*; aquele que é o MX para um certo domínio de *e-mail*.
2. encadeamento de *proxies* com *relays* abertos, onde os *spams* são entregues por SMTP a um servidor de correio real, com seu domínio próprio que, entretanto, recebe correio endereçado a outros domínios que não o seu;
3. encadeamento de *proxies* com máquinas da rede que não são servidores SMTP verdadeiros, mas que possuem instalado algum software para se comportarem como servidores de correio, com vistas a serem exploradas explicitamente pelo *spammer*;
4. encadeamento de *proxies* abertos, quando o *spammer* abusa dois ou mais *proxies* abertos em sequência.

Em todos esses casos, a identificação do *spammer* se torna quase impossível do ponto de vista do destinatário, uma vez que pelo menos um *proxy* aberto foi usado para ocultar a origem real do *spam*. A Figura 1 ilustra graficamente cada uma dessas situações. É importante ressaltar que, em teoria, o encadeamento de *N proxies* abertos pode ser combinado com a entrega da mensagem a um *relay* aberto ou *bot*, no último passo antes da entrega ao servidor final.

2. Trabalhos Relacionados

Existem diversos trabalhos que caracterizam estratégias de disseminação de *spams* sob o ponto de vista do abuso dos recursos de rede. Em geral, esses trabalhos coletam dados a partir de infraestruturas desenvolvidas para capturar uma forma específica de disseminação de *spams* e portanto focam em apenas uma etapa específica do caminho pelo qual as mensagen trafegam, como *botnets* [Lee et al. 2007], abuso a *relays* abertos [Pathak et al. 2008] e *logs* de servidores de e-mail [Li and Hsieh 2006].

No caso de trabalhos que analisam *logs* de servidores de e-mail, apenas a última máquina a que o *spammer* se conectou é analisada, já que as mensagens coletadas dessa

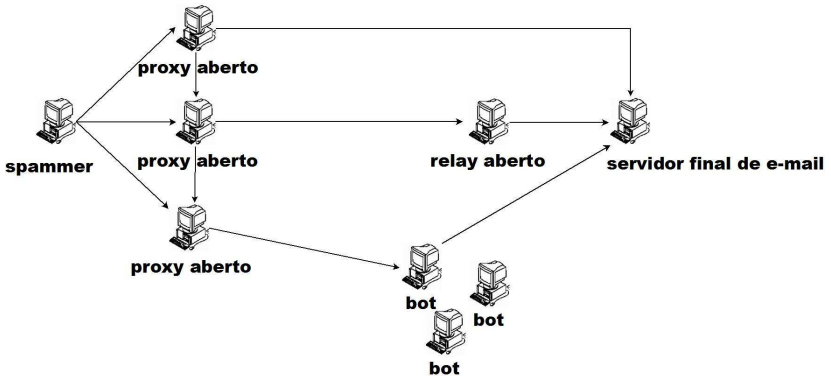

Figura 1. estratégias de encadeamento de máquinas para envio de *spams*

forma não permitem uma análise aprofundada do caminho pelo qual a mensagem passou antes de ser entregue, porque os cabeçalhos SMTP não são confiáveis e são facilmente forjados pelos *spammers*.

Existem trabalhos que analisam conexões estabelecidas por *honeypots* de baixa-interatividade, mas eles focam na análise das características das origem dos abusos aos sensores, como o país de origem das conexões, as portas abusadas nos sensores e distribuição de endereços IP [Calais et al. 2008a, Calais et al. 2008b, Steding-Jessen et al. 2008], sem analisar o destino das conexões. A abordagem deste artigo é diferente porque a origem e o destino das conexões estabelecidas com os *honeypots* são consideradas em conjunto, o que permite aumentar o conhecimento sobre os diferentes caminhos percorridos pelos *spams* antes de serem entregues aos destinatários.

Alguns trabalhos mencionam a criação de cadeias de máquinas para envio de *spams* como algo possível [Boneh 2004, Andreolini et al. 2005, Oudot 2003], mas eles não demonstram e caracterizam efetivamente esses comportamentos. Não conhecemos nenhum trabalho que identifique e quantifique tais abusos, e este artigo pretende preencher essa necessidade.

3. Metodologia de Caracterização

A metodologia proposta para analisar o encadeamento de máquinas para envio de *spam* através da Internet brasileira consiste em três etapas bem definidas. A primeira etapa refere-se à coleta dos dados, realizada a partir de *honeypots* de baixa interatividade. Em seguida, separamos as máquinas de destino das conexões intermediadas pelos *honeypots* em conexões a servidores de correio finais, *proxies* abertos e máquinas de usuários finais infectadas (como *bots*) ou *relays* abertos. Finalmente, caracterizamos e quantificamos como cada endereço IP de origem abusa cada um desses grupos em termos de número de conexões, número de conexões por máquina abusada e volume e duração dos abusos. As duas primeiras etapas são descritas a seguir, enquanto a terceira é discutida juntamente com os resultados na próxima seção.

