ANÁLISE DO TRÁFEGO DE SPAM COLETADO AO REDOR DO MUNDO

PEDRO HENRIQUE BRAGIONI LAS CASAS

ANÁLISE DO TRÁFEGO DE SPAM COLETADO AO REDOR DO MUNDO

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: DORGIVAL GUEDES NETO

Belo Horizonte

Fevereiro de 2013

Bragioni Las Casas, Pedro Henrique
L337a Análise do Tráfego de Spam Coletado ao Redor do Mundo / Pedro Henrique Bragioni Las Casas. — Belo Horizonte, 2013
xxi, 124 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Dorgival Guedes Neto

1. Spam. 2. Botnet. I. Título.

CDU 519.6*22 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Análise do tráfego de spam coletado ao redor do mundo

PEDRO HENRIQUE BRAGIONI LAS CASAS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. DORGIVAL OLAVO GUEDES NETO - Orientador
Departamento de Ciência da Computação - UFMG

DRA. CRISTINE HOEPERS
CGI.br

PROF. HUMBERTO TORRES MARQUES NETO
Instituto de Ciências Exatas e Informática - PUC/MG

DR. KLAUS STEDING -Jessen
CGI.br

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 06 de março de 2013.

Resumo

Diversos esforços têm sido feitos para se criar uma visão abrangente do tráfego de *spam*. Entretanto, observações em pontos isolados da Internet estão sempre limitadas por fatores de localidade espacial. Esta dissertação pretende acrescentar uma dimensão a essa análise ao contrastar amostras de tráfego de spam coletadas simultaneamente em diferentes pontos. Além disso, neste trabalho objetiva-se também avaliar o fator tempo no tráfego de *spam*, e os impactos causados por ele.

Nossas análises indicam que fatores como localização e conectividade têm impacto sensível sobre o tráfego observado, porém certas características, como perfis das mensagens enviadas por diferentes protocolos, endereços de origem e padrões de teste dos spammers se repetem ao redor do mundo. Identificamos também que o tráfego de *spam* é muito variável ao longo do tempo, apresentando padrões diferentes em momentos distintos.

Abstract

Several efforts have been pursued to create a comprehensive view of spam traffic. However, observations at isolated points of the Internet are always limited by factors of spatial locality. This dissertation aims to add a dimension to this analysis by contrasting samples of spam traffic collected simultaneously at different points. Furthermore, this study aims to evaluate the time factor in the spam traffic, and the impacts caused by it.

Our analyses indicate that factors such as location and connectivity have significant impact on the observed traffic, but certain features, such as profiles of messages sent by different protocols, source addresses and test patterns from spammers repeat themselves around the world. We also identified that the spam traffic varies considerably over time, with different patterns in different times.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Ainda hoje, *spam* é um dos grandes problemas presentes na Internet. Relatórios recentes apontam que cerca de 75% de todos os e-mails que trafegam na rede são *spam* [Symantec, 2011]. Devido ao baixo custo para se disseminar esse tipo de mensagem, além da alta rentabilidade (Rao & Reiley [2012] mostram que a margem de lucro é de mais de 100:1 com relação aos gastos), esse tipo de mensagens indesejada é cada vez mais propagada. Com isso, estudos mostram que o abuso causado pelos *spams* acarretam na ordem de 20 bilhões de dólares de prejuízo anualmente às empresas e à sociedade em geral [Sipior et al., 2004; Rao & Reiley, 2012].

Além do caráter indesejado, *spams* são muitas vezes relacionados com o envio de *phishing*, mensagens utilizadas para obtenção de dados pessoais com objetivos ilícitos, além da propagação de códigos maliciosos (*malwares*), como cavalos de tróia, vírus e *worms* [Newman et al., 2002], tornando-os ainda mais nocivos para a rede e seus usuários. O combate ao *spam* é caracterizado pela constante evolução das técnicas de detecção dessas mensagens, uma vez que as ferramentas utilizadas pelos *spammers* se mostram cada vez mais sofisticadas, reduzindo a eficácia dos filtros anti-*spam* e a rastreabilidade dos *spammers* [Goodman et al., 2007]. Um exemplo disso é o crescimento na utilização de máquinas infectadas por *malwares*, como os *bots*, para o envio de spam e *phishing* [Xie et al., 2008a], permitindo que o *spammer* permaneça no anonimato.

Mesmo com o desenvolvimento de técnicas de combate, é necessário um esforço contínuo para entender a forma de atuação dos *spammers* ao enviar mensagens indesejadas pela Internet, devido à sua natureza evolutiva. Essa evolução acontece tanto no modo como *spammers* disseminam suas mensagens pela rede, buscando maximizar o volume de mensagens enquanto mantêm sua identidade oculta, quanto na forma como compõem o conteúdo das mesmas [Pu, 2006]. Logo, analisar o comportamento dos *spammers* ao longo do tempo permite identificar padrões e características que po-

dem ajudar na evolução das ferramentas de combate ao *spam*. Por exemplo, o tráfego gerado por máquinas participantes de *botnets* tende a apresentar características similares. Sendo assim, o entendimento desses padrões pode ser útil para identificar *bots* e as redes de que fazem parte.

Para entender o comportamento dos *spammers* na rede, realizamos nesta dissertação a caracterização do tráfego de *spam* utilizando dados coletados por *honeypots* de baixa interatividade localizados em redes intermediárias. O ponto principal da análise aqui realizada está no fato dos dados serem coletados em diversos pontos distintos da Internet, fornecendo assim visões diferentes do problema e não apenas uma visão proveniente de apenas uma fonte de dados, como visto em diversos trabalhos anteriores. Enquanto Silva et al. [2011] avaliaram o impacto de diferentes configurações dos *honeypots* na coleta de *spams*, aqui consideramos a influência da posição geográfica na forma com que o tráfego de *spam* incide na máquina. Com isso, pretendemos adicionar uma dimensão às análises existentes, ao contrastar amostras de um grande volume de tráfego, oriundos de diversas localidades, reduzindo distorções comuns em coletas restritas a um único local.

Além disso, avaliamos também a questão temporal inerente ao tráfego de *spam*. Dado seu comportamento de constante mudança, visando ludibriar as técnicas de combate, o tráfego gerado pelos *spammers* tende a sofrer alterações ao longo do tempo. Para avaliarmos esse fator, utilizamos uma carga de trabalho que se estendeu por 176 dias, em que foram coletados cerca de 1 bilhão e 800 milhões de mensagens enviadas a partir de 180 mil endereços IP. A coleta foi feita através de 8 *honeypots* de baixa interatividade, localizados em 7 *country codes* distintos.

Os resultados indicam que algumas características, como endereços de origem e padrões de teste utilizados pelos *spammers*, se repetem nas diversas localidades observadas. Além disso, mostraram que o comportamento das campanhas de *spam* que aparecem nos *honeypots* possui grande influência no tráfego gerado, causando, por exemplo, grandes variações na quantidade de endereços IP observados.

Considerando-se o tráfego de *spam* enviado através de conexões a *open relays* SMTP, os resultados evidenciaram que esse relaciona-se diretamente com as mensagens de teste enviadas pelos *spammers*. Observou-se também que os transmissores responsáveis por esse tráfego enviam poucas mensagens cada um, com padrões ao longo do dia, sugerindo participação em *botnets*. Já no tráfego gerado pelos demais protocolos (SOCKS e HTTP), os atacantes enviam mensagens continua e regularmente, sugerindo a utilização de infra-estruturas dedicadas.

Durante o período de análise, um dos *honeypots* teve seu endereço IP trocado, o que nos permitiu observar o comportamento associado ao aparecimento de um novo

honeypot na rede. Os resultados evidenciam que o comportamento dos *spammers* ao descobrirem uma máquina vulnerável é de começar o envio de mensagens utilizando o protocolo SOCKS indiscriminadamente, enquanto o início do envio de mensagens utilizando SMTP possui restrições.

Uma das principais contribuições realizadas nesta dissertação está no fato de que mostramos que aspectos como o tipo do tráfego de *spam* que incide um *honeypot*, além de sua intensidade, estão mais relacionados aos aspectos de conexão da máquina, como a qualidade da rede em que ela se encontra, do que a sua localização.

Por fim, uma contribuição importante deste trabalho é com relação ao entendimento do tráfego de *spam* ao longo do tempo. Através de nossas análises identificamos que o tráfego gerado pelos *spammers* é muito variável, apresentando diferentes padrões em momentos distintos da análise. Esses resultados são importantes para que novas técnicas de combate ao *spam*, entendendo a instabilidade desse tráfego ao longo do tempo, consigam evoluir junto ao *spam* e apresentar sucesso ao detectá-lo.

O restante desta dissertação está organizado da seguinte forma: os trabalhos relacionados são discutidos no capítulo 2; o capítulo 3 descreve a metodologia de caracterização e o capítulo 4 apresenta os resultados mais relevantes. Finalmente, o capítulo 5 apresenta algumas conclusões e sugere trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

Considerando-se o foco desta dissertação, os trabalhos relacionados podem ser divididos em duas partes: a primeira dedica-se a analisar o contexto de *spam* e as estratégias de combate à estes. Em seguida, são apresentados trabalhos relacionados ao uso de *botnets* para disseminação de *spams*.

2.1 Caracterização do Tráfego de *Spams*

Diversos trabalhos estudam características apresentadas pelos *spams*, uma vez que peculiaridades presentes nas mensagens podem ser fundamentais para detecção dos mesmos e na identificação daqueles que os enviam.

Gomes et al. [2007] analisaram uma carga de trabalho composta por mensagens eletrônicas de usuários de uma universidade brasileira. Naquele trabalho foram destacadas diversas características capazes de diferenciar *spams* de mensagens legítimas como, por exemplo, o tamanho médio das mensagens, em que as legítimas são, em média, de seis a oito vezes maiores que *spams*, mesmo ambos apresentando uma distribuição Log-Normal. Além disso, os autores derivaram modelos representando a taxa de chegada de *spams* que mostram 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. Em uma extensão daquele trabalho, os mesmos autores indicaram que o tráfego legítimo de correio eletrônico apresenta menor entropia que o tráfego gerado pelos *spammers*, os quais, geralmente, enviam e-mails indistintamente para os seus alvos [Gomes et al., 2009]. Ao contrário desses trabalhos, cujos dados são provenientes de um único servidor de destino, esta dissertação utiliza conjuntos de dados provenientes de servidores intermediários localizados em diferen-

tes posições geográficas, objetivando entender a influência da localidade no tráfego de *spam*.

Alguns trabalhos focam em entender o conteúdo gerado por *spams*. Kim & Choi [2008] caracterizaram o tráfego de *spam* a partir de dados da camada de aplicação coletados em servidores de correio eletrônico de destino, mostrando que, em sua grande maioria, *spams* são baseados em texto, o que faz com que seu tamanho seja pequeno. Os resultados indicam também que o intervalo entre chegadas de *spams* é bem inferior ao intervalo entre e-mails legítimos (menor que 5 segundos em 95% dos casos). Dhinakaran & Lee [2007] caracterizam os *spams* baseando-se na divisão daqueles que enviam anexos e dos que utilizam apenas texto na mensagem. Com isso, eles mostram que a grande maioria dos *spams* não possuem anexo e em consequência, apresentam tamanho pequeno, sendo menor que 3KB, confirmando a afirmação de [Kim & Choi, 2008]. Assim como nesses trabalhos, em nossos resultados mostramos que mais de 90% dos *spams* coletados são menores que 1KB.

Guerra et al. [2008b] propõem uma metodologia de caracterização de campanhas de *spams* baseadas em seu conteúdo. O trabalho utiliza árvores de padrões frequentes para geração das campanhas e mostra como os responsáveis pelas campanhas tentam ocultá-las. O método de construção de campanhas proposto naquele trabalho foi utilizado nesta dissertação. Através dos resultados, os autores mostraram que cerca de 86% das mensagens são constituídas por HTML e que mais de 96% do total de mensagens apresenta pelo menos uma URL em seu conteúdo. Com relação ao modo de disseminação das mensagens pelas campanhas, é apresentado que os principais protocolos utilizados são SOCKS e HTTP, assim como nessa dissertação. Estendendo esse trabalho, Totti et al. [2012] realizaram a caracterização temporal da disseminação das campanhas de *spams*. Os resultados obtidos mostraram que as mensagens de uma campanha são distribuídas como rajadas ao longo do tempo, com períodos de atividade curtos, seguidos por longos períodos de inatividade.

Caracterizando a natureza evolutiva do problema de *spam*, Pu [2006] analisa a evolução dos *spammers* na forma com que suas mensagens são construídas. O autor conseguiu mostrar que, ao longo do tempo, determinadas técnicas de ofuscação deixam de ser usadas, muitas vezes em virtude de mudanças no ambiente, como a correção de alguma falha de segurança. Por outro lado, algumas técnicas se mantêm por mais tempo.

Considerando apenas as propriedade de rede do tráfego de *spam*, temos o trabalho de Ouyang et al. [2011], que mostrou que métricas derivadas de um único pacote de rede ou baseadas apenas em propriedades de fluxos não são eficientes para classificação de *spams* por si só, mas que a combinação desses conjuntos de métricas aumenta a

eficácia da classificação de *spams*.

Richard Clayton [2006] propôs um projeto chamado *SpamHINTS*, que visa desenvolver técnicas para, através da análise de pacotes do protocolo SMTP, inferir padrões indicativos de atividades de envio de *spam*. Venkataraman et al. [2007] estudaram a eficácia de utilizar o comportamento histórico de endereços de IP para prever se um e-mail é legítimo ou *spam*, uma vez que, de acordo com o trabalho, e-mails legítimos são provenientes de endereços IP's cujo tempo de vida é duradouro, enquanto *spams* provêm de IP's efêmeros.

Outros esforços para detecção de *spam* incluem o algoritmo proposto por Sperotto et al. [2009], que analisa as características do tráfego de rede, como tempo de inatividade e quantidade de picos no fluxos de pacotes SMTP. Os autores utilizaram dados provenientes da rede de uma universidade alemã, e informações de listas de bloqueio DNS para validar o algoritmo proposto. Além disso, Taveira & Duarte [2008] propuseram um mecanismo anti-*spam* baseado na reputação e autenticação dos usuários tentando minimizar os falso positivos ao classificar os e-mails. Beverly & Sollins [2008] apresentam uma técnica de detecção de *spam* que utiliza características da camada de transporte. O método, chamado *SpamFlow*, utiliza *Support Vector Machines* (SVM) para classificar as mensagens. Sanchez et al. [2011] utilizam uma técnica de bloqueio de *spam* baseada na distinção entre máquinas de usuários finais e máquinas de servidores de e-mail legítimos. O algoritmo foi motivado pelas observações dos autores de que a maior parte das atividades de *spam bots* são realizadas por máquinas de usuários finais, em detrimento de servidores de e-mail. Nesta dissertação, objetivamos entender melhor as características de rede do tráfego de *spam*, para auxiliar o desenvolvimento de ferramentas de detecção, como nos trabalhos citados anteriormente.

Como mostrado por esses trabalhos, é extremamente importante o entendimento da natureza do *spam*, principalmente por ser um problema em constante evolução. Dessa forma, esta dissertação apresenta uma visão global desse tráfego, em pontos intermediários da rede, o que permite que administradores entendam a natureza do tráfego que passa por suas redes, auxiliando na melhoria das técnicas de combate ao *spam*.

2.2 Caracterização do Comportamento de *Spammers* e *Botnets* no Envio de *Spams*

Botnets têm sido utilizados em larga escala para o envio de *spams*. Com a utilização de inúmeros *bots* distribuídos, *spammers* enviam milhares de mensagem em um período

curto de tempo. Relatórios recentes mostram que 88% do total de *spams* enviados são provenientes dessas redes [Symantec, 2011]. Desta forma, o estudo de como essas redes trabalham e como é feita a disseminação de *spams* através delas é extremamente importante para o melhor entendimento do tráfego de *spam* e para o desenvolvimento de novas estratégias de combate.

Com base nesse tema, Stone-Gross et al. [2011] destacaram o gerenciamento de campanhas de *spam* enviadas por *botnets* do ponto de vista do *botmaster*, indivíduo responsável pela *botnet* e por sua manutenção. Analisando um conjunto de dados reais de mais de 2,5 TB de dados, o trabalho mostrou que apenas 30% de todo o tráfego foi efetivamente entregue, com as falhas sendo causadas pela utilização de endereços de e-mail inválidos, identificação por *blacklists* e *timeout* das conexões. Sob o ponto de vista econômico, os autores estimaram que o lucro dos *spammers* observados, que fornecem *spam* como um serviço, tenha sido entre US$1,7 milhões e US$ 4,2 milhões, de junho de 2009 a outubro de 2010.

John et al. [2009] analisaram *botnets* que produzem tráfego de spam tanto recebido quanto enviado por uma rede universitária norte-americana. Aquele trabalho mostrou que apenas 6 *botnets* enviaram 79% dos *spams* recebidos por aquela rede universitária. Segundo os autores, *botnets* podem ser classificadas em pelo menos duas categorias, sendo uma caracterizada por *botnets* que enviam o maior número de *spams* possível, enquanto a outra é formada por *botnets* que possuem um controle de fluxo e, desse modo, enviam *spams* moderadamente, sem saturar a rede a qual se encontram. Nos resultados da dissertação, mostramos que, assim como nesse trabalho, pequenos grupos de IP's são responsáveis por grande parte dos *spams* recebidos. Por exemplo, apenas um Sistema Autônomo, composto por 159 endereços IP's, foi responsável por cerca de 50% das mensagens coletadas durante o período analisado.

Ramachandran & Feamster [2006] determinaram características de tráfego na camada de rede que são comuns a *spammers*. Utilizando um conjunto de dados contendo mais de 10 milhões de mensagens de e-mail classificadas como *spam*, os autores mostraram que a grande maioria delas são provenientes de pequenas porções de endereços IP's, fato comprovado por Kokkodis & Faloutsos [2009]. Além disso, os autores mostraram que a maior parte dos *spams* recebidos são disseminados através de *bots*, que enviam apenas uma pequena fração de mensagens por um pequeno período de tempo. Baseando-se nas diferenças encontradas entre *spammers* e usuários legítimos através de características da camada de rede, Hao et al. [2009] propuseram um método de detecção de *spammers*. Através dos resultados encontrados, os autores mostraram que uma grande fração dos endereços de IP's de *spammers* são provenientes de poucos ASes, o que também pôde ser visto nos resultados encontrados nesta dissertação. Além

disso, foi observado nesse trabalho que a atividade de usuários legítimos durante o dia é concentrada em algumas horas, com pico por volta das 10 horas da manhã, enquanto a atividade de *spammers* é espalhada por todo o dia. Nesta dissertação, mostramos que a atividade dos *spammers* utilizando os protocolos HTTP e SOCKS se mantêm constante durante todo o dia, enquanto a atividade utilizando SMTP possui picos durante o dia.

Guerra et al. [2009] caracterizaram o encadeamento de conexões para o envio de *spam*. Os autores mostraram que *spammers* encadeiam o abuso a *open proxies* com abusos a *open relays*, máquinas infectadas ou outros *open proxies*. Como conclusão, foi mostrado que *spammers* tentam se conectar poucas vezes a cada máquina abusada, além de enviar poucas mensagens de cada uma delas, de forma que a detecção da máquina seja difícil. Em um trabalho posterior [Guerra et al., 2010], os mesmos autores caracterizaram *spammers* a partir de listas de destinatários. Segundo os autores, 72% dos destinatários de um endereço IP também é utilizado por outro endereço. Essa sobreposição de lista de destinatários pode indicar que esses endereços IP's pertencem a um mesmo grupo, que age coordenadamente.

Considerando a origem do *spam* na rede, Las-Casas et al. [2012b] propuseram uma ferramenta de detecção de *spammers* que analisa o tráfego de saída de um provedor. Utilizando métricas como número de mensagens enviadas durante um determinado período de tempo, tamanho médio das transações SMTP e tempo médio entre chegada de mensagens, os autores foram capazes de identificar *spammers* com eficácia elevada e baixas taxas de falso positivo e falso negativo. Nesta dissertação também avaliamos características como número de mensagens e tamanho médio das mensagens, porém, diferente do trabalho de Las-Casas et al. [2012b], em que os dados são coletados na origem, aqui utilizamos informação do tráfego de *spam* provenientes de máquinas localizadas em redes intermediárias. Estendendo esse trabalho, os autores caracterizaram o impacto temporal no método de detecção de *spammers* proposto [Las-Casas et al., 2012a]. Nesse trabalho foi mostrado que, mesmo durante um período de 28 dias, o método se mantém estável, e que o uso de aprendizado ativo para seleção do conjunto de treino do método resulta em ganho de desempenho de 8% ao classificar usuários legítimos.

Em outro trabalho relacionado, Schatzmann et al. [2009] propuseram a detecção de *spammers* no nível de sistemas autônomos (AS), coletando e combinando as visões locais de múltiplos servidores de e-mail destinatários. Nesse contexto, essa dissertação pode ajudar na busca por mais informações sobre as características desse tipo de tráfego.

Xie et al. [2008b] utilizaram o conteúdo da mensagem para identificar campanhas

de *spam* e os *botnets* responsáveis por essas campanhas. Naquele trabalho, os autores propuseram um sistema chamado *AutoRE*, que agrupa os *spams* em campanhas utilizando as URLs das mensagens. A premissa do trabalho é que *spams* enviados por *botnets* tendem a ser similares em seu conteúdo. Os resultados obtidos pelos autores mostram que *bots* são largamente espalhados pela Internet, e que não possuem padrões de envio distintos de servidores normais quando observados individualmente, o que sugere que a detecção de *bots* individuais continua sendo um desafio. Zhuang et al. [2008] também utilizaram o conteúdo das mensagens para clusterizá-las em campanhas e, em seguida, uni-las para assim identificar *botnets*. Seus resultados mostraram que cerca de 50% das *botnets* possuem mais de mil *bots* e que 80% das *botnets* utilizam menos da metade dos *bots* de cada vez.

Neste trabalho realizamos uma caracterização do tráfego de *spam* que, diferente de trabalhos anteriores, possui uma visão global, evitando assim possíveis distorções por se considerar um ponto único da rede. Acreditamos que este trabalho pode ser útil para a evolução dos trabalhos que consideram o combate ao *spam*, uma vez que nele é mostrado como o tráfego de *spam* se comporta em diferentes locais e quais padrões e características são importantes para a contínua identificação das mensagens indesejadas.

Capítulo 3

Metodologia

Antes de podermos avaliar os resultados, é importante entender como a análise foi desenvolvida. A metodologia seguida para análise dos dados inicia-se com a coleta das mensagens de *spam* e das informações de abuso de rede associadas a cada uma. Em seguida, os dados coletados foram processados e então analisados com relação aos fatores temporais e às campanhas de *spam*.

3.1 Coleta de Dados

A captura e coleta dos dados utilizados neste trabalho foi realizada através de uma arquitetura composta por um conjunto de sensores que implementam *honeypots* de baixa interatividade [Provos & Holz, 2007]. *Honeypots* são recursos computacionais utilizados para coletar, analisar, bloquear ou desviar ataques e/ou abusos provenientes da rede visando algum serviço ou sistema. O *honeypot* oferece o serviço, aparentando ser legítimo, para dessa forma atrair o seu alvo e ser abusado e/ou atacado pelo mesmo e, então, registrar sua atividade. Por estar em um ambiente controlado, esses ataques não extrapolam o ambiente pré-determinado no *honeypot*.

Os *honeypots* podem ser divididos em duas categorias, que refletem a forma como se relacionam com seu alvo: baixa e alta interatividade. *Honeypots* de baixa interatividade apenas emulam os serviços a serem abusados, não sendo implementações reais. Por outro lado, *honeypots* de alta interatividade fornecem um sistema completo para uso do alvo, contendo sistemas operacionais, aplicações e serviços reais. Esses sistemas são protegidos e possuem limitações para impedir que outros sistemas sejam atacados a partir do *honeypot*. Neste trabalho utilizamos *honeypots* de baixa interatividade para captura dos dados.

