IDENTIFICAÇÃO E CARACTERIZAÇÃO DE CAMPANHAS DE *SPAM* A PARTIR DE *HONEYPOTS*

PEDRO HENRIQUE CALAIS GUERRA

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE CAMPANHAS DE *SPAM* A PARTIR DE *HONEYPOTS*

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Wagner Meira Jr.

Belo Horizonte

16 de junho de 2009

Calais Guerra, Pedro Henrique
Identificação e Caracterização de Campanhas de *Spam* a partir de *Honeypots* / Pedro Henrique Calais Guerra. — Belo Horizonte, 2009
xii, 55 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Wagner Meira Jr.

1. Mineração de Dados. 2. Spam. 3. Segurança de Redes. I. Título.

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais, Claret e Sueli, pelo apoio durante toda a vida e por darem condições plenas para que eu realizasse o curso.

Aos meus orientadores, Wagner Meira Jr. e Dorgival Guedes, pela orientação e paciência com meus eternos *drafts* incompletos :-). Considero um privilégio ter sido orientado pelos dois simultaneamente e ter aprendido com a personalidade, conhecimento e experiência particulares de cada um.

Ao CGI.br, no nome de Klaus Steding-Jessen, Cristine Hoepers e Marcelo Chaves, pelos dados fornecidos, imprescindíveis para a execução do trabalho. As sugestões, críticas e comentários de vocês foram fundamentais para o desenvolvimento da dissertação.

À Universidade Federal de Minas Gerais e ao seu Departamento de Ciência da Computação, por ser responsável pela minha formação nos últimos seis anos e pelo provimento da infraestrutura e ambiente propício ao desenvolvimento de pesquisas de qualidade.

Finalmente, agradeço aos colegas do laboratório e-Speed pelo companheirismo e companhia durante os dois anos de curso, em especial ao Douglas, pelo apoio durante e execução do trabalho.

Resumo

Este trabalho apresenta uma metodologia para caracterização de estratégias de disseminação de *spams* a partir da identificação de campanhas. Para entender com profundidade como *spammers* abusam os recursos da rede e constróem suas mensagens, uma análise agregada das mensagens de *spam* não é suficiente. O agrupamento de mensagens de *spam* em suas respectivas campanhas permite revelar comportamentos que não poderiam ser percebidos ao considerar o conjunto de mensagens como um todo. Este trabalho propõe uma técnica para identificação de campanhas de *spam* baseada na construção de uma Árvore de Padrões Frequentes, capaz de capturar os invariantes no conteúdo das mensagens e detectar mensagens que diferem apenas por características ofuscadas e variadas aleatoriamente por *spammers*. A técnica foi capaz de agrupar um conjunto de 350 milhões de mensagens em 57.851 campanhas distintas. Em seguida, essas campanhas foram caracterizadas em termos de seus conteúdos e da forma como exploram recursos da rede. A partir da aplicação de algoritmos de mineração de regras de associação, foi possível determinar co-ocorrência de atributos das campanhas que revelam diferentes estratégias de disseminação de *spams*. Em particular, foram determinadas relações significativas entre a origem do *spam* e a forma como ele é disseminado na rede, entre sistemas operacionais e tipos de abuso e na forma como *spammers* encadeiam abusos entre máquinas na rede para entregar mensagens enquanto mantém anonimato. Os dados utilizados no trabalho foram coletados a partir de *honeypots* de baixa-interatividade que emulam *proxies* e *relays* abertos, comumente abusados por *spammers*. A coleta dos dados por esses emuladores estabeleceu uma visão do tráfego de *spams* antes que as mensagens fossem entregues aos destinatários, o que permitiu a determinação das diferentes estratégias de entrega de mensagens empregadas por *spammers*.

Abstract

This work presents a methodology for the characterization of spamming strategies based on the identification of spam campaigns. To deeply understand how spammers abuse network resources and obfuscate their messages, an aggregated analysis of spam messages is not enough. Grouping spam messages into campaigns is important to unveil behaviors that cannot be noticed when looking at the whole set of spams collected. We propose a spam identification technique based on a frequent pattern tree, which naturally captures the invariants on message content and detect messages that differ only due to obfuscated fragments. The technique was able to group 350 million messages into 57,851 distinct campaigns. After that, we characterize these campaigns both in terms of content obfuscation and exploitation of network resources. Our methodology includes the use of attribute association analysis: by applying an association rule mining algorithm, we were able to determine co-occurrence of campaign attributes that unveil different spamming strategies. In particular, we found strong relations between the origin of the spam and how the network was abused, between operating systems and types of abuse and patterns that describe how spammers chain machines over the Internet to conceal their identities. Data was collected from low-interaction honeypots emulating open proxies and open relays, traditionally abused by spammers. The data collected from these emulators created a vantage point of spams from inside the network, before the messages were delivered to recipients, and that allowed the determination of the different strategies adopted by spammers to deliver their messages.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 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 & 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.

O fato do custo de envio de *e-mails* ser muito baixo, comparado ao da correspondência convencional, serve como incentivo ao uso do correio eletrônico para o envio de *e-mails* comerciais não-solicitados em grandes quantidades (Cerf, 2005). Além disso, o *spam* tem sido um meio usual para enviar mensagens relacionadas com a obtenção de dados pessoais com objetivos ilícitos (*phishing*) e para disseminação de códigos maliciosos (Milletary, 2005). Devido à proporção atingida pelo problema, várias abordagens técnicas têm sido utilizadas para tratá-lo, como, por exemplo, a adoção de recomendações para configuração de sistemas de *e-mail* (Lindberg, 1999; Killalea, 2000), uso de filtros de conteúdo de mensagens como o *Spam Assassin* (SpamAssassin, 2007) e listas de bloqueio (Cook et al., 2006). Ao mesmo tempo, observa-se um aumento da sofisticação dos *softwares* de envio de *spam*, o que torna as técnicas de bloqueio existentes menos eficientes e a rastreabilidade do *spammer* mais difícil (Cranor & LaMacchia, 1998; Milletary, 2005). Um exemplo desse fenômeno é o crescimento na utilização de máquinas infectadas por códigos maliciosos, como os *bots*, para o envio de *spam* e *phishing*, permitindo que o *spammer* permaneça no anonimato (Milletary, 2005). De fato, o *spamming* é uma atividade de natureza evolutiva, tanto na forma como os *spammers* constróem o conteúdo das mensagens (Pu & Webb, 2006) quanto no modo como disseminam suas mensagens pela rede, buscando maximizar o volume de mensagens que enviam enquanto mantêm sua identidade oculta. A interminável batalha entre *spammers* e anti-*spammers* é uma característica marcante do problema do *spam*, conhecida como *spam arms race* (Paulson, 2005), em que ambos evoluem ao mesmo tempo tentando se sobrepor à força do outro. Por isso, um esforço contínuo para entender como *spammers* geram, distribuem e disseminam suas mensagens pela Internet é necessário, para manter ou mesmo melhorar a efetividade das

técnicas de combate ao *spam*.

O objetivo principal desta dissertação é propor uma metodologia de caracterização de estratégias de disseminação de *spams* e apresentar um estudo de caso real que aplica a metodologia proposta. Por *estratégia de disseminação de spams*, considera-se qualquer recurso utilizado pelo *spammer* para maximizar o alcance de suas mensagens, reduzindo a probabilidade de que a mensagem seja retida em filtros anti-*spam* e de que ela seja identificada e rastreada. Por exemplo, a inserção de termos aleatórios no corpo da mensagem é uma estratégia de disseminação de conteúdo que visa dificultar a ação dos filtros anti-*spam*, enquanto o comprometimento de máquinas de usuários para disseminar *spams* a partir delas é uma estratégia de entrega das mensagens na rede. A partir da identificação das estratégias mais utilizadas, pretende-se contribuir para aumentar o conhecimento sobre como *spammers* atuam para construir e disseminar suas mensagens na rede, e, com isso, fomentar o desenvolvimento de técnicas de combate ao *spam* mais efetivas e robustas.

O eixo principal da metodologia é a identificação das campanhas de *spam* que são disseminadas na rede. Em geral, *spammers* disfarçam e variam o conteúdo que enviam de maneira sistemática e automatizada, inserindo trechos aleatórios no corpo da mensagem e nos *links* nela contidos (Sophos.com, 2004). O objetivo é evitar ao máximo o envio de mensagens idênticas, pois isso facilitaria a detecção de sua atuação. Sem a capacidade de identificar grupos de mensagens que tiveram origem em um texto base comum (ou um *template*, que define uma campanha de *spam*), não é possível isolar o tráfego gerado por diferentes *spammers*. Sendo assim, para analisar o comportamento de *spammers* na rede é necessário identificar os grupos de mensagens que correspondam a uma mesma campanha, neutralizando o impacto da ofuscação das mensagens e das estratégias de disseminação, para então estudar as características de exploração da rede de cada um desses grupos de mensagens, que, embora possam conter *links* e textos distintos, possuem um objetivo em comum, como anunciar um produto específico.

A identificação das campanhas de *spam* é considerada um passo essencial para a identificação de estratégias de *spamming* por várias razões. Primeiramente, a identificação das campanhas cria novas dimensões que podem ser analisadas e correlacionadas, como duração e volume dos abusos. A análise agregada dos dados é limitada no sentido de determinar estratégias de atuação, por esconder relações entre as grandezas e por apontar comportamentos médios que não são necessariamente representativos. Ao agrupar as mensagens em suas respectivas campanhas, é possível caracterizar como *cada spammer* disseminou as suas mensagens. Além disso, dependendo da estratégia de coleta de mensagens aplicada, o volume de mensagens coletado pode ser muito grande e processá-las pode ser custoso ou mesmo inviável. O agrupamento de mensagens em campanhas trata esse problema ao criar um critério de sumarização dos dados, reduzindo o volume de informação a ser processado, mas mantendo as informações-chave de cada abuso. Finalmente, a identificação de campanhas de *spam* neutraliza o efeito do volume variável de mensagens associadas a cada campanha, que pode esconder comportamentos frequentes que ocorrem apenas em campanhas pequenas.

Este trabalho propõe uma técnica para identificação de grupos de mensagens associadas à uma mesma campanha de *spam*, descrita no Capítulo 4.1. A técnica é baseada na extração de características relevantes de cada mensagem de *spam* e na inserção dessas características em uma estrutura de dados do tipo Árvore, que consegue detectar as características invariantes entre as mensagens e que definem cada campanha de *spam*. A técnica proposta é mais geral que outras abordagens propostas na literatura, que se apóiam em características específicas de tecnologias correntes empregadas por *spammers*, como a inclusão de imagens nas campanhas. Além disso, a técnica determina padrões de ofuscação sem pré-definir os padrões a serem procurados entre as mensagens, o que a permite resistir à evolução inerente ao *spamming*.

A partir da identificação das campanhas, é proposta uma metodologia de caracterização das estratégias de disseminação de *spams*, baseada na detecção de invariantes e padrões de co-ocorrência das estratégias de disseminação em cada grupo de mensagens associadas a uma única campanha. Esses invariantes e padrões representam comportamentos dos *spammers* e podem apoiar a definição de critérios para a detecção, identificação e minimização dos impactos do *spam*. As relações determinadas neste trabalho confirmaram alguns comportamentos conhecidos na literatura e revelaram outras estratégias de disseminação de *spams* que, até então, não haviam sido documentadas em trabalhos científicos.

A efetividade da metodologia foi demonstrada aplicando-a a uma coleção de mensagens de *spam* capturadas durante doze meses por *honeypots* de baixa interatividade (Provos & Holz, 2007), configurados de modo a simular computadores atuando como *proxies* e *relays* abertos (Steding-Jessen et al., 2008). Os *honeypots* coletam as mensagens antes que elas cheguem aos destinatários, permitindo análises do comportamento de *spammers* que não seriam possíveis em *logs* de servidores de *e-mail*, por exemplo.

A dissertação está organizada em cinco capítulos, além desta Introdução. O Capítulo 2 lista os trabalhos relacionados e explica como eles são complementados por este trabalho. No Capítulo 3 é apresentada a arquitetura de coleta dos dados e números gerais dos *spams* coletados. Em seguida, é apresentada a técnica projetada para agrupar mensagens de *spam* em campanhas e os resultados da caracterização, no Capítulo 4, e, especificamente, a análise dos encadeamentos de máquinas utilizadas por *spammers* para disseminar mensagens, no Capítulo 5. Finalmente, as conclusões do trabalho são apresentadas no Capítulo 6, bem como os trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

A Revisão Bibliográfica deste trabalho foi dividida em duas partes: a primeira é dedicada às estratégias que visam identificar e detectar campanhas de *spam*. Em seguida, é apresentado um sumário dos principais trabalhos que caracterizam a atuação de *spammers* para construir e disseminar suas mensagens.

2.1 Identificação de Campanhas de *Spam*

A literatura apresenta uma série de trabalhos cujo objetivo é identificar grupos de mensagens que, embora potencialmente possam conter textos, imagens ou *links* diferentes, correspondem a um mesmo abuso.

SpamScatter (Anderson et al., 2007) é uma técnica que determina campanhas de *spam* a partir das páginas referenciadas pelas mensagens. A técnica, batizada pelos autores de *image shingling*, consiste em computar a similaridade entre as imagens de cada uma dessas páginas, a partir da ideia de que as páginas apontadas pela mesma campanha de *spam* devam apresentar conteúdos bastante semelhantes. Cada página é decomposta em segmentos e as páginas que compartilham um grande número de segmentos são agrupadas. A metodologia adotada pelos autores é similar à apresentada nesta dissertação, no sentido de que as campanhas são identificadas e então caracterizadas. No entanto, o trabalho foca na infraestrutura de hospedagem das páginas dos *spams*, enquanto esta dissertação é voltada para a caracterização dos abusos à infraestrutura da rede e padrões de geração de conteúdo dos *spams*.

Uma outra técnica relevante para identificar campanhas de *spam* é o algoritmo I-Match (Kolcz & Chowdhury, 2007), que agrupa mensagens que sejam duplicatas uma das outras desde que elas compartilhem os termos de um dicionário pré-definido (*Lexicon*). Esse dicionário pode ser obtido, por exemplo, a partir da seleção de palavras representativas de um conjunto pré-classificado de *spams*. A representatividade de cada palavra pode ser medida a partir de métricas como o *idf* (*Inverse Document Frequency*) de cada palavra. Determinado o *Lexicon*, duas mensagens de *spam* serão duplicatas uma da outra se compartilharem o mesmo conjunto de palavras no *Lexicon*, e esse conjunto representa a assinatura das mensagens. Os

autores atentam para o fato de que uma aplicação direta da identificação de mensagens duplicatas é a própria identificação de *spams*, a partir da propriedade inerente ao *spam* de ser enviado em grandes quantidades em um curto período de tempo (Chowdhury & Alspector, 2004), o que não acontece com mensagens convencionais.

Existem trabalhos que identificam campanhas de *spam* a partir da similaridade das imagens inseridas no conteúdo das mensagens (Wang et al., 2007; Mehta et al., 2008), apoiando-se em características como tamanho das imagens, saturação e histogramas das cores, formato de codificação das imagens e metadados. Esses trabalhos foram motivados, principalmente, pelo grande crescimento de estratégias adotadas por *spammers* que utilizam imagens e evitam escrever textos nas mensagens, a fim de dificultar a atuação dos filtros baseados em classificação de texto.

Além do texto e imagens das mensagens, a similaridade entre as URLs contidas nas mensagens também já foi explorada para agrupar mensagens em campanhas. Nesse caso, as URLs contidas nos *spams* são quebradas em fragmentos e mensagens que compartilhem uma determinada proporção do número de fragmentos são agrupadas. Em (Yeh & Lin, 2006), Os autores utilizaram também o conhecimento de algumas estratégias comumente adotadas por *spammers*, como a variação aleatória nos parâmetros CGI das URLs, e a conversão dos domínios das URLs para endereços IP, para que as URLs cujos domínios foram ofuscados fossem mapeadas para o mesmo endereço IP.

A principal diferença entre esses trabalhos e esta dissertação, é que, no caso deste trabalho, a identificação das campanhas não é o objetivo principal, mas um passo para que a atuação de *spammers* possa ser caracterizada e compreendida. Em termos da estratégia utilizada para agrupar as mensagens, pode-se dizer que a técnica proposta nesta dissertação é mais geral e menos dependente de estratégias específicas adotadas por *spammers*, conforme ficará claro no Capítulo 4.

2.2 Caracterização do Comportamento de *Spammers*

Os trabalhos que estudam a atuação de *spammers* podem ser divididos em duas categorias: trabalhos que caracterizam o comportamento de rede e trabalhos que estudam a forma como *spammers* geram o conteúdo de suas mensagens.

Na primeira categoria, um trabalho bastante citado é o de Ramachandran & Feamster (2006), que estudam como *spammers* exploram a infraestrutura da Internet para enviar suas mensagens, incluindo as faixas de endereços IP mais usadas para se enviar *spam* e tipos de abuso mais comuns. Entre outras conclusões, os autores destacam o fato de que as mensagens de *spam* são enviadas de faixas muito restritas de endereços IP. São mostradas também algumas estatísticas em relação à origem das mensagens, como os sistemas operacionais mais comuns e os sistemas autônomos (*ASes*) que enviam mais mensagens. O trabalho apresentado nesta dissertação também caracteriza como a rede é abusada por *spammers*, mas, ao invés de tratar as mensagens como um conjunto único, as mensagens são primeiramente agrupadas em

campanhas e então é analisado como cada grupo de endereços IP disseminou cada campanha.

O artigo de Li e Hsieh (Li & Hsieh, 2006) é outro trabalho recente que analisa as estratégias de disseminação de *spams*. Os autores agruparam as mensagens pela URL contida no conteúdo e analisaram a estrutura do grafo que representa as relações entre endereços IP e URLs em que uma aresta entre um IP e uma URL significa que o IP enviou uma mensagem referenciando a URL. Eles analisam algumas propriedades desse grafo, com destaque para o surgimento de grandes grupos de IPs que enviam mensagens referenciando a mesma URL. De certa forma, os autores unificaram mensagens em suas respectivas campanhas, porém, considerando apenas os casos em que as URLs de cada mensagem eram idênticas. Porém, a técnica apresentada nesta dissertação agrupa mensagens em campanhas mesmo que as URLs contidas em cada uma não sejam idênticas.

Uma outra corrente de trabalhos de caracterização de *spams* foca em entender o conteúdo gerado por *spammers*. Esses trabalhos verificam características das mensagens como o tipo das mensagens, a presença de *links*, imagens e anexos (Dhinakaran et al., 2007) e mostram propriedades das URLs inseridas pelos *spammers* no corpo das mensagens (Georgiou et al., 2008), como os domínios mais populares e a dinamicidade dos endereços.

Por fim, (Gomes et al., 2007) apresentou uma extensa caracterização de cargas de trabalho de *spams*. Os autores derivaram modelos para representar a taxa de chegada de *spams* e o tamanho das mensagens, apresentando comparações com cargas de trabalho de *e-mails* legítimos. Os autores mostraram, por exemplo, que enquanto o envio de mensagens legítimas exibe padrões temporais diários e semanais característicos, com picos em determinados momentos do dia e da semana, o envio de *spam* não exibe nenhuma diferença significativa em termos de volume ao longo do período analisado. Uma outra diferença interessante entre *spams* e mensagens legítimas apontada pelos autores é que, apesar de ambos apresentarem uma distribuição Log-Normal no tamanho médio das mensagens, as mensagens legítimas são, em média, de seis a oito vezes maiores.