3.1. Coleta de Dados

A captura das mensagens de *spam* analisadas foi realizada utilizando-se 10 *honeypots* de baixa interatividade, instalados em redes brasileiras de banda larga de 5 operadoras dife-

rentes (cabo e ADSL). O *spam* capturado era periodicamente coletado por um servidor central, responsável também pela monitoração dos *honeypots* [Steding-Jessen et al. 2008].

Os *honeypots* foram configurados de modo a simular computadores com *proxies* e *mail relays* abertos, tradicionalmente abusados para o envio de *spam* e para a realização de outras atividades maliciosas [Krawetz 2004]. Um *spammer* que tentasse abusar de um desses *honeypots* para o envio de *spam* seria levado a acreditar que teve sucesso em enviar suas mensagens, embora nenhum *spam* fosse efetivamente entregue.

A captura de mensagens utilizou o *software* `Honeyd` [Provos and Holz 2007] em conjunto com subsistemas de emulação de SMTP e *proxies* HTTP e SOCKS desenvolvidos para esse fim. Qualquer máquina que se conectasse à porta 25 de um dos *honeypots* teria a impressão de estar interagindo com um servidor SMTP configurado como *open relay*, pronto a repassar as mensagens. Já máquinas que se conectassem a portas tradicionais de *proxies* abertos seriam levados a acreditar que suas conexões a servidores SMTP externos seriam bem-sucedidas. Todas as transações efetuadas pelos subsistemas do `Honeyd` foram armazenadas em *logs* com informações como data e hora, IP de origem da atividade e protocolo que foi abusado no *honeypot*. Caso o ataque fosse a um *proxy*, registrou-se também o IP e porta de destino pretendidos. Todas as mensagens SMTP observadas, seja por terem sido entregues ao *relay* ou por terem passado pelos *proxies*, foram armazenadas com endereços de destino e conteúdo.

Ao todo, foram processadas cerca de 350 milhões de mensagens durante um período de 12 meses. A Tabela 1 apresenta a porcentagem das conexões associadas a cada tipo de serviço emulado pelos *honeypots*.

Tabela 1. número de mensagens enviadas por tipo de conexão

conexões a serviços nos *honeypots*		serviço pretendido na máquina destino	
Tipo de serviço	**Porcentagem**	**Tipo de serviço**	**Porcentagem**
proxy HTTP	61,9%	*relay* (porta 25)	99,6%
proxy SOCKS	36,8%	outros portas	0,4 %
relay (porta 25)	1,3%		

Neste trabalho, consideramos a análise das conexões através do *proxy* HTTP e ao *relay* SMTP dos *honeypots*. Não consideramos as conexões SOCKS porque a maioria delas foi estabelecida a partir da versão 4 do protocolo, que não registra o nome das máquinas de destino das conexões, o que era necessário para nossas análises. Isso significa que 36,8% das conexões não foi considerada, mas o volume restante ainda é significativo.

Quanto ao destino do encadeamento, vemos que quase todas as conexões efetuadas através dos *proxies* foram destinadas à porta 25 da máquina de destino, indicando que normalmente os *spammers* utilizaram um nível de encadeamento desse tipo antes de se conectar através de SMTP para continuar a entrega das mensagens.

3.2. Identificação de Máquinas Participantes de Encadeamentos

A fim de compreender o objetivo do encadeamento observado em cada caso em que os *proxies* HTTP foram explorados para estabelecer conexões com máquinas com o protocolo SMTP, precisamos diferenciar essas máquinas destino entre aquelas que seriam

servidores de correio finais, *relays* e máquinas de usuários de alguma forma infectadas para oferecer o mesmo serviço.

Uma propriedade que diferencia a conexão a servidores legítimos (finais ou *relays* abertos) do encadeamento com *bots* é a presença de nomes de máquinas que tragam claramente uma relação com o serviço de correio. Servidores, em geral, são representados por nomes bem definidos e únicos, como `mail.ufmg.br`, `mx.uol.com.br` ou `mta-v1.mail.vip.tp2.yahoo.com`.

Um critério básico considerado neste trabalho para se afirmar que uma máquina alvo é um servidor final é verificar se seu endereço está mapeado para o mesmo domínio dos endereços encontrados nos destinatários dos *spams*. Este seria o caso (1) discutido na introdução.