Entre as principais vantagens do *honeypot* está a interação direta com o responsável pelo ataque e/ou abuso, possibilitando a captura de informações importantes sobre o abuso, e assim, a análise e avaliação profunda e real do problema. Além disso, não há incidência de falsos positivos, uma vez que todo o tráfego direcionado ao *honeypot* é, por natureza, anômalo ou malicioso. Por outro lado, existe a desvantagem do *honeypot* fornecer uma visão limitada dos ataques, uma vez que apenas aqueles diretamente direcionados a ele são registrados. Para minimizar este problema, utilizamos diversos *honeypots*, localizados em pontos distintos do planeta. Com isso, temos diferentes visões de como o tráfego de *spam* se comporta ao redor do mundo. Outra desvantagem está na possibilidade dos atacantes perceberem que a máquina sendo abusada é um *honeypot*, o que poderia comprometer o processo de coleta. Para tentar identificar se a máquina abusada está realmente repassando as mensagens, atacantes utilizam mensagens de teste. Essas mensagens são devidamente tratados pelos *honeypots*, como será explicado posteriormente, o que dificulta a identificação desses pelos *spammers*.

Os *honeypots* utilizados emulam *open proxies* e *open mail relays*, máquinas na rede que são tradicionalmente abusadas por *spammers*. *Proxy* é um servidor intermediário que realiza conexões em nome de outros clientes. Um *open proxy* permite acesso de qualquer origem a qualquer IP e porta de destino. O objetivo dos *spammers* ao utilizar um *open proxy* é dificultar sua identificação, uma vez que a máquina abusada enxergará apenas o *proxy* que está re-encaminhando o abuso, e burlar listas negras, uma vez que o *proxy* estaria entregando as mensagens e não outro *host* que já poderia estar em uma lista negra. Os protocolos considerados neste trabalho para conexão *proxy* são HTTP e SOCKS. O *proxy* HTTP funciona como *cache*, na requisição de documentos na Web pelo cliente. Quando esse *proxy* é mal-configurado, ele aceita intermediar conexões iniciadas com o comando CONNECT, e então podem ser usados para conectar a um servidor de correio. O protocolo SOCKS foi desenvolvido exatamente para viabilizar serviços de *proxies* de conexão, especialmente para máquinas que estejam atrás de um *firewall*. Ambos os protocolos são abusados por *spammers*, que os usam para entregar mensagens indesejadas, sem revelar sua identidade.

Open relays são servidores de correio eletrônico utilizando o protocolo SMTP que, seja por decisão, má configuração, ou até mesmo infecção por algum software malicioso, permitem que qualquer cliente se conecte a ele e o utilize para enviar *e-mails* a servidores de terceiros. Como o remetente e o destinatário utilizando o *open relay* não necessitam ser usuários cadastrados no servidor em questão, *spammers* abusam dessa máquina para enviarem mensagens indesejadas pela rede. Assim como o *open proxy*, os *spammers* utilizam o *open relay* para anonimizar sua origem. Para tal, eles manipulam o cabeçalho SMTP da mensagem, dificultando que o destinatário consiga

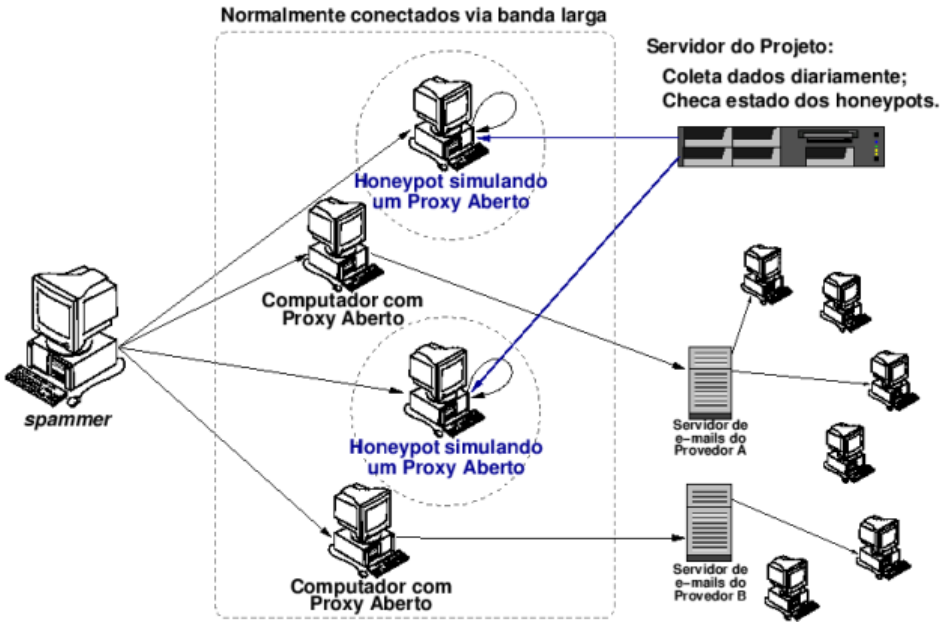


Figura 3.1. Arquitetura dos *honeypots* abusados pelos *spammers*.

descobrir a real origem do *spam*.

A figura 3.1, criada por Steding-jessen et al. [2007], representa a arquitetura utilizada para captura e coleta dos dados. Através dela é possível verificar como os *honeypots* emulando *open proxy* e *open relay* funcionam. Os *spammers*, a procura de anonimato, buscam por máquinas que apresentam alguma dessas características, para poder abusá-las. Ao se conectar a um dos *honeypots*, estes se comportam como se aceitassem os seus comandos. Entretanto, ao invés da mensagem ser repassada ao seu destino final, ela é armazenada e posteriormente transferida para um servidor de coleta. A captura das mensagens foi feita com o sistema *spamsinkd*, desenvolvido por Klaus Steding-Jessen, técnico do Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (CERT.br) [CERT.br, 2013]. Esse sistema é um *daemon* de rede que atua capturando as mensagens de *spam* e armazenando-as em *mailboxes*.

Para cada conexão recebida pelo *spamsinkd*, são registradas diversas informações relacionadas, como o endereço IP de origem, endereço IP de destino (nos casos de *proxy* SOCKS e HTTP) , porta TCP abusada nos *honeypots*, o protocolo utilizado (SMTP, HTTP ou SOCKS) e a data e a hora da conexão. Para fins de padronização, utilizou-se GMT como fuso-horário. Outras informações coletadas incluem também o provável sistema operacional associado a cada origem. Para cada endereço IP de origem e destino, foi registrado o Sistema Autônomo (AS), o prefixo de rede e o *Country Code* (CC) associados. Finalmente, o sistema armazena o cabeçalho e o conteúdo completo da mensagem que teria sido enviada.

A coleta do *spam* armazenado pelos *honeypots* foi feita, em períodos regulares, por um servidor central. O mecanismo de coleta, desenvolvido por Marcelo Chaves, técnico do CERT.br, foi implementado utilizando-se cópia remota e o programa de sincronização *rsync*, através de um túnel SSH criptografado. As mensagens coletadas por esse servidor estão armazenadas em *mailboxes*. Cada *mailbox* refere-se às mensagens enviadas de um endereço IP de origem para um determinado *honeypot*, para cada dia. A organização desses *mailboxes* foi feita através de diretórios. Inicialmente, cada diretório refere-se à uma parte do IP. Percorrendo-se esses diretórios, localizam-se então os diretórios relativos aos *honeypots*, e em seguida, os diretórios referentes a cada dia. Por fim, encontram-se os *mailboxes* que contém as mensagens. Além do armazenamento da mensagem, cada *spam* possui também um registro em um arquivo de *log*. Esses arquivos são organizados por *honeypot* e por dia. As informações utilizadas nesta dissertação foram recuperadas através dos arquivos de *log*. Quando necessário, pudemos voltar aos *mailboxes* para obter mais detalhes das mensagens.

Nenhuma das mensagens de *spam* coletadas pelos *honeypots* foi efetivamente entregue aos destinatários. Para evitar que os *spammers* percebessem que essas máquinas eram *honeypots*, foi preciso fazê-los acreditar que as mensagens foram entregues com sucesso. Para tal, o *honeypot* simulava uma comunicação bem-sucedida com o alvo indicado pelo *spammer*, porém as mensagens nunca eram enviadas ao destinatário. A única exceção trata *spammers* que enviam mensagens de teste como forma de validar a máquina sendo abusada. Ao longo da implementação dos *honeypots* para coleta dos *spams*, foram identificados alguns padrões que são claramente utilizados pelos *spammers* como forma de testar a máquina sendo abusada. O padrão identificado consiste em incluir o endereço da máquina alvo do abuso no assunto ou corpo da mensagem. Essas mensagens são destinadas para um ou mais endereços anônimos de destino usados pelo *spammer*. Quando esse padrão é identificado pelo *honeypot*, a mensagem é repassada ao seu destinatário real. Além disso, a mensagem é armazenada com os demais *spams* e um registro da mesma é criado em um arquivo de *log* específico, contendo informações como o endereço IP de origem, a porta e o protocolo abusado. Esse processo é realizado pelo sistema *spamtestd*, também desenvolvido por Klaus Steding-Jessen, técnico do CERT.br. O objetivo dessas mensagens é a validação da máquina sendo abusada, de forma a confirmar que o abuso está sendo realizado com sucesso.

Nesta dissertação foram utilizados dados provenientes de oito *honeypots*, localizados em diferentes posições ao redor do globo. Dentre os oito *honeypots* utilizados, dois se encontram em redes brasileiras (BR-01 e BR-02), enquanto os restantes se localizam em diferentes *country codes*: AU-01 (Austrália), AT-01 (Áustria), EC-01 (Equador), NL-01 (Holanda), TW-01 (Taiwan) e UY-01 (Uruguai). Um outro *honeypot*, CL-01

(Chile), que era inicialmente considerado, não pôde ser utilizado em nossa análise devido a problemas em sua conexão e limitações em sua banda. Os *honeypots* possuem *hardwares* distintos e localizam-se em redes de qualidades diferentes. Como mencionado anteriormente, a distribuição dos *honeypots* por diferentes pontos da rede mundial teve por objetivo permitir uma visão ampla do tráfego de *spam* ao redor da Internet para, com isso, reduzir distorções comuns em coletas restritas a uma única localidade.

Consideramos o período de 176 dias, entre 09/05/2012 e 31/10/2012. Esse período foi escolhido por ser o maior intervalo recente em que a coleta de todos os *honeypots* ocorreu com poucas interrupções. Antes e depois daquele período, falhas de energia ou reconfigurações de rede causaram a indisponibilidade de alguns dos *honeypots* por períodos maiores. Cabe ressaltar que esses acontecimentos fazem parte do mundo real e todos os *honeypots* estão sujeitos a eles. Ao todo, durante aquele período, foram coletados quase 1 bilhão e 800 milhões de mensagens, enviadas a partir de mais de 180 mil endereços IP distintos. Mais detalhes sobre o tráfego coletado são apresentados no capítulo 4.

3.2 Processamento dos dados

Inicialmente, os dados coletados passaram por uma etapa de pré-processamento. Essa etapa objetiva facilitar e tornar mais eficiente a análise dos dados. Em seguida, realizamos a análise temporal dos dados. Por fim, fizemos a análise dos dados através de campanhas de *spam*.

3.2.1 Pré-processamento

No momento de captura das mensagens, elas são armazenadas em *mailboxes*, além de serem gerados registros em arquivos de *log* para cada dia e para cada *honeypot*. Como forma de facilitar e tornar a recuperação mais eficiente, os arquivos de *log* foram pré-processados e as informações das mensagens foram armazenadas em uma base de dados.

Durante o pré-processamento, ao percorrer os os arquivos de *log* do período, as seguintes informações foram identificadas para cada uma das mensagens:

- endereço IP de origem;
- prefixo de rede do endereço;
- sistema autônomo (AS);

- *country code* (CC) associado ao IP de origem;
- protocolo utilizado na conexão;
- porta utilizada para conexão;
- sistema operacional (SO) utilizado para o envio;
- código(s) retornado(s) pela lista de bloqueio *Spamhaus Zen*;
- tamanho da mensagem;
- data/hora do recebimento da mensagem.

Como resultado do pré-processamento, todas essas informações são armazenadas em tabelas em um banco de dados *MySQL*. Cada tabela contém as informações relacionadas a um *honeypot*. Isso torna simples o processamento de consultas sobre comportamentos de cada *honeypot*, bem como outras características como, por exemplo, por endereço IP.

Além do pré-processamento das mensagens coletadas, realizamos também o pré-processamento das mensagens de teste. Assim como as demais mensagens, essas também possuem arquivos de *log* organizados para cada dia e para cada um dos *honeypots*. Em seu processamento, foram coletadas as seguintes informações:

- endereço IP de origem;
- protocolo utilizado na conexão;
- porta utilizada para conexão.

Esses dados foram armazenados em tabelas distintas, por *honeypot*, na mesma base de dados *MySQL* citada anteriormente.

Cabe ressaltar que, se necessário, podemos retornar aos dados pré-processados e recuperar as informações originais caso seja relevante contabilizar outros elementos que não foram preservados pelo pré-processamento ou outros detalhes sobre um evento de interesse que tenha sido identificado.

3.2.2 Análise das séries temporais

A partir do pré-processamento foram derivadas métricas informativas, capazes de auxiliar na análise e caracterização do tráfego coletado. Através dessas métricas, geramos dados agregados por protocolo, possibilitando analisar as diferenças apresentadas pelo

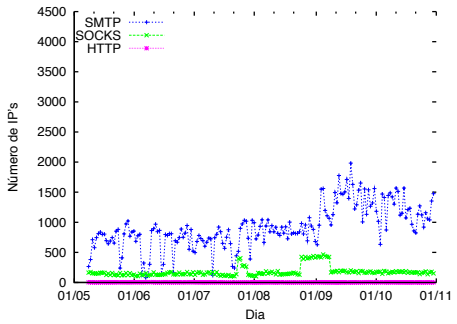
spam enviado pelos métodos distintos. Além disso, foi possível analisar temporalmente as características de cada protocolo, bem como de cada um dos *honeypots*, o que evidenciou as diferenças e semelhanças associadas à localização da máquina abusada.

Essa análise se deu em quatro etapas distintas. A primeira foi feita através da construção de gráficos representado o comportamento do tráfego de *spam* ao longo dos 176 dias, tanto sob o ponto de vista geral, quanto individual, para cada *honeypot*. Para tal, foram analisadas métricas como as mensagens por dia, número de endereços IP de origem distintos, quantidade de prefixos de rede e sistemas autônomos identificados, além do volume de dados propagado em cada dia. Todas essas métricas foram geradas por protocolo. A figura 3.2(a) mostra um exemplo de série temporal. Essa análise permite identificar variações no tráfego ao longo do tempo, e ainda comparar o comportamento de cada protocolo, além das variações apresentadas devido à localização do coletor. Em seguida, verificamos a correlação entre as séries temporais dos atributos de cada *honeypot*, além da correlação das características entre os *honeypots*. Esses resultados nos permitiram identificar que, como esperado, o volume do tráfego relaciona-se fortemente com o número de mensagens. Observamos também que há uma forte relação entre as características das mensagens de teste para os *honeypots*, com valores próximos de 1,0 na maioria dos casos. As outras correlações encontradas não apontaram resultados relevantes.

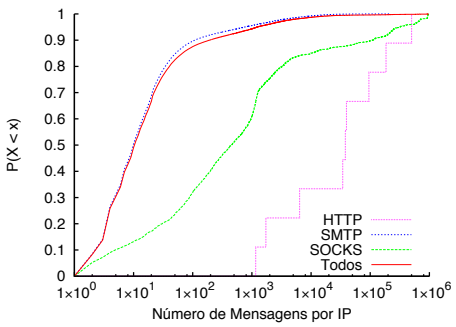
Na segunda etapa, geramos gráficos contendo funções de distribuição acumulada (CDFs) tanto para o número de mensagens por endereço IP, como também para o tamanho das mensagens. Esses gráficos foram gerados para todos os *honeypots*. Um exemplo é mostrado na figura 3.2(b). A partir dessas CDFs é possível entender a distribuição do número de mensagens entre os transmissores, bem como do tamanho da mensagem e, com isso, encontrar padrões distintos nos diferentes protocolos.

A terceira etapa consistiu na geração de gráficos representando a atividade dos prefixos de rede e dos sistemas autônomos no período analisado, como apresentado na figura **??**. Esse processo consistiu em encontrar todos os prefixos de rede e ASes em cada *honeypot* e, então, verificar os dias em que cada um deles esteve ativo. Através desses dados, geramos gráficos em que o eixo y representa cada prefixo de rede ou AS e o eixo x representa cada dia avaliado. Quando um prefixo/AS y se encontra ativo em determinado dia x, um ponto é inserido em (x,y). Com isso, podemos identificar o perfil de atividade dos prefixos/ASes em determinados momentos ajudando a entender melhor a geração do tráfego de *spam*.

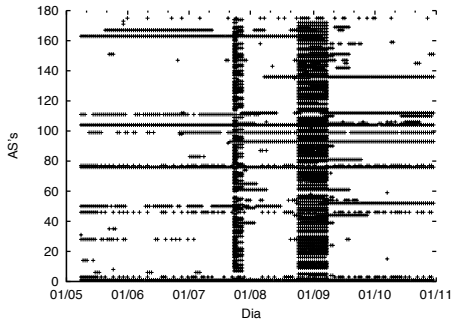
Por fim, realizamos a análise considerando a hora do dia em que os *spams* foram recebidos. Nessa etapa, consideramos as mensagens e os endereços IP responsáveis pelas mesmas para cada protocolo, a cada hora do dia e geramos gráficos do volume



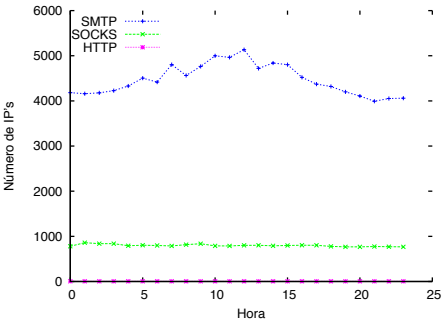
(a) Série temporal dos Endereços IP do *honeypot* UY-01



(b) CDF do número de mensagens por IP do *honeypot* UY-01



(c) Distribuição dos ASes utilizando SOCKS no *honeypot* UY-01



(d) Número de endereços IP por hora do dia no *honeypot* UY-01

Figura 3.2. Exemplo dos gráficos utilizados na análise dos resultados

por hora para os diferentes *honeypots* (figura 3.2(d)). Essa análise visou identificar padrões circadianos no tráfego.

3.2.3 Campanhas de *spam*

Além da análise citada anteriormente, realizamos também a análise dos dados baseado em campanhas de *spam*. Uma campanha de *spam* é um conjunto de mensagens que possuem um objetivo comum e uma mesma estratégia de disseminação [Guerra et al., 2008a].

Durante a transmissão dessas mensagens, *spammers*, na tentativa de dificultar a sua identificação, introduzem técnicas de ofuscação como, por exemplo, a inserção de trechos de texto aleatórios. Com isso, a identificação através de métodos convencionais das mensagens que têm mesma origem se torna um desafio. O objetivo das campanhas é minimizar os efeitos dessas técnicas de ofuscação e detectar as mensagens de uma mesma origem.

A idéia central da técnica para identificação das campanhas é a de que um *spam-*

mer, ao disseminar uma campanha, mantém algumas partes da mensagem estáticas, enquanto outras seções do conteúdo são alteradas de forma sistemática e automatizada. Por exemplo, muitas vezes o *spammer* altera apenas o assunto da mensagem, porém todas as outras características se mantêm as mesmas.

Neste trabalho utilizamos o algoritmo de agrupamento *FPCluster*, desenvolvido por Pires et al. [2012] para, explorando essas propriedades apresentadas pelos *spams*, agrupar as mensagens em campanhas, com base em diversos atributos e identificar as estratégias de ofuscação utilizadas. Esse algoritmo constrói uma árvore de padrões frequentes (*FP-Tree*) para organizar as mensagens e identificar atributos invariantes das mesmas. Nessa árvore, os atributos de uma mesma mensagem são inseridos em um mesmo caminho e os atributos mais frequentes são inseridos nos níveis mais altos da árvore. Dessa forma, mensagens que compartilhem um prefixo comum (ou seja, uma seqüência ordenada de atributos) compartilham um caminho na árvore, o que permite que menos espaço seja gasto para representar a mesma informação. A ofuscação pode ser observada na árvore no ponto onde há um repentino aumento no número de filhos de um nó.

Para extração das campanhas, executa-se um algoritmo de corte baseado no tamanho das sub-árvores. Define-se uma frequência f e corta-se as sub-árvores a partir do nó que possui frequência maior ou igual. Com isso, cada campanha será uma dessas sub-árvores encontradas. O objetivo desse corte é maximizar os atributos comuns entre as mensagens, diferenciando as campanhas somente por aqueles que realizam a ofuscação das mensagens e, ao mesmo tempo, não permitindo que haja segmentação de mensagens de uma mesma campanha [Guerra et al., 2008a; Totti et al., 2012; Pires et al., 2012].

O processamento das campanhas desenvolvido por Pires et al. [2012] é realizado diariamente, a partir da coleta e armazenamento das mensagens capturadas, usualmente realizado em ciclos de 24 horas. Devido à periodicidade da coleta, as campanhas se limitam a um dia. Como a caracterização aqui realizada foi feita considerando-se um período longo (176 dias), definimos um processo de fusão das campanhas para que essas pudessem ser identificadas quando se estendessem por vários dias. Além disso, fundimos as campanhas dos vários *honeypots*, para conseguirmos identificar uma mesma campanha em *honeypots* distintos e entender a influência da localização na recepção das diferentes campanhas.

O processo de fusão das campanhas é realizado comparando o conjunto de atributos de cada uma. Esse conjunto de atributos é o caminho da raiz da *FPTree* ao ponto de ramificação (corte). Caso duas campanhas possuam o mesmo conjunto de atributos, independente da ordem, elas serão consideradas como sendo uma mesma campanha. A

ordem não é considerada por ser dependente da frequência de cada atributo, que pode variar durante o período, a cada dia.

A etapa inicial desse processo consistiu em listar a assinatura (atributos presentes na campanha) de todas as campanhas presentes em todos os dias e em todos os *honeypots*. Nessa lista, ordenamos todas as assinaturas e unificamos todas aquelas que se mostraram iguais. A partir dessa nova lista, geramos um mapa contendo todas as campanhas e um identificador para cada uma delas. Esse método é conservador, por não unir campanhas que difiram por qualquer atributo, ou que tenham assinaturas contidas em outras maiores (o que poderia ocorrer, se certas variações não fossem observadas em certos dias).

Percorremos então todas as campanhas em todos os dias e em todos os *honeypots* e fornecemos um identificador baseado no mapa gerado para cada uma delas. Dessa forma, conseguimos identificar campanhas que se estendem por vários dias e que aparecem em vários *honeypots*, dado que elas possuem o mesmo identificador.

Para a geração das campanhas utilizamos os dados do dia 09/05/2012 ao dia 31/08/2012. Esse período foi escolhido por ser o maior período disponível para processamento de todos os *honeypots*. Sem aplicar a técnica de fusão, foram encontradas 573.945 campanhas, valor que foi reduzido para 195.722 campanhas distintas, ao se realizar o agrupamento descrito.

Capítulo 4

Resultados

Os resultados encontrados na análise do tráfego de *spam* são apresentados e discutidos nesse capítulo. Inicialmente apresentamos uma visão geral dos dados, seguida pela análise de elementos particulares, além da comparação dos diferentes *honeypots*. Observações interessantes como, por exemplo, a análise do comportamento das mensagens de teste e a discussão sobre o redescobrimento de um *honeypot* são mostradas posteriormente. Por fim, apresenta-se a análise baseada em campanhas e um método visando agrupar os endereços IP similares através de suas campanhas em comum.

4.1 Visão Geral

A fim de entender melhor o tráfego de *spam* como um todo, em um primeiro momento analisamos os dados coletados pelos *honeypots* de forma agregada. Com isso, é possível visualizar de maneira geral o que acontece com relação aos *spams* que trafegam pela rede em todo o mundo.