Um interessante tópico de pesquisa relacionada a *spam* é a caracterização da natureza evolutiva do problema. O trabalho de Pu e Webb (Pu & Webb, 2006) apresenta algumas análises acerca da evolução temporal dos *spammers* no que se refere ao uso de técnicas para construir suas mensagens. Essas técnicas foram computadas a partir das características identificadas nas mensagens pelo filtro *Spam Assassin* (SpamAssassin, 2007). Os autores mostraram que, ao longo do tempo, algumas técnicas de ofuscação deixam de ser usadas, muitas vezes em virtude de mudanças no ambiente, como a correção de alguma falha de segurança nos programas clientes de *e-mail*. Por outro lado, outras estratégias conseguem persistir por mais tempo.

A principal diferença entre os trabalhos aqui apresentados e esta dissertação é que, em geral, os trabalhos anteriores avaliam as mensagens como um todo ou individualmente, não havendo uma abordagem sistemática de entendimento das estratégias de disseminação de campanhas. O trabalho possivelmente mais semelhante à esta dissertação foi apresentado no SIGCOMM 2008 (Xie et al., 2008). Os autores apresentaram uma técnica para identificar

campanhas de *spam* a partir da derivação de expressões regulares que capturam o padrão de ofuscação de URLs gerado a partir de cada campanha. Em seguida, características das campanhas foram descritas, como o curto tempo pelo qual cada URL permanece ativa e acessível e o fato de que as campanhas são enviadas de muitas origens distintas.

Capítulo 3

Coleta de Dados

Este Capítulo descreve a metodologia aplicada para coletar mensagens de *spam* analisadas no trabalho e as informações de abuso da rede associadas a cada uma. Em seguida, são apresentados números gerais da coleta realizada.

3.1 Infraestrutura de Coleta de Dados

A arquitetura de coleta de dados incluiu um conjunto de sensores que implementam *honeypots* de baixa-interatividade (Provos & Holz, 2007). Genericamente, *honeypots* são recursos computacionais dedicados a serem sondados, atacados ou comprometidos, em um ambiente que permita o registro e controle dessas atividades.

Entre as vantagens da implantação de *honeypots* para estudos de segurança, podemos citar, principalmente, a baixa incidência de falsos positivos (todo o tráfego direcionado aos *honeypots* é, potencialmente, malicioso ou anômalo) e a capacidade de capturar comportamentos novos e inesperados (Jakobsson & Myers, 2006). Por outro lado, as principais desvantagens são a de que, em geral, *honeypots* fornecem uma visão limitada de ataques e abusos, registrando apenas aqueles que foram direcionados explicitamente a eles, e o fato de que originadores de abusos podem perceber que a máquina abusada é um *honeypot*, o que comprometeria a relevância e aplicabilidade do conceito.

Neste trabalho, foram utilizados dados coletados de *honeypots* de baixa-interatividade, que constituem uma categoria de *honeypots* que implementam apenas emulações dos serviços a serem abusados, e não implementações reais (Jakobsson & Myers, 2006).

Esses *honeypots* implementam emuladores de *proxies* abertos e *relays* abertos, que são máquinas na rede tradicionalmente abusadas por *spammers*. Um *proxy* é um servidor que atua como intermediário, realizando conexões em nome de outros clientes. Um *proxy* aberto permite conexões de qualquer origem para qualquer endereço IP e porta de destino e permitem ao *spammer* esconder sua origem, pois a máquina abusada enxergará apenas o *proxy* como originador do abuso (Oudot, 2003). Dois protocolos comumente utilizados para estabelecimento de conexões *Proxy* são o HTTP e o SOCKS. Um *Proxy* HTTP permite que o cliente

requisite um documento na Web e o *Proxy* procura pelo documento em seu *cache*. Se encontrado, o documento é retornado imediatamente. Caso contrário, o *Proxy* busca o documento no servidor remoto, entrega-o ao cliente e salva uma cópia no seu *cache*. Isso permite uma diminuição na latência, já que o servidor *Proxy*, e não o servidor original, é acessado, proporcionando ainda uma redução do uso de banda. Quando esse *Proxy* HTTP é mal-configurado, ele aceita intermediar conexões para, por exemplo, disseminar *spams*. O protocolo SOCKS é um outro protocolo de Internet projetado para permitir que aplicações cliente-servidor utilizem transparentemente o serviço de uma rede, mesmo que elas estejam atrás de um *firewall*, e ele também é abusado por *spammers* de forma similar para entregar mensagens de *spam* enquanto mantém o seu anonimato. O número de *proxies* abertos disponíveis na Internet é significativo, e existem sítios na Internet que vendem listas de *proxies* abertos. As ferramentas de *bulk mail* utilizadas por *spammers* também contam com funcionalidades para enviar as mensagens por meio desses *proxies* (Stern, 2008).

Os *honeypots* também implementaram emulações de *relays* abertos, que, ao contrário de um servidor de *e-mail* (*relay*) corretamente configurado, permitem a entrega de mensagens de qualquer origem para qualquer destinatário (Boneh, 2004). Por serem servidores de correio eletrônico onde o remetente e o destinatário não são usuários do servidor em questão, os *relays* abertos também são abusados por *spammers* para entrega de *spams* na rede. De acordo com o protocolo SMTP, mensagens legítimas são enviadas entre o servidor de *e-mail* do emissor da mensagem e o servidor do destinatário, passando, no meio do caminho, por servidores intermediários conhecidos como *mail relays*. A cada *mail relay* (servidor de correio) pelo qual a mensagem passa, uma linha `Received:` é adicionada ao cabeçalho da mensagem e, portanto, o cabeçalho da mensagem identifica toda a cadeia de máquinas pelo qual a mensagem passou, desde que ele não seja manipulado. Dessa forma, *spammers* obtém anonimato manipulando a lista `Received:` do cabeçalho SMTP e enviando os *spams* por meio de *relays* abertos, para que o receptor da mensagem não consiga descobrir a real origem do *spam*.

Cada um desses serviços são emulados em determinadas portas TCP, que são então abusadas por *spammers*. As portas emuladas estão listadas na Tabela 3.1 (Steding-Jessen et al., 2008):

Tabela 3.1. Portas emuladas pelos *honeypots*

Protocolo	**Portas TCP**
SMTP	25
HTTP	80, 81, 2282, 3128, 3332, 3382, 3802, 4480, 5490, 6588, 8000, 8080, 8090, 11120, 57123, 63809, 65506
SOCKS	559, 1029, 1080, 1202, 1813, 1978, 1979, 2280, 2425, 3127, 3380, 3800, 4471, 4777, 4894, 5748, 6042, 7531, 9938, 10000, 10001, 10232, 11117, 15859, 19086, 24971, 24972, 24973, 30021, 30022, 35612, 38994, 40934, 41457, 57123, 63808

A Figura 3.1, extraída de (Steding-Jessen et al., 2008), exibe como os *proxies* e *relays* abertos (emulados) implantados coletam mensagens. Um *spammer*, buscando esconder sua

origem real, procura máquinas na rede que envie mensagens em seu nome, sejam *proxies* abertos ou servidores de *e-mail* mal-configurados (*relays* abertos). *Spammers* mais sofisticados, inclusive, conseguem encadear várias máquinas em sequência. Por exemplo, um *spammer* pode abusar vários *proxies* abertos em sequência, seguido de um abuso a um *relay* aberto, para finalmente, entregar a mensagem ao servidor de correio da vítima. Outra possibilidade é o encadeamento de *proxies* abertos com *bots*, que são máquinas de usuários finais na rede que são infectadas e, controladas remotamente, se tornam meios para disseminação de ataques, furto de dados e envio de *spam*. As estratégias de encadeamento de máquinas para disseminação de *spams* são discutidas no Capítulo 5.

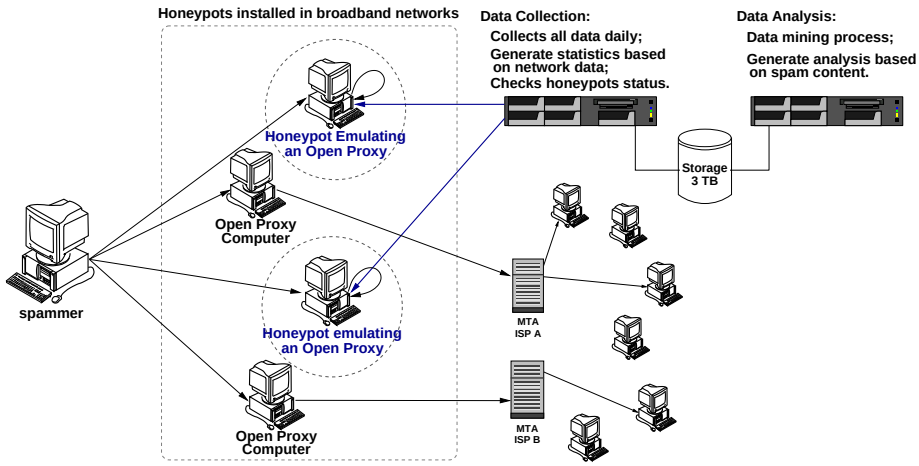

Figura 3.1. *Honeypots* abusados por *spammers*

A implantação dos *honeypots* não fez parte desta dissertação e foi conduzida por uma equipe de técnicos do Comitê Gestor da Internet no Brasil (CGI.br) (Steding-Jessen et al., 2008), que utilizou os subsistemas de emulação SMTP e HTTP implementados pelo *software* de código aberto `Honeyd` (Provos & Holz, 2007), complementados pela implementação de um emulador SOCKS *Proxy*, em suas versões 4 e 5.

Foram implantados 10 *honeypots* em 5 redes de banda larga brasileiras, tanto em redes a cabo quanto redes ADSL (Steding-Jessen et al., 2008). Os *honeypots* não coletam os dados no destino final dos *spams*, como em *spam traps* e *logs* de servidores de *e-mail*. Ao capturar os *spams* a partir do abuso a *proxies* e *relays* abertos, foi possível estabelecer uma visão de "dentro" da rede, antes das mensagens serem entregues. Essa visão é fundamental para a metodologia proposta neste trabalho para determinação de estratégias de disseminação de *spams*.

Para cada conexão direcionada aos módulos do `Honeyd`, foram registrados a data e a hora da conexão, o endereço IP de origem, endereço IP de destino, porta TCP abusada nos *honeypots* e porta TCP que o *spammer* tentou abusar no destino da conexão, assim como o protocolo utilizado (HTTP, SOCKS ou SMTP). O `Honeyd` também registrou o Sistema Operacional associado a cada IP de origem em cada conexão TCP, obtido a partir de téc-

nicas de *fingerprinting* passivo (Provos & Holz, 2007). O cabeçalho e conteúdo completo das mensagens que os *spammers* tentaram entregar por meio dos *honeypots* também foram armazenados.

Para cada endereço IP de origem e destino, foi obtido o *Sistema Autônomo* (AS) e o *Country Code* (CC) associado, a partir do mapeamento organizado pela norma ISO 3166 (ISO, 2006).

Todo *honeypot* precisa, necessariamente, oferecer mecanismos de contenção do tráfego de saída. No caso dos *honeypots* implantados para a coleta dos dados considerados neste trabalho, nenhuma mensagem era efetivamente entregue aos destinatários. Para evitar que os *spammers* notassem que as máquinas abusadas eram *honeypots* (Krawetz, 2004), era necessário fazê-los acreditar que as mensagens eram entregues com sucesso. Porém, as mensagens eram guardadas localmente e nunca entregues aos destinatários. Por isso, a arquitetura proposta pelos desenvolvedores do projeto (Steding-Jessen et al., 2008) contou com recursos para entregar apenas mensagens específicas de teste enviadas por *spammers*. Essas mensagens possuíam algumas características bem definidas, como o endereço IP do *honeypot*, a porta e o protocolo abusado no campo *assunto* da mensagem.

3.2 Estatísticas Gerais dos Dados Coletados

A Tabela 3.2 exibe uma visão geral dos dados considerados neste trabalho, a partir de um ano de coleta em 10 *honeypots* emulando *proxies* e *relays* abertos. Neste período, 350,5 milhões de mensagens foram consideradas. O número de mensagens únicas (cerca de 32 milhões) e de URLs únicas (mais de 6 milhões) também é expressivo.

Tabela 3.2. Visão geral dos dados analisados

Característica	**Valor**
Período	08/07/2006 à 23/06/2007
Endereços IP	160.291
Sistemas Autônomos	2.557
Mensagens	350.565.583
Mensagens únicas	32.111.981
Spams com URLs	318.881.218 (91%)
URLs únicas	6.413.148

A Tabela 3.3 mostra o número de mensagens enviadas por cada serviço emulado provido pelos *honeypots*. Fica claro que abusos aos serviços de *Proxy* (HTTP e SOCKS) são muito mais frequentes que os abusos à porta 25 (SMTP), que responde por menos de 2% dos abusos. No entanto, as proporções são diferentes quando verificadas as portas que são alvo dos *spammers* na máquina de destino das conexões que são intermediadas pelos *honeypots*. Nesse caso, a grande maioria das conexões (99,6%) são destinadas à porta de *relay* das máquinas.

Esses abusos originaram-se de 154 *Country Codes* distintos, e observa-se uma grande concentração em apenas 5 países: Taiwan, China, Estados Unidos, Canadá e Japão, conforme

Tabela 3.3. Número de mensagens enviadas por tipo de abuso

Serviço abusado nos *honeypots*		Serviço abusado na máquina destino	
Tipo de serviço	**Porcentagem**	**Tipo de serviço**	**Porcentagem**
proxy HTTP	61,9%	*relay* (porta 25)	99,6%
proxy SOCKS	36,8%	outras portas	0,4 %
relay (porta 25)	1,3%		

Tabela 3.4. *Country Codes* (CC) mais frequentes de origem dos *spams*

#	CC	*mensagens*	%
01	TW	253.099.364	72,28
02	CN	53.027.016	15,14
03	US	21.160.642	6,00
04	CA	5.165.242	1,46
05	JP	4.540.936	1,29
06	KR	3.573.585	1,00
07	HK	3.216.452	0,91
08	UA	1.426.348	0,41
09	DE	829.705	0,23
10	BR	655.796	0,19

a Tabela 3.4. Apenas Taiwan é responsável por 72,28% das mensagens enviadas, e poucos abusos se originam de máquinas instaladas na própria rede brasileira (0,16%).

Esses números globais, embora forneçam uma ideia geral dos *spams* que trafegam na rede brasileira, não são suficientes para explicar e determinar o comportamento dos *spammers*, pois os dados são agregados. Em decorrência do grande volume de dados coletados, existem correlações e padrões implícitos entre as grandezas consideradas que não são visíveis ao se tratar a coleção de mensagens como um todo, o que motivou o projeto da metodologia de caracterização de estratégias de disseminação de *spams* proposta neste trabalho a partir de técnicas de mineração de dados, descrita no próximo Capítulo.

Capítulo 4

Estratégias de Disseminação de Campanhas de *Spam*

Neste Capítulo, é apresentada a metodologia de caracterização de estratégias de disseminação de campanhas de *spam*, assim como os resultados que emergiram da aplicação da metodologia ao conjunto de dados descrito no Capítulo 3. A metodologia divide o processo de caracterização em um procedimento de duas partes. Primeiro, são gerados perfis para cada campanha, determinando as características compartilhadas pelos *spams* que formam cada uma delas. Em seguida, os invariantes identificados entre as campanhas determinam diferentes estratégias de disseminação de *spams*. As estratégias apresentadas foram divididas em duas categorias: estratégias de geração de conteúdo das campanhas, apresentadas na Seção 4.2, e padrões de abuso da rede, apresentados na Seção 4.3.

4.1 Identificação de Campanhas de *Spam* utilizando Árvore de Padrões Frequentes

Após a coleta dos dados, o passo seguinte da metodologia proposta nesta dissertação correspondeu à determinação de grupos de mensagens que possuam o mesmo objetivo e uma mesma estratégia de disseminação, compondo uma *campanha de spam*. Ao agrupar mensagens em suas respectivas campanhas, o objetivo é minimizar os efeitos das técnicas de ofuscação empregadas por *spammers*, que sistematicamente alteram o conteúdo das mensagens enviadas, seja no corpo da mensagem ou no assunto (Sophos.com, 2004). As Figuras 4.1, 4.2 e 4.3 ilustram a importância de se agrupar as mensagens em campanhas como etapa importante para fins de caracterização de estratégias de atuação de *spammers*. A Figura 4.1 representa o estado original dos dados após coletados, em que cada endereço IP (à esquerda) envia mensagens diferentes, devido ao esforço de ofuscação de cada *spammer*. Por conta disso, o comportamento de cada *spammer* é fragmentado e análises das estratégias para disseminação dos abusos são limitadas. Como passo intermediário para solucionar este problema, organiza-

se as mensagens em grupos (campanhas), conforme ilustrado na Figura 4.2. Dessa forma, é possível isolar o tráfego associado a cada abuso e caracterizá-los em termos de volume, duração e dispersão geográfica, entre outros critérios (Figura 4.3).

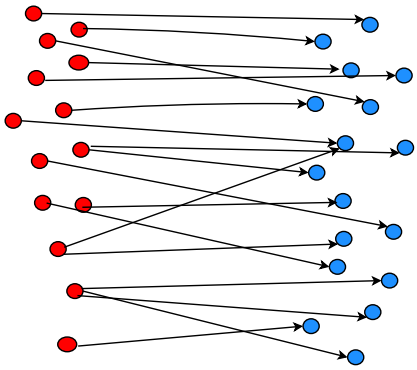

Figura 4.1. Mensagens de *spam* disseminadas por endereços IP: comportamento fragmentado

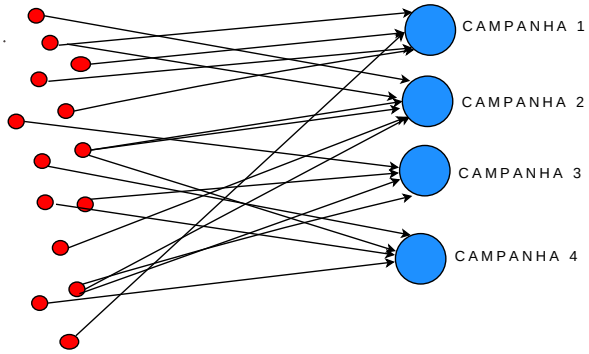

Figura 4.2. Agrupamento de mensagens em campanhas

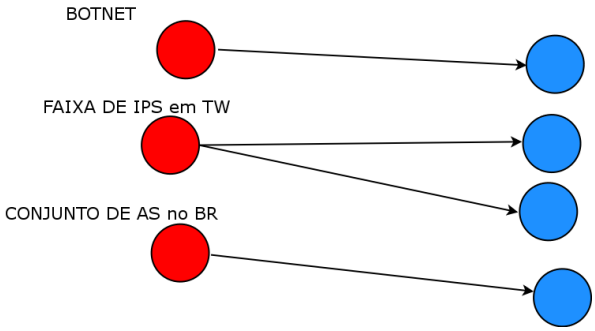

Figura 4.3. Análise do comportamento de cada campanha de *spam*

A premissa básica da técnica para identificar campanhas proposta neste trabalho é a de que um *spammer*, ao disseminar uma campanha, mantém estáticas algumas partes e fragmentos da mensagem e altera outras seções do conteúdo, de forma sistemática e automatizada, a partir das ferramentas de *Bulk Mail* (Stern, 2008). Essas ferramentas oferecem recursos para

personalização e ofuscação das mensagens, como a inclusão de textos aleatórios no cabeçalho e corpo do *e-mail* (Figura 4.4). Por exemplo, cada mensagem de uma campanha pode conter termos ligeiramente diferentes no campo *assunto*, embora outros termos estejam sempre presentes, já que o *spammer* precisa manter o assunto legível e um nível excessivo de ofuscação pode reduzir a probabilidade da mensagem ser lida. Outras formas de geração automatizada de campanhas com conteúdo diferente incluem a inserção do nome da vítima no texto e a inserção de fragmentos aleatórios nas URLs contidas no corpo da mensagem, a fim de prevenir que elas sejam identificadas e bloqueadas.