No caso de máquinas de usuários finais, na grande maioria dos provedores banda larga os nomes associados a máquinas de usuários normalmente apresentam algum tipo de padrão fixo complementado com uma parte variável usada para diferenciar cada máquina, como uma parte do endereço IP ou simplesmente um numerador. Por exemplo, clientes do provedor de serviços norte-americano Verizon são identificados na rede pelo formato `static-<IP>.<LOCAL>.dsl-w.verizon.net`. As máquinas de clientes do provedor HINET (em Taiwan) são identificadas segundo o formato `<IP>.HINET-IP.hinet.net`. Dessa forma, grandes conjuntos de máquinas alvos de encadeamento através dos *proxies* que compartilham um padrão comum tendem a ser grupos de usuários, indicando máquinas sendo abusadas por *malware* que as faz agir como *mail relays*.

Com base nas observações anteriores desenvolvemos uma técnica para diferenciar encadeamento a servidores de abusos a máquinas de usuários. Primeiramente, os nomes dos domínios dos endereços de destino das mensagens e os nomes das máquinas de destino do encadeamento de conexões pelo *proxy* são quebrados em *tokens*. Por exemplo, `mail.ufmg.br` é formado pelos fragmentos `mail`, `ufmg` e `br`. Já `<IP>.veloxzone.com.br` seria dividido em `<IP>`, `veloxzone`, `com` e `br`. Em seguida, os *tokens* extraídos são rotulados quanto a seu tipo (domínio de e-mail ou máquina de destino) e inseridos em uma estrutura de dados conhecida como árvore de padrões frequentes [Tan et al. 2005, Calais et al. 2008a]. A inserção na árvore é feita de forma que os *tokens* mais frequentemente encontrados no conjunto de dados surgem nos níveis mais altos e as características infrequentes ou aleatórias ficam nos níveis mais baixos da árvore (próximos às folhas). Dessa forma, máquinas de usuários em redes de grandes provedores, pelos seus nomes baseados em padrões comuns, compartilham caminhos na árvore para as partes fixas do padrão e são separadas apenas pelas suas características aleatórias em um mesmo nível. Como as alterações são pouco frequentes, elas ficam nas folhas e endereços com padrões comuns terminam em uma sub-árvore bem definida, com um grande número de irmãos.

4. Caracterização de Encadeamentos de Conexões para Envio de Spams

Nesta seção apresentamos os resultados da aplicação da metodologia proposta a dados coletados na Internet brasileira. A Tabela 2 exibe uma visão geral dos dados coletados. Observamos que, durante o período de 15 meses em que *spams* foram coletados, foram armazenadas mais de 230 milhões de mensagens que seriam enviadas por meio do encadeamento de conexões utilizando o *proxy* HTTP para se conectar à porta 25 da máquina

seguinte na cadeia. Foram estabelecidas 89,8 milhões de conexões com os *honeypots*, resultando, em média, em 2,6 mensagens enviadas por conexão. Essas conexões foram originadas de 93.757 endereços IP distintos, que, conforme mostrado em um trabalho anterior [Calais et al. 2008a], em geral são relacionados com a origem real dos *spammers*. É essa relação estreita entre esses IPs e o *spammer* exatamente que leva ao uso de *proxies* antes da conexão a um computador usando SMTP para garantir que suas identidades permaneçam ocultas para o destinatário.

Tabela 2. Dados relativos ao encadeamento através dos *proxies* HTTP

período de análise	08/07/2006 a 23/06/2007
mensagens enviadas	230.109.671
conexões observadas	89.836.643
endereços IP de origem distintos	93.757
destinatários distintos das mensagens	$3{,}2 \times 10^9$
domínios de destinatários distintos	6.710.121
endereços IP de destino distintos	459.218

4.1. Evidências de Encadeamento de Máquinas para Envio de *Spam*

Enquanto as conexões observadas foram destinadas a quase 460 mil máquinas distintas, mais de 6,7 milhões de domínios de e-mail de destinatários das mensagens foram observados durante o período analisado (Tabela 2). A observação de que o número de máquinas às quais as conexões se destinam é muito menor que o número de domínios de e-mail aos quais as mensagens seriam entregues leva a uma das conclusões fundamentais deste artigo: **uma porção significativa das mensagens não se destina diretamente aos servidores de correio de destino após passarem pelos proxies abertos na Internet brasileira**. Caso contrário, o número máquinas de destino distintas deveria ser próximo do número de domínios de e-mail distintos ou até superior, dado que alguns domínios de e-mail são mapeados em mais de um servidor MX. Por exemplo, mensagens destinadas a `yahoo.com.tw` podem ser entregues a `mta-v1.mail.vip.tp2.yahoo.com`, `mta-v2.mail.vip.tp2.yahoo.com` e `mta-v2.mail.vip.tpe.yahoo.com`. Tal comportamento confirma que grande parte dos *spams* que são disseminados pela infraestrutura da Internet brasileira são encadeados com outras máquinas antes de serem entregues ao destino final.