A tabela 4.1 oferece uma visão geral dos dados coletados pelos oito *honeypots*. No período de 176 dias, quase 1 bilhão e 800 milhões de mensagens foram coletadas, oriundas de endereços associados a 141 *country codes* distintos, cerca de 59% dos *country codes* definidos. Do total de mensagens, 67,78% foram enviadas utilizando o protocolo SOCKS, 16,90% usando HTTP e apenas 15,32% fazendo uso de SMTP. É interessante notar que, das aproximadamente 303 milhões de mensagens utilizando o protocolo HTTP, mais de 239 milhões seriam enviadas pelo *honeypot* TW-01 que, conforme será discutido posteriormente, possui características bastante distintas dos demais.

Apesar das mensagens observadas terem sido originandas de 141 *country codes* distintos, alguns poucos CCs foram responsáveis pela enorme maioria de todo o *spam*

Tabela 4.1. Visão Geral dos dados coletados em todos os *honeypots*.

	SMTP(%)	**SOCKS(%)**	**HTTP(%)**	**Total**
Mensagens (milhões)	274,52 (15,32%)	1214,74 (67,78%)	303,02 (16,90%)	1.792,29
Endereços IP	161.398 (89,53%)	18.801 (10,43%)	3.692 (2,05%)	180.262
Prefixos de rede	13.574 (93,22%)	1.343 (9,22%)	168 (1,25%)	14.561
Sistemas Autônomos (AS)	2.355 (95,73%)	330 (13,41%)	45 (1,83%)	2.460
***Country Codes* (CC)**	138 (97,87%)	65 (46,09%)	13 (9,21%)	141
Volume de tráfego (TB)	0,98 (13,97%)	4,35 (61,59%)	1,72 (24,44%)	7,06
Msgs/endereço (milhares/IP)	1,70	64,61	82,07	9,94
Volume/endereço (MB/IP)	6,40	242,61	490,16	41,08
Volume/msg (KB/msg)	3,85	3,84	6,11	4,23

recebido. A tabela A.2 mostra aqueles que enviaram mais mensagens indesejadas ao longo do período. Como é possível verificar, os 10 *country codes* que mais enviaram mensagens são responsáveis por mais de 92% do total de *spam* recebido. Mais de 60% das mensagens recebidas foram oriundas de endereços associados ao CC US. Considerando o *country code* PH, podemos notar características diferentes do US. Esse CC apresenta um número menor de endereços IP provenientes de apenas 9 ASes distintos e, ainda assim, é o segundo a mais enviar *spams*. Assim como PH, os *country codes* JP, KR e IT também apresentam poucos endereços IP. Já os *country codes* CN e TW são o terceiro e quarto, respectivamente, com maior número de mensagens e apresentam uma quantidade muito elevada de transmissores mas um número relativamente baixo de ASes. Os demais apresentam características próximas a US com relação a endereços IP e prefixos de rede. Esses resultados sugerem que o envio de *spams* está concentrado em alguns poucos *country codes*.

Tabela 4.2. 10 Country Codes que mais enviaram mensagens em todos os *honeypots*.

	Mensagens	**Endereços IP**	**Prefixos's**	**ASes**
US	1.087.783.265 (60,69%)	2.365 (1,31%)	1.185 (8,14%)	388 (15,77%)
PH	230.230.162 (12,85%)	171 (0,09%)	45 (0,31%)	9 (0,37%)
CN	91.543.044 (5,11%)	76.197 (42,27%)	5.456 (37,47%)	76 (3,09%)
TW	83.561.691 (4,66%)	64.842 (35,97%)	239 (1,64%)	29 (1,18%)
BR	66.234.708 (3,70%)	4.203 (2,33%)	1.155 (7,93%)	151 (6,14%)
JP	39.139.575 (2,18%)	217 (0,12%)	82 (0,56%)	33 (1,34%)
RU	20.622.747 (1,15%)	4.196 (2,32%)	1.004 (6,89%)	401 (16,30%)
KR	13.195.943 (0,74%)	346 (0,19%)	210 (1,44%)	50 (2,03%)
IN	11.947.247 (0,67%)	17.667 (9,80%)	1.562 (10,73%)	77 (3,13%)
IT	8.115.800 (0,45%)	220 (0,12%)	109 (0,75%)	26 (1,06%)

Analisando-se os dados coletados, apresentados na figura A.3, observamos 2.460

Sistemas Autônomos (ASes) distintos. Entretanto, os 10 ASes que enviaram o maior número de *spam* recebidos pelos *honeypots* são responsáveis por mais de 80% do total das mensagens. Apenas o AS que mais enviou mensagens, proveniente de US, responde por cerca de 50% do total de *spams* recebidos. Nesse Sistema Autônomo foram observados apenas 159 endereços IP, que disseminaram quase 900 milhões de mensagens nos 176 dias analisados. Esse AS fornece serviços de *hosting* e *cloud computing*, sendo seus endereços IP estáticos, o que evidencia que esses atacantes utilizam máquinas especializadas para enviar as mensagens de *spam*. Esse dado confirma que um grupo pequeno de transmissores é responsável pela maior parte do *spam* encontrado na rede conforme mostrado por John et al. [2009]. Com relação aos outros ASes, podemos notar que a grande maioria possui menos de 100 endereços IP distintos. Porém, dois deles são bastante diferentes dos demais, com mais de 50 mil transmissores cada. Esses ASes fornecem serviço de DSL, sendo seus endereços IP dinâmicos, o que leva a crer que tais atacantes são *spambots*. Além disso, eles são provenientes de TW e CN, que são *country codes* que também possuem quantidade elevada de endereços.

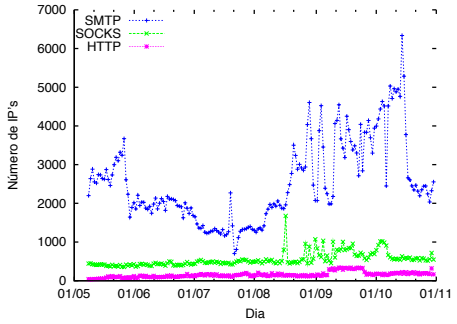
Tabela 4.3. 10 Sistemas Autônomos que mais enviaram mensagens em todos os *honeypots*.

	Mensagens	**Endereços IP**	**Prefixos**	**Volume (GB)**	**CC**	**Tipo**
10297	896.990.985 (50,05%)	159 (0,09%)	4 (0,03%)	2.714,84 (37,54%)	US	Host/Cloud
9299	147.271.014 (8,22%)	74 (0,04%)	19 (0,13%)	340,70 (4,71%)	PH	Misto
29802	143.452.043 (8,00%)	23 (0,01%)	2 (0,01%)	397,46 (5,50%)	US	Host
6648	78.075.302 (4,36%)	50 (0,03%)	10 (0,06%)	178,45 (2,45%)	PH	DSL/Bus.
3462	70.511.218 (3,93%)	61.598 (34,17%)	108 (0,74%)	275,70 (3,81%)	TW	DSL
4134	56.936.786 (3,18%)	53.972 (29,94%)	4.256 (29,23%)	1.858,41 (25,70%)	CN	DSL
2497	15.674.950 (0,87%)	26 (0,01%)	8 (0,05%)	73,80 (1,02%)	JP	Cloud
4713	12.916.038 (0,72%)	25 (0,01%)	12 (0,08%)	28,98 (0,40%)	JP	Cloud
27699	11.238.638 (0,63%)	1.191 (0,66%)	49 (0,33%)	39,60 (0,55%)	BR	DSL/Bus.
23650	10.707.705 (0,60%)	38 (0,02%)	27 (0,19%)	113,93 (1,58%)	CN	?

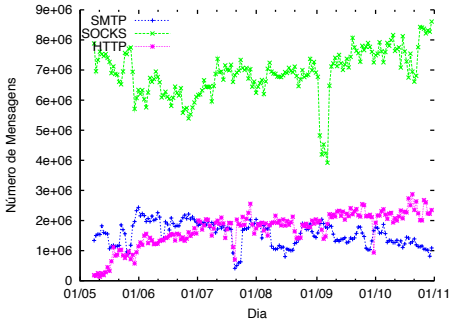
4.2 Comportamentos por protocolo

Observando-se a tabela 4.1, é possível notar alguns elementos interessantes do comportamento dos *spammers*. As máquinas que fazem uso do protocolo SMTP para envio de *spam* representam alto percentual do total de endereços IP (quase 90%), porém retratam um fração baixa do número de mensagens (apenas 15,32%), sendo o protocolo com menor número de *spams*. Por outro lado, dos 180.262 endereços IP que enviaram *spam*, apenas 10,42% foram responsáveis por mensagens utilizando o *proxy* SOCKS, entretanto esses representam cerca de 68% do total de mensagens. Já os transmissores utilizando *proxy* HTTP constituem apenas 2% do total e enviaram quase 17%

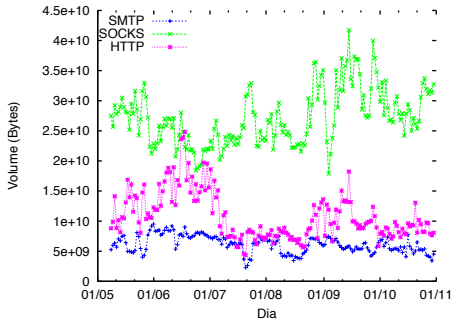
do total de mensagens coletadas. Através do número de mensagens por endereço IP, podemos notar que os transmissores utilizando esses protocolos enviaram alto número de mensagens durante o período de 176 dias analisados, sendo uma média de 64 mil mensagens/endereço IP para o protocolo SOCKS e 84 mil mensagens/endereço IP para HTTP, o que representa 43 vezes mais mensagens que aqueles que usaram SMTP. Com isso, é possível concluir que os protocolos SOCKS e HTTP são usados para envio "por atacado" de mensagens, ou seja, enviam um grande volume de mensagens, levando a crer que utilizam infraestrutura específica para disseminação de *spam*. Já o protocolo SMTP é usado por máquinas que enviam um número menor de mensagens, indicando que esse protocolo é, possivelmente, utilizado por *bots*.



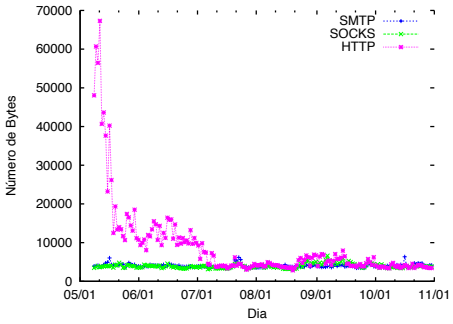
(a) Endereços IP ao longo do tempo



(b) Mensagens ao longo do tempo



(c) Volume (bytes) ao longo do tempo



(d) Tamanho das mensagens ao longo do tempo

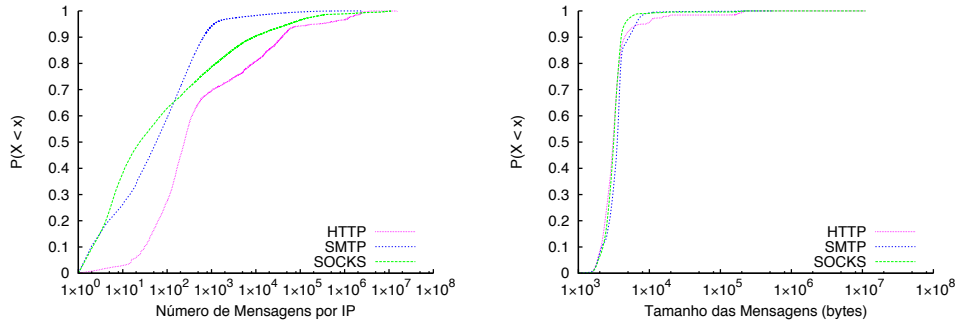
Figura 4.1. Dados agregados por protocolo ao longo do tempo.

Uma análise temporal dos dados agregados por protocolo indica comportamentos relativamente estáveis em termos de mensagens enviadas por cada protocolo, conforme pode ser visto na figura 4.1(b). O volume total enviado usando-se cada protocolo, mostrado na figura 4.1(c), também segue um comportamento bastante semelhante, apesar de uma oscilação maior no período de maio a junho. Já a figura 4.1(d) mostra que o tamanho médio das mensagens enviadas por SOCKS e SMTP permaneceu relativamente estável durante todos os dias, enquanto as mensagens enviadas por HTTP variaram ao

longo do tempo, com algumas de suas mensagens bem maiores que as demais, principalmente no início do período, causando a oscilação observada na figura 4.1(c). Conforme mostrado na tabela 4.1, a média de tamanho das mensagens utilizando HTTP é quase duas vezes maior que dos demais protocolos, basicamente devido ao período de maio e junho. Inspecionamos essas mensagens maiores e observamos que elas foram enviadas por apenas 5 endereços distintos e todos provenientes do *country code* CN, e foram recebidas por todos os *honeypots*, com exceção do BR-01 e TW-01. Naquele período, BR-01 havia trocado de endereço recentemente e ainda não estava sendo abusado pelos *spammers* em geral, recebendo mensagens por esse protocolo de apenas 11 endereços distintos. Já TW-01, como discutido posteriormente, tem um comportamento diferenciado com relação a tráfego oriundo de CN. Logo, o aparecimento dessas pode estar relacionado a algumas campanhas específicas originadas em CN.

Com relação ao comportamento do número de endereços IP ao longo do tempo, podemos perceber que os protocolos SOCKS e HTTP se mantêm estáveis, com apenas alguns picos em alguns determinados dias para o protocolo SOCKS (fato ocasionado pelo surgimento de novas campanhas, discutido posteriormente). Além disso, esses protocolos apresentam número relativamente baixo de endereços IP a cada dia, sendo em média 155 e 549 para HTTP e SOCKS, respectivamente. Considerando a série temporal do número de transmissores que utilizam SMTP, percebemos um comportamento mais variável. Há um aumento do número de endereços IP no fim de maio, seguido por um período de queda, nos meses de junho e julho, e um período instável no mês de setembro. Já no final do período analisado, foi possível perceber uma forte queda na quantidade desses transmissores.

A figura 4.2 apresenta dados sobre a distribuição de mensagens por endereços de origem e por tamanho para cada protocolo, ressaltando as diferenças entre eles. A figura 4.2(a) mostra que os endereços que utilizam protocolos diferentes possuem comportamentos distintos com relação ao número de mensagens enviadas. Enquanto aproximadamente 60% dos endereços IP utilizando SMTP e SOCKS enviaram menos de 100 mensagens ao longo dos 176 dias, apenas 25% dos atacantes utilizando HTTP enviam menos que isso. Entretanto, mesmo os protocolos SMTP e SOCKS apresentando características próximas considerando endereços IP que enviaram poucas mensagens, eles se mostram diferentes quando observamos os transmissores que enviam mais mensagens. Ao passo que mais de 20% dos atacantes enviando *spam* através de SOCKS enviaram mil mensagens ou mais, apenas 5% daqueles que utilizam SMTP disseminaram essa quantidade. Além disso, pouco mais de 1% dos transmissores SMTP enviou mais que dez mil mensagens, enquanto cerca de 10% dos SOCKS e 20% dos HTTP ultrapassaram esse valor.



(a) Número de mensagens por endereço de origem

(b) Tamanho das mensagens

Figura 4.2. Distribuições acumuladas para mensagens por protocolo.

Com relação ao tamanho das mensagens enviadas usando cada protocolo (fig. 4.2(b)), o perfil de 80% das mensagens enviadas por cada protocolo são bastante semelhantes, sendo as mensagens enviadas por SMTP apenas ligeiramente maiores. Entretanto, apenas 3% das mensagens enviadas usando SOCKS tinham mais de 5 KB, enquanto eram 8% e 10% das mensagens enviadas por SMTP e HTTP, respectivamente. Já no caso de HTTP, pouco mais de 1,5% das mensagens enviadas com aquele protocolo eram maiores que 100 KB, enquanto apenas 0,4% e 0,2% das mensagens utilizando SOCKS e SMTP, respectivamente, foram maiores que esse valor.

Além disso, apenas uma fração pequena das máquinas utilizou mais de um protocolo para enviar *spam* durante o período (a soma dos números de endereços IP distintos que foram observados usando cada protocolo é só cerca de 2% superior ao total de endereços distintos). Esses fatores nos permitem afirmar que as máquinas que utilizam cada tipo de protocolo têm comportamentos diferentes, indicando origens (*spammers*) diferentes:

- máquinas que usam *proxies* abertos (HTTP e SOCKS) enviam mensagens em grande volume, o que sugere o uso de uma infraestrutura especializada para o envio de *spam*;

- máquinas que enviam *spam* usando SMTP tendem a enviar bem menos mensagens, mas seu conjunto ainda é responsável por uma parcela significativa do tráfego, o que sugere a formação de *botnets*.

4.3 Análise dos diversos locais de coleta

A visão geral dos dados fornece informações interessantes sobre comportamentos que se mantêm entre os *honeypots*. Entretanto, analisar cada um dos *honeypots* individualmente nos permite identificar características não possíveis de serem observadas visualizando os dados como um todo. Dado que os *honeypots* estão em localizações distintas, as características do material coletado em cada ponto pode apresentar variações.

Os gráficos apresentados nessa seção tiveram sua escala normalizada, uma vez que avaliamos ser melhor para a comparação entre os dados dos *honeypots*. Por conta disso, alguns gráficos tiveram suas curvas próximas do eixo x. Além disso, algumas figuras serão repetidas para facilitar a leitura do texto.

4.3.1 O *honeypot* AT-01

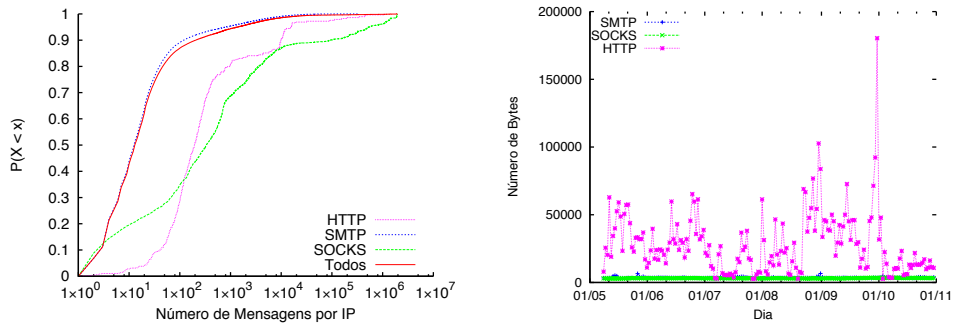
Tabela 4.4. Visão geral dos dados coletados no *honeypot* AT-01.

	SMTP	**SOCKS**	**HTTP**	**Total**
Mensagens (milhões)	21,64 (10,46%)	181,87 (87,90%)	3,40 (1,64%)	206,91
Endereços IP	55.329 (96,22%)	2.196 (3,82%)	289 (0,49%)	57.499
Prefixos de rede	7.659 (95,59%)	473 (5,90%)	34 (0,42%)	8.012
Sistemas Autônomos (ASes)	1.849 (96,40%)	184 (9,59%)	11 (0,57%)	1.918
***Country Codes* (CC)**	128 (98,46%)	49 (37,69%)	6 (4,62%)	130
Volume de bytes (GB)	69,57 (10,17%)	532,31 (77,83%)	82,09 (12,00%)	683,98

Como é possível notar na tabela 4.4, o protocolo SOCKS é responsável pelo maior número de mensagens e por grande parte do volume de tráfego nesse *honeypot*. Esse protocolo responde por quase 88% do total de mensagens e 78% do volume total. Em contrapartida, apenas 3,8% dos endereços IP utilizaram-no.

A figura 4.3(a) mostra que os endereços IP utilizando SOCKS enviam muitas mensagens. Diferente dos demais *honeypots*, em que o transmissores que usam HTTP tendem a enviar mais mensagens, aqui aqueles que usam SOCKS enviam um número maior. Por exemplo, mais de 30% dos atacantes usando *proxy* SOCKS enviaram mais de mil mensagens, enquanto apenas cerca de 20% daqueles que usam *proxy* HTTP enviam mais que esse valor. Já o protocolo SMTP apresenta número elevado de transmissores, mas um percentual relativamente baixo do total de mensagens (10,46%), mostrando que seus endereços IP enviam poucas mensagens, fato comprovado na figura 4.3(a).

Considerando o volume dos dados, na tabela 4.4 mostramos que o protocolo SOCKS é responsável pelo maior percentual (78%), enquanto os demais protocolos possuem valores próximos (10,17% e 12,00% para SMTP e HTTP, respectivamente).



(a) CDF das Mensagens por IP

(b) Tamanho das Mensagens ao longo do tempo

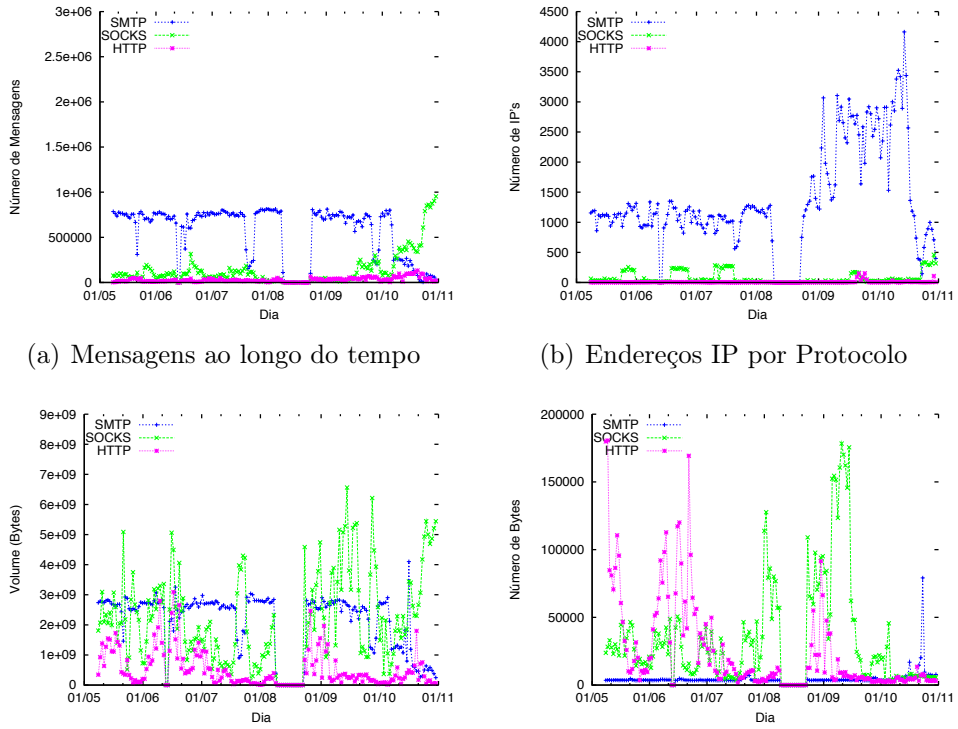
Figura 4.3. Características do *honeypot* AT-01.

Através da figura 4.3(b) podemos entender que, pelo fato do protocolo HTTP possuir mensagens com tamanho médio elevado, mesmo possuindo um número muito inferior de mensagens, o volume de seu tráfego chega a ser maior que do protocolo SMTP, que apresenta mensagens pequenas durante todo o período.

4.3.2 O *honeypot* AU-01

Nesse *honeypot*, o protocolo SMTP responde pela maior parte dos *spams* recebidos e dos endereços de IP responsáveis por esses, como pode ser visto nas figuras 4.4(a) e 4.4(b). Através da série temporal dessas duas características, podemos observar que durante a maior parte do período, o protocolo SMTP prevalecem com relação os demais. Porém, a partir do dia 07 de outubro, há uma queda brusca no tráfego SMTP e, em consequência, um aumento no número de mensagens utilizando o protocolo SOCKS. Com isso, o número de mensagens e endereços IP, assim como o volume do tráfego (figura 4.4(c)) passou a ser maior para o protocolo SOCKS.