A técnica desenvolvida nesta dissertação para identificação de campanhas de *spam* explora essas propriedades das ferramentas de *Bulk Mail* e é composta de duas etapas. Na primeira etapa, são extraídas características relevantes de cada mensagem de *spam*. Essas características são:

1. idioma
2. *layout*
3. composição (tipo) da mensagem (HTML, texto, imagem e combinações)
4. URLs contidas no corpo da mensagem
5. assunto

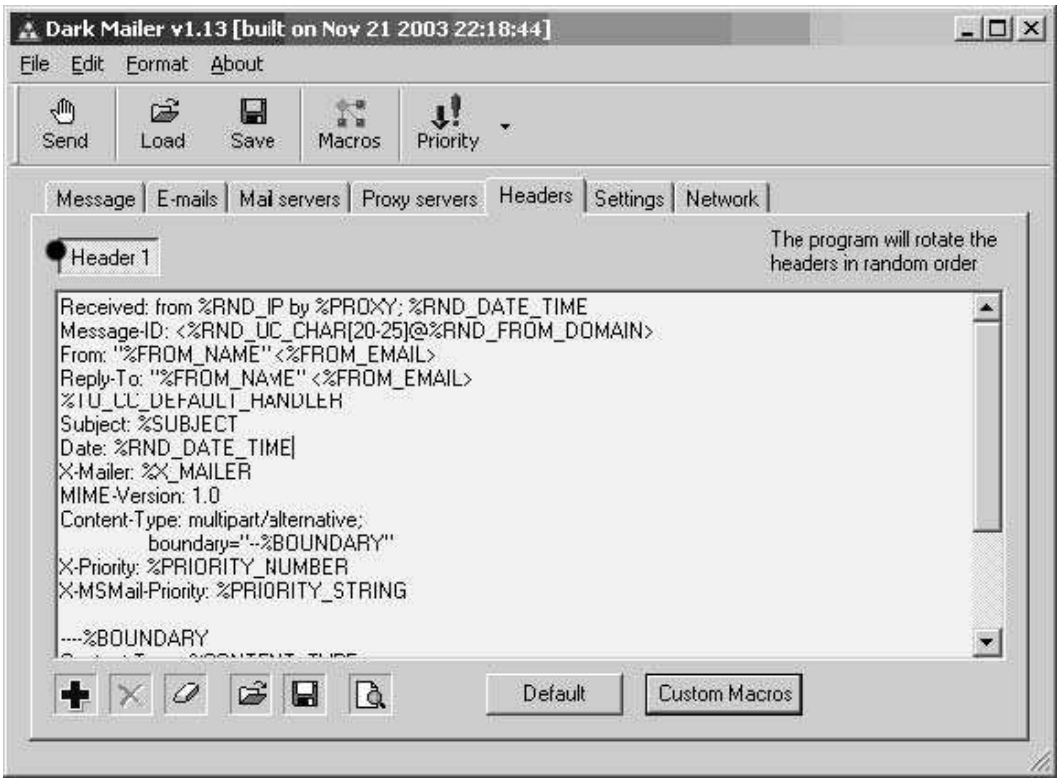

Figura 4.4. Típica ferramenta de *Bulk Mailer*

A seguir descreveremos em mais detalhes como essas características foram extraídas das mensagens.

A determinação do idioma de cada mensagem foi implementada a partir de uma técnica descrita na literatura que se baseia na computação de *n-gramas* (Cavnar & Trenkle, 1994). A técnica se baseia no fato de que cada idioma possui um conjunto representativo de *n-gramas*, que são sequências de caracteres de tamanho N. O algoritmo extrai *n-gramas* de textos pré-classificados de cada idioma e monta um modelo de classificação que considera os *n-gramas* mais frequentes em cada idioma. Quando um texto novo precisa ser classificado, a sua sequência de *n-gramas* é comparada com as sequências de cada idioma e o idioma da mensagem é aquele que possui a sequência característica de *n-gramas* mais similar. A técnica apresenta índices de acerto bastante satisfatórios, variando entre 80,0% e 99,8% (Cavnar & Trenkle, 1994). Além de ser uma característica importante para separar as mensagens de campanhas diferentes, o idioma também cumpre um papel crucial para determinar as estratégias de disseminação dos *spams*, já que ele é um indicador do público-alvo dos *spams*.

O *layout* da mensagem é uma codificação que mapeia as propriedades de formatação da mensagem para uma sequência de caracteres, a partir da proposta de Claudiu Musat (Musat, 2006). Por exemplo, uma mensagem com duas linhas de texto, seguidas de uma URL e duas linhas em branco seria mapeada para o *layout* TTUBB. O *layout* é um invariante importante a ser considerado com vistas ao objetivo de agrupar mensagens em suas respectivas campanhas, já que é uma prática comum entre os *spammers* manter o aspecto geral de cada mensagem intacto, como ilustrado na Figura 4.5.

O assunto da mensagem também se mostra um invariante importante, pois, embora *spammers* insiram componentes aleatórios nessa parte da mensagem, a fim de evitar que sejam bloqueados, é importante que eles mantenham alguns termos fixos e legíveis, para que haja alguma chance da mensagem ser efetivamente lida pelos destinatários. O assunto da mensagem é quebrado em fragmentos (separados por espaço em branco), para que esses trechos do campo possam ser identificados.

Além do idioma, composição (tipo), *layout* e assunto, a informação das URLs contidas nas mensagens também desempenha um papel crucial para agrupar as mensagens em suas respectivas campanhas (Yeh & Lin, 2006). Cada URL encontrada nas mensagens é quebrada em fragmentos (separados por "/", "." e "?") que são considerados características independentes.

A partir das características extraídas de cada mensagem, o segundo passo da técnica prevê a construção de uma árvore de padrões frequentes, também chamada *FP-Tree* (Tan et al., 2005). A *FP-Tree* é uma estrutura de dados proposta inicialmente para minerar conjuntos frequentes em grandes bases de dados eficientemente (Han et al., 2004), por meio de uma representação compacta da base de dados em uma estrutura de dados do tipo árvore. Para representar uma base de dados transacional de forma compacta a partir de uma Árvore de Padrões Frequentes, os seguintes passos devem ser seguidos:

1. percorrer a base de dados uma vez, para computar a frequência de cada item;

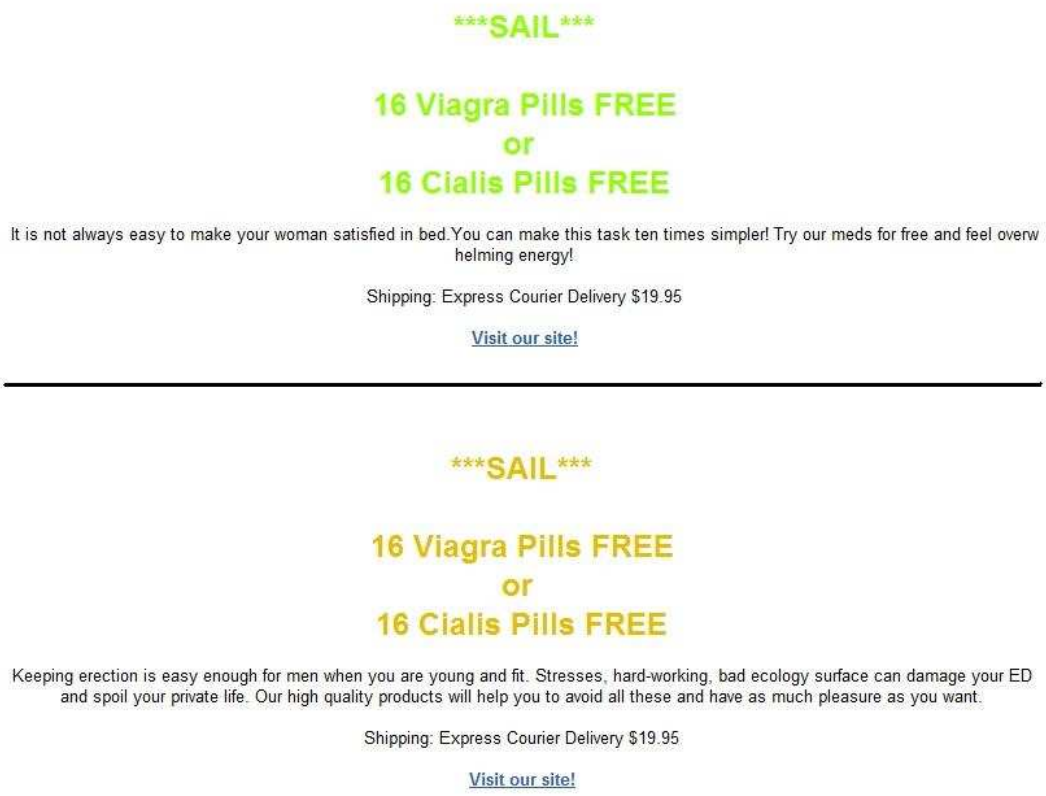

Figura 4.5. Mensagens de *spam* de uma mesma campanha, com texto diferente e *layout* idêntico (BTBTTTBTTBTBU)

2. percorrer a base de dados e ordenar os itens de cada transação em ordem decrescente, de acordo com a frequência de ocorrência obtida no passo 1;

3. criar uma árvore de padrões frequentes. Nessa árvore, os itens de uma mesma transação são inseridos em um mesmo caminho na árvore e os itens mais frequentes são inseridos nos níveis mais altos na árvore. Dessa forma, transações que compartilhem um prefixo comum (ou seja, uma sequência ordenada de itens) compartilham um caminho na árvore, o que permite que menos espaço seja gasto para representar a mesma informação.

A Tabela 4.1 ilustra essa abordagem a partir de uma amostra de uma base de dados hipotética, extraída de (Han et al., 2004), e a listagem dos itens de cada transação ordenados pela frequência de cada um. A frequência de cada item é apresentada na Tabela 4.2 e, na árvore resultante (Figura 4.6), as transações 100 e 200 compartilham o mesmo caminho na árvore, devido ao prefixo frequente (f,c,a) presente nas duas transações. Por sua vez, as transações 300, 400 e 500 compartilham o item frequente *b*.

A representação de uma base de dados transacional em uma árvore do tipo *FP-Tree* tende a ser compacta em relação à representação original porque cada prefixo é armazenado

Tabela 4.1. Amostra de uma base de dados hipotética

transação	Itens	Itens ordenados
100	b, c, m, p, f, a	f, b, a, m, p
200	a, b, c, f, m, x	f, b, a, m, x
300	b, h, j, o	b, o, h, j
400	b, c, k, s	b, c, k, s
500	b, c, e, l	b, c, e, l

Tabela 4.2. Frequência de cada item na base de dados hipotética

#	Item	Frequência
1	f	450
2	b	400
3	a	300
4	c	200
5	m	180
6	p	160
7	o	100
8	h	80
9	j	70
10	e	60
11	l	40
12	k	40
13	s	20
14	x	10

uma única vez na árvore, representando conjuntos de itens que são frequentes na base de dados. Essa propriedade da árvore inspirou sua aplicação como uma técnica de agrupamento de mensagens de *spams* em suas respectivas campanhas, conforme será descrito a seguir.

A árvore de padrões frequentes foi adaptada para que cada transação considerada seja o conjunto de características extraídas de cada mensagem. Dessa forma, cada nó da árvore, com exceção da raiz, representa uma característica extraída das mensagens de *spam* que é compartilhada pelas sub-árvores abaixo. Cada caminho na árvore representa conjuntos de características que co-ocorrem nas mensagens, em ordem decrescente de frequência das ocorrências. Como resultado, duas mensagens que possuem muitas características frequentes em comum (como idioma, tipo e *layout*) e diferem apenas por características infrequentes vão compartilhar um mesmo caminho na árvore. A raiz da árvore de padrões frequentes é um nodo vazio, criado para separar sub-árvores que não possuem nenhuma característica em comum.

Os fragmentos aleatórios tradicionalmente inseridos por *spammers* no corpo das mensagens fazem com que o número de elementos nos níveis mais baixos da árvore aumente significativamente, e este é o ponto em que as campanhas são delimitadas, isto é, todas as mensagens que pertencem às sub-árvores anteriores ao aumento significativo no número de nós do caminho são agrupadas na mesma campanha. A Figura 4.7 exibe a distribuição acumulada

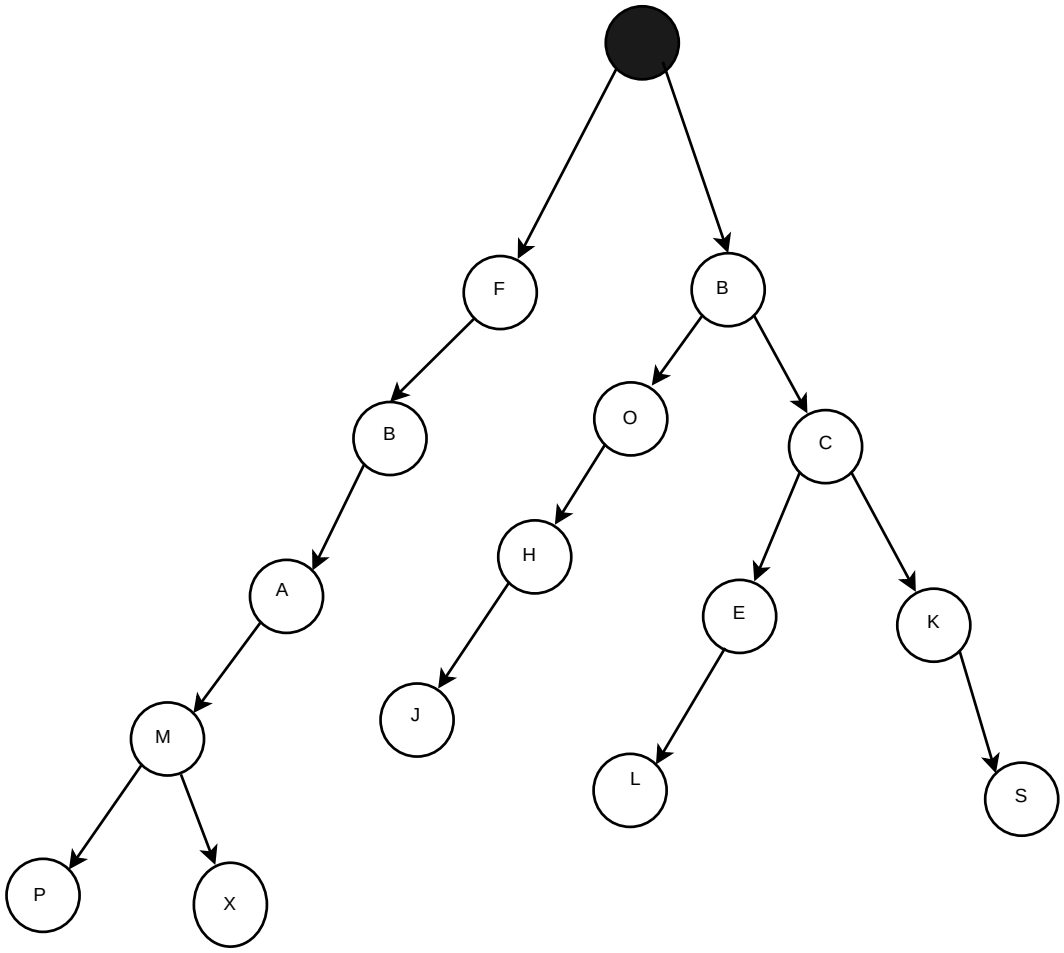

Figura 4.6. Árvore de Padrões Frequentes para a amostra da Base de Dados

da frequência de ocorrência de cada característica extraída no conjunto de dados analisado. Pode-se notar que cerca de 90% das características ocorrem menos de 10 vezes, em meio a 350 milhões de mensagens analisadas, o que comprova o esforço de ofuscação e inserção de termos aleatórios empreendido por *spammers* e que suporta a estratégia proposta.

A Tabela 4.3 exibe algumas características extraídas de mensagem de *spam* hipotéticas. As duas primeiras mensagens são bastante parecidas; no entanto, elas possuem termos diferentes no campo *assunto* e no parâmetro da URL. No entanto, o número de características comuns entre as duas mensagens é grande: a composição (HTML), o idioma (inglês), termos no campo *assunto* ("Buy", "Viagra" e "Now"), o *layout* e vários fragmentos da URL ("http", "www", "buyviagraonline" e "com"). Esses 10 invariantes agrupam as duas mensagens na mesma campanha, apesar das diferenças em um fragmento da URL e no assunto. Como o *spammer* gera vários fragmentos diferentes nessas posições (remetendo ao *e-mail* e nome de cada vítima), essas características serão infrequentes em relação aos invariantes listados anteriormente, e portanto, serão inseridos nas folhas da árvore. Os invariantes podem ser entendidos como a assinatura da campanha, pois são características que discriminam uma campanha de outras, que possuirão sequências de invariantes distintas.

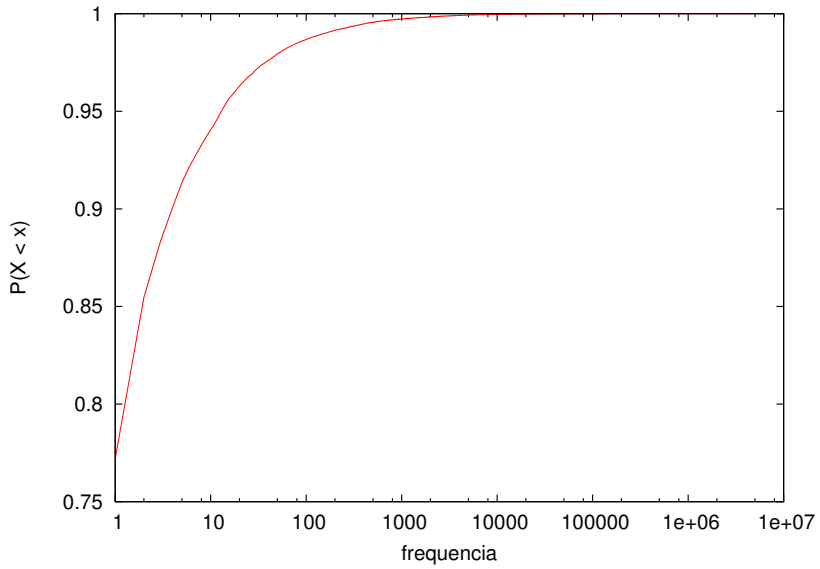

Figura 4.7. Frequência de ocorrência de cada característica (CDF)

Tabela 4.3. Exemplos de características extraídas de mensagens de *spam*

Mensagem	Atributo	Valor
1	Tipo	HTML
	Idioma	inglês
	Assunto	Pedro, Buy Viagra Now!
	Layout	TTTUBBB
	URL	`http://www.buyviagraonline.com?email=pedro@dominio.com`
2	Tipo	HTML
	Idioma	inglês
	Assunto	João, Buy Viagra Now!
	Layout	TTTUBBB
	URL	`http://www.buyviagraonline.com?email=joao@dominio.com`
3	Tipo	HTML + GIF
	Idioma	português
	Assunto	Compre seu Rolex!
	Layout	BBBTTTUUB
	URLs	`http://www.xmw.rolex.cjb.net?DFG567safaf@` e `www.rolex.com`
4	Tipo	HTML + GIF
	Idioma	português
	Assunto	Compre seu Rolex!
	Layout	BBBTTTUUB
	URLs	`http://www.qmt.rolex.cjb.net?DDGGH66554A` e `www.rolex.com`

No caso das mensagens 3 e 4, um padrão similar é detectado. As duas mensagens compartilham características frequentes, sendo ambas escritas em português, compostas por um anexo HTML e uma imagem GIF e o mesmo *layout*, além do assunto e uma URL em comum (`www.rolex.com`). Portanto, elas também serão agrupadas em uma mesma campanha.