O comportamento em que *spams* destinados a muitos domínios de e-mail são entregues a menos endereços IP do que o esperado se a entrega fosse sempre ao servidor final é recorrente. A Figura 2 mostra a distribuição acumulada do número de domínios alvo de *spam* diferentes encontrados nas mensagens entregues a cada endereço IP de destino. Mais de 50% dos IPs de destino recebem mensagens para mais de dois domínios diferentes, mais de 10% recebem mensagens para mais de dez domínios e alguns destinos recebem mensagens para mais de 100 domínios distintos, o que reforça que esses endereços representam máquinas que não são os servidores de destino das mensagens e estão apenas intermediando a entrega.

4.2. Classificação dos Tipos de Destino do Encadeamento

A Tabela 3 exibe o número de máquinas abusadas como destino das conexões encadeadas através dos *proxies* HTTP e o número de conexões associadas a cada um dos tipos

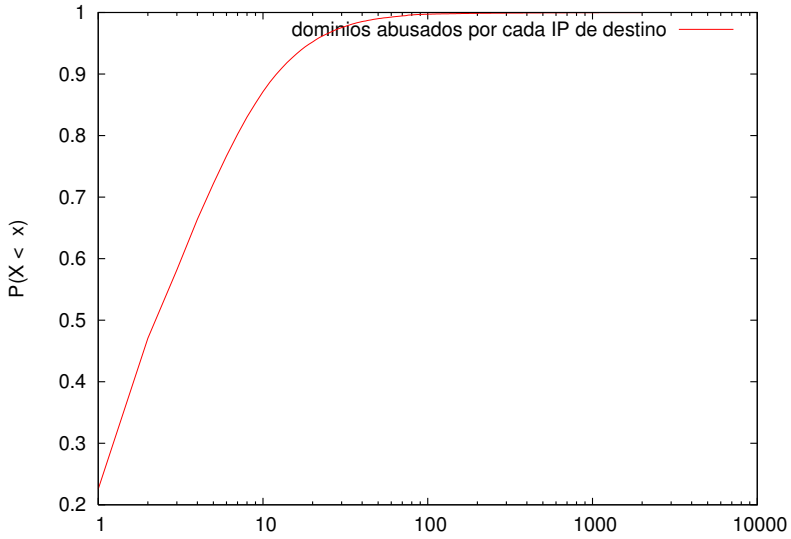

Figura 2. Número médio de domínios diferentes encontrados nos destinatários das mensagens entregues a cada IP de destino diferente (CDF)

de abuso identificados após a aplicação da técnica apresentada na Seção 3.2. Embora o número de servidores de correio legítimos seja próximo do número de máquinas de usuários finais, percebe-se que a maioria das conexões (72,3 milhões) são destinadas aos primeiros. Cerca de metade das conexões aos servidores de correio finais foi destinada a servidores do serviço de correio do *Yahoo!* (veja a Tabela 4), o que é coerente com a alta popularidade do domínio `yahoo.com.tw` entre os alvos preferidos dos *spams* encontrados em nosso conjunto de dados. Uma pequena fração das conexões observadas indica ainda um nível de encadeamento mais alto: foram observadas 342 mil tentativas de conexão a outras máquinas em portas associadas a serviços de *proxy*, ao invés de buscarem conexão direta por SMTP. É importante notar que, apesar de pequeno (0,4%, conforme a Tabela 1), esse valor pode estar subestimado, pois ao contrário dos encadeamentos para a porto 25, esses pedidos eram abortados simulando uma mensagem de erro no *honeypot*. A criação de uma cadeia de *proxies* abertos torna o rastreamento do *spammer* muito mais difícil, porque mesmo que os *logs* de cada *proxy* estejam disponíveis, é necessário reconstruir todo o caminho de requisições [Andreolini et al. 2005].

Tabela 3. Número de máquinas e conexões por tipo de destino

número de máquinas de usuários finais	192.507
número de servidores de correio finais	222.957
número de proxies abertos	7.102
número de conexões destinadas a máquinas de usuários finais abusadas	11.585.078
número de conexões destinadas a servidores de correio	72.335.130
número de conexões a *proxies* abertos	341.669

Por outro lado, determinamos 894 grupos de máquinas de usuários finais que claramente não são servidores de e-mail. A Tabela 5 lista os 15 grupos de endereços IP mais comumente visados como destino das conexões à porta 25 encadeadas pelos *honeypots*. Chama a atenção o número expressivo de máquinas que fazem parte da rede do provedor HINET, de Taiwan. O resultado indica que Taiwan participa intensamente nas três