Um comportamento apresentado por esse *honeypot* diferente dos demais está no volume do tráfego ocasionado pelo protocolo SOCKS. Aqui, mesmo apresentando um número relativamente pequeno de mensagens (19,00%), o volume de seu tráfego foi bastante considerável (45,55%), sendo maior que os dos demais protocolos. Esse fato ocorre porque as mensagens que utilizam SOCKS observados nesse *honeypot* apresentam tamanho médio elevado durante todo o período, como pode ser observado na figura 4.4(d). É importante dizer que, nos dias 13 e 14 de junho, além do período de 10 a 23 de agosto, não houve captura dos dados, ocasionando em vales nas curvas, como mostrado nos gráficos presentes na figura 4.4.



(a) Mensagens ao longo do tempo

(b) Endereços IP por Protocolo

(c) Volume (bytes) ao longo do tempo

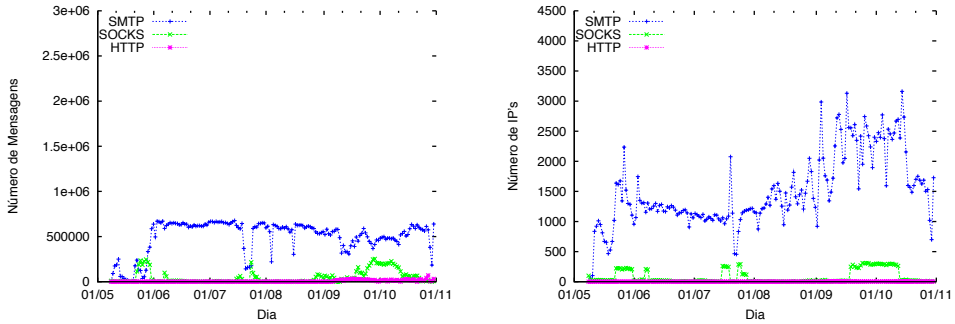
(d) Tamanho das Mensagens ao longo do tempo

Figura 4.4. Séries temporais por protocolo do *honeypot* AU-01.

4.3.3 O *honeypot* BR-01

O *honeypot* BR-01 apresenta o menor número de mensagens entre todos aqueles analisados. Isso pode ser explicado pelo fato de estar posicionado em uma rede de baixa qualidade, o que limita o tráfego incidente a ele e, além disso, o seu endereço IP foi modificado no primeiro dia da coleta, o que fez com que poucas mensagens fossem observadas nos primeiros dias. O protocolo responsável pela maior parte das suas mensagens é o SMTP, que responde por mais de 90% do total e compreende quase todos os endereços IP (98%). Como pode ser visto nas figuras 4.5(a) e 4.5(b), o número de mensagens enviadas utilizando o protocolo SMTP, bem como o número de endereços IP são maiores que os demais protocolos durante todo o período. Já a série temporal dos transmissores que enviam *spam* através de SOCKS é bem inferior à do protocolo SMTP, porém é constante, com alguns picos (ocasionado por campanhas, como será explicado posteriormente). Um desses picos, mais duradouro que os demais (25 dias, de 19/09 a 13/10) incide também no aumento do número de mensagens e do volume de tráfego desse protocolo. Como pode ser observado na figura 4.6(b), nesse período praticamente todos os prefixos desse protocolo que foram detectados na coleta esti-

veram ativos. Esse gráfico apresenta os prefixos ativos durante o intervalo analisado. Nesse gráfico, o eixo y apresenta os prefixos e o eixo x os dias analisados. Um ponto é adicionado em (x,y) quando um prefixo y estiver ativo no dia x. O surgimento de novas campanhas, possivelmente com a participação da grande maioria dos prefixos, é responsável por esse fato.

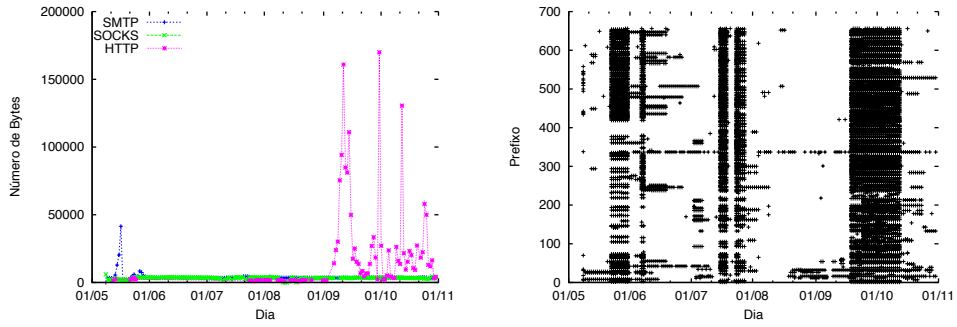


(a) Mensagens ao longo do tempo (b) Endereços IP ao longo do tempo

Figura 4.5. Séries temporais por protocolo do *honeypot* BR-01.

Com relação ao tráfego relacionado ao protocolo HTTP, apenas 11 endereços IP's distintos (0,01% do total), foram responsáveis por todas as suas mensagens, que representam 1,10% do total. Esses endereços são provenientes dos *country codes* CN, TW e US, sendo que 99,85 do total das mensagens enviadas por esses transmissores são vindas de CN. Porém, mesmo sendo um número pequeno de transmissores e um valor relativamente baixo de mensagens ao longo do período analisado, o volume de tráfego ocasionado por esse protocolo foi relevante, com cerca de 25 GB, equivalente a 7,5% do tráfego total. Esse fato é explicado pelo aparecimento de mensagens que usam HTTP e possuem tamanho elevado. A figura 4.6(a) mostra que, a partir do dia 02 de setembro, o tamanho médio dessas mensagens aumentou significativamente, ocasionando então o aumento do volume desse tráfego.

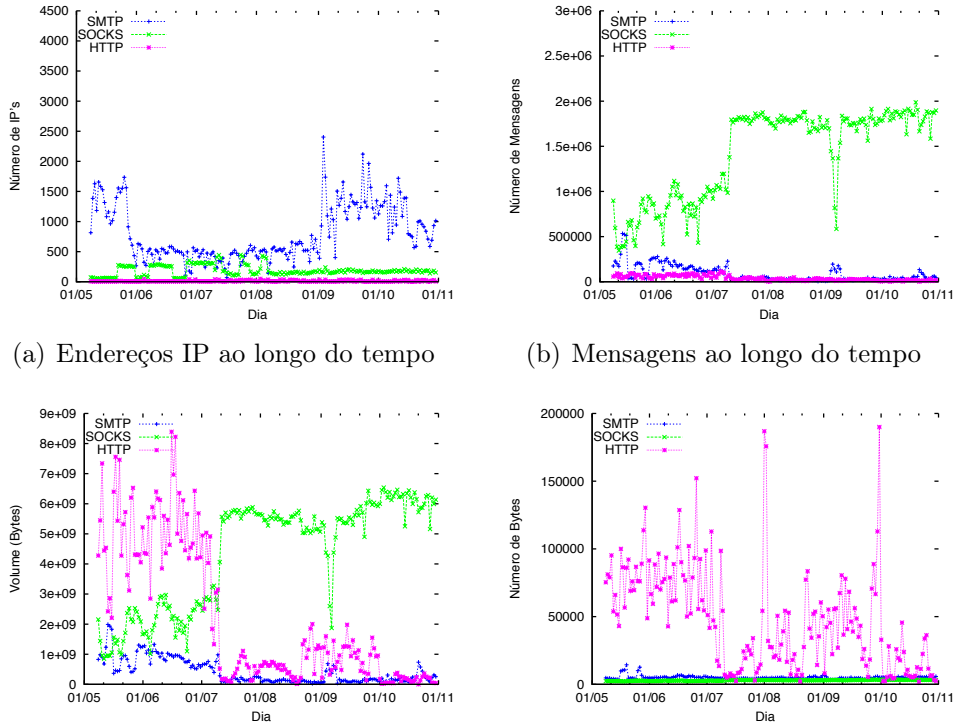
Considerando os Sistemas Autônomos (ASes) de origem dos endereços IP responsáveis pelos *spams* recebidos nesse *honeypot*, verificamos que dentre os dez que mais enviaram mensagens, seis deles se encontraram no CC BR. Além disso, verificando os *country codes* de origem que mais enviaram mensagens, pudemos observar que o principal foi também o BR, responsável por 15% das mensagens incidentes nessa máquina, o que indica uma localidade entre origens e máquinas atacadas nesse caso.



(a) Tamanho das Mensagens ao longo do tempo

(b) Prefixos de rede usando SOCKS

Figura 4.6. Tamanho das mensagens e distribuição dos prefixos ao longo do tempo do *honeypot* BR-01.



(a) Endereços IP ao longo do tempo

(b) Mensagens ao longo do tempo

(c) Volume (bytes) ao longo do tempo

(d) Tamanho das mensagens ao longo do tempo

Figura 4.7. Séries temporais por protocolo do *honeypot* BR-02.

4.3.4 O *honeypot* BR-02

Nesse *honeypot*, mais de 96% dos transmissores utilizaram o protocolo SMTP. Porém, esses são responsáveis por um número pequeno de mensagens (5,71% do total), além de um volume de tráfego baixo (6,12%). Em contrapartida, os grandes responsáveis pelo

tráfego de *spam* coletado nesse *honeypot* utilizam *proxy* SOCKS. Tendo sido gerado por apenas 3,53% do total de endereços IP's, o tráfego enviado através de SOCKS representa mais de 91% de todas as mensagens e mais de 710,5 GB (63% do volume). Como pode ser visto, os atacantes utilizando esse protocolo enviam um número elevado de mensagens, sendo que cerca de 50% enviaram mais de mil mensagens. Já considerando o protocolo SMTP, cujo transmissores tendem a enviar número baixo de mensagens, mais de 85% desses enviaram menos de 100 mensagens ao longo dos 176 dias observados.

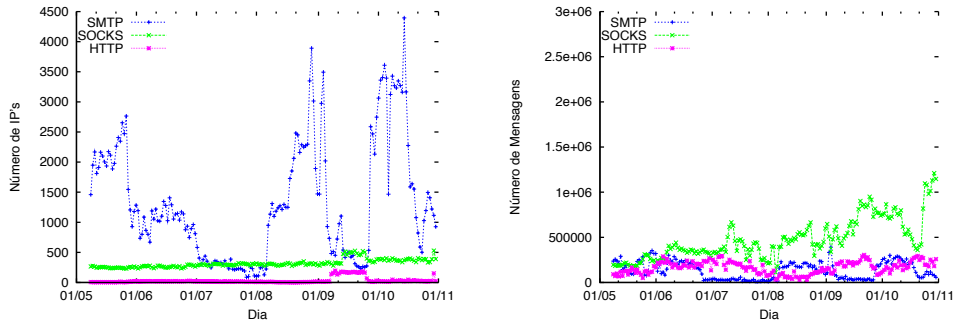
Durante todo o período de coleta, o número de mensagens utilizando SOCKS foi maior que aquelas usando HTTP. Porém, no início do período, mesmo existindo um número maior de mensagens enviadas através de *proxy* SOCKS, o volume do tráfego HTTP foi maior, com média de 4,68 GB e 1,88 GB por dia, para HTTP e SOCKS, respectivamente. Isso se deve pelo fato dos *spams* utilizando HTTP serem maiores que aqueles enviados através de SOCKS, cuja média do tamanho tende a ser baixa e estável durante todo o período. Entretanto, no dia 11 de julho, houve um grande aumento no número de *spams* efetuados através de SOCKS e, em contrapartida, uma diminuição das mensagens que usaram HTTP. Com isso, o volume do protocolo SOCKS aumentou consideravelmente fazendo com que o tráfego de HTTP diminuísse consideravelmente, com média de 5,14 GB por dia, enquanto o segundo teve média de 0,58 GB por dia.

4.3.5 O *honeypot* EC-01

Tabela 4.5. Visão Geral dos dados coletados no *honeypot* EC-01

	SMTP	**SOCKS**	**HTTP**	**Total**
Mensagens (milhões)	23,44(17,24%)	83,51 (61,42%)	29,02 (21,34%)	135,97
Endereços IP's	105.795 (93,32%)	6.907 (6,09%)	2.597 (2,29%)	113.360
Prefixos de rede	8.399 (90,67%)	1.050 (11,33%)	78 (8,42%)	9.263
Sistemas Autônomos (AS)	1.551 (92,82%)	287 (17,17%)	22 (13,17%)	1.671
***Country Codes* (CC)**	124 (97,64%)	62 (48,82%)	7 (5,51%)	127
Volume de tráfego (GB)	143,01 (10,56%)	757,63 (55,96%)	453,26 (34,48%)	1.353,91

Mesmo o *honeypot* EC-01 não apresentando um número total de mensagens elevado (135 milhões de mensagens) quando comparado aos demais, o volume de tráfego total é bastante alto, com 1.353 GB, sendo o maior dentre todos os *honeypots*. Esse fato pode ser explicado pelo perfil das mensagens recebidas por essa máquina. Enquanto a média global do tamanho das mensagens é 3,85 KB, 3,84 KB e 6,11 KB para os protocolos SMTP, SOCKS e HTTP respectivamente, aqui esses valores são 6,40 KB, 9,51 KB e 16,37 KB, ou seja, consideravelmente maiores que os valores globais apresentados pelos dados coletados.



(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo

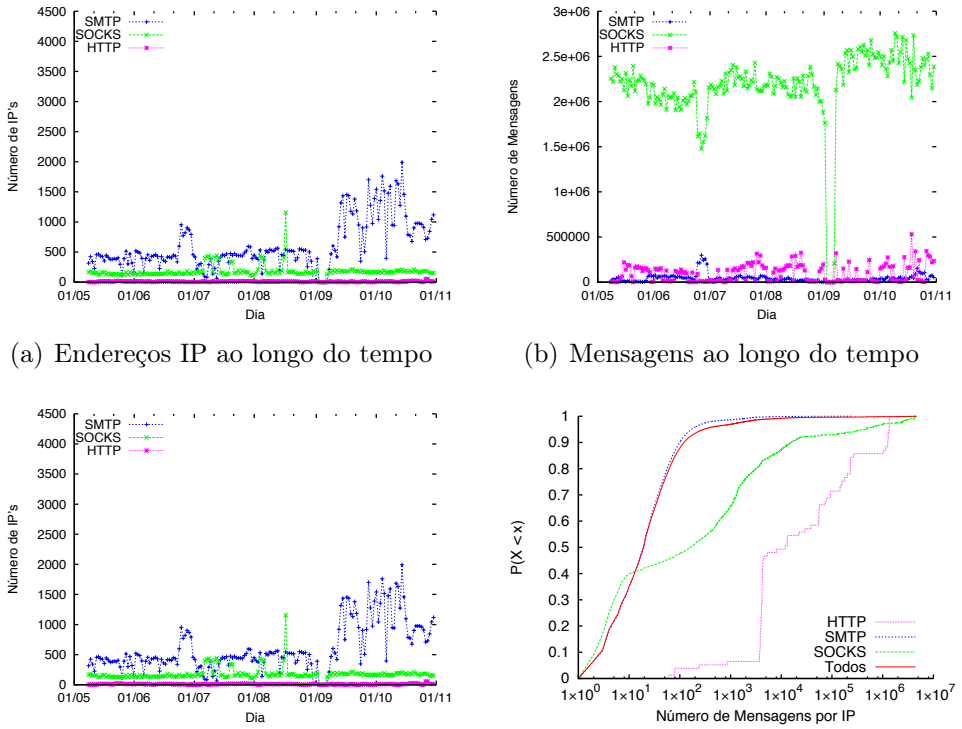
Figura 4.8. Séries temporais por protocolo do *honeypot* EC-01.

Nesse *honeypot*, o comportamento temporal das mensagens que utilizam o protocolo HTTP é relevante, chegando a ter, em alguns momentos, número de mensagens maior que o protocolo SMTP, como pode ser visto na figura 4.8(b). Já considerando o número de transmissores, basicamente por todo o período houveram mais endereços IP utilizando SMTP que os demais protocolos, como mostrado na figura 4.8(a). Porém, entre os dias 05 e 28 de setembro, houve uma forte queda no número de atacantes que usam o protocolo SMTP. Como será explicado na seção 4.5, esse fato está fortemente ligado às mensagens de teste. Nos momentos com alta incidência de endereços IP enviando mensagens de teste, o tráfego SMTP se eleva, enquanto nos períodos em existem poucos desses, o número de mensagens utilizando SMTP diminui, como no período citado. A queda no número de transmissores utilizando SMTP pode ter ocasionado no aumento dos demais protocolos.

4.3.6 O *honeypot* NL-01

O *honeypot* NL-01 apresenta um número elevado de mensagens (406 milhões), sendo que a maior parte delas foi enviada utilizando o protocolo SOCKS (93,91%). Entretanto, mesmo apresentando alto número de mensagens, esse protocolo possui número baixo de endereços IP (apenas 5% do total). Através da figura 4.9(d) podemos perceber que os transmissores utilizando SOCKS enviaram muitas mensagens no período, com cerca de 35% desses atacantes enviando mais de mil mensagens. Como esse *honeypot* está localizado em uma rede de alta qualidade, isso pode ter favorecido para o envio de muitas mensagens por esses transmissores. Já para o protocolo SMTP, que representa 95% dos transmissores e apenas 2% das mensagens, verificamos que cada atacante envia poucas mensagens, sendo que cerca de 90% foi responsável por menos de 100 *spams* nos 176 dias. O protocolo HTTP possui apenas 77 endereços IP enviando mensagens, mas

que são responsáveis por mais de 17 milhões dessas (4,22%) e 51 GB (3,95%). Através das CDF da figura 4.9(d) observamos que esses atacantes enviam muitas mensagens, com aproximadamente 30% deles tendo disseminado mais de cem mil *spams*.



(a) Endereços IP ao longo do tempo

(b) Mensagens ao longo do tempo

(c) Volume (bytes) ao longo do tempo

(d) CDF das Mensagens por Endereço IP

Figura 4.9. Séries temporais por protocolo do *honeypot* NL-01.

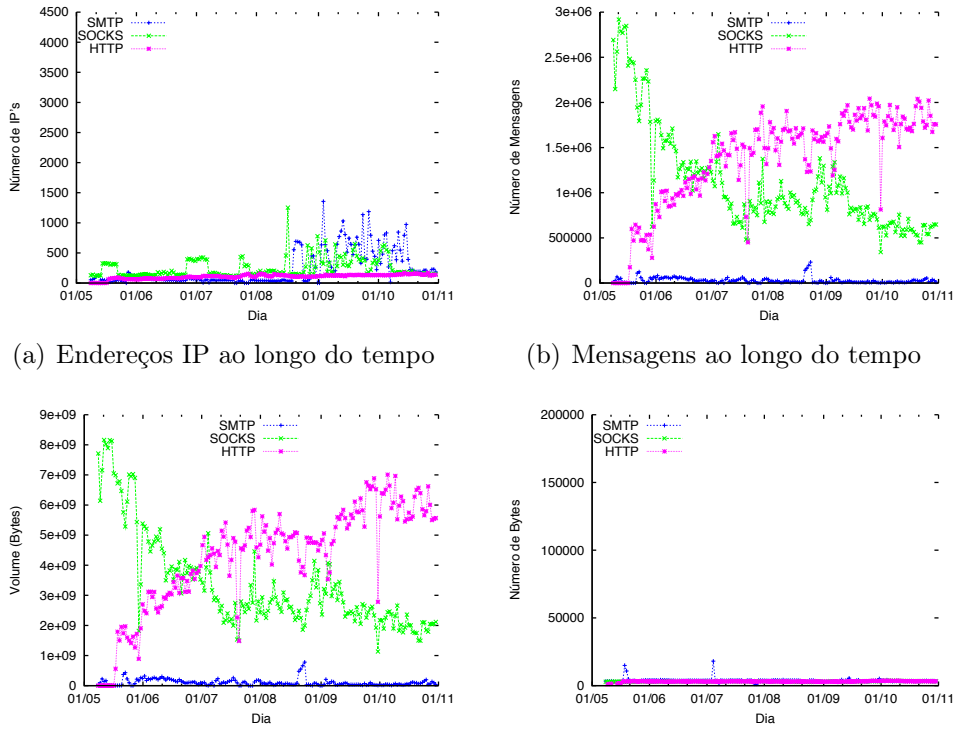
Observando as séries temporais por protocolo (figuras 4.9(a), 4.9(b) e 4.9(c)) verificamos que, durante todo o período, tanto o número de *spams* quanto o volume do tráfego do protocolo SOCKS é maior que o dos demais protocolos, com exceção dos dias 03, 04, 05 e 06 e setembro, em que não houve coleta dos dados. Com relação à curva do número de endereços IP, observa-se que há mais transmissores utilizando SMTP em praticamente todos os dias. Nos quatro primeiros meses de coleta, essa curva se manteve estável, com poucas variações em alguns dias. Porém, após um período de quatro dias em que o *honeypot* ficou impossibilitado de realizar a captura dos dados, o número desses atacantes aumentou consideravelmente. Entretanto, mesmo com esse aumento, o número de mensagens utilizando esse protocolo teve pouca variação. A causa dessa situação foi provavelmente devido ao fato do volume de dados utilizando SOCKS ser elevado, ocupando quase totalmente a banda disponível. Dessa forma, mesmo aumentando-se o número de endereços IP, não seria possível aumentar significativamente o número de *spams* recebidos.

Analisando os *country codes* de origem dos *spams*, observamos que das 406 milhões de mensagens, cerca de 344 milhões, ou seja, 85%, é proveniente do CC US. Além disso, o tráfego gerado pelos endereços IP oriundos desse CC são responsáveis por 1.030 GB, quase 80% do volume total.

4.3.7 O *honeypot* TW-01

Conforme mencionado anteriormente, o perfil que mais difere dos demais foi o do *honeypot* TW-01. Apesar de apresentar um número relativamente pequeno de transmissores (31.809), menor que todos os demais, o número de mensagens é muito alto, com mais de 440 milhões de mensagens, sendo o maior de todos separadamente.

Outro aspecto que diferencia TW-01 dos demais é o tráfego HTTP. Enquanto os outros 7 *honeypots* receberam poucas mensagens utilizando esse protocolo ao longo do período, no TW-01 houve mais de 239 milhões de *spams* enviados dessa forma. Esse número representa mais de 79% do total de mensagens utilizando HTTP recebidas por todos os *honeypots*. Todas essas mensagens recebidas pelo TW-01 foram enviadas a partir de apenas 1.008 endereços IP, sendo 65% provenientes do próprio TW.



(a) Endereços IP ao longo do tempo

(b) Mensagens ao longo do tempo

(c) Volume (bytes) ao longo do tempo

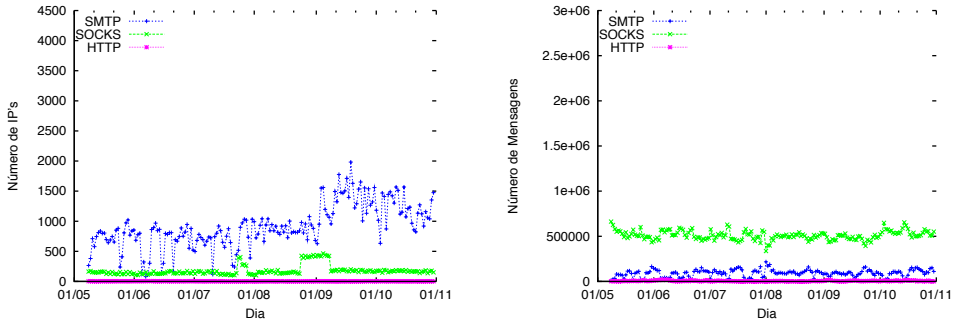
(d) Tamanho das mensagens ao longo do tempo

Figura 4.10. Características do *honeypot* TW-01.

Ainda com relação ao número de *spams* recebidos, o tráfego SMTP foi muito pequeno, representando apenas 1,14% do total. Já o tráfego SOCKS é bastante significativo, sendo utilizado por mais de 12.200 endereços IP e totalizando mais de 195 milhões de mensagens (44,48%). Outro fator interessante desse *honeypot* está no número de campanhas presente em seu tráfego. Durante o período de 176 dias, quase 97 mil campanhas distintas foram observadas no TW-01, o que representa quase 50% do total de campanhas presentes em todos os *honeypots*. Além disso, TW-01 é o único *honeypot* cujo tamanho das mensagens se mantém estável para todos os três protocolos (com excessão de dois pequenos picos das mensagens utilizando SMTP), como mostrado na figura 4.10(d).