Pode-se citar três vantagens principais do método proposto para agrupar mensagens de *spam* em suas respectivas campanhas. A principal delas é que a técnica consegue se adaptar

à evolução dos *spammers*, no sentido de que os padrões de ofuscação não são pré-definidos; ao contrário, eles são determinados naturalmente: em nenhum momento é especificado quais atributos devem estar presentes nos primeiros níveis ou nas folhas da árvore. Essa propriedade da estratégia de identificação de campanhas é fundamental, pois seria inviável considerar todas as possíveis combinações de ofuscação. Por exemplo, *spammers* capazes de manipular sub-domínios conseguem inserir termos aleatórios em qualquer posição da URL, como pode ser observado nas mensagens 3 e 4 apresentadas na Tabela 4.3. Além disso, a árvore é extensível, pois novas características podem ser adicionadas à árvore sem necessidade de alterar o algoritmo básico de funcionamento da técnica. Por exemplo, embora termos do texto do corpo de cada mensagem não tenham sido utilizados neste trabalho, nada impede que a árvore de padrões frequentes seja estendida para considerá-los.

Uma outra característica relevante da técnica é o fato de que a árvore de padrões frequentes não se limita a detectar as campanhas, mas ela também descreve como o conteúdo de cada campanha foi gerado, a partir da distribuição de características entre os níveis da árvore. Por exemplo, a partir da árvore construída com as mensagens da Tabela 4.3, é possível determinar que a campanha associada às mensagens 1 e 2 foi gerada a partir da inserção de termos aleatórios no assunto e no parâmetro da URL e pela fixação das demais características. No caso da campanha resultante das mensagens 3 e 4, determinou-se que o *spammer* ofuscou o próprio domínio de uma das URLs da mensagem, o que não seria determinado por outras técnicas (Xie et al., 2008).

Por fim, é importante mencionar que a abordagem proposta é escalável, porque as mensagens não são comparadas par-a-par, o que geraria um algoritmo quadrático no número de mensagens. Sendo m o número de mensagens a serem processadas e c o número de características consideradas em cada mensagem, o custo do algoritmo é o custo de computar as frequências de cada característica presente em cada mensagem ($O(m.c)$), ordenar as freqûencias ($O(c.m)*log(c.m)$) e inserir as características extraídas das mensagens na árvore ($O(c*m)$).

Entre as técnicas de detecção de duplicatas de documentos descritas na literatura, a abordagem com árvore de padrões frequentes se encaixa na categoria de métodos baseados em assinatura, já que a sequência de características na árvore define um identificador único para cada campanha. A abordagem, no entanto, não é sensível a textos aleatórios, o que é um problema comum de tais técnicas (Kolcz & Chowdhury, 2007). Ao contrário, a inserção de textos aleatórios por parte do *spammer* ajuda a definir um padrão de ofuscação que caracteriza a campanha em questão. Pode-se citar, ainda, que a árvore de padrões frequentes é uma abordagem mais elegante e genérica para o problema de determinação de campanhas de *spam*, por não se apoiar em nenhuma particularidade de tecnologias atualmente empregadas por *spammers*.

A seguir, serão apresentadas as estratégias de disseminação de *spams* a partir das campanhas determinadas com a técnica proposta. As estratégias foram divididas em dois grupos: estratégias de geração de conteúdo das campanhas (Seção 4.2) e estratégias de disseminação

das campanhas na rede (Seção 4.3).

4.2 Estratégias de geração de conteúdo

Após a aplicação da técnica para identificação de campanhas baseada em uma árvore de padrões frequentes no conjunto de dados descrito no Capítulo 3, 57.851 campanhas distintas foram identificadas. A redução é de cerca de duas ordens de grandeza; seis milhões de URLs únicas reduziram-se em seis dezenas de milhares de campanhas.

A Figura 4.8 exibe a função da distribuição acumulada do número de mensagens associadas a cada campanha. A maioria das campanhas é pequena, embora exista um número significativo de campanhas que disseminaram mais de 100.000 mensagens cada uma. Algumas dessas campanhas são exibidas na Figura 4.9 por meio de uma pequena porção da Árvore de Padrões Frequentes resultante, gerada com o *software* Pajek (Nooy et al., 2004). Na Figura, cada cor representa um atributo diferente (ex: *layout*, idioma, fragmento de URL) e o diâmetro de cada nodo é proporcional ao logaritmo da frequência da característica nas mensagens, evidenciando a presença dos invariantes nos primeiros níveis. Por exemplo, no caso de um *spammer* que envia mensagens com URLs que seguem o formato `www.domain.com?parameter=random`, todos os fragmentos aleatórios são inseridos na quinta posição da URL, o que caracteriza um padrão bem definido de ofuscação.

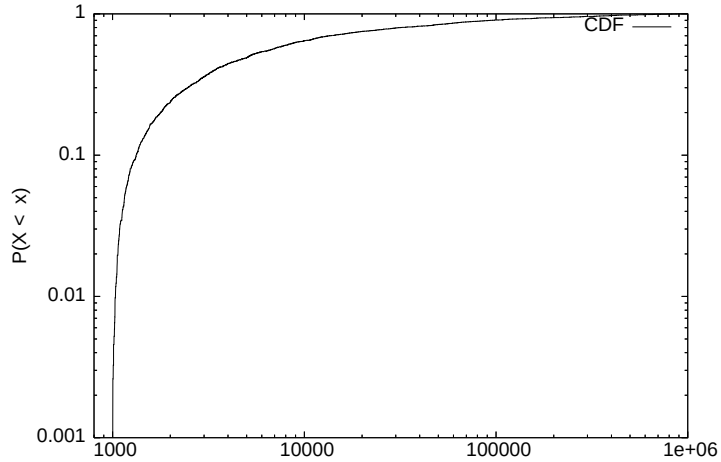

Figura 4.8. Número de mensagens em cada campanha (CDF)

A Figura 4.9 exibe três agrupamentos densos, no meio da Figura, que são campanhas que compartilham alguns atributos, até o nível 7 na árvore. Essas campanhas foram diferenciadas por padrões diferentes de ofuscação, determinados pela sequência de cores a partir desse nível. Essas três campanhas variaram alguns de seus atributos aleatoriamente, o que gerou o grande número de nodos nos níveis mais baixos da árvore para essas campanhas.

A Figura 4.10 exibe outras duas ramificações da árvore, e cada uma representa uma campanha diferente. Em ambas, é possível notar que cada campanha é detectada devido a dois tipos de padrões: uma sequência de invariantes, que são características compartilhadas

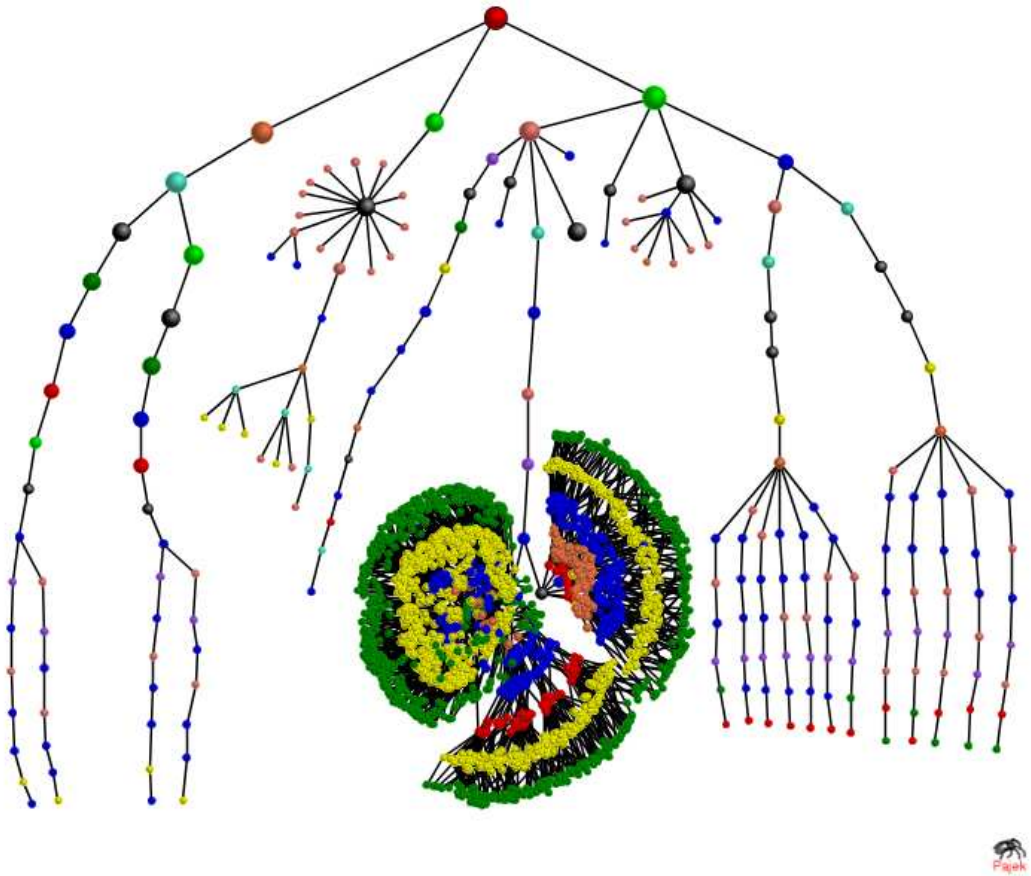

Figura 4.9. Diferentes campanhas identificadas pela Árvore de Padrões Frequentes

por todas as mensagens da campanha, e um conjunto de características ofuscadas que são instâncias diferentes de um mesmo atributo (representado por cada cor).

A fim de ilustrar como cada campanha explorou a ofuscação dos diversos atributos, determinou-se o número de instâncias diferentes para cada um dos cinco atributos considerados no trabalho (idioma, tipo da mensagem, *layout* e fragmentos do assunto e URLs). Um sumário é apresentado na Tabela 4.4. Nota-se que o tipo da mensagem e o idioma constituem-se em importantes invariantes; apenas 16 tipos (e combinações) e 21 idiomas distintos foram encontrados entre os *spams* analisados. O *layout* da mensagem também se mostra como uma importante característica para agrupar mensagens em campanhas, já que apenas 259.933 formatações distintas foram encontradas no conjunto de 350 milhões de mensagens.

Tabela 4.4. Número de Instâncias por Atributo

Característica	**Número de Instâncias**
Tipo de Mensagem	16
Idioma	21
Layout	259.933
Fragmentos do assunto	600.121
Fragmentos de URL	18.967.160

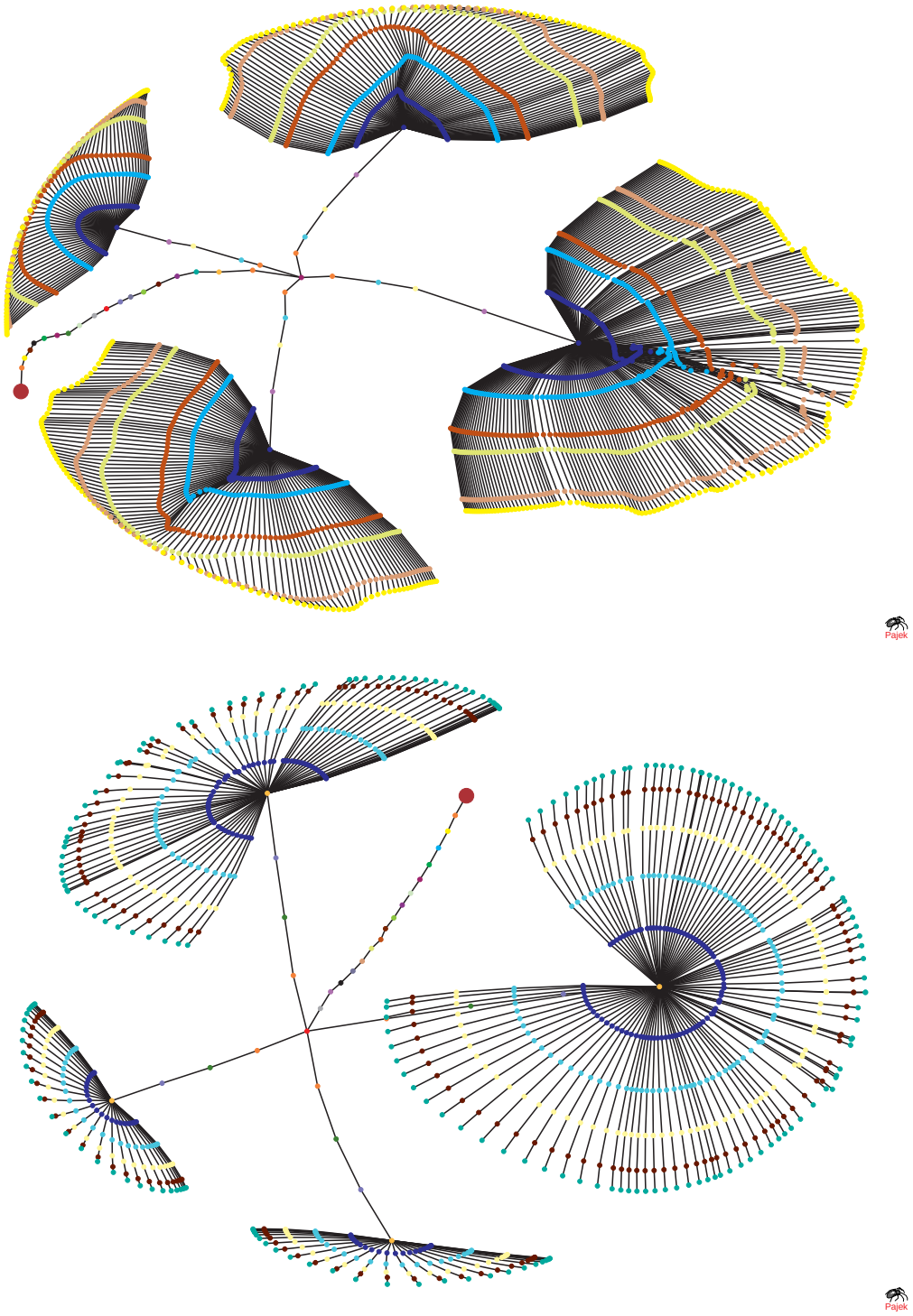

Figura 4.10. Campanhas identificadas pela Árvore de Padrões Frequentes

A análise da Árvore de Padrões Frequentes mostra que o idioma, em geral, é o primeiro atributo utilizado para separar as campanhas, o que já seria esperado, dado que comumente mensagens de uma mesma campanha são escritas no mesmo idioma. Em seguida, o tipo da mensagem e *layout* são usualmente encontrados na árvore e, normalmente, fragmentos de

URL são as características menos discriminativas, encontradas nas folhas. No entanto, elas não são menos importantes, pois ajudam a compor o padrão de ofuscação associado a cada campanha.

A partir da árvore, pode-se identificar, qualitativamente, três tipos de campanhas em termos de ofuscação de URLs: campanhas estáticas, campanhas com sub-campanhas e campanhas com ofuscação aleatória. Na amostra da Árvore de Padrões Frequentes exibida na Figura 4.9, esses três tipos podem ser observados. Campanhas fixas são aquelas em que o *spammer* mantém todas as características de suas mensagens fixas, incluindo URLs. Em geral, essas URLs são *links* curtos com nomes significativos e legíveis, como *buydvds.com*. Na árvore, essas campanhas possuem poucos nós, com profundidade 3 ou 4. Uma estratégia diferente corresponde à venda de diferentes produtos na mesma campanha, o que gera um conjunto de URLs em que cada URL corresponde a um produto diferente de um mesmo *web site*. Por exemplo, *dvd1.htm, dvd2.htm e dvd3.htm* são produtos diferentes associados à mesma campanha. Desde que o *spammer* mantenha outras partes da URL e o *layout* da mensagem fixo, essas mensagens distintas serão agrupadas na mesma campanha porque o fragmento da URL que especifica o produto é infrequente se comparado com as outras características da mensagem. Esse caso pode ser observado, na árvore, pelos caminhos em ambos os extremos da figura. Finalmente, a terceira categoria de campanha é aquela em que *spammers* ofuscam as URLs automaticamente por meio da inserção de fragmentos aleatórios. Os três agrupamentos no centro da Figura ilustram esse caso.

Em relação à composição das mensagens, nota-se que cerca de 80% das campanhas são constituídas de um único bloco HTML (Figura 4.11), enquanto 10% são formadas por um elemento HTML e uma imagem GIF. Embora menos frequentes, também são observadas campanhas formadas por anexos HTML e texto puro (5%), JPEG e HTML (4%), e mensagens constituídas apenas por um elemento de texto puro (3%).

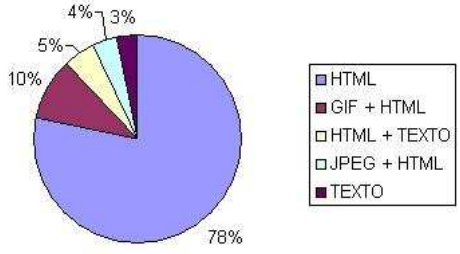

Figura 4.11. Composição das campanhas de *spam*, em porcentagens

As campanhas de *spam* também foram caracterizadas quanto ao número de URLs que foram inseridas nas mensagens associadas a essas campanhas. A Figura 4.12 mostra um histograma que exibe quantas URLs estavam contidas em cada campanha de *spam* detectada pela Árvore de Padrões Frequentes. Pode ser observado que uma pequena parcela das campanhas não possui nenhuma URL, enquanto a maior parte das campanhas possui entre 1 e 6 URLs. Chama a atenção o fato de que algumas campanhas possuem uma quantidade

expressiva de URLs, chegando a quase duas dezenas.