Tabela 4. conexões a servidores e-mail oficiais mais frequentes

host de destino	número de conexões	%
mta-v1.mail.vip.tp2.yahoo.com	19.513.600	26,98
mta-v2.mail.vip.tp2.yahoo.com	8.022.648	11,09
mta-v2.mail.vip.tpe.yahoo.com	4.733.704	6,54
mta-v1.mail.vip.tpe.yahoo.com	2.349.188	3,25
mx.seed.net.tw	1.689.263	2,34
mxs.pchome.com.tw	1.193.082	1,65
mta-v21.level3.mail.yahoo.com	526.417	0,73
mx1.url.com.tw	503.718	0,70
mx3.url.com.tw	499.783	0,69
mx5.url.com.tw	484.759	0,67

etapas de entrega de mensagens: *spammers* originam *spams* [Calais et al. 2008a] a partir de máquinas em TW, encadeiam abusos a *proxies* abertos com máquinas de usuários finais em TW e entregam as mensagens a destinatários em TW. A lista dos domínios mais abusados incluem outros ISPs em TW, como `seed.net.tw` e `isl.net.tw`. Entre os grupos de máquinas que foram abusadas por *spammers*, destacam-se também grandes provedores de servidores dedicados (*dedicated hosting*) e *datacenters*, como `ev1servers.net` e `secure-server.net`. Essas máquinas, possivelmente, são casos de servidores mal-configurados, que aceitam conexões de qualquer origem.

Cerca de um terço (36%) dos grupos de máquinas abusadas estão localizadas nos Estados Unidos. Isso indica que um caminho comum dos *spams* é partirem de máquina em TW, explorarem um proxy aberto no Brasil, serem direcionados a máquinas nos Estados Unidos (provavelmente, *bots* com um servidor SMTP ativado) e só então serem entregues ao destinatário final, quase sempre em TW. Nossa técnica para encontrar grupos de máquinas determinou máquinas abusadas associadas a outros 68 *Country-Codes*, liderados por GB, JP e BR, além de US e TW. Os principais ISPs encontrados em nosso estudo coincidem com uma lista de provedores de acesso que hospedam o maior número de máquinas infectadas no planeta, o que comprova que ***spammers* encadeiam *proxies* abertos e máquinas de usuários finais infectadas, instalados em redes de banda larga**. Entendemos que o motivo que leva os *spammers* a intermediarem a conexão com máquinas de usuários finais por meio de *proxies* abertos é que é importante para o originador do abuso manter sua identidade secreta, mesmo que a máquina abusada seja uma máquina de usuário final. Por exemplo, existem *honeypots* projetados especialmente para fazerem parte de uma *botnet*, e nesse caso, a identidade do *spammer* seria revelada [Boneh 2004].

Esses resultados mostram que, apesar de *botnets* corresponderem a maior parte do *spam* que circula no planeta [Ramachandran and Feamster 2006] e por isso receberem grande atenção dos pesquisadores, o combate aos *proxies abertos* continua a ser necessário.

Embora ainda não consigamos diferenciar os abusos a máquinas de usuários configuradas como *relays* abertos da entrega de mensagens por meio de máquinas infectadas por programas maliciosos/*bots*, acreditamos que a maior parte dessas máquinas são, de fato, membros de *botnets* que são configuradas para encaminhar *spams*. Os argumentos nesse sentido são o fato de que abusos a *relays* abertos têm sido cada vez mais raros [Leyden 2003] e, ainda, que máquinas de usuários finais usualmente não possuem servidores de e-mail instalados. No entanto, abusos a *relays* abertos ainda acontecem.

Tabela 5. Número de máquinas nos principais grupos (ISPs) detectados

ISP/domínio	*Country-Code*	número de máquinas (IPs)
<IP>.HINET-IP.hinet.net	TW	15.045
<IP>.ev1servers.net	US	1.417
rrcs-<IP>.central.biz.rr.com	US	1.228
<IP>.static.isl.net.tw	TW	1.191
0.Red-<IP>.staticIP.rima-tde.net	ES	1.022
<IP>.seed.net.tw	TW	966
<IP>.ptr.us.xo.net	US	882
<IP>.dsl.scrm01.pacbell.net	US	877
ip-<IP>.ip.secureserver.net	US	849
<IP>.dynamic.hinet.net	TW	746
c-<IP>.hsd1.nj.comcast.net	US	735

Isso pode ser facilmente verificado nos dados coletados, onde encontramos alguns poucos casos onde um *spammer* se utilizou do *proxy* de um *honeypot* para encadear uma conexão ao *mail relay* de outro *honeypot* do conjunto.

A Figura 3 considera apenas os endereços IP de origem que usaram os *honeypots* para se conectar às máquinas destino identificadas anteriormente como máquinas de usuários finais. Para aquelas máquinas de origem, a figura mostra a relação entre o número de conexões realizadas e o número de IPs de destino diferentes aos quais as conexões foram direcionadas. O gráfico, em escala logarítmica no eixo X, sugere uma relação linear entre o número de conexões estabelecidas e o número de máquinas abusadas por endereço IP. O coeficiente de correlação entre as duas grandezas é de 90%, o que indica que existe uma forte relação entre elas. A correlação das mesmas variáveis para o abuso a servidores de correio finais é de apenas 65% (gráfico não apresentado), o que reflete o fato de que alguns domínios de *e-mail* são mais populares que outros e recebem mais conexões. No caso dos abusos à máquinas de usuários finais, todas são consideradas igualmente pelos *spammers*.