O *honeypot* TW-01 está localizado em uma rede de alta velocidade e de qualidade superior à de todos os demais coletores. Esse perfil de rede parece beneficiar *spammers* que usam transmissão "por atacado", talvez porque esses transmissores tendem a permanecer ativos por mais tempo e se beneficiam da estabilização do algoritmo de controle de congestionamento de TCP. Dessa forma, eles teriam uma condição privilegiada para aproveitar a banda disponível, dificultando o acesso a máquinas de menos capacidade que enviam apenas poucas mensagens periodicamente. Provavelmente pelo mesmo motivo, aumentos significativos no número de transmissores usando SOCKS observados durante o período em alguns momentos não tiveram grande impacto sobre o tráfego observado, pois os novos transmissores se viam em desvantagem ao concorrer pela banda disponível com os transmissores pesados já existentes.

4.3.8 O *honeypot* UY-01



(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo

Figura 4.11. Séries temporais por protocolo do *honeypot* UY-01.

O *honeypot* UY-01 coletou cerca de 106 milhões de *spams*, enviados por quase 52 mil endereços IP. Esse baixo número de mensagens coletadas pode ser explicado pelo

fato dessa máquina estar localizada em uma rede de baixa qualidade. Desse total de endereços, 96% refere-se ao protocolo SMTP, sendo responsável por 15% das mensagens. Já os 3,65% que respondem pelo protocolo SOCKS enviaram mais de 84% de todo o *spam* naquele *honeypot*. Como pode ser visto na figura 4.11, o protocolo SMTP possui mais endereços IP que os demais em todos os dias analisados. Já quanto ao número de mensagens, o protocolo SOCKS foi maior durante todo o período. Com relação ao protocolo HTTP, observa-se números irrisórios de endereços IP na figura 4.11(a). Esse protocolo contabilizou apenas 9 endereços distintos (0,01%), provenientes de CN e HK, porém esses foram responsáveis por cerca de 900 mil mensagens, indicando uma média de 100 mil mensagens por endereço IP, valor bastante elevado.

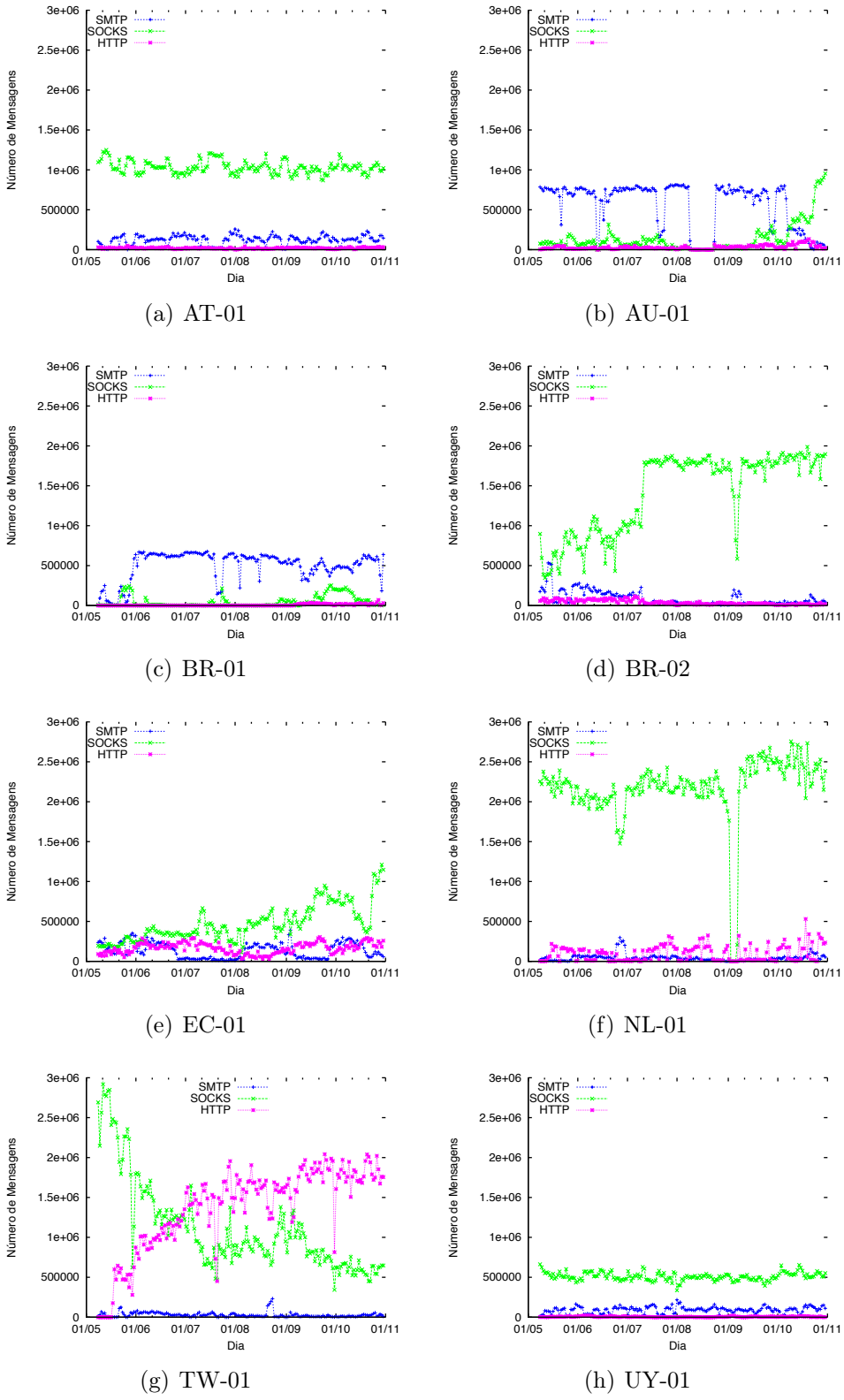
4.4 Comparação entre os *honeypots*

Após pontuarmos detalhes relacionados a cada um dos *honeypots*, nessa seção apresentamos a comparação dos dados coletados por cada um deles, mostrando tanto características comuns quanto pontos divergentes.

4.4.1 Tráfego de *spam* por protocolo

O tráfego apresentado por todos os *honeypots* apresenta variações ao longo do tempo. Como será mostrado posteriormente, a principal causa da variação do número de mensagens e número de endereços é o aparecimento e encerramento das campanhas. Entretanto, mesmo havendo variação, é possível perceber características gerais em cada *honeypot*, como mostrado nas figuras 4.12 e 4.13. Observando o AT-01, por exemplo, percebemos que o tráfego majoritário presente é causado por SOCKS, bem como nos *honeypots* BR-02, NL-01 e UY-01. Já o *honeypot* AU-01 possui maior parte do tráfego causado por SMTP, assim como o BR-01, enquanto EC-01 possui um equilíbrio no tráfego gerado por cada protocolo, com um aumento no número de mensagens utilizando SOCKS no final do período analisado. O *honeypot* TW-01, conforme discutido anteriormente, possui características distintas dos demais, apresentando maior número de *spams* que abusam de *proxies* HTTP.

Com relação ao número de endereços de origem, todos eles apresentam quantidade pequena de transmissores utilizando SOCKS que, conforme observado anteriormente, enviam um número alto de mensagens. Além disso, possuem um alto valor de transmissores explorando SMTP, sendo que em todos os *honeypots*, durante praticamente todos os dias são encontrados mais atacantes desse tipo, com exceção do TW-01 que, conforme pode ser visto na figura 4.10(a), possui menos transmissores utilizando SMTP



(a) AT-01 (b) AU-01

(c) BR-01 (d) BR-02

(e) EC-01 (f) NL-01

(g) TW-01 (h) UY-01

Figura 4.12. Número de mensagens ao longo do tempo para todos os *honeypots*.

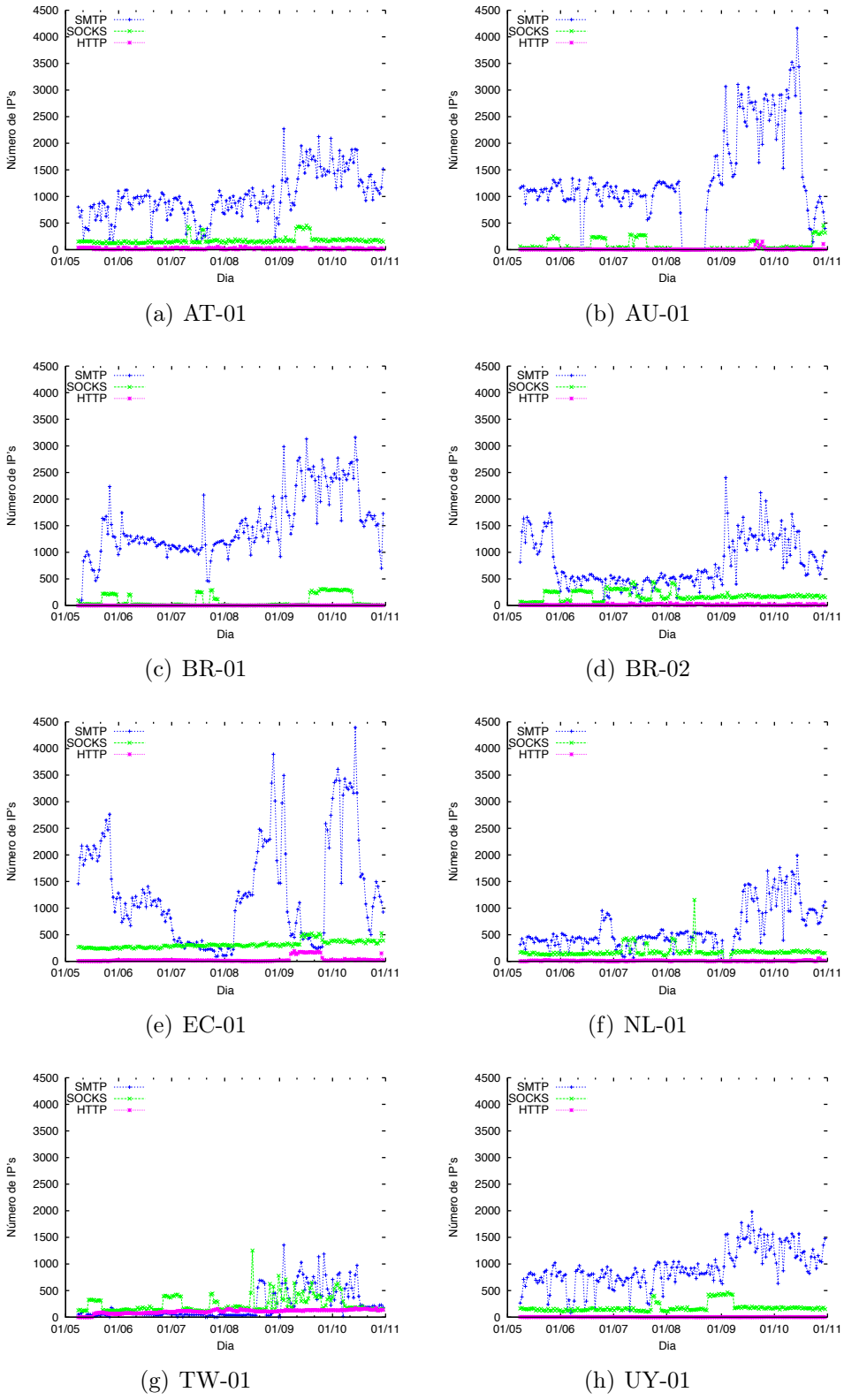
que aqueles que utilizam SOCKS no início do período e um equilíbrio entre esses protocolos no final da análise. Já o número de endereços IP de origem que abusam a máquina através de HTTP é quase irrisório, sendo que apenas os *honeypots* EC-01 e TW-01 foram atacados por mais de mil desses.

Através da tabela 4.6 podemos observar que os *honeypots* que coletaram o maior número de mensagens (TW-01 e NL-01), com quase 850 milhões de mensagens considerando-se o valor total dos dois, estão entre aqueles com menor número de endereços IP. TW-01 é o *honeypot* que tem menos transmissores distintos, com apenas 31 mil deles, e NL-01 possui o terceiro menor valor, com quase 65 mil atacantes tendo abusado dele. Já as máquinas que menos receberam *spams* (BR-01, UY-01 e AU-01, respectivamente, que somando todas as suas mensagens não alcançam o valor coletado por NL-01), possuem os maiores valores de endereços IP de origem distintos, sendo que todos eles receberam mensagens de mais de 100 mil transmissores. Com isso, podemos concluir que o número de mensagens recebidas não está diretamente relacionado à quantidade de endereços IP distintos. Isto ocorre devido ao perfil dos transmissores incidentes no *honeypot*. Por exemplo, BR-01 teve mais de 100 mil IP's porém, desses, 98% utilizam o protocolo SMTP que, conforme mostrado, enviam poucas mensagens. Outro exemplo está no *honeypot* NL-01 que, como já apresentamos na seção 4.3.6, possui atacantes utilizando SOCKS com perfil de envio muito intenso. Como consequência, mesmo não possuindo número elevado de endereços IP, coletou um alto número de mensagens. Uma possível explicação para isso está no fato de transmissores monopolizarem a banda em redes de melhor qualidade, como BR-02, TW-01 e NL-01, portanto haverão poucos atacantes enviado muitas mensagens.

Tabela 4.6. Visão Geral de todos os *honeypots*.

	AT-01	**AU-01**	**BR-01**	**BR-02**	**EC-01**	**NL-01**	**TW-01**	**UY-01**
Msgs (10^6)	206,91	127,13	97,24	271,28	135,97	406,58	440,33	106,72
End. IP (10^3)	57,50	100,54	100,26	79,74	113,36	64,92	31,80	51,70
Prefixos (10^3)	8,01	10,94	10,01	7,09	9,26	6,25	6,43	7,94
ASes (10^3)	1,91	2,09	2,06	1,47	1,67	1,28	1,21	1,90
Country Codes	130	137	132	121	127	119	114	128
Volume (GB)	683,98	782,72	341,27	1.127,72	1.353,91	1.290,80	1.271,00	380,28
Msgs/end. (10^3/IP)	3,60	1,26	0,96	3,40	1,20	6,26	13,84	2,06
Vol./end. (MB/IP)	12,18	7,97	3,49	14,48	12,23	20,36	40,92	7,53
Vol./msg (KB/msg)	3,46	6,45	3,67	4,35	10,44	3,33	3,03	3,74

Considerando-se o volume do tráfego coletado por cada *honeypot*, observamos também diferenças. Enquanto TW-01, NL-01 e BR-02, que receberam muitas mensagens, apresentam volume total bastante elevado, os *honeypots* BR-01, UY-01, AT-01 e AU-01 possuem volume de tráfego menor, tendo enviado menos mensagens que os



(a) AT-01 (b) AU-01

(c) BR-01 (d) BR-02

(e) EC-01 (f) NL-01

(g) TW-01 (h) UY-01

Figura 4.13. Número de endereços IP por protocolo.

demais. Uma exceção é o *honeypot* EC-01, que possui o maior volume dentre todos os *honeypots*, ultrapassando os 1.300 GB. Esse fato pode ser explicado pelo perfil das mensagens recebidas, que apresentam tamanhos maiores que nos demais *honeypots*, conforme discutido na seção 4.3.5.

4.4.2 Distribuição dos endereços IP de origem entre os honeypots

A tabela 4.7 apresenta o número de endereços IP em comum entre cada par de *honeypots*, em relação ao total de IP's do *honeypot* da linha. Analisando-a, percebemos que alguns *honeypots* possuem muitos endereços em comum, como AT-01 e BR-01: 87% dos endereços presentes em AT-01 também estão presentes no BR-01, correspondendo a 50% dos endereços ali observados nesse último (BR-01 foi contactado por mais máquinas). Outro exemplo é entre os *honeypots* AT-01 e AU-01, que possuem 48.465 endereços em comum, que representam 84% dos endereços vistos em AT-01. Esse alto número de origens em comum entre os *honeypots* indica que *spammers* tendem a distribuir seu tráfego por uma grande variedade de pontos intermediários, independente de localidade. Com isso, algumas das características do tráfego presente em diferentes pontos tendem a ser similares, como observado anteriormente, uma vez que parte dos transmissores são comuns e, em consequência, as mensagens de *spam* observadas tendem a ser semelhantes entre eles.

Tabela 4.7. Percentual de endereços IP em comum entre os *honeypots*.

	AT-01	**AU-01**	**BR-01**	**BR-02**	**EC-01**	**NL-01**	**TW-01**	**UY-01**	**Total**
AT-01	-	84%	87%	76%	64%	70%	20%	61%	57.499
AU-01	48%	-	72%	58%	62%	52%	17%	44%	100.547
BR-01	50%	72%	-	66%	63%	53%	16%	44%	100.264
BR-02	53%	74%	83%	-	69%	62%	17%	48%	79.746
EC-01	32%	55%	56%	48%	-	38%	11%	28%	113.360
NL-01	61%	81%	82%	76%	67%	-	20%	55%	64.923
TW-01	37%	55%	50%	43%	38%	40%	-	35%	31.809
UY-01	68%	85%	86%	74%	62%	69%	22%	-	51.709

Ainda com relação à tabela 4.7, podemos notar que o *honeypot* TW-01 é bastante distinto dos demais com relação aos endereços IP que abusaram dele. Além de registrar um número menor de endereços IP de origem (apenas 31.809), esses endereços não aparecem em altas proporções nos demais *honeypots*. O *honeypot* que apresenta maior número de endereços em comum com o TW-01 é o AU-01, com 55% dos 31.809 endereços do primeiro sendo observados no segundo também. Porém, em média, cerca de 43% dos endereços observados em TW-01 aparecem nos outros *honeypots*, valor

baixo quando comparado aos valores dos demais. Esse baixo percentual de endereços IP em comum com os demais *honeypots* talvez explique a diferença apresentada por seu tráfego com relação àquele das demais máquinas de captura.

Além disso, como forma de verificar se o número de endereços IP em comum entre cada par de *honeypot* é realmente maior que o esperado, computamos o número de endereços IP esperados em cada máquina, caso a associação deles a cada *honeypot* fosse aleatória (uniformemente distribuídos). Então, verificamos o número de endereços IP em comum esperado para cada par de máquinas. Como mostrado na tabela 4.8, observamos que, com exceção dos *honeypots* EC-01 e TW-01, o número de endereços IP em comum para cada par de coletores é notavelmente maior que o esperado, sendo muitas vezes maior que 50%. Isso sugere que os remetentes de *spam* enviam mensagens para máquinas específicas, provavelmente descobertas por informações externas, como por exemplo, compartilhamento de informações por *botnets*.

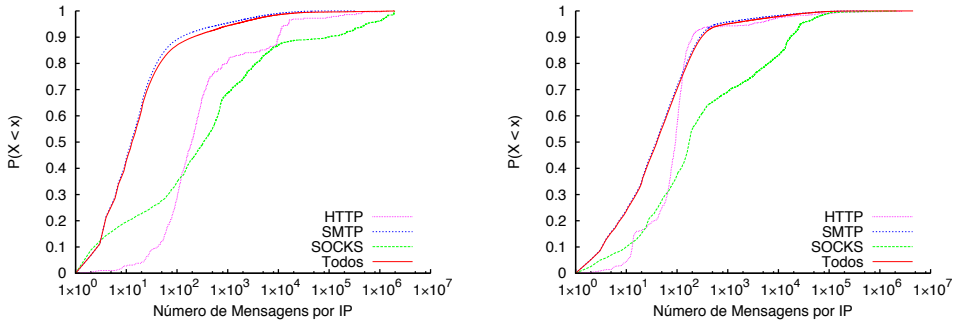
Tabela 4.8. Percentual da diferença entre o número de endereços IP em comum entre os *honeypots* e a distribuição aleatória de endereços IP para cada *honeypot*.

	AT-01	**AU-01**	**BR-01**	**BR-02**	**EC-01**	**NL-01**	**TW-01**	**UY-01**
AT-01	-	51%	56%	72%	2%	94%	13%	113%
AU-01	50%	-	29%	31%	-1%	44%	-4%	53%
BR-01	57%	29%	-	49%	0%	47%	-9%	53%
BR-02	33%	66%	49%	-	10%	72%	-4%	67%
EC-01	0%	-1%	1%	9%	-	6%	-38%	-2%
NL-01	91%	45%	47%	72%	7%	-	13%	92%
TW-01	16%	-1%	-10%	-3%	-40%	11%	-	21%
UY-01	113%	52%	55%	67%	-1%	92%	23%	-

4.4.3 Número de mensagens enviadas por endereço IP

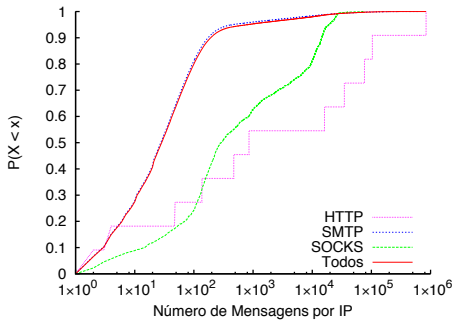
Tendo em vista o número de mensagens enviadas por cada endereço IP percebemos que, quando consideramos todos os transmissores, os *honeypots* possuem curvas próximas. Os *honeypots* AU-01 e EC-01, por exemplo, apresentam cerca de 70% dos endereços IP enviando menos de 100 mensagens ao longo do período, como pode ser visto nas figuras 4.14(b) e 4.14(e). Já os *honeypots* AT-01, BR-02 e NL-01 têm cerca de 30% dos endereços IP enviando 10 mensagens ou menos, e aproximadamente 85% sendo responsáveis por menos de 100 mensagens de *spam*. O mesmo ocorre quando são analisados os atacantes que enviam mensagens utilizando SMTP. Como pode ser visto na figura 4.14, as CDF's representando esse protocolo são bastante semelhantes entre os *honeypots*.

Ao analisar os demais protocolos, notamos diferenças significativas entre os *honeypots*. Com relação ao número de *spams* enviado pelos atacantes que fazem uso de

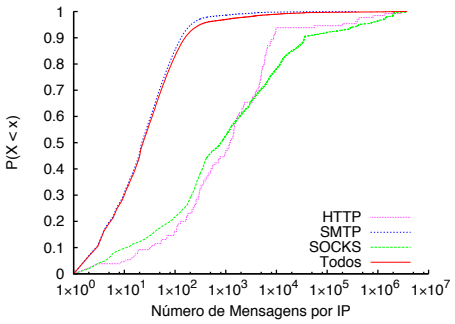


(a) AT-01

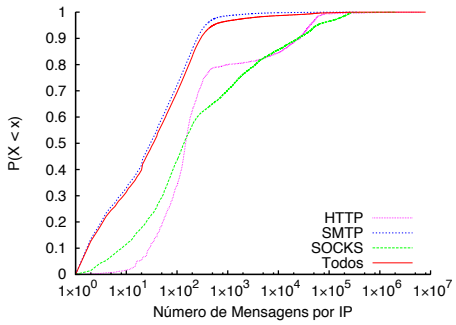
(b) AU-01



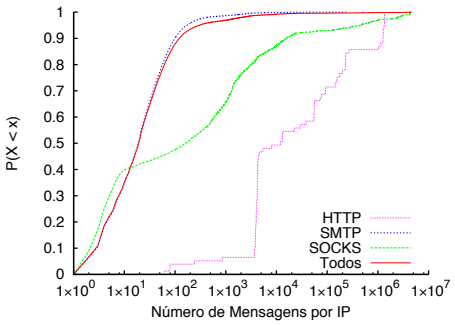
(c) BR-01



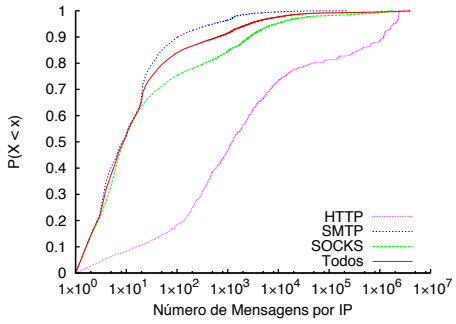
(d) BR-02



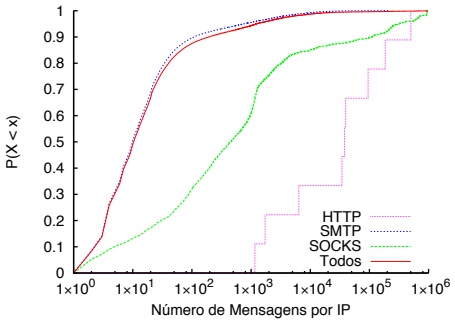
(e) EC-01



(f) NL-01



(g) TW-01



(h) UY-01

Figura 4.14. CDF do número de mensagens por IP de todos os *honeypots*.