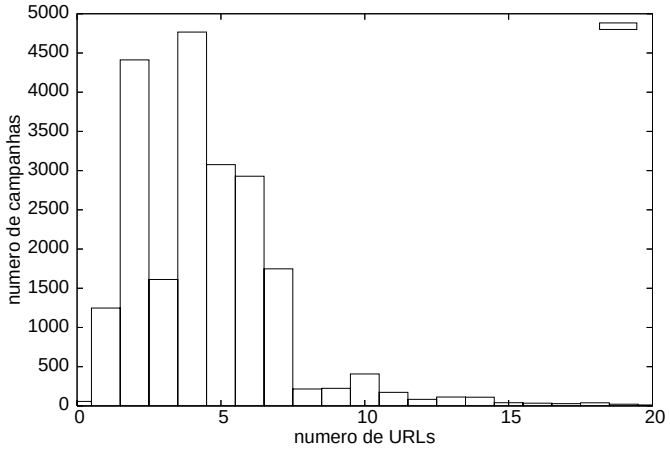

Figura 4.12. Número de URLs em cada campanha de *spam*

4.3 Estratégias de abuso à rede

Para caracterizar as estratégias de disseminação de *spams* na rede, buscou-se padrões e invariantes que descrevessem o comportamento dos *spammers* ao distribuir as mensagens de suas campanhas. Para tal, três grandes grupos de critérios foram considerados: a origem de cada abuso, o destino e o tipo de abuso. Os critérios origem e destino dos abusos consideram endereços IP, endereços de *e-mail* de origem e destino dos *spams* e também características como Sistemas Autônomos (*ASes*), países (*Country Codes*) e ISPs. O tipo de abuso corresponde, basicamente, aos protocolos de *Proxy* (HTTP e SOCKS) e *Relay* (SMTP) emulados pelos *honeypots* e descritos no Capítulo 3.

Como o protocolo que o *spammer* utilizou para abusar os *honeypots* consiste em um critério fundamental de análise das estratégias de disseminação de *spams* apresentada nesta seção, as definições de cada um serão recapituladas:

1. **Abuso a *proxies* abertos:** *proxies* são máquinas na rede que atendem a requisições repassando os dados a outros servidores. Um usuário (cliente) conecta-se a um servidor *Proxy*, requisitando algum serviço, como um arquivo, conexão, sítio da Web, ou outro recurso disponível em outro servidor. Quando mal-configurados, esses *proxies* se tornam *proxies abertos*, e são utilizados por *spammers* para entregar mensagens de forma anônima. Os dois protocolos de *proxy* emulados pelos *honeypots* que foram utilizados para coletar os dados considerados neste trabalho são o HTTP e o SOCKS.

2. **Abuso a *relays* abertos:** Na Internet, os servidores de correio, como `mail.ufmg.br` e `mail.hotmail.com`, aceitam se comunicar apenas com clientes que sejam donos de contas nos provedores e instituições em que estão hospedados. Quando esses servidores

são mal-configurados, eles aceitam entregar mensagens de qualquer origem para qualquer destino, tornando-se *relays* abertos e permitindo a disseminação de *spams* por meio deles. Os servidores de *e-mail* ficam ativos na porta TCP 25 (protocolo SMTP), e, portanto, as mensagens que chegam na porta 25 dos *honeypots* correspondem aos casos de *spammers* tentando abusar *relays* abertos.

A partir desses critérios, analisou-se o comportamento dos *spammers* na rede, isto é, a forma como cada campanha foi disseminada na rede em termos de distribuição geográfica, tipo de abuso, número de máquinas utilizadas para disseminá-la e os encadeamentos entre elas, que são estudados separadamente no Capítulo 5.

4.3.1 Números gerais

A fim de entender como *spammers* abusam os recursos da rede, combinou-se a análise dos valores totais coletados com a informação derivada das campanhas e aplicação de técnicas de mineração de dados.

A Tabela 4.5 exibe uma comparação dos números gerais para os três tipos de abuso registrados pelos *honeypots* (*proxies* HTTP e SOCKS, e *relays* abertos). Para cada tipo de abuso, são mostrados o número de mensagens que *spammers* tentaram entregar e o número de origens únicas dos abusos, para três granularidades distintas: endereços IP, Sistemas Autônomos (*ASes*) e *Country Codes* (CCs). As porcentagens não totalizam 100% porque algumas máquinas abusaram os *honeypots* de mais de uma forma.

Tabela 4.5. Números gerais dos abusos observados

Métrica	**Abuso: *Proxy* HTTP**	**Abuso: *Proxy* SOCKS**	**Abuso: *Relay* Aberto**
Mensagens	209.081.788 (59,7 %)	136.920.440 (38,9 %)	4.563.355 (1,3 %)
IPs únicos	70.769 (44,1 %)	64.803 (40,4 %)	64.318 (40,1 %)
ASes únicos	96 (3,8 %)	477 (18,7 %)	2467 (96,4 %)
CCs únicos	14 (9,9 %)	14 (9,9 %)	142 (100 %)
Mensagens / IP	2953	2098	62
Mensagens / AS	2.177.083	285.115	1.824

Observando a Tabela 4.5, percebe-se que abusos aos *honeypots* como *relays* abertos (emulando servidores de *e-mail* mal-configurados) são relativamente raros, correspondendo a apenas 1,3% do total de abusos. No entanto, as proporções são bastante diferentes ao se verificar a distribuição das origens dos abusos: mensagens que abusam os *proxies* HTTP e SOCKS são originadas de menos *ASes* (e países) que aquelas que abusam *relays* abertos, que são originadas de todo o mundo (142 países). Mesmo sendo responsáveis por um pequeno volume de mensagens, os abusos a *relays* abertos correspondem a 96,4% dos *ASes* únicos que abusaram os *honeypots* durante o período analisado. Além disso, máquinas abusando *relays* abertos enviam muito menos mensagens ao longo do tempo: enquanto um AS abusando *relays* abertos enviou menos de 2.000 mensagens, os *ASes* que abusaram HTTP e SOCKS enviaram mais de 2.000.000 e 280.000 mensagens em média, respectivamente.

Essa preferência pelo abuso a *proxies* pode ser explicada pela necessidade dos *spammers* de esconder melhor a origem dos abusos. Se *spammers* contatassem diretamente os servidores

de *e-mail* dos destinatários, suas origens seriam mais facilmente rastreadas, já que mesmo os *relays* abertos iriam gravar o endereço IP da conexão anterior, como comportamento padrão de todo servidor de *e-mail*.

A princípio, as diferenças observadas entre abusos a *proxies* e *relays* parecem indicar estratégias de disseminação de mensagens distintas, originadas de campanhas diferentes, mas isso não foi o observado. De fato, 90% das campanhas foram originadas apenas por abusos a *proxies* HTTP e SOCKS; no entanto, 10% das campanhas abusaram os *honeypots* tanto como *proxy* quanto como *relay*. A Figura 4.13 exibe, para essas campanhas, a função de distribuição acumulada do número de endereços IP observados associados a cada tipo de abuso, em cada campanha. No caso dos abusos a *proxies*, as campanhas se originam de um conjunto mais concentrado de endereços. Cerca de 50% das campanhas são disseminadas a partir de apenas 10 fontes que abusam *proxies* HTTP/SOCKS, enquanto 80% se originam de mais de 10 endereços IP (e 40% se originam de mais de 100 IPs) quando os abusos são direcionados à porta 25 dos *honeypots* (*relay*).

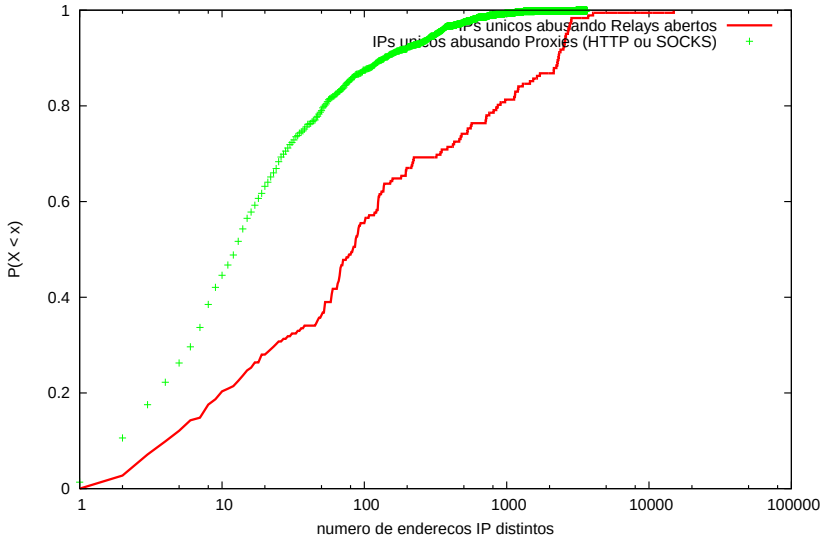

Figura 4.13. Número de endereços IP distintos abusando *proxies* e *relay* em cada campanha

4.3.2 Correlações entre atributos das campanhas

Na busca de evidências que explicariam essas diferenças relacionadas a abusos a *proxies* e *relays*, quatro diferentes características presentes em cada campanha foram correlacionadas: o tipo de abuso (*Proxy* HTTP, *Proxy* SOCKS e *Relay*), o *Country Code* de origem da mensagem, o *Country Code* de destino do *spam* e o idioma em que a mensagem foi redigida. O idioma foi obtido a partir da técnica baseada em *n-gramas* mencionada no Capítulo 4.1; o CC de origem foi obtido a partir das tabelas de alocação de endereços IP dos Registros Regionais de Internet; o país alvo das mensagens foi obtido a partir do endereço IP associado ao domínio extraído dos endereços de *e-mail* dos destinatários das mensagens. Para esta análise, não

foram considerados os endereços de *e-mail* no domínio `.com` (como `yahoo.com` e `gmail.com`), já que os usuários dessas contas de *e-mail* podem estar localizados em qualquer parte do mundo. Isso significa que 5% das mensagens foram desconsideradas na análise.

A partir dessas informações, foi aplicado um algoritmo de mineração de regras de associações nos dados de cada campanha. A análise de associação de atributos é uma técnica clássica da área de mineração de dados que objetiva determinar relações interessantes e previamente desconhecidas em grandes bases de dados (Tan et al., 2005). As relações determinadas são descritas na forma de *regras de associação*, segundo o formato A → B. Essa regra sugere que existe uma forte relação entre os atributos A e B, e, em geral, associa-se a cada regra algumas métricas que medem o quanto ela é interessante. Neste trabalho, considerou-se três métricas comumente analisadas:

1. **[suporte]**. O suporte de um conjunto de itens Z, Sup(Z), representa a porcentagem de transações da base de dados que contém os itens de Z. O suporte de uma regra de associação A → B, Suporte(A → B), é dado por Suporte(A U B).

2. **[confiança]**. Já a confiança desta regra, Confiança(A → B), representa, dentre as transações que contêm A, a porcentagem de transações que também contém B.

3. **[*lift*]**. O *lift* é uma medida da importância de uma regra, dada pela razão entre a confiança da regra e a confiança esperada, caso os atributos A e B fossem independentes. Um valor de *lift* superior a 1 indica que os atributos A e B aparecem juntos mais frequentemente do que o esperado; para um *lift* inferior a 1, a co-ocorrência é menos frequente que o esperado. O caso neutro ocorre quando o *lift* é 1.

Algumas das regras mais interessantes resultantes da aplicação do algoritmo estão listadas na Tabela 4.6. Na Tabela, a regra 1 exibe o abuso mais frequentemente observado no conjunto de dados estudado: 23,8 % dos abusos são relacionados a mensagens escritas em chinês abusando *proxies* HTTP. Além disso, 86% das mensagens com essas características foram enviadas de endereços IP alocados ao *Country Code* TW.

As regras 2 a 5 indicam que *spams* escritos em chinês também são originados de BR, AR, GB e PT, porém, abusando *relays* abertos. No caso do *Country Code* AR, a regra 6 mostra que 95% dos *spams* enviados daquele país abusando *Proxy* SOCKS foram escritos em espanhol. As regras 7 e 8 indicam que CN envia *spams* em chinês tanto por meio de *proxies* quanto de *relays* abertos.

As regras 9 a 11 mostram os abusos mais comuns relacionados a mensagens enviadas do *Country Code* US. Enquanto endereços IP alocados para os Estados Unidos enviam *spams* em chinês por meio de *relays* (regra 9) e em inglês por meio de *proxies* (regra 10), máquinas hospedadas no país também enviam *spams* em chinês por meio de *proxies* HTTP e SOCKS (regra 11), o que é diferente do observado para os outros países.

A partir desses resultados, pode-se concluir que existe uma forte relação entre o tipo de abuso, a origem e o destino dos *spams*. Enquanto TW envia mensagens predominantemente

Tabela 4.6. Regras de associação - origem, destino, idioma e tipo de abuso

Regra	Antecedente (se)	Consequente (então)	Suporte	Confiança	*Lift*
1	Idioma: chinês, abuso: HTTP	Origem: TW	23,8 %	86,0 %	1,1
2	Origem: BR	Idioma: chinês, abuso: *relays*, destino: TW	0,02 %	46,7 %	3,8
3	Origem: AR	Idioma: chinês, abuso: *relays*, destino: TW	0,01 %	76,7 %	4,5
4	Origem: GB	Idioma: chinês, abuso: *relays*, destino: TW	0,02 %	81,9 %	3,1
5	Origem: PT	Idioma: chinês, abuso: *relays*, destino: TW	0,01 %	43,0 %	2,3
6	Origem: AR, Abuso: SOCKS	Idioma: espanhol	0,01 %	95,0 %	4,3
7	Origem: CN, Abuso: HTTP	Idioma: chinês, destino: TW	7,5 %	84,0 %	1,3
8	Origem: CN, Abuso: *relays*	Idioma: chinês, destino: TW	6,3 %	78,0 %	1,1
9	Origem: US	Idioma: chinês, Abuso: *relays*, destino: TW	0,8 %	59 %	1,0
10	Origem: US, abuso: HTTP/SOCKS	Idioma: inglês	3,1 %	56 %	1,4
11	Origem: US, abuso: HTTP/SOCKS	Idioma: chinês	1,1 %	31 %	0,9

abusando *proxies* HTTP e SOCKS direcionadas a destinatários em TW (em chinês), a maior partes dos outros *country codes* (BR, AR, PT e outros 139 CCs) também envia mensagens em chinês, porém, abusando *relays* abertos. A única exceção é US, que também envia *spams* em chinês por meio de *Proxies* HTTP e SOCKS, o que pode indicar uma estratégia diferente de disseminação de *spams*.

As regras de associação, analisadas em conjunto com a Tabela 4.5, sugerem que *proxies* HTTP e SOCKS são abusados diretamente por *spammers*, isto é, os endereços IP que abusaram os *honeypots* representam a máquina do próprio *spammer*. Essa hipótese é reforçada pelo número concentrado de endereços IP que originam esses abusos e a coincidência entre o idioma dos *spams* e o idioma associado ao *Country Code* de origem. Por outro lado, abusos a *relays* abertos são originados de todas as partes do mundo. Na Tabela 4.6, foram apresentadas apenas algumas regras ilustrando o envio de *spams* em chinês por meio de *relays* abertos, para alguns países, mas, na verdade, esse padrão se estende para todos os outros países do conjunto de dados analisado. Esses casos podem corresponder a *proxies* HTTP e SOCKS abusados, ou uma estrutura mais organizada, mantida indiretamente sob o controle de *spammers*, como *botnets* (Cooke et al., 2005).

Os resultados em relação a abusos originados do *Country Code* AR (Argentina) ilustram essas diferenças claramente. 76,7% de todos os *spams* originados de AR estão escritos em chinês, atingem destinatários em TW e abusam *relays* abertos. Por outro lado, 95% dos *spams* originados de AR abusando SOCKS estão em espanhol.

As correlações obtidas confirmam que a maior parte das mensagens (72%) é enviada por TW, redigida em chinês e dirigida a domínios de destino associados a TW, como `yahoo.com.tw` e `hinet.net`. O segundo padrão mais frequente corresponde a mensagens originadas em TW redigidas em inglês, porém, destinadas a domínios em TW (16%). Essa relação retrata a importância de considerar as características das mensagens em conjunto para se entender seus objetivos, pois, mesmo escritas em inglês, as mensagens são destinadas a TW e a destinatários da Ásia. Esses *spams*, inclusive, contém URLs com *links* para páginas em chinês. O que acontece, em muitos casos, é que os *spams* originados por TW contém em seu corpo fragmentos de textos em inglês, aleatórios, que visam confundir os filtros *anti-spam*.

Os números gerais apresentados no Capítulo 3 e nesta seção, juntamente com as regras

de associação exibidas na Tabela 4.6 e outras estatísticas geradas a partir do conjunto de dados analisados permitem sumarizar os fluxos dos *spams* que trafegam na rede brasileira da seguinte forma:

1. A maior parte das mensagens que *spammers* tentaram enviar por meio dos *honeypots* vem de fora do Brasil (cerca de 99%);

2. A maior parte das mensagens são originadas de TW e CN (chinês e inglês) e US (inglês);

3. 92% das campanhas originadas no Brasil são enviadas, ao mesmo tempo, de TW. A maioria é enviada, simultaneamente, de vários outros países do mundo. Esse é um claro exemplo de como a identificação das campanhas (unificação das mensagens) constitui-se em um passo importante para entender as estratégias de disseminação dos *spams*;

4. No conjunto de 350 milhões de mensagens analisadas, apenas em 956.772 (0,27%) foram encontrados destinatários brasileiros, com domínio `.com.br`. Essas mensagens estavam redigidas, em sua maior parte, em chinês e inglês;

5. As mensagens de teste, que são enviadas pelos *spammers* para verificar se as máquinas abusadas estão ativas, são originadas em sua grande maioria dos *Country Codes* TW e US. Não foram encontradas mensagens de teste originadas de endereços IP de origem brasileiros;

6. Embora o campo *From:* das mensagens seja um campo sobre o qual o *spammer* detém total controle, e que permite que ele insira qualquer endereço de *email*, válido ou não, observamos que, em 83% das mensagens, o endereço no *From:* contém um domínio que mapeou para o *Country Code* TW. Isso indica que, intencionalmente, o *spammer* quer fazer parecer para a vítima que o produto/serviço anunciado está relacionado a TW.

7. Foi encontrada apenas cerca de uma dezena de campanhas de *spam* em português em todos os dados analisados.

Todas essas evidências nos permitem concluir, que, a partir da visão dos dados coletados nos sensores, **o Brasil é difusor de *spam***, isto é, a sua infraestrutura é abusada pelos outros países para envio de *spams*.

As técnicas de minerações de dados que aplicamos aos dados coletados nos indicaram que as mensagens de *spam* trafegam no Brasil de duas formas bem distintas:

1. A partir de *spammers* que abusam *proxies* abertos no Brasil, enviando mensagens em grande quantidade. A origem das mensagens está associada a seu idioma e domínios de destino. Esses abusos são originados, primordialmente, de TW e CN, contendo mensagens escritas em chinês e inglês e são direcionadas a destinatários da Ásia;

2. A partir de *spammers* que abusam *relays* abertos no Brasil, enviando mensagens em quantidade bastante reduzida. A origem das mensagens não está associada ao idioma nem aos domínios de destino. Essas mensagens são originadas de toda a parte do mundo. Essas correlações ilustram o potencial das minerações de dados no sentido de revelar relacionamentos até então desconhecidos: dado que a maioria absoluta das abusos são direcionados a *proxies* abertos, os abusos aos *relays* abertos ficam desapercebidos, embora revelem um padrão importante de comportamento.

Os *spams* originados no Brasil enquadram-se nesse último caso. 99,9% das mensagens originadas de IPs brasileiros foram enviadas através de abusos de *relays* abertos (porta 25/TCP). Esse resultado é bastante diferente do comportamento mais comum verificado nos dados, em que os abusos a *proxies* abertos correspondem à absoluta maioria das conexões, como detalhado no Capítulo 3.