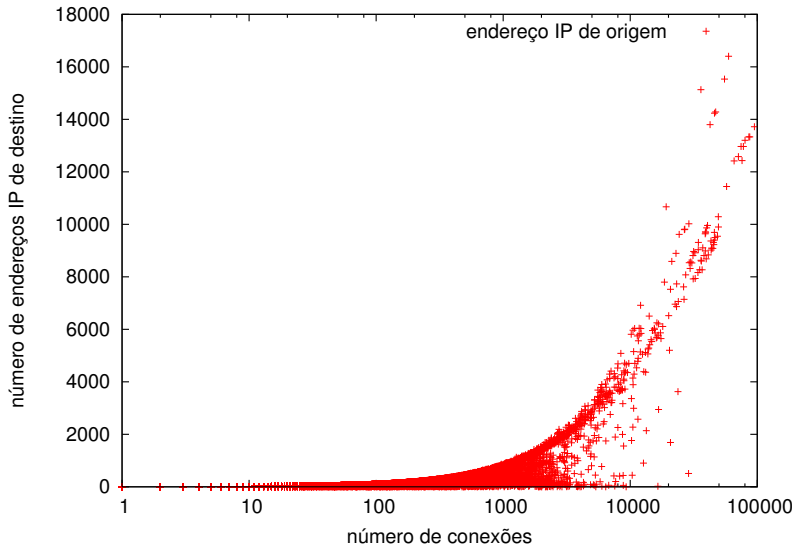

Figura 3. número de conexões estabelecidas x número de IPs de destino que seriam supostamente abusados por cada endereço IP de origem

4.3. Impacto do Uso de Encadeamentos para Disseminação de *Spams*

Após identificar e separar os destinos das conexões em grupos (servidores de correio finais, máquinas de usuários finais e *proxies* abertos), investigamos como cada *spammer* (representado por um endereço IP de origem) abusa esses grupos. Em particular, procuramos analisar como a quantidade de máquinas abusadas e a intensidade com que cada máquina é abusada afeta o volume de mensagens que o *spammer* entrega e por quanto tempo ele persiste os abusos.

A Figura 4, em escala *log-log*, verifica a correlação, para cada IP de origem, entre o número de máquinas de destino diferentes contatadas e o volume de mensagens enviado por aquela origem. Apesar do espalhamento observado, o coeficiente de correlação é significativo (72%). Nota-se que apenas *spammers* que dispõem de listas de máquinas de usuários finais de tamanhos superiores a 10.000 elementos conseguiram enviar mais de 1 milhão de *spams*.

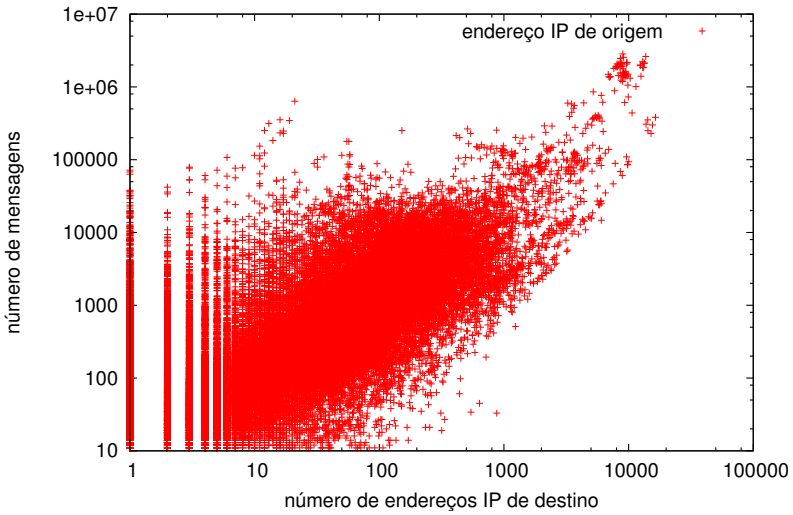

Figura 4. número de endereços IP de destino que seriam supostamente contatados por cada IP de origem x volume de mensagens enviadas

Ao contrastar o número de endereços IP de destino abusados por endereço IP de origem e o número de dias pelo qual esse IP enviou *spams* (Figura 5), fica claro que apenas *spammers* que contam com infraestrutura para abusar milhares de endereços IP de destino conseguem longevidade suficiente para enviar mensagens por vários meses. A maior parte dos IPs de origem permanece ativo por menos de dois meses.