SOCKS, BR-01 e BR-02, por exemplo, contam com apenas 20% dos endereços IP enviando menos de 100 mensagens, enquanto as máquinas EC-01 e NL-01 possuem cerca de 45%, e TW-01 chega a ter 75% enviando menos que esse valor. Os *honeypots* AT-01, BR-02, NL-02 e UY-01 têm cerca de 10% de seus endereços IP enviando mais de cem mil mensagens durante o período, já AU-01, TW-01 possuem apenas 1%. Já BR-01 não possui nenhum transmissor com essa quantidade de mensagens, sendo que aquele com maior valor enviou cerca de 77 mil *spams* nos 176 dias avaliados.

Como mostrado na figura 4.14, o protocolo HTTP possui diferenças ainda maiores. Enquanto 90% dos endereços IP do *honeypot* AU-01 enviaram 200 mensagens ou menos, TW-01 conta com cerca de 75% de transmissores que enviaram mais que esse valor e NL-01 tem 97% de seu total ultrapassando essa marca. Essas máquinas citadas possuem endereços IP muito intensos, enviando quantidades muito altas de *spam* através desse protocolo. Mais de 20% dos atacantes em TW-01 enviaram mais de cem mil mensagens. Já para NL-01, esse número aumenta para 30%.

4.4.4 Distribuição dos *Country Codes* de origem

Como os *honeypots* se localizam em diferentes posições do globo, torna-se interessante observar a origem dos transmissores dos *spams* recebidos por cada um deles, para verificar se a localização influencia no recebimento de mensagens de diferentes origens. A tabela 4.9 mostra os 10 *country codes* que mais enviaram mensagens para cada um dos *honeypots*. Em todos os *honeypots*, os *country codes* que mais abusaram são basicamente os mesmos, alterando pouco entre eles e, como seria de se esperar, estão todos entre os mais frequentes no global (tab. A.2).

Tabela 4.9. Percentual do total de mensagens por *country codes* em cada *honeypots*.

#	AT-01	AU-01	BR-01	BR-02	EC-01	NL-01	TW-01	UY-01
01	US (72%)	CN (22%)	BR (15%)	US (71%)	TW (24%)	US (85%)	US (68%)	US (68%)
02	PH (14%)	TW (12%)	CN (13%)	PH (16%)	BR (20%)	PH (10%)	PH (23%)	PH (14%)
03	CN (2%)	BR (11%)	US (9%)	CN (3%)	CN (20%)	JP (2%)	JP (4%)	CN (4%)
04	JP (2%)	US (7%)	TW (7%)	JP (3%)	US (8%)	TW (1%)	TW (3%)	BR (2%)
05	BR (1%)	RU (6%)	RU (7%)	TW (2%)	IT (3%)	CN (1%)	CN (0,25%)	JP (2%)
06	TW (1%)	KR (3%)	KR (4%)	BR (1%)	KR (2%)	BR (0,18%)	BR (0,16%)	TW (1%)
07	RU (1%)	IN (3%)	IN (3%)	HK (0,36%)	RU (2%)	IN (0,08%)	TH (0,10%)	RU (1%)
08	KR (0,37%)	HK (2%)	GB (3%)	KR (0,30%)	HK (2%)	TH (0,07%)	IN (0,07%)	KR (1%)
09	IN (0,34%)	UA (2%)	UA (2%)	IN (0,29%)	IN (2%)	RU (0,06%)	RU (0,05%)	IN (1%)
10	HK (0,33%)	GB (2%)	NL (2%)	TH (0,28%)	MY (1%)	AE (0,06%)	GE (0,05%)	UA (0,41%)

Entretanto, a proporção de mensagens enviadas por cada *country code* se altera em cada um. Dos oito *honeypots*, em cinco deles mais de 68% das mensagens recebidas

são provenientes de US. Já nos outros três *honeypots*, a origem dos *spams* se distribui entre os vários *country codes*. Além disso, naqueles três, o número de mensagens concentrado nos dez principais CCs é significativamente inferior ao dos demais, indicando uma maior distribuição da origem dos *spammers* nesses casos. No caso do AU-01, por exemplo, CN é o que mais o abusou, porém representa apenas 22% do total de mensagens, enquanto no NL-01, US é responsáveis por mais de 85% dos *spams*. Outro ponto interessante é que mesmo os dois *honeypots* brasileiros estando próximos geograficamente, a distribuição dos *country codes* originários das mensagens recebidas por eles é muito distinta. No BR-02, US representa mais de 70% dos *spams*, enquanto no BR-01 representa apenas 9%. Já as mensagens provenientes do próprio Brasil equivalem a mais de 15% dos *spams* recebidos pelo BR-01, sendo o *country code* a mais abusar aquela máquina. Por outro lado, no BR-02 o *country code* BR é apenas o sexto, com cerca de 1% do total de mensagens recebidas.

Essa diferença de perfil chama a atenção pelo fato dos *honeypots* BR-01 e EC-01 estarem instalados em redes de "pior qualidade", segundo relatos dos membros da nossa equipe que acompanham essas máquinas ao longo do tempo. Apesar de todos os coletores terem uma limitação de banda de entrada de 1 Mbps, a capacidade máxima dos links de acesso às redes onde estão aquelas máquinas é mais baixa que a do restante, além de serem conexões que ao longo do tempo se mostraram mais instáveis, com maiores taxas de perda, etc. Isso sugere que a semelhança entre essas máquinas (e sua diferença para as demais) pode ser mais devido à sua conectividade que à sua posição na rede. Nesse sentido, TW-01 também se destaca, por ser o coletor na rede considerada de maior banda e qualidade.

Entretanto, fatores geográficos ainda podem desempenhar um papel na escolha dos *spammers*. Chama a atenção, por exemplo, o fato de o *honeypot* TW-01 (Taiwan) ser o que apresenta menor volume de tráfego originado de CN (China). A razão para essa diferença não está clara e exigiria uma análise mais abrangente das condições (inclusive políticas) daquela região.

4.4.5 Distribuição dos Sistemas Autônomos de origem

Como foi mostrado anteriormente, apenas 10 Sistemas Autônomos distintos são responsáveis por mais de 80% de todo o tráfego de *spam* recebido pelos *honeypots*. Dessa forma, seria interessante entender o quanto cada AS contribui para as mensagens coletadas em cada uma das máquinas.

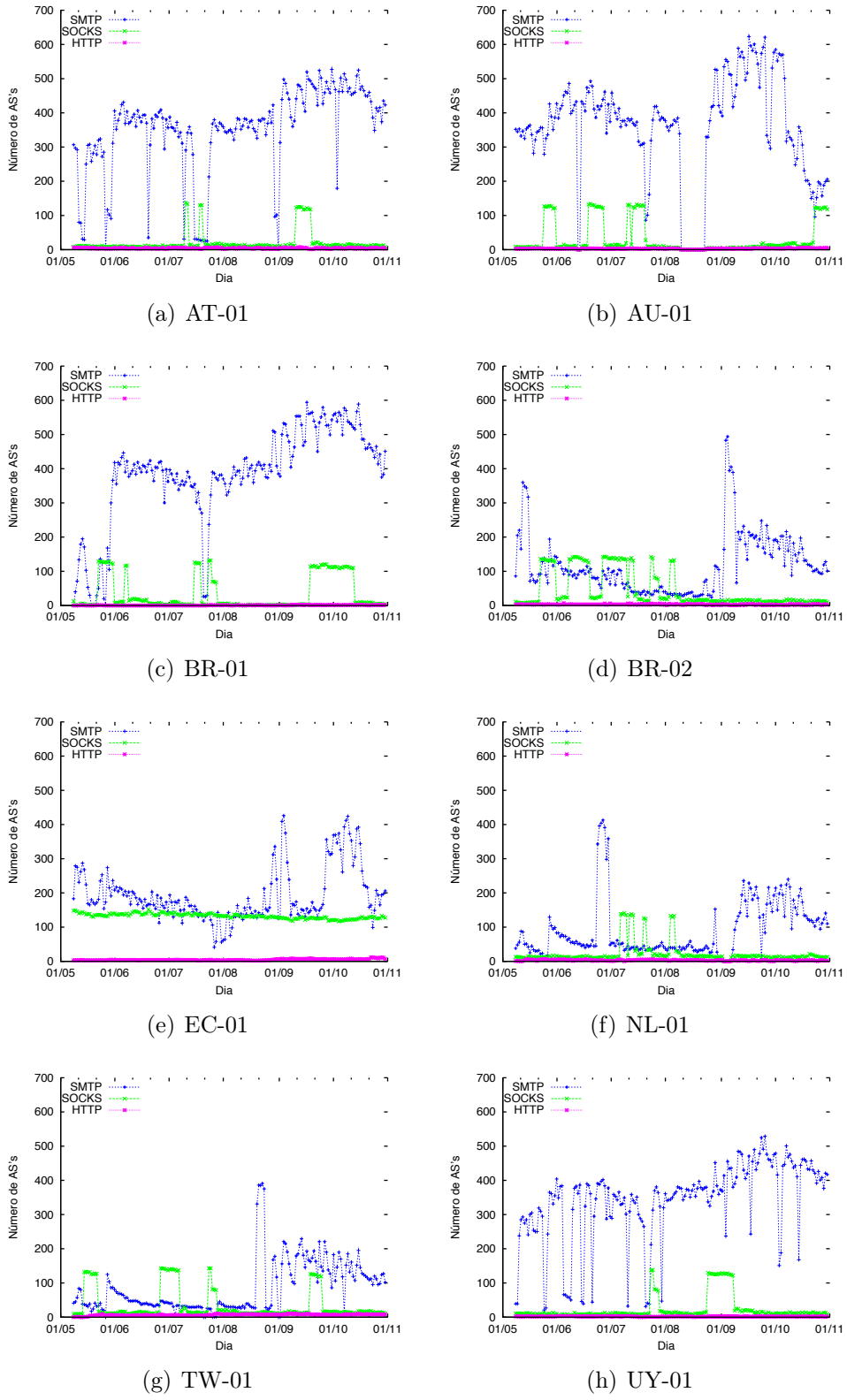
A tabela 4.10 mostra os 10 ASes que mais enviaram mensagens em cada *honeypot* durante o período. Observando-a, podemos verificar que, assim como na análise dos

Tabela 4.10. Percentual do total de mensagens por Sistemas Autônomos mais ativos em cada *honeypot*.

#	AT-01	AU-01	BR-01	BR-02	EC-01	NL-01	TW-01	UY-01
01	10297 (62%)	4134 (13%)	3462 (7%)	10297 (56%)	3462 (19%)	10297 (72%)	10297 (59%)	10297 (58%)
02	9299 (9%)	3462 (8%)	4134 (6%)	29802 (12%)	4134 (14%)	29802 (11%)	9299 (15%)	9299 (8%)
03	29802 (9%)	4837 (3%)	28573 (3%)	9299 (11%)	27699 (4%)	9299 (6%)	29802 (9%)	29802 (8%)
04	6648 (5%)	23650 (2%)	4837 (2%)	6648 (4%)	8167 (3%)	6648 (3%)	6648 (8%)	6648 (5%)
05	4134 (1%)	28573 (2%)	18881 (2%)	4134 (2%)	23650 (3%)	3462 (1%)	3462 (3%)	4134 (2%)
06	3462 (1%)	18881 (2%)	27699 (2%)	3462 (2%)	3269 (3%)	21788 (1%)	2497 (2%)	3462 (1%)
07	2497 (0,7%)	27699 (2%)	4230 (2%)	4713 (1%)	18881 (3%)	4134 (0,7%)	4713 (0,9%)	2497 (0,7%)
08	4713 (0,5%)	38186 (2%)	16276 (1%)	2497 (0,7%)	4230 (2%)	4713 (0,7%)	21788 (0,6%)	4725 (0,6%)
09	4725 (0,5%)	24164 (1%)	10429 (1%)	21788 (0,7%)	28573 (2%)	2497 (0,7%)	9658 (0,4%)	4837 (0,5%)
10	21788 (0,4%)	9924 (1%)	8167 (1%)	4725 (0,5%)	24164 (2%)	4725 (0,4%)	4780 (0,3%)	4713 (0,4%)

country codes, existem dois padrões distintos na distribuição dos *honeypots*. As máquinas AT-01, BR-02, NL-01, TW-01 e UY-01 possuem grande parte de suas mensagens provenientes de apenas um AS (10297). Enquanto os demais *honeypots* possuem maior distribuição nos sistemas autônomos responsáveis pelas mensagens recebidas. Por exemplo, o AS que mais enviou mensagens em BR-01 totaliza apenas 7% das mesmas. Essa diferença na distribuição das mensagens pelos ASes pode ser vista no comportamento apresentado pela série temporal do número de mensagens para cada *honeypot*. Todas as máquinas presentes no primeiro grupo citado receberam mais mensagens utilizando SOCKS (ou HTTP, no caso do TW-01), sendo a maior parte delas provenientes do AS 10297. Já os *honeypots* do segundo grupo coletaram um valor maior de mensagens que fazem uso de SMTP (EC-01 possui um equilíbrio no tráfego entre os protocolos). Isto pode ter sido devido à ausência dos principais sistemas autônomos utilizando esse protocolo, tornando o tráfego do protocolo SMTP maior que dos demais.

É interessante notar que, mesmo possuindo um tráfego bastante distinto dos demais, o *honeypot* TW-01 mostra como principal AS o mesmo das demais máquinas, representando cerca de 59% do total de suas mensagens. Para explicar esse fenômeno, observamos com maiores detalhes o AS 10297. A tabela 4.11 mostra como o tráfego desse AS se distribui. Enquanto nas máquinas AT-01, BR-02 e UY-01 esse sistema autônomo enviou mais de 99,99% de mensagen sutilizando SOCKS, nos *honeypots* NL-01 e TW-01 houve distribuição das mensagens enviadas entre *proxy* SOCKS e *proxy* HTTP. O *honeypot* NL-01 recebeu mais de 8 milhões de mensagens desse AS utilizando HTTP, que representaram cerca de 2,8% do total das mensagens enviadas por esse AS àquela máquina. Os outros 97,2% foram disseminadas através de SOCKS (houve ainda uma quantidade irrisória de mensagens SMTP). Já o comportamento observado no *honeypot* TW-01 é bastante diferente dos demais. Dos 259 milhões de



(a) AT-01 (b) AU-01

(c) BR-01 (d) BR-02

(e) EC-01 (f) NL-01

(g) TW-01 (h) UY-01

Figura 4.15. Número de Sistemas Autônomos por protocolo.

spams, mais de 60% utilizou o protocolo HTTP, enquanto apenas 40% utilizou SOCKS. Pelo menos 83% de todos os endereços IP daquele AS em TW-01 estão presentes nos demais *honeypots* citados. Analisando esses transmissores em comum, observamos que um mesmo IP enviou mensagens utilizando HTTP no TW-01 e SOCKS nos demais. Um desses endereços, por exemplo, enviou 2.423.507 *spams* através de *proxies* HTTP e 1.536.030 por *proxies* SOCKS, e no AT-01 enviou 1.947.166 mensagens utilizando somente SOCKS.

Tabela 4.11. Número de mensagens enviadas pelo Sistema Autônomo 10297.

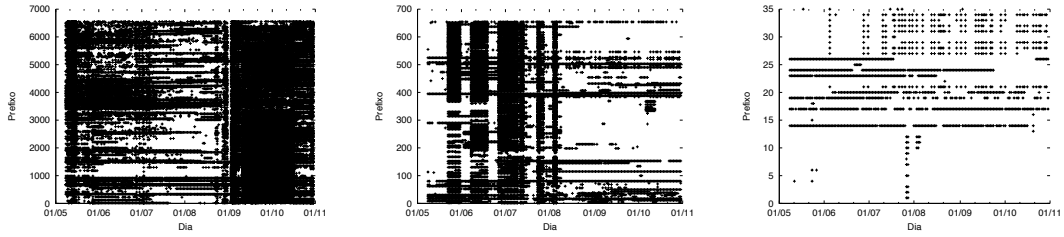
	AT-01	**AU-01**	**BR-01**	**BR-02**	**EC-01**	**NL-01**	**TW-01**	**UY-01**
SMTP (10^3)	11,7	55,9	47,0	4,9	17,7	2,7	5,1	9,4
SOCKS (10^3)	128.472,2	0,2	1,9	152.737,4	0,4	285.657,9	102.520,0	62.229,4
HTTP (10^3)	0,3	1,3	0,3	0,6	2,5	8.242,5	156.977,2	0,2

Através das séries temporais de todos os Sistemas Autônomos, apresentadas na figura 4.15, para cada *honeypot*, podemos verificar que durante praticamente todo o período o número de ASes utilizando o protocolo SMTP é maior que dos demais. Assim como nas curvas para o número de endereços IP por dia, existem picos, com aumento dos ASes utilizando SOCKS, em determinados momentos. Com exceção desses períodos, os valores para esse protocolo são bastante estáveis. Na figura 4.15(e), por exemplo, pode-se ver que o comportamento desse protocolo é praticamente constante em todos os dias. Já o número de ASes que utilizam HTTP é bastante baixo durante o período observado.

4.4.6 Análise temporal dos prefixos de rede

Com o objetivo de entender a distribuição do tráfego originado em cada prefixo de rede presente nos *honeypots*, geramos gráficos indicando os dias em que cada prefixo esteve ativo durante a coleta. Com isso, podemos entender se existem períodos com alta atividade dos prefixos de rede.

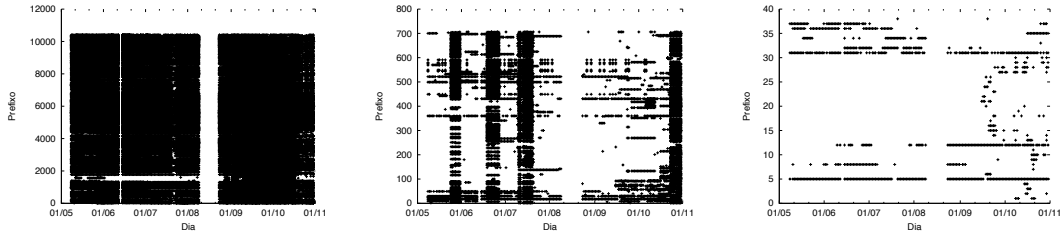
Para o tráfego de *spam* gerado pelo protocolo SOCKS, no *honeypot* BR-02, mostrado na figura 4.16(b), nota-se que no início da coleta, nos meses de junho e julho, houve momentos em que praticamente todos os prefixos de rede observados estiveram ativos. Nesses intervalos, ocorreram picos no número de endereços IP que, conforme será explicado na seção 4.8.1, foram ocasionados pelo surgimento de novas campanhas. Por essa figura (4.16(b)) pode-se observar também que na segunda metade da coleta vários prefixos estiveram inativos. Nesse mesmo período, houve um grande aumento do número daqueles utilizando SMTP, como pode ser visto na figura 4.16(a). Esse fato



(a) SMTP (b) SOCKS (c) HTTP

Figura 4.16. Distribuição temporal dos prefixos de rede incidentes no *honeypot* BR-02.

está diretamente relacionado ao aparecimento de novos endereços IP enviando mensagens de teste através desse protocolo, como será mostrado adiante. Nesse protocolo, até o mês de setembro os prefixos de rede mostraram atividades distintas, sendo que alguns se mantiveram ativos durante esse período, enquanto outros estiveram ativos em determinados momentos e inativos em outros.



(a) SMTP (b) SOCKS (c) HTTP

Figura 4.17. Comportamento temporal dos prefixos de rede incidentes no *honeypot* AU-01. A falha nos dias 13 e 14 de junho e entre 10 e 23 de agosto é devido ao *honeypot* ter ficado fora do ar.

Diferente do comportamento dos prefixos de rede no BR-02, o *honeypot* AU-01 possui a maior parte dos prefixos de rede que utilizaram SMTP ativos durante todo o período. A figura 4.17(a) retrata esse comportamento. Já aqueles que utilizam o protocolo SOCKS apresentaram atuação semelhante ao citado anterior, com intervalos de grande inatividade, seguidos por rajadas em que a grande parte mostrou-se ativo.

Considerando-se os prefixos que fazem uso de *proxy* HTTP, foi possível perceber que a grande maioria deles são observados esporadicamente, com poucos se mantendo ativos durante todo o período, conforme pode ser visto nas figuras 4.16(c) e 4.17(c).

4.5 Análise das mensagens de teste recebidas pelos *honeypots*

Uma característica interessante encontrada em todos os *honeypots* é o padrão de mensagens de teste recebidas. Essas mensagens são geradas pelos *spammers* periodicamente e contêm uma a identificação da máquina que está aparentemente sendo utilizada para entregar o *spam* (nesse caso, um dos *honeypots*). Elas são identificadas no tráfego coletado e recebem tratamento especial, sendo as únicas mensagens que o *honeypot* realmente entrega, periodicamente, para garantir que o *spammer* ache que seu objetivo está sendo alcançado e continue a utilizar a nossa infraestrutura.

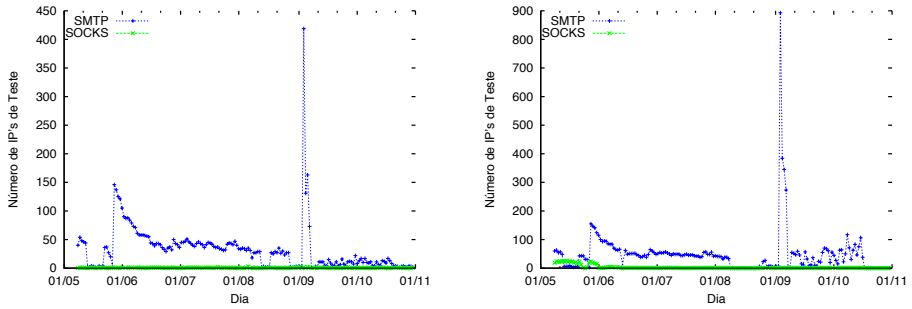
Como pode ser visto na figura 4.18, as curvas que representam o número de endereços IP que enviaram mensagens de teste são muito semelhantes entre todos os coletores, com correlações extremamente elevadas (muito próximas de 1,0, para a maioria dos casos). A tabela 4.12 mostra os endereços em comum entre pares de *honeypots*. Por ela, percebe-se que, além das curvas serem semelhantes, os *honeypots* apresentam muitos endereços em comum entre si como, por exemplo, nos *honeypots* AT-01 e BR-01, em que 1.071 IP's puderam ser identificados em ambos, representando 73% do total de IP's do AT-01.

Em todos os *honeypots* houve um pico de endereços IP enviando mensagens de teste no início do mês de junho, seguido por uma queda gradual no número de IP's. No mês de setembro ocorreu outro pico (com exceção de NL-01, que não teve coleta nesse período), ainda maior que o anterior e de queda abrupta. A tabela 4.13 mostra o percentual de endereços em comum nos pares de *honeypots*. Mesmo o pico tendo ocorrido em todos as máquinas, o percentual apresentado é próximo ao apresentado considerando-se todo o período.

Tabela 4.12. Percentual de endereços IP que enviaram mensagens de teste em comum entre os *honeypots*.