A Figura 4.14 mostra, para cada campanha identificada, o número de endereços IP de origem dos abusos aos *honeypots* e países de origem associados a esses mesmos abusos à porta 25 (*relay)* nessas campanhas. O gráfico reforça a dispersão desses abusos, que chegam a vir de várias dezenas de países do mundo, na mesma campanha. Por outro lado, os abusos a *proxies* abertos são originados de um número muito mais restrito de IPs. Um indicativo de que a origem do abuso a *proxies* é muito mais concentrada é o fato de que 95% das campanhas são originadas, cada uma, de no máximo dois *ASes* que exploram *proxy*. O fato dos IPs de origem que abusam *proxies* estarem, em sua maioria, associados a poucos *ASes* (e, até mesmo, a uma única rede /24) em cada campanha nos leva a acreditar que esses abusos se originam de grupos de máquinas controlados pelo *spammer* dedicadas à atividade de *spamming*.

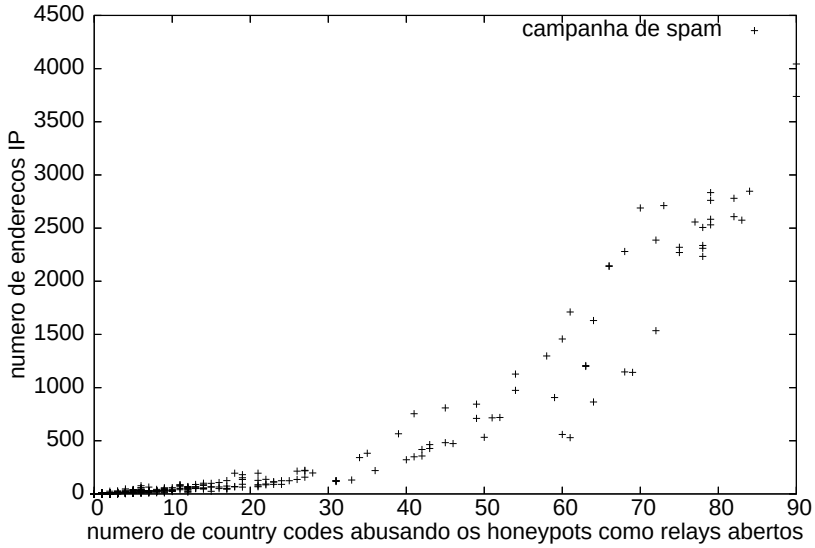

Figura 4.14. IPs distintos x Países distintos abusando *relays* abertos, em cada campanha de *spam*.

4.3.2.1 Diferenças de comportamento de *spammers* de acordo com o sistema operacional

Para cada endereço IP que abusou os *honeypots* registrou-se o Sistema Operacional associado, por meio de técnicas de *fingerprinting* passivo (Provos & Holz, 2007). A Tabela 4.7 lista a frequência de ocorrência dos sistemas operacionais no conjunto de dados analisado. Apesar de não ter sido possível determinar o sistema operacional associado à maior parte das mensagens, fica claro que o sistema operacional Windows é mais comum que sistemas como Linux, FreeBSD e Solaris.

Tabela 4.7. Sistemas Operacionais mais frequentes das máquinas de origem dos *spams*

#	Sistema Operacional	*mensagens*	%
01	Desconhecido	209.319.751	59,9
02	Windows	138.802.847	39,9
03	Linux	2.160.574	0,01
04	Solaris	17.622	0,0053
05	FreeBSD	16.931	0,0051
06	Outros	3.573.585	0,01

Novamente, um algoritmo de mineração de regras de associação foi aplicado a fim de determinar as diferentes estratégias associadas à adoção de sistemas operacionais por parte de *spammers* e máquinas abusadas. Os principais padrões encontrados estão listados na Tabela 4.8. A Tabela revela alguns padrões significativos envolvendo os sistemas operacionais das máquinas que abusaram os *honeypots* e o tipo de abuso associado a cada uma. As regras de associação 1 a 3 mostram que máquinas configuradas com os sistemas operacionais Linux, FreeBSD e Solaris abusam os *honeypots* principalmente como *relays* abertos, na maior parte dos abusos observados. O valor alto para o *lift* em todos os casos (superior a 8,0) indica que as chances de se observar abusos a *relays* abertos é muito maior quando as mensagens se originam de máquinas instaladas com Linux, FreeBSD e Solaris, mesmo que esses sistemas operacionais correspondam a menos de 3% do total de fluxos observados.

Tabela 4.8. Regras de associação – sistemas operacionais e tipos de abuso

Regra	**Antecedente (se)**	**Consequente (então)**	**Suporte**	**Confiança**	***Lift***
1	SO: Linux	Abuso: *Relay* aberto	1,3 %	97,0 %	8,0
2	SO: FreeBSD	Abuso: *Relay* aberto	0,7 %	100 %	8,2
3	SO: Solaris	Abuso: *Relay* aberto	0,6 %	100 %	8,2
4	SO: Windows	Abuso: *Relay* aberto	4,1 %	7 %	0,6
5	SO: Windows	Abuso: HTTP	7,1 %	62 %	0,9
6	SO: Windows	Abuso: SOCKS	15,3 %	31 %	1,2
7	SO: Desconhecido	Abuso: HTTP	49,8 %	72 %	1,0
8	SO: Desconhecido	Abuso: SOCKS	16,1 %	26 %	1,0

As regras 4 a 6 mostram que os sistemas Windows são usualmente utilizados para abusar *proxies* SOCKS (com 31% de confiança) e HTTP (com 62% de confiança). No conjunto de dados analisado, mais da metade dos sistemas operacionais das máquinas que abusaram os *honeypots* não puderam ser identificados (regras 7 e 8) pelas técnicas de *fingerprinting*

aplicadas. Como as proporções de abusos a HTTP e SOCKS são similares àquelas observadas para o sistema Windows, muitas dessas conexões podem estar associadas com Windows Vista, que era um sistema operacional novo à época da coleta dados e poderia não ter uma assinatura capturável ainda.

O fato de que a quase totalidade dos abusos a *proxies* abertos é originada por máquinas configuradas com Windows é algo esperado, ao relacionarmos com a conclusão anterior de que o abuso aos *proxies* é associado à origem real: os *bulk mailers*, que são ferramentas desenvolvidas, especificamente, para a atividade de disseminação de *spam*, principalmente para Windows, e não para sistemas Unix. Isso explica o fato de não serem observados abusos a *proxies* abertos originados de máquinas configuradas com sistemas Unix, como Linux e Solaris. Embora não seja possível comprovar essa hipótese, uma possibilidade é que as máquinas Unix abusando os *honeypots* sejam, na verdade, *proxies* abertos, ou seja, máquinas já abusadas por *spammers* e que compõem, juntamente com os *honeypots*, uma cadeia de máquinas para disseminação de *spams*.

4.3.3 Representatividade dos Resultados

Uma questão que emerge a partir da análise dos resultados é a representatividade dos dados coletados, isto é, determinar se os *spams* coletados e suas características são uma amostra fiel dos *spams* que circulam pela infraestrutura da Internet brasileira. Uma possibilidade levantada durante as análises era a de que os abusos provenientes de Taiwan às portas de *Proxy* podiam atuar de forma tão intensa que impediam que outros países enviassem seus *spams* através dos sensores. Foram feitas três verificações que indicam que os *spams* analisados representam uma amostra representativa dos abusos a *proxies* e *relays* abertos na rede brasileira:

1. A proporção das conexões rejeitadas por cada tipo de abuso (HTTP, SOCKS e *relay*) que são registradas pelo `tcpdump` é bastante similar à mesma proporção computada para as mensagens coletadas;

2. Em um experimento em que apenas conexões de endereços IP brasileiros eram aceitas, o número de mensagens coletadas mostrou-se extremamente baixo. Isso pode indicar que não há uma demanda reprimida de *spammers* brasileiros que, sob a perspectiva dos sensores, não estariam conseguindo agir devido à intensidade das conexões vindas da região da Ásia.

3. O comportamento dos *spammers* é similar em todos os *honeypots* implantados:

    - Para todos os sensores, TW, CN e US são os *country codes* que mais enviam *spams* abusando *proxies* (HTTP e SOCKS). Em todos os sensores, TW, CN e US, juntos, respondem por cerca de 95% dos *spams*;

- Em todos os sensores, TW e CN abusam *proxies* disseminando campanhas redigidas em chinês. o *Country Code* US dissemina campanhas em inglês e chinês em todos os sensores;
- No caso dos abusos a *relay* (porta 25), a diversidade na origem do *spam* e a dissociação da origem com seu destino também é observada para todos sensores (com exceção dos sensores 01 e 09). As Tabelas 4.9 e 4.10 exibem a diversidade de CCs verificada para cada sensor quanto às portas associados aos abusos a *proxies* (HTTP e SOCKS) e ao *relay*, respectivamente. Os resultados são coerentes entre os sensores. Os sensores 01 e 09 eram máquina mais instáveis e que ficaram ativas por menos tempo, o que pode explicar

Tabela 4.9. Número de *Country Codes* de origem encontrados para os abusos a portas de *Proxy* (HTTP e SOCKS) em cada sensor

sensor	**número de CCs de origem**
01	8
02	5
03	7
04	7
05	10
06	8
07	8
08	6
09	2
10	6

Tabela 4.10. Número de *Country Codes* de origem encontrados para os abusos à porta 25 (*Relay* aberto) em cada sensor

sensor	**número de CCs de origem**
02	39
03	82
04	42
05	67
06	98
08	67
10	123

Embora os resultados sejam consistentes, não é possível generalizá-los como uma amostra fiel de todos os *spams* que circulam na Internet. Primeiramente, todos os *honeypots* foram implantados em redes brasileiras, o que não permitiu verificar em detalhes como outros países são abusados. O fato dos sensores se localizarem em redes brasileiras pode, inclusive, ser a razão principal pela qual poucas campanhas em português sejam observadas nos dados: já que o objetivo do *spammer* ao abusar *proxies* e *relays* abertos é ocultar sua origem,

ele pode preferir escolher máquinas localizadas em faixas de endereços IP que não sejam do seu próprio país, dificultando ainda mais seu rastreamento. Além disso, *spams* enviados somente a partir de *botnets* sem passar por *proxies* ou *relays* abertos não são observados. Enquanto é possível afirmar que o Brasil é difusor de *spam*, não há convicção para se afirmar que o Brasil não é originador de *spam*, pelo fato de que os sensores estavam implantados apenas em redes brasileiras. Para investigar essa possibilidade, seria necessário implantar *honeypots* em redes fora do Brasil e verificar se campanhas de *spam* em português são observadas.

Capítulo 5

Encadeamento de Máquinas para Disseminação de *Spams*

Uma das principais preocupações dos *spammers* é esconder sua origem real, inclusive em termos da localização na rede. Isso acontece por dois motivos. Primeiramente, a atividade de disseminar *spams* é considerada ilegal em muitos países, forçando os *spammers* a manterem anonimato. Além disso, se *spammers* enviassem *spams* diretamente de suas máquinas, eles seriam facilmente bloqueados pelos servidores de *e-mail* (Boneh, 2004), pois um grande volume de mensagens originadas de uma mesma fonte atrairia a atenção dos administradores de rede.

Como resultado da constante batalha entre *spammers* e anti-*spammers*, os *spammers* mais sofisticados conseguem combinar diferentes técnicas para ocultação de identidade e distribuição dos abusos, criando *cadeias de máquinas* para disseminação de *spams* e tornando seu rastreamento mais difícil ou mesmo impossível. A definição de uma *cadeia* corresponde à sequência de conexões utilizadas para encaminhar o conteúdo de uma campanha de *spam* até que ela seja entregue por uma conexão SMTP. Algumas dessas possíveis cadeias estão ilustradas na Figura 5.1, já exibida no Capítulo 3:

A Figura ilustra os encadeamentos caracterizados nesta dissertação, que, genericamente, enquadram-se em quatro casos:

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 *relays* verdadeiros, mas que possuem instalado algum software para se comportarem como servidores de correio, com vistas a serem exploradas explicitamente pelo *spammer*. Podem ser máquinas infectadas

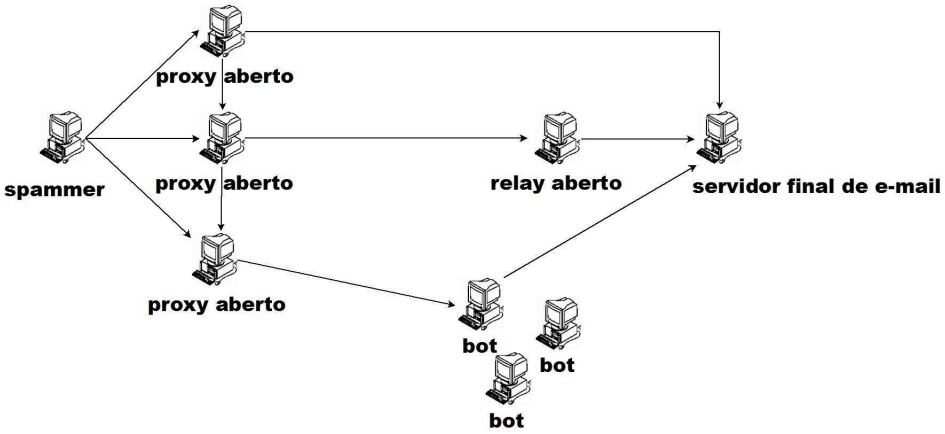

Figura 5.1. Cadeias de máquinas para envio de *spams*

com um *malware* que tenha apenas essa finalidade ou mesmo *bots* (máquinas infectadas capazes de se propagar automaticamente e que permitem que a máquina infectada seja controlada remotamente (CERT.br, 2009; McCarty, 2003));

4. encadeamento de *proxies* abertos, quando o *spammer* abusa dois ou mais *proxies* abertos em sequência.

Todos esses encadeamentos podem ser observados desde que um dos *honeypots* participe da cadeia. Por isso, algumas cadeias não podem ser observadas, como aquelas em que a máquina do *spammer* entrega a mensagem para um *bot* e este a entrega para a máquina do usuário, sem passar por *proxies* abertos e *relays* abertos.

A arquitetura de coleta de *spams* implantada permite determinar sempre a máquina que se conectou ao *honeypot* (que pode ser de um *spammer* ou de uma máquina já abusada) e a máquina que seria alvo da conexão estabelecida pelo *honeypot*, que pode ser um servidor de *e-mail*, um outro *proxy* ou *relay* aberto ou mesmo uma máquina de usuário final. É importante ressaltar que essas conexões não são efetivamente estabelecidas, embora o *spammer* tenha a impressão de que conseguiu entregar mensagens por meio dos *proxies* e *relay* emulados.

Embora algumas dessas cadeias tenham sido reportadas pela comunidade de pesquisa sobre *spams* (Boneh, 2004; Andreolini et al., 2005) e discutidas informalmente em listas de discussão e fóruns técnicos, a caracterização dessas cadeias ainda é limitada na literatura científica, e a análise do comportamento de rede dos *spammers* desempenhada nesta dissertação incluiu a identificação e quantificação de alguns tipos dessas cadeias.

Ao observar as origens e os destinos das conexões estabelecidas com os *honeypots* e os próximos passos executados pelos *spammers* no processo de encadeamento de máquinas e a identificação das mensagens associadas a cada campanha, é possível adquirir conhecimento sobre as sequências de máquinas abusadas por cada um. Os trabalhos que caracterizam a forma como *spammers* abusam os recursos de rede, em geral, coletam dados a partir de uma estratégia de disseminação de *spams* específica, como *botnets* (Kreibich et al., 2008; Lee et al.,

2007), *spam traps* (Gansterer & Ilger, 2007) e *relays* abertos (Pathak et al., 2008). Por isso, eles focam em uma etapa específica do caminho na rede percorrido pelas mensagens até serem entregues aos destinatários. No caso de trabalhos que analisam *logs* de servidores de *e-mail* (Li & Hsieh, 2006; Gomes et al., 2004), apenas a última máquina abusada pelo *spammer* antes de ser entregue ao servidor é analisada, já que *spams* coletados dessa forma não permitem o estudo de cadeias de máquinas, pois os cabeçalhos SMTP podem ser facilmente forjados pelos *spammers*. Mesmo trabalhos que analisam conexões *Proxy* estabelecidas a *honeypots* focam na análise das características dos abusos que chegam aos sensores, como a origem e a distribuição de endereços IP dos abusos (Steding-Jessen et al., 2008). Existem trabalhos que mencionam a criação de cadeias de máquinas com o objetivo de disseminar *spams* como algo tecnicamente possível (Boneh, 2004; Andreolini et al., 2005; Oudot, 2003), mas eles não caracterizam e efetivamente demonstram esses comportamentos.

Para analisar o encadeamento de máquinas, é necessário separar os abusos às máquinas de destino das conexões intermediadas pelos *honeypots* (de acordo com a percepção do *spammer*) em abusos a servidores de *e-mail* finais, *proxies* abertos, *relays* abertos e máquinas de usuários finais. A partir da análise desses abusos no contexto das campanhas de *spam*, foi possível determinar as diferentes estratégias adotadas por *spammers* para encadear conexões entre máquinas na rede.

5.1 Identificação de Tipos de Cadeias

Para entender as cadeias observadas nos casos em que *spammers* abusaram *proxies* para estabelecer conexões com outras máquinas, é necessário diferenciar as máquinas-alvo das conexões entre aquelas que são servidores finais de *e-mail*, aquelas que são *relays* abertos e máquinas de usuários finais que foram infectadas e passaram a atuar como *relays*.

Os *honeypots* não foram configurados para tentar identificar os tipos das máquinas-alvo dos abusos durante a coleta dos dados (verificando, por exemplo, se elas eram listadas como servidores MX associados aos domínios de *e-mail*). Por isso, foi necessária a definição de uma heurística para classificar as máquinas de destino das conexões.

A heurística projetada para classificar os destinos das conexões assume que, em geral, observa-se que os grandes servidores de *e-mail* são representados por nomes únicos e bem definidos, como `mail.hotmail.com` e `mta-v1.mail.vip.tp2.yahoo.com`. A partir dessa observação, utilizou-se o nome das máquinas como critério para diferenciar servidores de *e-mail* finais e máquinas de usuários finais. Os servidores são máquinas cujos nomes incluem prefixos como `mail`, `smtp` e `mta`, por exemplo. Embora possa haver falsos positivos (nomes de máquinas de usuário com esses prefixos) e falsos negativos (servidores de *e-mail* com outros nomes), os resultados obtidos foram aceitáveis.

Para identificar as cadeias que envolvem máquinas de usuários finais, usou-se o fato de que os provedores de serviços (ISPs) em geral assinalam nomes para as máquinas de seus clientes que combinam partes fixas com uma parte variável que diferencia cada máquina,

normalmente um identificador numérico ou o próprio endereço IP que foi assinalado para a máquina. Por exemplo, clientes do provedor norte-americano Verizon são identificados na rede no formato `static-<IP>.<LOCATION>.dsl-w.verizon.net`. Já as máquinas sob responsbilidade do provedor HINET (em Taiwan) são nomeados segundo o formato `<IP>.HINET.-IP-.hinet.net`. Essa característica dos nomes de máquinas de usuários remete à ideia de que existem partes invariantes e partes variadas nos padrões de nomes assinalados por cada ISP, e, em vista disso, a Árvore de Padrões Frequentes foi novamente aplicada para diferenciar servidores de *e-mail* de máquinas de usuários.