A Figura 6 mostra que *spammers* que conseguem enviar mensagens por muitos meses são os mesmos que estabelecem, em média, poucas conexões a cada uma das máquinas que abusam. Essa observação indica que os *spammers* mais bem-sucedidos são aqueles que conseguem distribuir mais os seus abusos e, então, passarem desapercebidos. O que limita o volume de mensagens que um *spammer* consegue entregar não parece ser a largura de banda a que eles têm acesso, mas a capacidade que eles têm de encadear suas mensagens através de muitos intermediários diferentes ao mesmo tempo. Essa característica tem sido identificada na literatura como uma característica marcante de *botnets* [SpamCop 2007] e impõe uma série de desafios no combate ao *spam* [Naraine 2007].

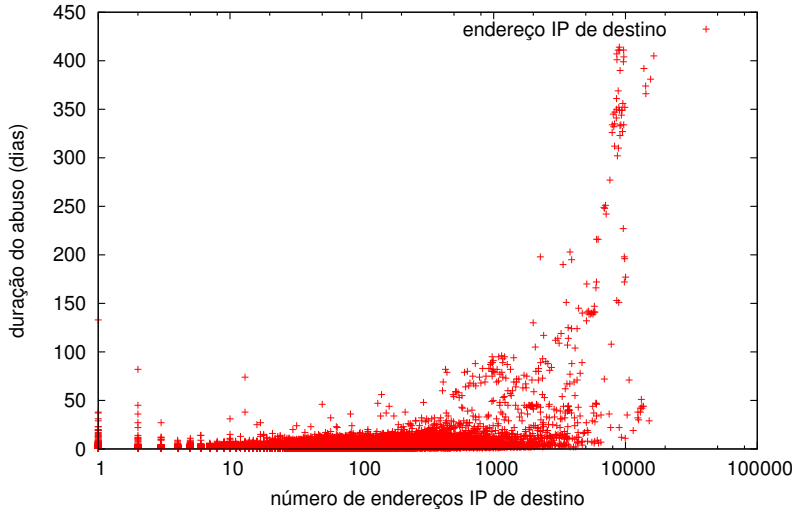

Figura 5. número de endereços IP de destino que seriam supostamente abusados por cada endereço IP de origem x número de dias pelo qual o IP de origem permanece ativo

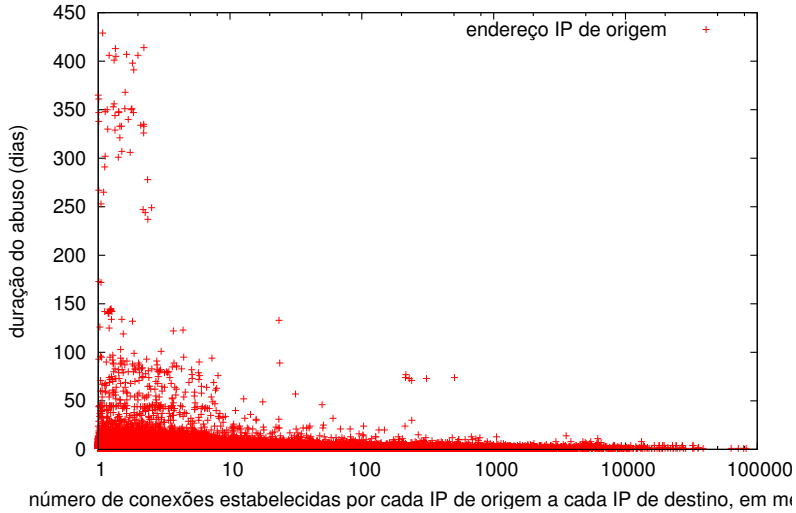

Figura 6. número de conexões que cada IP de origem estabelece a cada IP de destino x número de dias pelo qual o IP de origem permanece ativo

Como muitas máquinas de usuários finais se conectam à rede por meio de endereços IP dinâmicos, a identificação dos abusos é ainda mais complicada.

5. Conclusão e Trabalhos Futuros

Neste trabalho, mostramos que *spammers* encadeiam abusos a *proxies* abertos na Internet brasileira com abusos a *relays* abertos, máquinas de usuários infectadas (que podem fazer parte de *botnets*) e outros *proxies* abertos, além dos próprios servidores de correio finais. O trabalho analisou conexões HTTP estabelecidas por *spammers* a *honeypots* de baixa interatividade implantados na rede brasileira. A principal contribuição deste artigo

é identificar e quantificar esses comportamentos, que, embora descritos na literatura como possíveis, ainda não haviam sido demonstrados. Acreditamos que, para os pesquisadores que desenvolvem técnicas para combate ao *spam*, é importante estar ciente dos múltiplos caminhos que as mensagens percorrem antes de serem entregues aos destinatários. O desenvolvimento de mecanismos baseados em reputação de nós e que consideram as características das conexões SMTP devem levar em consideração que o encadeamento de máquinas para entrega dos *spams* é algo comum.