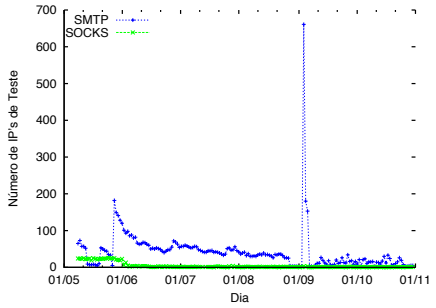
	AT-01	**AU-01**	**BR-01**	**BR-02**	**EC-01**	**NL-01**	**TW-01**	**UY-01**	**Total**
AT-01	-	71%	73%	65%	61%	41%	52%	55%	1.475
AU-01	28%	-	44%	46%	37%	19%	27%	30%	3.727
BR-01	38%	58%	-	48%	48%	27%	35%	39%	2.842
BR-02	36%	64%	51%	-	43%	27%	32%	37%	2.676
EC-01	28%	44%	42%	37%	-	19%	26%	29%	3.174
NL-01	55%	64%	69%	66%	55%	-	41%	52%	1.097
TW-01	54%	70%	71%	61%	59%	32%	-	54%	1.418
UY-01	51%	70%	69%	61%	58%	36%	48%	-	1.601

Analisando essa ocorrência, onde ocorre aumento significativo dos endereços IP enviando mensagens de teste utilizando SMTP, e correlacionado-os com a série tem-

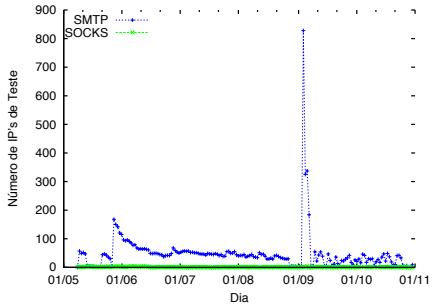


(a) AT-01

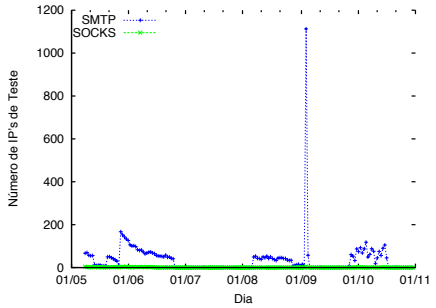
(b) AU-01



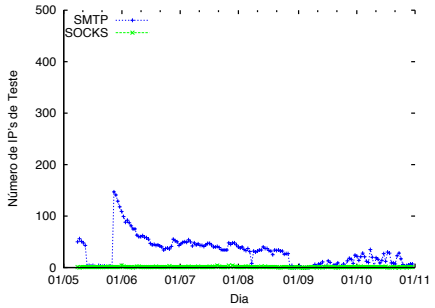
(c) BR-02



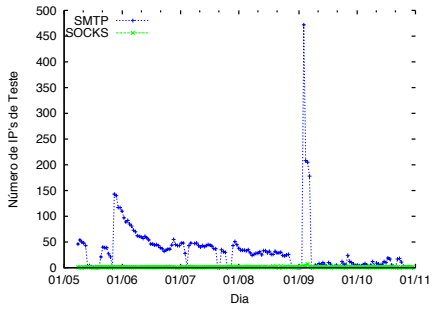
(d) BR-01



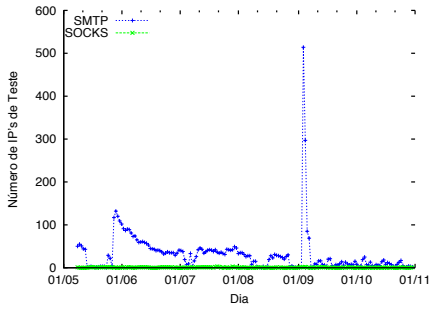
(e) EC-01



(f) NL-01



(g) TW-01



(h) UY-01

Figura 4.18. Número de endereços IP que enviaram mensagens de teste. O *honeypot* NL-01 não estava ativo nos dias em que ocorreu o pico mostrado nos demais gráficos.

poral que apresenta a quantidade de endereços IP diariamente para esse protocolo, percebemos que a segunda também possui aumento em seus valores. Foi possível verificar que, exatamente no dia em que ocorre o pico dos transmissores de mensagens de teste (04 de setembro), é desencadeado um aumento de endereços IP, que segue até o fim do período analisado, como pode ser visto na figura 4.13.

No *honeypot* EC-01, por exemplo, podemos notar grandes semelhanças entre o comportamento do número de endereços IP com a quantidade daqueles que enviam mensagens de teste. A figura 4.18(h) mostra que em dois momentos do período analisado (em julho e setembro) há uma diminuição muito grande no número de transmissores de mensagens de teste através de SMTP. Esses intervalos são exatamente aqueles em que há um vale na quantidade de atacantes SMTP, como pode ser visto na figura 4.13(e) (página 40). Já quando ocorre o pico de endereços IP enviando mensagens de teste no início de setembro, é possível perceber que o valor total de transmissores SMTP nesse período também aumenta consideravelmente.

Portanto, através desses fatos, podemos afirmar que o tráfego de *spam* gerado pelo protocolo SMTP relaciona-se fortemente com suas mensagens de teste. Isso ocorre porque, quando atacantes confirmam o funcionamento de uma máquina sendo abusada, como o fazem através dessas mensagens, eles tendem a comunicar a outros endereços IP a existência dessa, principalmente no caso de *botnets* que, conforme mostramos anteriormente, relacionam-se com o protocolo SMTP.

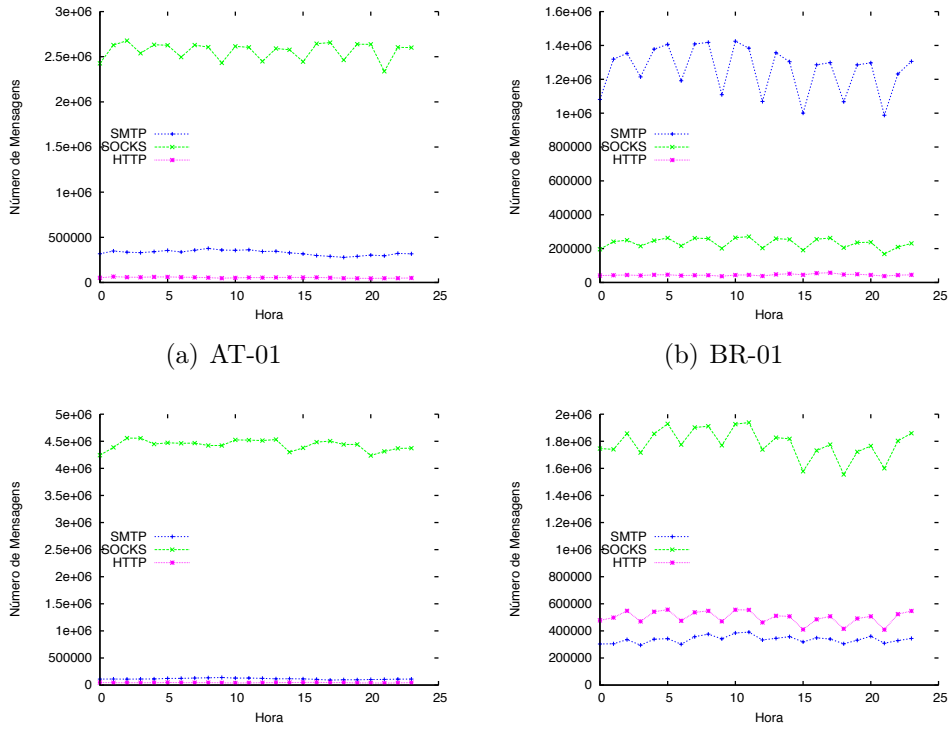
Tabela 4.13. Percentual de endereços IP que enviaram mensagens de teste em comum entre os *honeypots* no pico de IP's nos dias 04, 05 e 06 de setembro. O *honeypot* NL-01 não estava ativo nesse período.

	AT-01	**AU-01**	**BR-01**	**BR-02**	**EC-01**	**NL-01**	**TW-01**	**UY-01**	**Total**
AT-01	-	69%	68%	55%	62%	x	53%	48%	637
AU-01	33%	-	59%	44%	51%	x	39%	40%	1.312
BR-01	36%	64%	-	45%	54%	x	41%	39%	1.210
BR-02	40%	66%	62%	-	59%	x	47%	44%	876
EC-01	35%	60%	58%	46%	-	x	40%	40%	1.124
NL-01	x	x	x	x	x	x	x	x	x
TW-01	46%	69%	68%	56%	60%	x	-	47%	742
UY-01	40%	68%	62%	51%	58%	x	46%	-	765

Foram encontrados 7.249 endereços IP enviando mensagens de teste em todos os *honeypots*. Desse total, 607 (8,4%) enviaram apenas mensagens de teste no período observado, enquanto os 91,6% restantes enviaram tanto mensagens de teste quanto *spams*. Outra característica desses endereços é que a maioria (70,0%) é proveniente de redes domésticas, de acordo com dados de *black lists* consultadas durante a coleta. Com base ainda nessas *black lists*, 33,6% são, garantidamente, máquinas infectadas

por algum tipo de *malware*[1]. Esses fatores indicam que, normalmente os testadores são máquinas infectadas presentes em redes domésticas e que, além de testarem o funcionamento da máquina abusada, continuam enviando *spams* em paralelo a essas mensagens.

4.6 Análise do *spam* por hora do dia



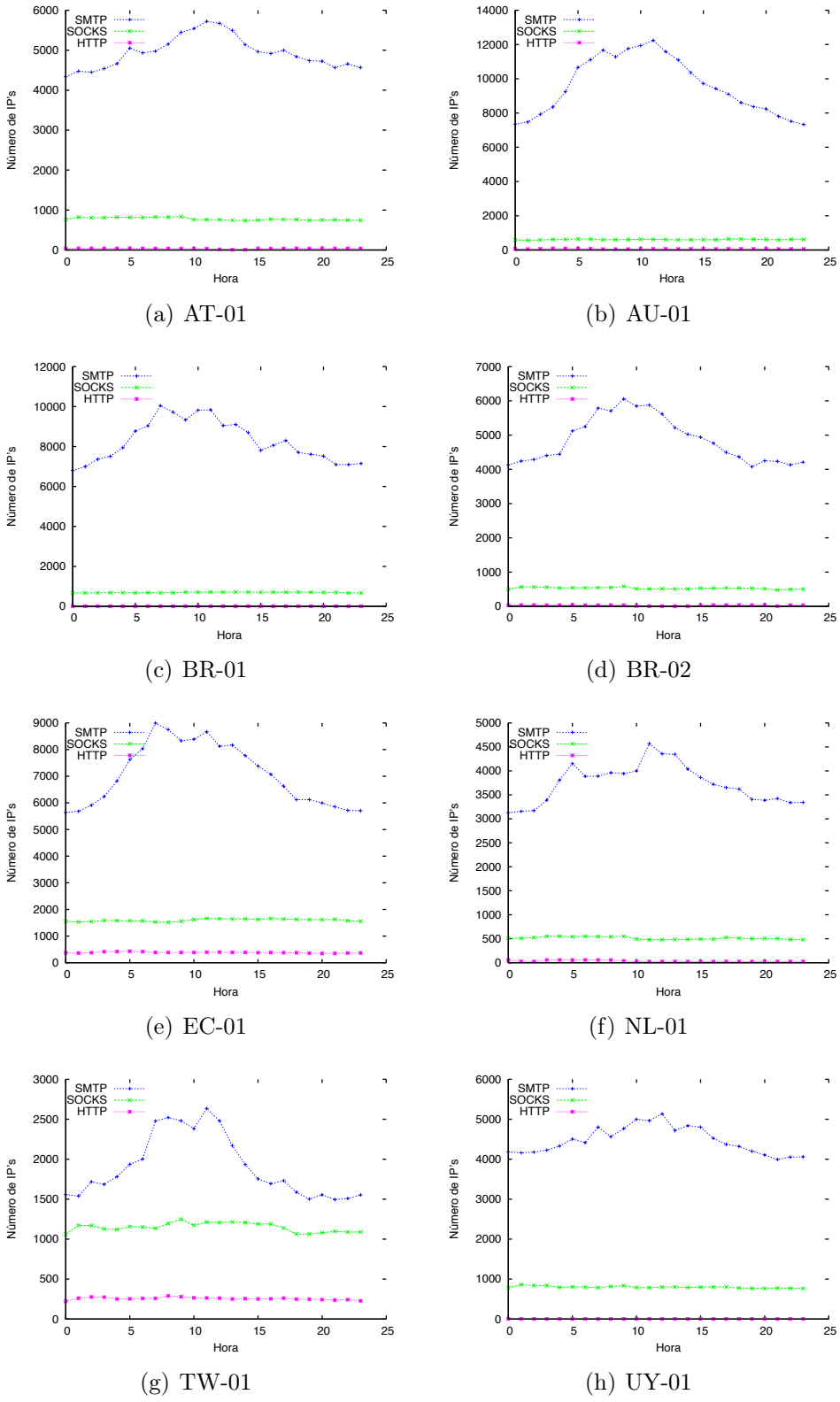
(a) AT-01 (b) BR-01

(c) BR-02 (d) EC-01

Figura 4.19. Número de mensagens por hora do dia.

Outra característica analisada foi o comportamento do tráfego de *spam* ao longo do dia. Para tal, utilizamos o fuso horário GMT para padronizar a hora encontrada em todas as mensagens. Registramos então o número de mensagens encontradas em cada hora do dia, para cada protocolo, além do número de transmissores distintos. Os dados utilizados nesse processo são referentes aos meses de setembro e outubro. A intuição por trás dessa análise está no fato de que transmissores utilizando os protocolos SOCKS e HTTP tendem a enviar inúmeras mensagens durante o dia, levando a crer que estão ativos por todo o período, enquanto aqueles referentes ao tráfego SMTP enviam poucas mensagens, indicando que isso ocorre paulatinamente, em momentos espaçados.

[1]http://www.spamhaus.org



(a) AT-01 (b) AU-01

(c) BR-01 (d) BR-02

(e) EC-01 (f) NL-01

(g) TW-01 (h) UY-01

Figura 4.20. Número de endereços IP por hora do dia.

Através da figura 4.19 podemos notar que o comportamento apresentado pelo número de *spams* recebidos ao longo do dia é estável, com pequenas alterações, principalmente relacionadas ao protocolo SOCKS.

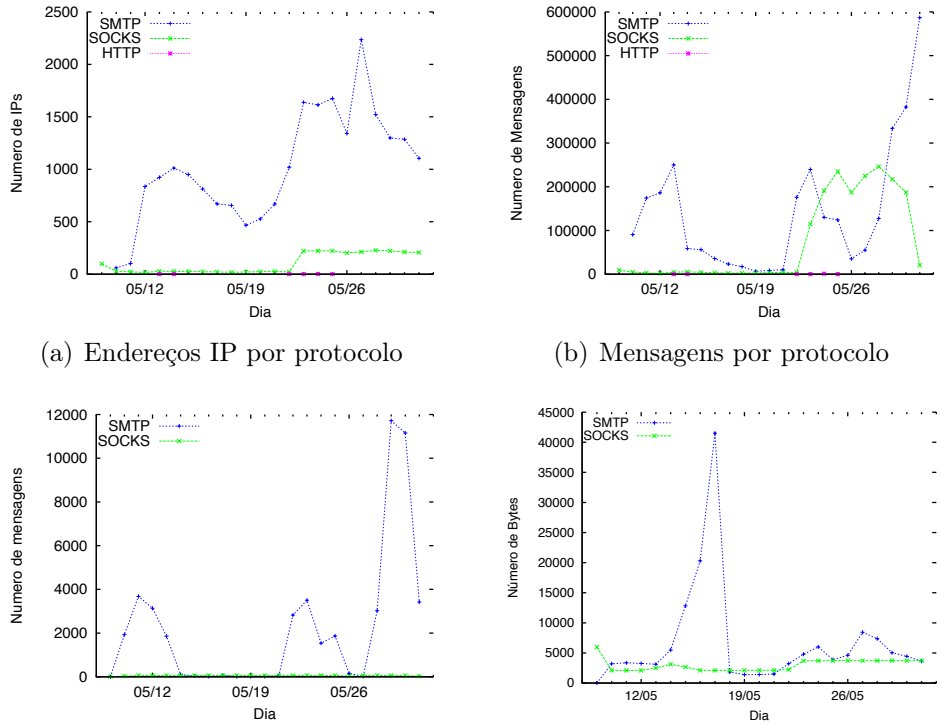
Porém, analisando os endereços IP ativos em cada hora do dia, conforme apresentado na figura 4.20, é possível observar que o protocolo SMTP apresenta um padrão diferente dos demais. Enquanto os protocolos SOCKS e HTTP continuam estáveis, assim como no número de mensagens por hora, a quantidade de transmissores SMTP se altera no decorrer do dia. O perfil apresentado por essa curva, semelhante em todos os *honeypots*, mostra um aumento no número de endereços IP em determinadas horas do dia. Esse fato torna ainda mais evidente que os transmissores enviando mensagens através de SMTP tendem a ser máquinas infectadas, atuando como *bots*, uma vez que tal comportamento sugere que as máquinas enviando essas mensagens estão ligadas apenas em certas horas do dia, indicando serem usuários domésticos infectados. Outra observação que pode ser feita é que, mesmo com a variação do número de endereços IP, o número de mensagens se mantêm constante, o que sugere que existem máquinas atuando de maneira dedicada para envio de *spam* através desse protocolo, cujo trabalho se intensifica nas horas com menor número de transmissores, provavelmente pela liberação de banda. Já para os demais protocolos, evidencia-se também que seus atacantes utilizam infraestruturas dedicadas, dado que seu tráfego é estável durante todo o dia.

4.7 Início do ataque a uma máquina vulnerável na rede

Por fatores externos ao projeto, o endereço IP de um dos *honeypots* foi alterado no início do nosso período de coleta. Apesar dessa máquina já estar em atividade há um longo período, a troca de seu endereço acabou simulando o aparecimento de uma nova máquina na rede. Com isso, foi possível analisar o processo de descobrimento de *proxies* e *mail relays* em uma máquina vulnerável por um *spammer*. Os gráficos da figura 4.21 mostram o comportamento do tráfego recebido por aquele *honeypot* logo após o início de uma operação usando o novo endereço.

Logo no primeiro dia após o surgimento da máquina, cerca de 100 endereços IP já começaram a abusar da máquina, enviando cerca de 9 mil mensagens. Todos esses transmissores abusaram da máquina através do protocolo SOCKS. Uma possível explicação para o seu surgimento é que *spammers* que utilizam SOCKS vasculham a rede à procura de *proxies* abertos e quando encontram já começam a enviar *spams*.

Já os transmissores que utilizam SMTP apareceram somente no segundo dia após o aparecimento do *honeypot*, coincidindo com o aparecimento da primeira mensagem de teste enviada por um *spammer*. Esse fato confirma o que foi visto anteriormente, mostrando que o tráfego SMTP está diretamente relacionado às mensagens de teste, e que os *spammers* que utilizam SMTP se valem de mensagens de teste para confirmar se a máquina a ser abusada realmente entrega as mensagens, para então iniciar a disseminação de suas mensagens.



(a) Endereços IP por protocolo

(b) Mensagens por protocolo

(c) Mensagens de teste por protocolo

(d) Tamanho das mensagens por protocolo

Figura 4.21. Tráfego após surgimento do *honeypot* BR-01.

Com relação aos endereços utilizando SMTP, a partir do segundo dia o número de mensagens de teste aumenta, coincidindo com o aumento do número de transmissores e do número de mensagens enviadas. A maioria dos transmissores que passaram a usar a máquina por SMTP nunca chegaram a fazer uma varredura da máquina antes do primeiro envio, nem usaram mensagens de teste, o que sugere que foram identificados da existência do *honeypot* por algum canal externo (provavelmente através dos canais de comando e controle de uma *botnet*). Quanto ao tráfego de *spam* gerado através de conexões HTTP, foi possível notar que, em um primeiro momento, suas mensagens foram inexistentes. Somente após 122 dias de análise, foi possível observar o surgimento de mensagens utilizando esse protocolo. Porém, esse tráfego continuou fraco compa-

rado aos demais, totalizando apenas 11 endereços IP distintos, responsáveis por cerca de um milhão de mensagens. Já considerando o tráfego utilizando SOCKS, verificamos que após o primeiro dia, o número de transmissores enviando *spams* com esse protocolo se estabiliza, com apenas alguns picos, em que o número de endereços aumenta consideravelmente, assim como o número de mensagens enviadas por eles. A seção a seguir discute esse comportamento com mais detalhes.

4.8 Análise baseada em campanhas

Para analisar mudanças do padrão de tráfego, recorremos em alguns casos à identificação de campanhas de *spam*. Conforme dito anteriormente, uma campanha é um conjunto de mensagens que possuem um objetivo comum e uma mesma estratégia de disseminação [Guerra et al., 2008a]. Para sua identificação, utilizamos o algoritmo de agrupamento *FPCluster*, desenvolvido por Pires et al. [2012], baseado na construção de árvores de padrões frequentes, usadas para extrair os padrões de agrupamento das mensagens.

Para a geração das campanhas foram considerados os dados do dia 09 de maio ao dia 31 de agosto. Esse período foi escolhido por ser o maior intervalo em que as campanhas estavam processadas para todos os *honeypots*. Inicialmente as campanhas foram geradas por dia, para cada *honeypot*. Em seguida, realizamos a fusão das campanhas, de modo a unificar campanhas iguais em dias ou *honeypots* diferentes. O procedimento seguido é descrito na metodologia (seção 4.8. página 57). Após realizar esse procedimento, 573.945 campanhas foram agrupadas em 195.722 campanhas de maior duração.

A figura 4.22 mostra o número de campanhas ativas em cada dia, para cada um dos *honeypots*. Através dela, verificamos que alguns *honeypots* como AT-01, EC-01 e UY-01 não possuem grandes variações com relação a essa característica. Já a máquina BR-01 iniciou recebendo mensagens de poucas campanhas pois, como dito anteriormente, ela obteve um novo endereço IP no dia em que iniciamos a coleta. Porém, após alguns dias esse valor aumentou, uma vez que novos atacantes descobriram a máquina e se tornou estável, com poucas variações. Os alvos NL-01 e TW-01 apresentaram maior número de campanhas que os demais em todos os dias e tiveram maiores variações ao longo do intervalo observado. Nesses *honeypots* as campanhas se mantiveram ativas por praticamente todo o período, como pode ser visto na figura 4.23(b), para o NL-01. Por fim, o *honeypot* BR-02 apresentou variação quando consideramos o início e o fim do período coletado. Nos primeiros dias de análise, um número baixo de campanhas

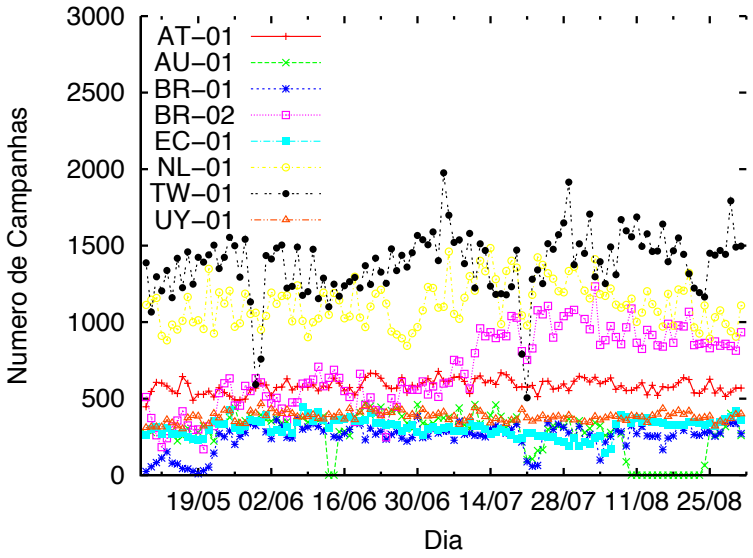
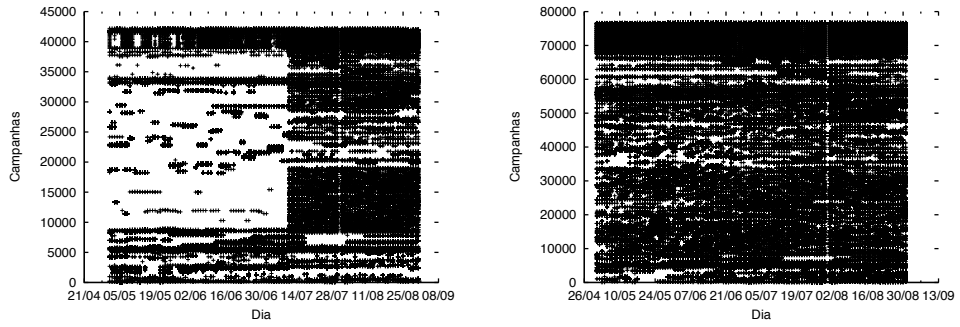


Figura 4.22. Série temporal das campanhas em todos os *honeypots*.

distintas foram encontradas nele. Entretanto, no dia 11 de julho houve um grande aumento nos valores dessa métrica. Esse dia é o mesmo em que observamos um grande aumento no número de mensagens que utilizam o protocolo SOCKS, como pode ser visto na figura 4.7(b) (página 31). Pela figura 4.23(a) é possível identificar esse mesmo fato. Essa figura apresenta a distribuição das campanhas ativas ao longo do tempo. Por ela, podemos notar que no início do período muitas campanhas estavam inativas e apenas no ponto citado essas campanhas se iniciaram, causando aumento no tráfego utilizando SOCKS.