O nome de cada máquina alvo de conexões *Proxy* HTTP foi quebrado em fragmentos em cada nível da hierarquia DNS. Por exemplo, `smtp1.google.com` seria quebrado nos fragmentos `smtp1`, `google` e `com`. Esses fragmentos foram, então, inseridos em uma Árvore de Padrões Frequentes, de modo que os fragmentos que compõem o nome de uma máquina definem um caminho na árvore e os fragmentos mais frequentes são encontrados nos níveis mais altos, e o termos infrequentes ou aleatórios ficam próximos às folhas. Foi registrado, também, quantas conexões utilizaram cada nome. Dessa forma, máquinas de usuários finais abusadas conectadas a grandes provedores compartilham a maior parte dos seus caminhos a partir da raiz da árvore, em decorrência das partes fixas no formato de seus nomes. Esses nomes diferem apenas por fragmentos que correspondem a seus identificadores únicos, em geral seus endereços IP ou parte deles. Como essas características são menos frequentes que as características fixas de cada provedor, esses nomes pertencentes a um mesmo provedor formam sub-árvores com um grande número de irmãos e um caminho na árvore em comum, até esses nós. A estratégia é bastante similar à proposta para agrupar mensagens em campanhas proposta no Capítulo 4.1.

5.2 Análise dos Encadeamentos de Máquinas para Disseminação de *Spams*

Esta seção apresenta os resultados da caracterização de cadeias de máquinas para disseminação de *spams*. A Tabela 5.1 exibe alguns dados relevantes para o entendimento do encadeamento de máquinas. Durante o período analisado, mais de 230 milhões de mensagens foram entregues pelos *spammers* aos emuladores de *Proxy* HTTP dos *honeypots*, por meio de cerca de 90 milhões de conexões, resultando, em média, em 2,6 mensagens entregues por cada conexão. Essas conexões foram originadas de 93.757 endereços IP, que direcionariam mensagens para 459.218 endereços IP de destino das conexões. Nas subseções seguintes, os principais resultados da caracterização são discutidos. As conexões SOCKS não foram consideradas nesta análise pois a maior parte das conexões desse tipo envolviam a versão 4 do protocolo, que não registra o nome da máquina de destino das conexões e portanto inviabilizava a aplicação da técnica de classificação de máquinas.

A seguir, serão apresentadas as principais conclusões da análise das origens e destinos das conexões (e, consequentemente, as cadeias de máquinas formadas) em cada campanha de *spam*.

Tabela 5.1. Visão geral das conexões direcionadas aos emuladores HTTP *Proxy* dos *honeypots*

conexões HTTP *Proxy*	89.836.643
endereços IP de origem únicos	93.757
destinatários únicos	3.2×10^9
domínios de *e-mail* únicos	6.710.121
endereços IP de destino únicos	459.218

5.2.1 Estabelecimento de cadeias que não terminam no servidor de destino

As conexões *Proxy* HTTP direcionadas aos *honeypots* seriam destinadas a cerca de 460 mil máquinas distintas. Por outro lado, mais de 6,7 milhões de domínios de *e-mail* únicos seriam alvos de *spams* no período analisado (Tabela 5.1). Como o número de domínios de *e-mail* é cerca de 15 vezes maior que o número de máquinas alvo das conexões, há uma evidência de que grande parte das conexões não terminam nos servidores de *e-mail* finais.

A Figura 5.2 exibe a distribuição acumulada do número de domínios de e-mail encontrados nas mensagens de *spam* que seriam entregues a cada endereço IP alvo de conexões. Mais de 50% dos endereços IP alvo de conexões receberam mensagens direcionadas a mais de dois domínios de *e-mail* distintos, e mais de 10% receberia mensagens endereçadas a mais de 10 domínios. Alguns endereços IP receberiam mensagens direcionadas a mais de 100 domínios, o que indica que essas máquinas não seriam o destino final das mensagens, mas apenas intermediários do processo de entrega de *spams*.

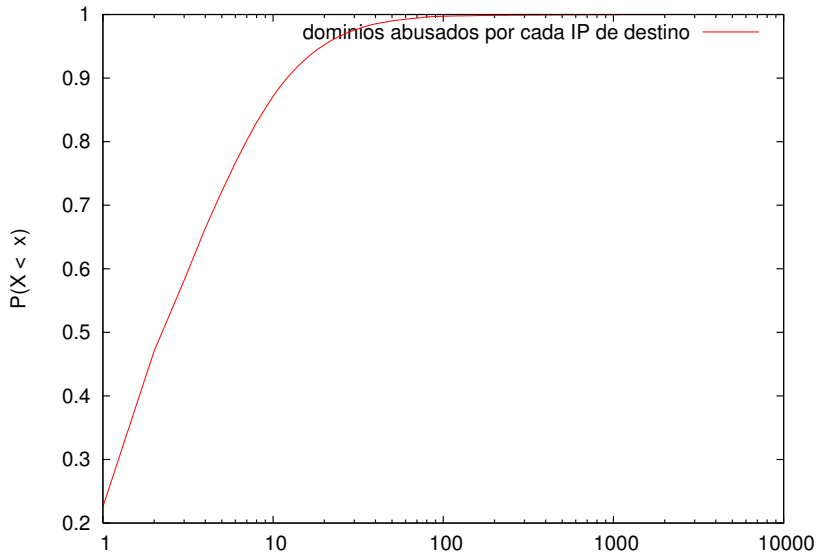

Figura 5.2. Número médio de domínios diferentes encontrados nos destinatários das mensagens entregues a cada IP de destino diferente (CDF)

5.2.2 Encadeamento de *proxies* abertos com máquinas de usuários finais infectadas

Após a aplicação da metodologia descrita na Seção 5.1, foram identificadas 94.480 máquinas que representam máquinas de usuários finais e que não são servidores de *e-mail*. Essas máquinas provavelmente são máquinas mal-configuradas ou máquinas infectadas por algum *malware* que as instruem a se comportar como *relays* abertos. Essas máquinas estão distribuídas entre 894 grupos de máquinas, que, grosso modo, correspondem a diferentes provedores (ISPs). A Tabela 5.2 lista os dez países que hospedam o maior número de máquinas infectadas. Mais de um terço dos grupos representam ISPs norte-americanos. Os principais grupos identificados estão listados na Tabela 5.3. Entre os provedores, também é possível identificar alguns grupos de máquinas associadas a serviços de *hosting* dedicado e serviços de *datacenter*, como os grupos `secureserver.net` e `ev1servers.net`. Não foi possível determinar se essas máquinas são casos de servidores mal-configurados, máquinas infectadas por *malware* ou ainda máquinas propositadamente configuradas por clientes *spammers*.

Tabela 5.2. Países que hospedam o maior número de máquinas de usuários finais que enviam *spams*

#	CC	número de endereços IP distintos (ISPs)	%
01	US	59.800 (351)	36,6
02	TW	38.925 (61)	23,8
03	CN	24.708 (19)	15,1
04	HK	6.880 (28)	4,2
05	GB	6.564 (59)	4,0
06	KR	5.925 (8)	3,6
07	JP	5.631 (48)	3,5
08	DE	5.627 (50)	3,4
09	BR	5.049 (37)	3,1
10	CA	3.958 (35)	2,4

Esses resultados indicam que, embora reportado em trabalhos anteriores que a maior parte do *spam* é enviado de máquinas infectadas e que *proxies* abertos não são mais comuns na Internet (Ramachandran & Feamster, 2006), combater *proxies* abertos ainda é necessário. A subestimação do impacto de *proxies* abertos pode ser decorrente do fato de que as observações em *logs* de servidores de *e-mail* apontam as máquinas de usuários finais dos países listados na Tabela 5.2 como o último passo para entrega da mensagem, na última linha `Received:` no cabeçalho SMTP das mensagens. No entanto, *proxies* abertos são um mecanismo comum para ocultação de identidade, conforme mostrado nas diferentes cadeias discutidas neste Capítulo. Dessa forma, o combate a *proxies* abertos, por meio de listas de bloqueio e configuração correta de máquinas ainda é um meio importante para combater diferentes estratégias de disseminação de *spams* que incluem *proxies* abertos em suas rotas, incluindo as situações em que *botnets* e outros tipos de máquinas infectadas são os últimos intermediários dos abusos.

Os resultados também mostram que o entendimento das cadeias de máquinas para envio

Tabela 5.3. Número de máquinas nos principais grupos (ISPs) abusados como *relays* abertos

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

de *spam* permite caracterizar as infraestruturas que estão nos passos anterior e seguinte às máquinas que efetivamente coletam os dados, aumentando o conhecimento sobre estratégias de disseminação de *spams*. Neste exemplo, foi investigada a distribuição geográfica de máquinas infectadas a partir da análise das conexões direcionadas a elas por intermédio dos *proxies* abertos emulados pelos *honeypots*.

5.2.3 Visão incompleta das campanhas

O encadeamento de máquinas torna a medição do comportamento dos *spammers* difícil; apenas as cadeias que incluem pelo menos um dos *honeypots* é observada, e isso pode explicar porque, em média, as campanhas identificadas neste trabalho são pequenas (90% das campanhas enviaram menos do que 5.000 mensagens), mesmo sendo reconhecido pela comunidade que campanhas de *spam* em geral atingem milhões de destinatários. Para verificar esse fenômeno, foi computado o número de mensagens que cada campanha de *spam* enviou às portas de *proxy* de cada um dos *honeypots*, em média. O resultado é exibido na Figura 5.3, que considerou apenas as campanhas que abusaram mais de um *honeypot*. Pode ser observado que *spammers* explicitamente enviam um volume pequeno de mensagens a cada *proxy* aberto. Como muitas campanhas enviam menos de 1.000 mensagens a cada *honeypot*, eles podem, na verdade, ter explorado centenas de outros *proxies* abertos na Internet para disseminar suas mensagens.

5.2.4 Intercalação de abusos a servidores de *e-mail* finais com abusos a *relays* abertos e máquinas infectadas em uma mesma campanha

Spammers nem sempre procuram *relays* abertos ou máquinas de usuários finais infectadas após abusarem *proxies* abertos. Entre as 89 milhões de conexões *Proxy* HTTP estabelecidas com os *honeypots*, uma porção significativa (72 milhões, ou 80,1% de todas as conexões) foram direcionadas a servidores de *e-mail*, ou seja, servidores apontados por um registro MX. Um

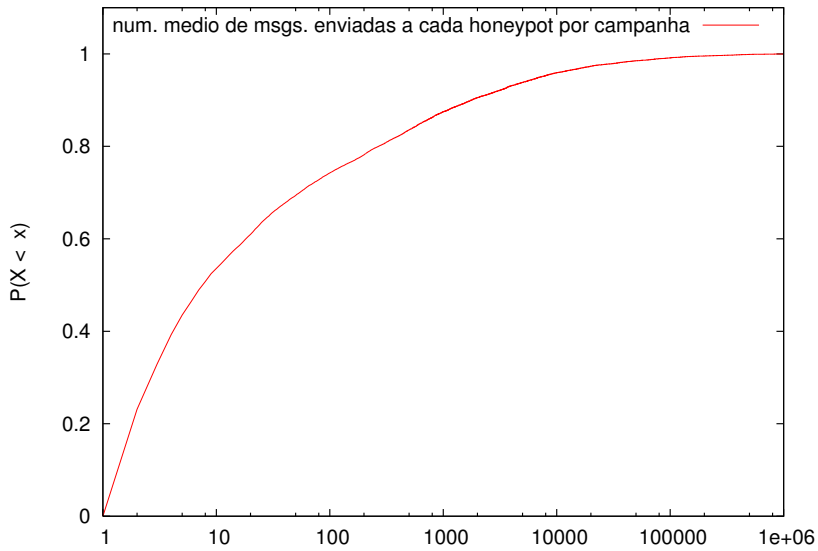

Figura 5.3. Número médio de mensagens enviadas a cada *honeypot* em cada campanha (CDF)

único MTA, `mta-v1.mail.vip.tp2.yahoo.com`, foi alvo de 19,5 milhões de conexões, o que reflete a alta popularidade do domínio `yahoo.com.tw` entre os *spammers* no conjunto de dados analisado.

As análises das cadeias no contexto das campanhas de *spam* mostrou que a maior parte dos *spammers* não toma uma decisão exclusiva entre abusar os servidores finais ou criar cadeias com *relays* abertos ou máquinas infectadas: 91% das campanhas exibem um comportamento híbrido, intercalando ambos os tipos de cadeia.

Este é um resultado típico que não é observado por *honeypots* que emulam apenas *relays* abertos ou servidores de *e-mail* isoladamente. O resultado complementa o que foi observado no trabalho de Pathak et al. (Pathak et al., 2008), que reportou que uma porção significativa das máquinas que abusaram os *relays* abertos implantados estava registrada em listas de bloqueio, isto é, em alguns momentos essas máquinas também abusaram diretamente servidores de *e-mail* finais e por isso foram registradas em tais listas.

5.2.5 Raridade das cadeias envolvendo *proxies* abertos

A partir das características das origens e destinos das conexões, analisou-se as conexões aos emuladores de *proxies* HTTP para verificar se eram estabelecidas cadeias de mais de um *proxy* aberto entre as campanhas de *spam*.

A grande maioria das conexões estabelecidas por meio dos *proxies* abertos dos *honeypots* foi direcionada à porta 25 da máquina seguinte; apenas 0,4% das conexões foram direcionadas a portas diferentes, o que sugere que a maioria dos *spammers* cria apenas um nível de encadeamento por meio de *proxies*. Essa observação é coerente com as características dos abusos direcionados aos *honeypots*, que, em sua maioria, são originadas de *Country Codes*

associados ao idioma e alvo dos *spams*, o que indica que os portas *Proxy* HTTP abusados pelos *spammers* foram as primeiras máquinas contatadas por eles, nesses casos. 97,4% das campanhas originaram-se apenas do *Country Code* associado ao idioma da campanha e o alvo dessas campanhas (na porta 25) também está localizado neste CC. A conclusão, então, é que cadeias de *proxies* abertos não são frequentes. Os casos de encadeamento entre *proxies* abertos, provavelmente, correspondem aos casos em que a origem da conexão não é relacionada com o idioma e o alvo do *spam*, e a origem provavelmente já é um *proxy* aberto abusado.

5.2.6 Impacto da dispersão dos abusos para disseminação de *spams*

Ao analisar as cadeias estabelecidas pelos *spammers* em cada campanha, procurou-se 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 5.4, 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 mais de 10.000 máquinas (sejam máquinas infectadas ou *relays* abertos) conseguiram enviar mais de 1 milhão de *spams*.

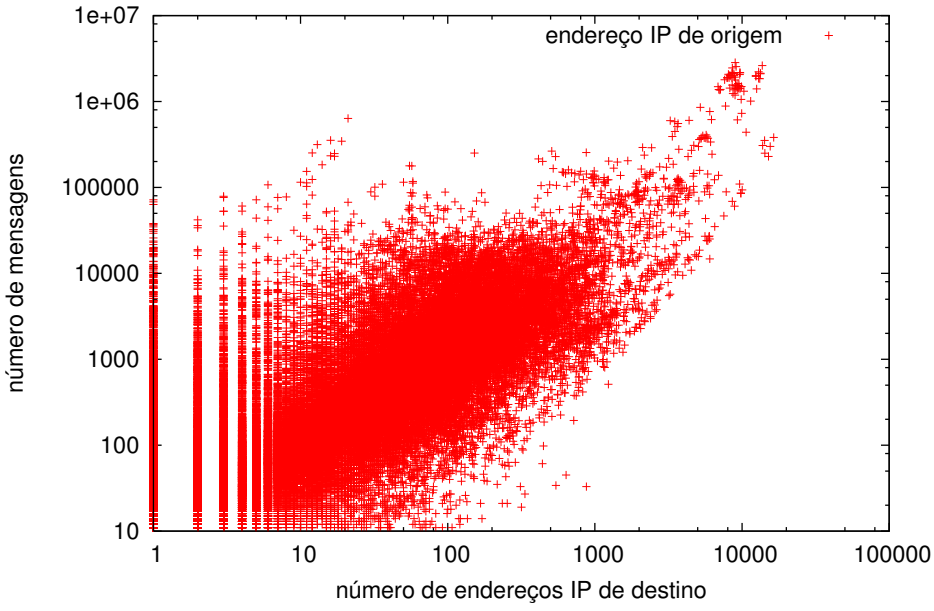

Figura 5.4. Número de endereços IP de destino contatados por cada IP de origem x volume de mensagens enviadas

Ao contrastar o número de endereços IP de destino abusados por cada endereço IP de origem e o número de dias pelo qual esse IP enviou *spams* (Figura 5.5), 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 endereços IP de origem permanece ativo por menos de dois meses.

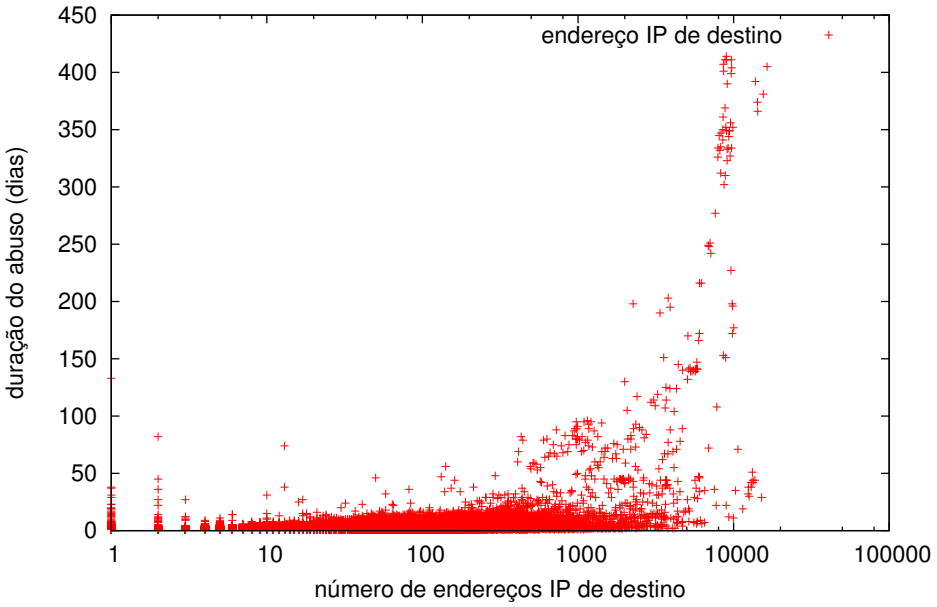

Figura 5.5. Número de máquinas de destino abusados por cada endereço IP de origem x número de dias que o IP de origem permanece ativo

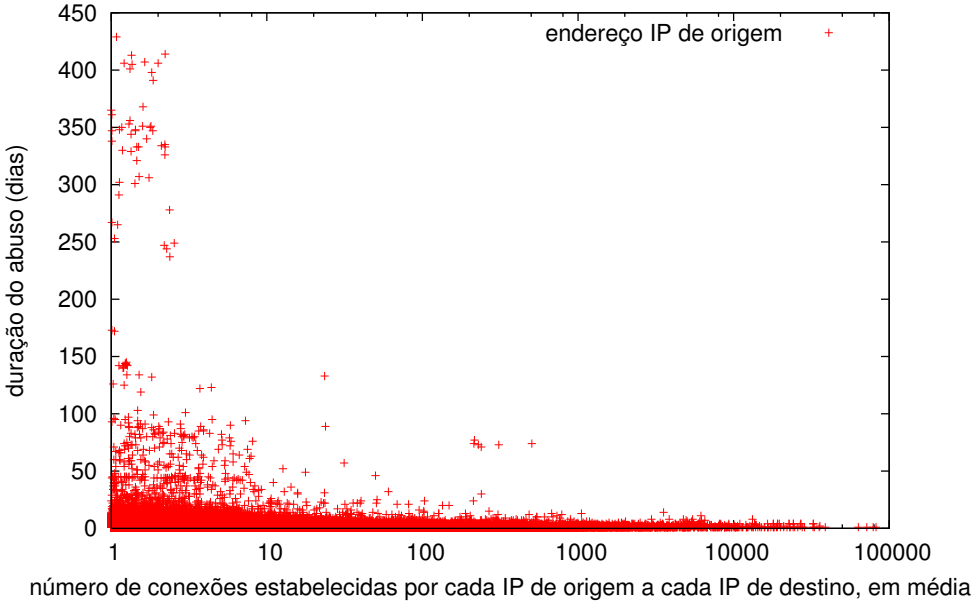

Figura 5.6. Número de conexões que cada IP de origem estabelece x número de dias que o IP de origem permanece ativo

A Figura 5.6 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, passam 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.