Nossos resultados mostram que os *spammers* tentam se conectar poucas vezes a cada máquina abusada e enviar poucas mensagens de cada uma delas, para que a detecção da atuação individual de cada máquina seja difícil. Os *spammers* que enviam maiores volumes de mensagens são aqueles que conseguem abusar o maior número de máquinas e enviar, em média, o menor número de mensagens por cada uma delas.

Como trabalhos futuros, pretendemos usar a informação das campanhas de *spam* disseminadas para entender melhor como cada máquina alvo das conexões HTTP é abusada e determinar, por exemplo, se há máquinas que são abusadas por um grande conjunto de *spammers* e outras de uso exclusivo de um grupo ou indivíduo. Esse conhecimento pode ajudar a identificar padrões que diferenciem o abuso a *relays* abertos de máquinas que fazem parte de *botnets*. Pretendemos, ainda, implantar *honeypots* em outros países e assim analisar a disseminação dos *spams* sob um ponto de vista global, a partir de uma arquitetura de coleta aprimorada que torne mais fácil relacionar as mensagens e as conexões que foram estabelecidas para enviar cada uma delas.

6. Agradecimentos

Este trabalho foi parcialmente financiado por NIC.br, CNPq, CAPES, FAPEMIG e FINEP.

Referências

Andreolini, M., Bulgarelli, A., Colajanni, M., and Mazzoni, F. (2005). Honeyspam: honeypots fighting spam at the source.

Boneh, D. (2004). The difficulties of tracing spam email. `http://www.ftc.gov/reports/rewardsys/expertrpt_boneh.pdf`.

Calais, P. H., Guedes, D., Jr., W. M., Hoepers, C., and Steding-Jessen, K. (2008a). Caracterização de estratégias de disseminação de spams. In *Anais do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*.

Calais, P. H., Pires, D., Guedes, D., Wagner Meira, J., Hoepers, C., and Steding-Jessen, K. (2008b). A campaign-based characterization of spamming strategies. In *Proceedings of the Conference on e-mail and anti-spam (CEAS)*.

Hayes, B. (2003). Spam, spam, spam, lovely spam. *American Scientist*, 91(3):200–204.

Krawetz, N. (2004). Anti-honeypot technology. *IEEE Security & Privacy*, 2(1):76–79.

Lee, W., Wang, C., and Dagon, D. (2007). *Honeynet-based Botnet Scan Traffic Analysis*, volume Volume 36.

Leyden, J. (2003). Open relay spam is dying out. `http://www.theregister.co.uk/2003/06/12/open_relay_spam_is_dying/`.

Li, F. and Hsieh, M.-H. (2006). An empirical study of clustering behavior of spammers and group-based anti-spam strategies. *Proceedings of the Third Conference on Email and Anti-Spam (CEAS). Mountain View, CA*.

Messaging Anti-Abuse Working Group (MAAWG) (2007). Email Metrics Program: Report #5 – First Quarter 2007. `http://www.maawg.org/about/MAAWG20071Q_Metrics_Report.pdf`.

Naraine, R. (2007). Is the botnet battle already lost? `http://www.eweek.com/print_article2/0,1217,a=191391,00.asp`.

Oudot, L. (2003). Fighting spammers with honeypots. `http://www.securityfocus.com/infocus/1747`.

Pathak, A., Hu, Y. C., and Mao, Z. M. (2008). Peeking into spammer behavior from a unique vantage point. In *LEET'08: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pages 1–9, Berkeley, CA, USA. USENIX Association.

Prakash, V. V. and O'Donnell (2005). Fighting spam with reputation systems. *Queue*, pages 36–41.

Provos, N. and Holz, T. (2007). *Virtual Honeypots: From Botnet Tracking to Intrusion Detection*. Addison-Wesley Professional, 1st edition. ISBN-13: 978-0321336323.

Pu, C. and Webb, S. (2006). Observed trends in spam construction techniques: a case study of spam evolution. *Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS)*.

Ramachandran, A. and Feamster, N. (2006). Understanding the network-level behavior of spammers. In *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 291–302, New York, NY, USA. ACM.

Sipior, J. C., Ward, B. T., and Bonner, P. G. (2004). Should spam be on the menu? *Commun. ACM*, 47(6):59–63.

SpamAssassin (2008). `http://spamassassin.apache.org`.

SpamCop (2007). Botnets. `http://forum.spamcop.net/scwik/BotNet`.

Steding-Jessen, K., Vijaykumar, N. L., and Montes, A. (2008). Using low-interaction honeypots to study the abuse of open proxies to send spam. *INFOCOMP Journal of Computer Science*, 7:44–52.

Tan, P., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co.

Whitworth, B. and Whitworth, E. (2004). Spam and the social-technical gap. *Computer*, 37(10):38–45.