5.2.7 Diferenças de dispersão entre abusos a *proxies* e *relays* abertos

Finalmente, investigou-se a correlação entre o tamanho médio das campanhas e a dispersão dos abusos às portas de *Proxy* (HTTP e SOCKS) e *Relay* (SMTP) dos *honeypots*. As Figuras 5.7 e 5.8 exibem os dois casos, agrupando as origens pelo *Country Code*. No eixo horizontal, é medida a dispersão da origem dos abusos e no eixo vertical, o tamanho médio das campanhas que foram disseminadas a partir de cada conjunto de emissores.

É possível observar que as campanhas que abusaram os *honeypots* como *proxies* abertos originam-se de 1 a 10 países (*Country Codes*) são usualmente grandes e enviam dezenas de milhares de *spams* (Figura 5.7), e elas também abusaram, em média, um número maior de *honeypots*. Por outro lado, as campanhas em que os abusos a *proxies* abertos originaram-se de mais de 40 CCs diferentes são pequenas e em média não abusaram mais de 2 *honeypots*. É interessante notar que, embora essas campanhas se originem de muitas fontes, elas abusaram os mesmos *honeypots*, o que sugere um alto nível de coordenação entre essas fontes de *spam*.

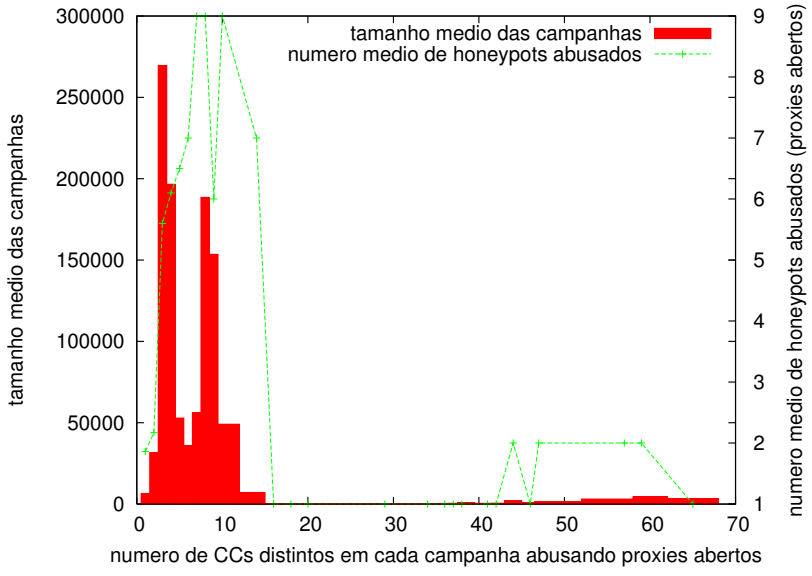

Figura 5.7. Tamanho médio das campanhas e número médio de *honeypots* abusados como *proxies*, por quantidade de *country codes* de origem

Ao considerar essas relações no caso dos abusos aos *relays* dos *honeypots*, o padrão é significativamente diferente (Figura 5.8). Dessa vez, mais *honeypots* são abusados e o tamanho

das campanhas aumenta, em média, conforme a origem dos abusos a *relays* abertos se torna menos concentrada.

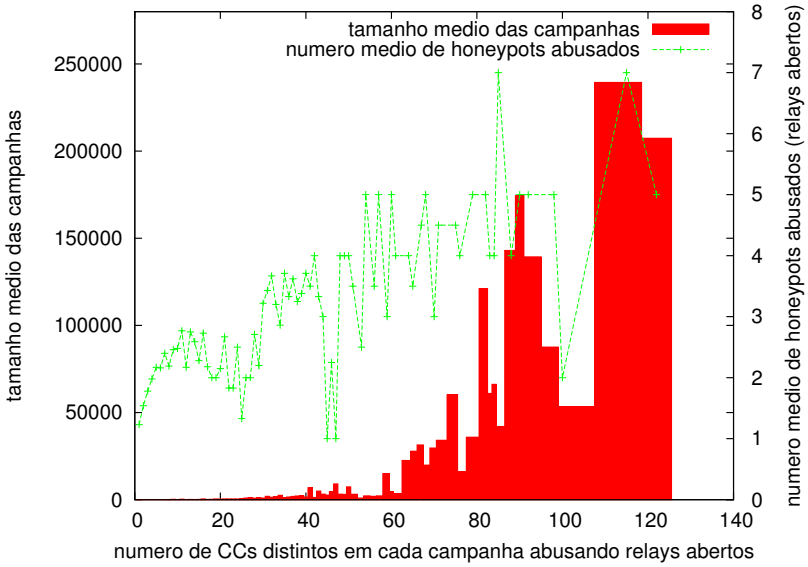

Figura 5.8. Tamanho médio das campanhas e número médio de *honeypots* abusados como *relays*, por quantidade de *country codes* de origem

Uma explicação para essas diferenças parte da observação de que esses abusos estão em pontos diferentes das cadeias de máquinas estabelecidas pelos *spammers*. Abusos a *relays* abertos e máquinas de usuários infectadas ocorrem no último estágio da cadeia e todos os *proxies* abertos que estão disseminando uma dada campanha direcionam seus abusos para os mesmos *relays* abertos e máquinas infectadas durante um mesmo período, como mencionado anteriormente. Isso explicaria o comportamento observado na Figura 5.8. No caso da Figura 5.7, uma etapa intermediária a cadeia é observada e quanto mais distribuída é a atividade do *spammers* neste passo, menos mensagens são observadas em cada *honeypot* que participa da cadeia. Na verdade, as campanhas pequenas observadas nesses cassos podem ser muito maiores que aquelas originadas de menos países. Como os *spammers* mais sofisticados conseguem distribuir mais os seus abusos, cada máquina sendo abusada por eles tem a impressão de que a campanha sendo disseminada é pequena.

Neste Capítulo, demonstrou-se alguns aspectos de estratégias de disseminação de campanhas de *spam* que podem ser observados a partir do entendimento de que os *honeypots* implantados para coleta de *spam* encaixam-se em uma cadeia de máquinas, intermediando conexões entre *spammers* e os servidores de *e-mail* alvo dos abusos. Essa análise complementa as estratégias descritas no Capítulo 4.

Capítulo 6

Conclusões e Trabalhos Futuros

Nesta dissertação, foi apresentada uma metodologia para caracterização de estratégias de disseminação de *spams*. O processo de análise se inicia com a extração de características essenciais das mensagens coletadas. A partir dessas características, as mensagens sumarizadas são processadas para se obter agrupamentos contendo as mensagens derivadas de uma mesma mensagem original por técnicas de ofuscação. As mensagens de cada agrupamento são, então, avaliadas em busca de correlações invariantes, na forma de características que co-ocorrem frequentemente. Dados os grandes volumes de dados e a necessidade de automação do processo de análise, técnicas de mineração de dados foram empregadas em cada etapa do processo.

Para agrupar as mensagens em campanhas, foi proposta uma técnica baseada na inserção de características relevantes extraídas das mensagens de *spam* (*layout*, idioma, assunto e fragmentos de URL). Dessa forma, as mensagens que compartilham um caminho comum na árvore e diferem por características infrequentes são agrupadas em campanhas.

A metodologia foi testada em um conjunto de dados de aproximadamente 350 milhões de mensagens coletadas por *honeypots* de baixa-interatividade implantados em redes brasileiras e que simulam *proxies* e *relays* abertos, comumente abusados por *spammers* para o envio de mensagens não-solicitadas.

A partir da identificação das campanhas de *spam*, foi aplicado um algoritmo de mineração de regras de associação para revelar padrões relevantes de *spamming*. Foi possível determinar que abusos a *proxies* HTTP e SOCKS originam-se de poucas máquinas e são fortemente correlacionados com o *Country Code* de origem da mensagem, o que sugere que tais abusos são originados pelos próprios *spammers*. Por outro lado, abusos a *relays abertos* são mais dispersos e se originam de muitas fontes simultaneamente, além de não guardarem relação com o idioma e destino do *spam*. A aplicação de um algoritmo de mineração de regras de associação aos dados das campanhas de *spam* também determinou relações entre sistemas operacionais e os tipos de abuso, indicando que os sistemas Linux e Solaris raramente são utilizados como origem dos abusos a *proxies* HTTP e SOCKS. Analisou-se também as cadeias de máquinas criadas por *spammers* para disseminar suas mensagens e que permitem aumentar o conhecimento sobre a forma como eles atuam, a partir das conexões intermediadas pelos

sensores. A análise dos encadeamentos mostrou que *spammers* encadeiam *proxies* abertos com *relays* abertos e máquinas de usuários na rede, e, portanto, combater *proxies* abertos ainda é necessário, mesmo com o crescimento no uso de *botnets* para a disseminação de *spams*.

Considera-se como principais contribuições do trabalho a proposição da metodologia de identificação de campanhas baseada em uma Árvore de Padrões Frequentes, bem como a escolha das características de cada mensagem considerada, e os padrões de comportamento identificados (Calais et al., 2008b,a, 2009a,b,c). Alguns desses padrões são novos e outros eram conhecidos, mas não haviam ainda sido demonstrados em trabalhos de cunho científico.

Como trabalhos futuros, pode-se citar diversas frentes de continuidade do projeto. A primeira é a validação da metodologia de identificação de campanhas de *spam* e uma análise comparativa com outras técnicas de detecção de campanhas, tanto em termos de acurácia e precisão na detecção das campanhas quanto na eficiência do uso de recursos computacionais. Considera-se, inclusive, a aplicação da árvore de padrões frequentes para determinar agrupamentos em outros domínios de aplicação, que não o *spam*. A segunda ramificação do trabalho é a aplicação da árvore de padrões frequentes como técnica para filtragem de *spams*: se mensagens legítimas forem inseridas na árvore, elas não formarão os padrões de ofuscação típicos ilustrados neste trabalho, e isso permitiria a distinção entre mensagens legítimas e não-solicitadas. Finalmente, pretende-se analisar as estratégias de disseminação de *spams* de forma *online*, ou seja, determinar padrões de *spamming* à medida que eles surgem. Essa etapa consiste em implementar uma versão incremental da árvore de padrões frequentes e estendê-la para possibilitar a identificação de padrões evolutivos e aspectos dinâmicos da disseminação de *spams*.

Referências Bibliográficas

Anderson, D. S.; Fleizach, C.; Savage, S. & Voelker, G. M. (2007). Spamscatter: Characterizing internet scam hosting infrastructure. In *USENIX Security*.

Andreolini, M.; Bulgarelli, A.; Colajanni, M. & Mazzoni, F. (2005). Honeyspam: honeypots fighting spam at the source. In *SRUTI'05: Proceedings of the Steps to Reducing Unwanted Traffic on the Internet on Steps to Reducing Unwanted Traffic on the Internet Workshop*, pp. 11--11, Berkeley, CA, USA. USENIX Association.

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

Cavnar, W. & Trenkle, J. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161--175, Las Vegas, US.

Cerf, V. G. (2005). Spam, spim, and spit. *Commun. ACM*, 48(4):39--43.

CERT.br (2009). Cartilha de segurança – glossário cert.br. `http://cartilha.cert.br/glossario/#b`.

Chowdhury, A. K. A. & Alspector, J. (2004). The impact of feature selection on signature-driven spam detection. *Proceedings of the 1st Conference on Email and Anti-Spam (CEAS). Mountain View, CA*.

Cook, D.; Hartnett, J.; Manderson, K. & Scanlan, J. (2006). Catching spam before it arrives: domain specific dynamic blacklists. In *ACSW Frontiers '06: Proceedings of the 2006 Australasian workshops on Grid computing and e-research*, pp. 193--202, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Cooke, E.; Jahanian, F. & McPherson, D. (2005). The zombie roundup: understanding, detecting, and disrupting botnets. In *SRUTI'05: Proceedings of the Steps to Reducing Unwanted Traffic on the Internet on Steps to Reducing Unwanted Traffic on the Internet Workshop*, pp. 6--6, Berkeley, CA, USA. USENIX Association.

Cranor, L. F. & LaMacchia, B. A. (1998). Spam! *Commun. ACM*, 41(8):74--83.

Dhinakaran, C.; Lee, J. K. & Nagamalai, D. (2007). Characterizing spam traffic and spammers. In *ICCIT '07: Proceedings of the 2007 International Conference on Convergence Information Technology*, pp. 831--836, Washington, DC, USA. IEEE Computer Society.

Gansterer, W. N. & Ilger, M. (2007). Analyzing uce/ube traffic. In *ICEC '07: Proceedings of the ninth international conference on Electronic commerce*, pp. 195--204, New York, NY, USA. ACM.

Georgiou, E.; Dikaiakos, M. D. & Stassopoulou, A. (2008). On the properties of spam-advertised url addresses. *J. Netw. Comput. Appl.*, 31(4):966--985.

Gomes, L. H.; Cazita, C.; Almeida, J. M.; Almeida, V. & Wagner Meira, J. (2004). Characterizing a spam traffic. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conf. on Internet measurement*, pp. 356--369, New York, NY, USA. ACM.

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

Guerra, P. H. C.; Guedes, D.; Jr., W. M.; Hoepers, C. & Steding-Jessen, K. (2008a). Caracterização de estratégias de disseminação de spams. In *26o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Rio de Janeiro, RJ.

Guerra, P. H. C.; Guedes, D.; Jr., W. M.; Hoepers, C.; Steding-Jessen, K. & Chaves, M. H. (2009a). Caracterização de encadeamento de conexões para envio de spams. In *27o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Recife, PE.

Guerra, P. H. C.; Pires, D.; Guedes, D.; Wagner Meira, J.; Hoepers, C.; Chaves, M. H. P. C. & Steding-Jessen, K. (2009b). Spamming chains: A new way of understanding spammer behavior. In *Proceedings of the 6th Conference on e-mail and anti-spam (CEAS)*, Mountain View, CA.

Guerra, P. H. C.; Pires, D.; Guedes, D.; Wagner Meira, J.; Hoepers, C. & Steding-Jessen, K. (2008b). A campaign-based characterization of spamming strategies. In *Proceedings of the 5th Conference on e-mail and anti-spam (CEAS)*, Mountain View, CA.

Guerra, P. H. C.; Pires, D.; Ribeiro, M. T.; Guedes, D.; Jr., W. M.; Hoepers, C.; Chaves, M. H. P. C. & Steding-Jessen, K. (2009c). Spam miner: A platform for detecting and characterizing spam campaigns. in: International conference on knowledge discovery and data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, França.

Han, J.; Pei, J.; Yin, Y. & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, 8(1):53--87.

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

ISO (2006). ISO 3166: Codes for the representation of names of countries and their subdivisions – Part 1: Country codes. `http://www.iso.org/iso/country_codes.htm`.

Jakobsson, . M. & Myers, S. (2006). *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. Wiley-Interscience.

Killalea, T. (2000). RFC 3013: Recommended Internet Service Provider Security Services and Procedures. `http://www.ietf.org/rfc/rfc3013.txt`.

Kolcz, A. & Chowdhury, A. (2007). Hardening fingerprinting by context. In *Proceedings of the 4th Conference on Email and Anti-Spam (CEAS). Mountain View, CA*.

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

Kreibich, C.; Kanich, C.; Levchenko, K.; Enright, B.; Voelker, G. M.; Paxson, V. & Savage, S. (2008). On the spam campaign trail. In *LEET'08: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pp. 1--9, Berkeley, CA, USA. USENIX Association.

Lee, W.; Wang, C. & Dagon, D. (2007). Honeynet-based botnet scan traffic analysis. In Lee, W.; Wang, C. & Dagon, D., editores, *Botnet Detection*, volume Volume 36, pp. 25--44. Springer Berlin / Heidelberg.

Li, F. & Hsieh, M.-H. (2006). An empirical study of clustering behavior of spammers and group-based anti-spam strategies. *Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS). Mountain View, CA*.

Lindberg, G. (1999). RFC 2505: Anti-Spam Recommendations for SMTP MTAs. `http://www.ietf.org/rfc/rfc2505.txt`.

McCarty, B. (2003). Botnets: big and bigger. *Security and Privacy, IEEE*, 1(4):87–90.

Mehta, B.; Nangia, S.; Gupta, M. & Nejdl, W. (2008). Detecting image spam using visual features and near duplicate detection. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pp. 497--506, New York, NY, USA. ACM.

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

Milletary, J. (2005). Technical trends in phishing attacks. Technical report, CERT Coordination Center, Carnegie Mellon University. `http://www.cert.org/archive/pdf/Phishing_trends.pdf`.

Musat, C. (2006). Layout based spam filtering. *Transactions on Engineering, Computing and Technology*.

Nooy, W. d.; Mrvar, A. & Batagelj, V. (2004). *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, New York, NY, USA.

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

Pathak, A.; Hu, Y. C. & 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*, pp. 1--9, Berkeley, CA, USA. USENIX Association.

Paulson, L. D. (2005). No quick fix for spam. *IT Professional*, 7(3):11--14.

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

Pu, C. & 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). Mountain View, CA.*

Ramachandran, A. & 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*, pp. 291--302, New York, NY, USA. ACM.

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

Sophos.com (2004). The Spam Economy: The Convergent Spam and Virus Threats. August 2004. `http://www.sophos.com/whitepapers/Sophos_spam-economy_wpus.pdf`.

SpamAssassin (2007). `http://spamassassin.apache.org`.

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

Stern, H. (2008). A survey of modern spam tools. *Proceedings of the 5th Conference on Email and Anti-Spam (CEAS). Mountain View, CA.*

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

Wang, Z.; Josephson, W.; Lv, Q.; Charikar, M. & Li, K. (2007). Filtering image spam with near-duplicate detection. In *Proceedings of the Fourth Conference on Email and Anti-Spam (CEAS). Mountain View, CA.*

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

Xie, Y.; Yu, F.; Achan, K.; Panigrahy, R.; Hulten, G. & Osipkov, I. (2008). Spamming botnets: signatures and characteristics. *SIGCOMM Comput. Commun. Rev.*, 38(4):171--182.

Yeh, C.-C. & Lin, C.-H. (2006). Near-duplicate mail detection based on url information for spam filtering. In *Information Networking. Advances in Data Communications and Wireless Networks*, pp. 842–851. Springer Berlin / Heidelberg.