A tabela 4.14 apresenta o percentual de campanhas em comum entre cada par de *honeypots* em relação ao total de campanhas do *honeypot* desta linha. Por exemplo, os *honeypots* AT-01 e NL-01 possuem 29.550 campanhas em comum, o que representa 80,52% do total das campanhas do AT-01. TW-01 possui um número pequeno de campanhas em comum com os demais *honeypots* e, em consequência, possui características muito distintas dos demais. Essas indicações nos levam a crer que *honeypots* com muitas campanhas em comum possuem muitas similaridades, o que não acontece nos casos em que o número de campanhas em comum é baixa.

Mais um fator interessante com relação ao impacto da rede em que o *honeypot* se encontra, é que aqueles localizados nas redes com maior qualidade possuem maior número de campanhas, enquanto os três identificados anteriormente como estando em redes de baixa qualidade (BR-01, EC-01 e UY-01) possuem um número muito menor de campanhas distintas. As campanhas observadas em alguns momentos têm impacto direto sobre o perfil de tráfego observado, à medida que essas se iniciam ou terminam,



(a) BR-02 (b) NL-01

Figura 4.23. Distribuição das campanhas ativas ao longo do tempo.

como será mostrado nas seções a seguir.

Tabela 4.14. Percentual de campanhas em comum entre os *honeypots*.

	AT-01	**AU-01**	**BR-01**	**BR-02**	**EC-01**	**NL-01**	**TW-01**	**UY-01**	**Total**
AT-01	-	4,14%	3,54%	43,55%	0,80%	80,52%	2,38%	4,03%	36.701
AU-01	32,79%	-	60,42%	29,08%	21,62%	16,43%	11,02%	23,67%	4.639
BR-01	32,06%	69,23%	-	24,33%	22,62%	15,98%	12,67%	23,46%	4.049
BR-02	39,10%	3,30%	2,41%	-	2,96%	80,68%	2,42%	2,37%	40.876
EC-01	7,10%	24,14%	22,05%	29,15%	-	9,24%	6,40%	6,64%	4.155
NL-01	40,28%	1,04%	0,88%	44,95%	0,52%	-	1,92%	1,83%	73.358
TW-01	0,90%	0,53%	0,53%	1,02%	0,27%	1,45%	-	15,59%	96.999
UY-01	6,33%	4,70%	4,07%	4,14%	1,18%	5,74%	64,71%	-	23.368

4.8.1 Aumento do número de endereços IP utilizando SOCKS

Nos gráficos referentes aos transmissores encontrados nos *honeypots* AT-01 e BR-02, (figs 4.13(a) e 4.13(d)), são perceptíveis períodos em que há um aumento no número de endereços IP que enviam mensagens utilizando SOCKS. Nesses períodos é visível um degrau na curva, que sobe no início do intervalo e volta ao seu valor mais baixo após alguns dias.

Para tentar explicar esse comportamento, focamos aqui no primeiro aumento ocorrido no *honeypot* BR-02, por volta do dia 23/05. Identificamos as principais campanhas do período e os endereços das máquinas que dela participavam. O resultado pode ser visto na tabela 4.15. Inicialmente, aquelas campanhas são praticamente inexistentes, com um número muito pequeno de transmissores; no dia em que ocorre o aumento do tráfego elas se passam a conter um grande número de endereços. Há claramente uma relação com o surgimento dessas novas campanhas, a participação de novos transmissores.

Tabela 4.15. Número de endereços IP das 10 maiores campanhas do dia 23/05 ao dia 31/05.

	22/05	**23/05**	**...**	**31/05**	**01/06**
Campanha 1	2	182	...	175	8
Campanha 2	2	170	...	174	9
Campanha 3	-	137	...	128	-
Campanha 4	-	137	...	130	-
Campanha 5	-	137	...	127	-
Campanha 6	-	137	...	129	-
Campanha 7	-	137	...	129	-
Campanha 8	-	137	...	129	-
Campanha 9	-	137	...	129	-
Campanha 10	-	137	...	127	-

4.8.2 Queda do número de transmissores utilizando SMTP

Como pode ser visto nas figuras 4.13(d), e 4.13(e) (página 4.13(d), durante o período de 26/05 e 01/06, houve uma queda visível no número de endereços IP que enviaram mensagens utilizando o protocolo SMTP em alguns *honeypots*. Para entender o motivo da queda, observamos as principais campanhas envolvidas nesse período. A tabela 4.16 contém as maiores campanhas no dia 26/05, dia em que se iniciou a queda, e nos dias seguintes. Como podemos observar, essas campanhas foram perdendo força, algumas chegando até a serem encerradas. Isso sugere que a queda se deveu ao término dessas campanhas.

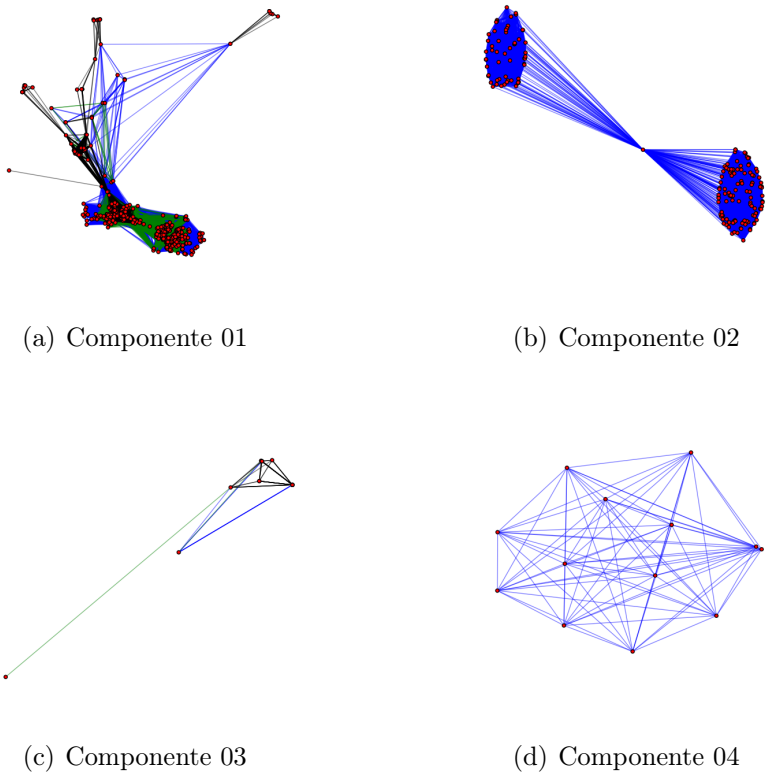
Tabela 4.16. Número de endereços IP das 10 maiores campanhas do dia 26/05.

	26/05	**27/05**	**28/05**	**29/05**	**30/05**	**31/05**	**01/06**
Campanha 1	469	476	29	31	20	9	4
Campanha 2	256	180	0	0	0	0	0
Campanha 3	237	217	123	63	1	28	6
Campanha 4	236	11	0	0	0	0	0
Campanha 5	220	144	0	0	0	0	0
Campanha 6	215	325	0	0	0	0	0
Campanha 7	200	94	32	0	0	0	0
Campanha 8	178	112	0	0	0	0	0
Campanha 9	140	1	0	0	0	0	0
Campanha 10	113	89	0	0	0	0	0

Outras análises semelhantes foram realizadas para outros pontos onde observamos mudanças bruscas no perfil do tráfego. Com base nessas análises, podemos concluir que eventos externos, como o início ou término de uma campanha, têm forte impacto sobre os detalhes do perfil de tráfego observado, devendo ser considerado em análises de tráfego desse tipo.

4.9 Identificação de grupos correlacionados através de campanhas

Através da análise das campanhas, detectamos a influência dessas na geração de tráfego de *spam* nos *honeypots*, bem como no surgimento de novos endereços IP abusando dessas máquinas. Com base nas observações descritas na seção anterior, desenvolvemos um método para, a partir das campanhas, conseguir identificar grupos de transmissores que sejam correlacionados. Para esse método, geramos grafos onde os vértices representam os endereços IP e as arestas entre dois vértices representam as campanhas em que os dois IP's participam juntas. Após o grafo gerado, aplicamos um algoritmo para obter os componentes conexos do grafo. Consideramos cada componente conexo como sendo um grupo distinto. A intuição dessa análise está no fato de que, como diferentes endereços IP compartilham campanhas, eles tendem a estar em um mesmo grupo de disseminação de *spam*, ou até mesmo fazerem parte de uma mesma *botnet*.



(a) Componente 01 (b) Componente 02

(c) Componente 03 (d) Componente 04

Figura 4.24. Grupos distintos encontrados no dia 09/05, para o *honeypot* AT-01.

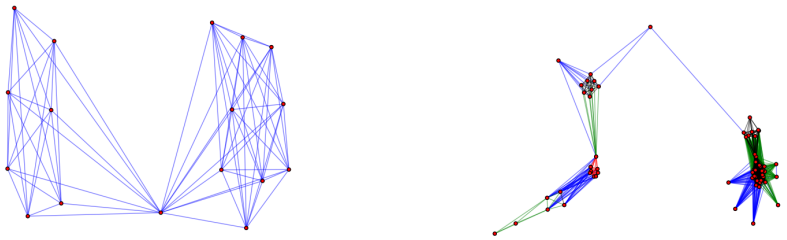
Como o número de campanhas em comum entre os endereços IP varia bastante, sendo que alguns possuem apenas uma campanha em comum, enquanto outros apre-

sentam mais de 100 campanhas, utilizamos cores para as arestas, de forma a indicar a força de relação entre cada par de atacante. Arestas azuis indicam que os endereços possuem apenas uma campanha em comum, verdes indicam de 2 a 10 campanhas, pretas são as que tem entre 11 e 50, enquanto as vermelhas são aquelas com mais de 50 campanhas de *spam* em comum.

A implementação de método, tanto para geração do grafo, quanto para obtenção dos componentes conexos, foi feita através da linguagem Python, utilizando a biblioteca NetworkX[2]. O posicionamento dos nós foi feito através do algoritmo Fruchterman-Reingold. A idéia do posicionamento é considerar a força entre dois nós. A geração foi feita para cada dia, em cada um dos *honeypots*.

Tabela 4.17. Campanhas que compõem o endereço IP 1.164.109.164 no dia 09/05, no AT-01.

Campanha	**Data/Hora**
Campanha 1	09/05 - 09:37:08
Campanha 1	09/05 - 09:37:29
Campanha 1	09/05 - 09:37:44
Campanha 2	09/05 - 09:47:45
Campanha 2	09/05 - 09:48:19
Campanha 2	09/05 - 09:49:05



(a) NL-01 - 27-05 (b) TW-01 - 23-05

Figura 4.25. Componentes com IP entre dois grupos distintos.

A figura 4.24 mostra alguns dos componentes obtidos para o dia 09/05, utilizando os dados provenientes do *honeypot* AT-01. Esse dia foi escolhido por apresentar os diferentes tipos de componentes encontrados na análise. Nele, foram gerados 9 grupos distintos. Na figura mostramos os quatro com maior número de endereços IP. Como pode ser visto, os grupos apresentam características distintas. O componente 1 (figura 4.24(a)) apresentou um número elevado de atacantes (720), e as características

[2]http://networkx.lanl.gov

apresentadas entre eles foram variadas. Enquanto alguns pares de endereços IP possuem apenas uma campanha em comum, outros têm entre 2 e 10 e alguns chegam a ter até 50 campanhas em comum, o que indica que esse componente pode ser subdividido em mais grupos. Assim como esse, o componente 3 (figura 4.24(c)) também apresenta características variadas quanto ao número de campanhas entre os endereços IP. No segundo componente, todos os pares de endereços apresentaram apenas uma campanha em comum. Dessa forma, podemos concluir que esse grupo é conciso, com comportamento estável. Porém, conforme mostrado na figura 4.24(b), os endereços que o compõem poderiam formar dois grupos distintos, dado que apenas um transmissor possui relacionamento com esses dois subgrupos. Esse problema foi recorrente em vários dias analisados e será melhor explicado a seguir. Por fim, o componente 4 apresenta um clique. Esse grupo é bastante estável, em que todos os endereços estão interligados e apresentam uma campanha em comum.

Tabela 4.18. Campanhas que compõem o endereço IP 222.186.43.206 no dia 27/05, no NL-01.

Campanha	**Data/Hora**
Campanha 1	27/05 - 00:00:01
⋮	⋮
Campanha 1	27/05 - 13:55:45
Campanha 2	27/05 - 14:12:29
⋮	⋮
Campanha 2	27/05 - 23:59:59

O problema de dois subgrupos, aparentemente distintos, que se interligam por meio de apenas um endereço IP ocorreu com frequência na análise. Na tentativa de explicar esse fato, observamos o comportamento das campanhas relacionadas a esses endereços. Para o primeiro exemplo, visto na figura 4.24(b), o transmissor recebe mensagens pertencentes a campanhas distintas em diferentes momentos. A tabela 4.17 mostra o comportamento das campanhas. Como pode ser visto, inicialmente são recebidas mensagens classificadas como sendo da campanha 1 e, após algum tempo, são recebidas somente mensagens pertencentes a outra campanha. Esse fato leva a crer que o endereço IP em questão estava alocado a uma máquina e no intervalo de tempo entre ass campanha, esse IP teria sido reassinalado para outro usuário. Portanto, para que os endereços IP fossem melhor agrupados, deveria-se considerar como sendo dois atacantes distintos, utilizando um mesmo endereço em períodos distintos. Diversos casos analisados apresentam o mesmo padrão, muitas vezes com intervalos de horas entre as campanhas. Outro exemplo desse padrão é mostrado na tabela 4.18, referente

ao componente mostrado na figura 4.25(a).

Tabela 4.19. Campanhas que compõem o endereço IP 173.45.94.35 no dia 23/05, no TW-01.

Campanha	**Data/Hora**
Campanha 1	23-05 - 00:43:35
Campanha 2	23-05 - 00:43:41
Campanha 1	23-05 - 00:43:49
Campanha 2	23-05 - 00:43:56
⋮	⋮
Campanha 2	23-05 - 00:44:15
Campanha 1	23-05 - 00:44:18
Campanha 2	23-05 - 00:44:19
⋮	⋮
Campanha 2	23-05 - 00:44:33
Campanha 1	23-05 - 00:44:36
Campanha 2	23-05 - 00:44:39
⋮	⋮
Campanha 2	23-05 - 00:44:55
Campanha 1	23-05 - 00:44:57

Por outro lado, outros grupos apresentam comportamentos mais complexos, como o da figura 4.25(b). A tabela 4.19 mostra que as campanhas do qual o endereço IP participa são simultâneas, ou seja, o transmissor enviou mensagens de *spam* participando de duas campanhas distintas de forma alternada ao longo do tempo. Nesse caso, esse atacante possui endereço IP estático, proveniente de um *hosting*, e participa das duas campanhas simultaneamente, o que explica tal comportamento. Outros casos apresentaram o mesmo problema, podendo ser explicados como esse, com o atacante participando de campanhas concomitantes, ou por conta do uso de NAT. Esses comportamentos podem explicar o problema apresentado nos componentes 1 e 2, mostrados na figura 4.24.

Capítulo 5

Conclusões

O combate ao *spam* é uma tarefa contínua, onde se busca sempre entender melhor a evolução dos *spammers* e suas técnicas de disseminação de mensagens. Este trabalho buscou estender o entendimento sobre padrões de comportamento usado por esses transmissores para enviar seu tráfego mantendo-se ocultos atrás de máquinas intermediárias. Para isso, utilizamos uma análise do tráfego de *spam* no nível da rede.

Diferentemente de outros trabalhos anteriores na área, utilizamos um conjunto de máquinas geograficamente dispersas pelo mundo, a fim de coletar tráfego de *spam* em diferentes pontos da Internet, além de visualizar o tráfego sob o ponto de vista intermediário da rede. Isso nos permitiu observar diversas características do tráfego e o impacto de sua localização. Além disso, fomos capazes de identificar outros fatores que impactam no tráfego de *spam* recebido pelos coletores, como a conectividade da máquina e as campanhas que a incidem. E ainda, pudemos verificar a influência do tempo na variação desse tráfego.

Dessa forma, acreditamos que nossas observações serão úteis para outros pesquisadores que busquem mais informações sobre a origem do *spam*, a fim de viabilizar novas pesquisas na área.

5.1 Principais resultados

Os resultados mais importantes encontrados nesta dissertação são listados abaixo:

- Algumas características do tráfego, como endereços de origem e, principalmente, o uso de mensagens de teste, se repetem em diferentes locais de coleta.
- O tipo do tráfego instantâneo, além de sua intensidade, são afetados pelo padrão de ocorrência das campanhas de *spam*. Por exemplo, alguns picos observados no

número de endereços IP utilizando SOCKS, além de quedas na quantidade de transmissores que usam SMTP são decorrentes do aparecimento e encerramento da atividade de determinadas campanhas.

- Transmissores que fazem uso de conexões SMTP enviam poucas mensagens e têm comportamento dependente das horas do dia, indicando serem pertencentes a *botnets*.

- Endereços IP que utilizam SOCKS e HTTP geram grandes volumes de tráfego, enviando muitas mensagens durante todo o dia, sugerindo que sejam infraestruturas dedicadas usadas para disseminação de *spam*.

- Atacantes que fazem uso do protocolo SOCKS provavelmente vasculham a rede a procura de *proxies* abertos e, assim que encontram, já iniciam seus ataques.

- Atacantes que enviam *spam* através de SMTP só começam a enviar mensagens após confirmação de que essa máquina realmente repassa suas mensagens, realizada através das mensagens de teste

- A distribuição das mensagens de teste possui forte correlação com o tráfego de *spam* enviado através do protocolo SMTP.

- Forte influência da qualidade da rede em que se encontra a máquina sendo abusada no tipo de *spam* a incidi-la e também na intensidade com o qual eles são enviados a ela. Dessa forma, verificamos que a conectividade das máquinas atacadas possui maior influência no tráfego de *spam* recebido do que fatores de localização.

- A compreensão do tráfego de *spam* ao longo do tempo e o entendimento de que o tráfego dessas mensagens varia muito com relação ao aspecto temporal. Nesta dissertação avaliamos 176 dias, e nos resultados é visível que o comportamento é muito variável, tornando-se difícil inferir padrões de caracterização regulares ao longo do tempo.

5.2 Trabalhos futuros

Como trabalhos futuros, pretendemos avançar nas análises com a observação do conteúdo específico das mensagens, para melhor entender a relação com localidades específicas, bem como realizar uma análise mais profunda dos padrões de tráfego de rede para confirmar as observações sobre o impacto da qualidade da conexão da rede atacada ao

restante da Internet. Desejamos analisar os sistemas autônomos de origem das mensagens recebidas, visando fortalecer o entendimento das características encontradas em nossas análises do tráfego de *spam*. Além disso, pretendemos ainda verificar o impacto da configuração das máquinas, por fatores como processamento e memória, no tráfego de *spam* recebido por elas.

Pretendemos também sistematizar o cruzamento das informações de cada um dos atacantes com dados relativos às campanhas e às características de rede, visando definir padrões recorrentes no tráfego ao longo do tempo. Além disso, desejamos também aprofundar no trabalho de identificação de *botnets* a partir dos grafos de campanhas.

Referências Bibliográficas

Beverly, R. & Sollins, K. (2008). Exploiting transport-level characteristics of spam. Em *Conference on Email and Anti-Spam*.

CERT.br (2013). SpamPots Project. `http://honeytarg.cert.br/spampots`. Acessado em: 19 de março de 2013.

Dhinakaran, C. & Lee, J. K. (2007). Characterizing Spam traffic and Spammers. Em *Proc. Int'l Conference on Convergence Information Technology*, Gyeongju, Coréia do Sul.

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

Gomes, L. H.; Cazita, C.; Almeida, J. M.; Almeida, V. & Jr., W. M. (2007). Workload Models of Spam and Legitimate E-mails. *Performance Evaluation*, 64(7-8):690–714.

Goodman, J.; Cormack, G. V. & Heckerman, D. (2007). Spam and the ongoing battle for the inbox. *Commun. ACM*, 50:24--33.

Guerra, P. H. C.; Guedes, D.; Meira Jr., W.; Hoepers, C. & Steding-Jessen, K. (2008a). Caracterização de estratégias de disseminação de spams. Em *Anais do SBRC 2008*, Rio de Janeiro, Brasil.

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

Guerra, P. H. C.; Pires, D. E. V.; Guedes, D.; Jr., W. M.; Hoepers, C.; Steding-Jessen, K. & Chaves, M. (2009). Caracterização do encadeamento de conexões para envio de spams. Em *Anais do SBRC 2009*, Recife, Brasil.

Guerra, P. H. C.; Ribeiro, M. T.; Guedes, D.; Jr., W. M.; Hoepers, C.; Steding-Jessen, K. & Chaves, M. H. (2010). Identificação e caracterização de spammers a partir de listas de destinatários. Em *Anais do SBRC 2010*, Gramado, RS.

Hao, S.; Syed, N. A.; Feamster, N.; Gray, A. & Krasser, S. (2009). Detecting Spammers with SNARE: Spatio-temporal Network-level Automatic Reputation Engine. Em *Proc. Usenix Security*.

John, J.; Moshchuk, A.; Gribble, S. D. & Krishnamurthy, A. (2009). Studying Spamming Botnets Using Botlab. Em *6th USENIX Symp. on Networked Systems Design and Implementation*, Boston, EUA.

Kim, J. & Choi, H. (2008). Spam Traffic Characterization. Em *Int'l Technical Conference on Circuits/Systems, Computers and Communications*, Shimonoseki City, Japão.

Kokkodis, M. & Faloutsos, M. (2009). Spamming botnets: Are we losing the war? ? Em *Proceedings of the 6th Conference on e-mail and anti-spam (CEAS)*.

Las-Casas, P. H. B.; Almeida, J. M.; Gonçalves, M. A.; Guedes, D.; Ziviani, A. & Marques-Neto, H. T. (2012a). Impacto da Evolução Temporal na Detecção de Spammers na Rede de Origem. Em *Anais do SBRC 2012*, Ouro Preto, Brasil.

Las-Casas, P. H. B.; Guedes, D.; Almeida, J. M.; Ziviani, A. & Marques-Neto, H. T. (2012b). SpaDeS: Detecting Spammers at the Source Network. *Computer Networks*.

Newman, M. E. J.; Forrest, S. & Balthrop, J. (2002). Email Networks and the Spread of Computer Viruses. *Physical Review E*, 66(3):035101.

Ouyang, T.; Ray, S.; Rabinovich, M. & Allman, M. (2011). Can network characteristics detect spam effectively in a stand-alone enterprise? Em *Proc. 12th Passive and Active Measurement Conference*.

Pires, D.; Totti, L.; Moreira, R.; Fazzion, E.; Fonseca, O.; Meira Jr., W.; Minardi, R. & Guedes, D. (2012). Fpcluster: an efficient out-of-core clustering strategy without a similarity metric. *Journal of Information and Data Management*, 3(2):132–141.

Provos, N. & Holz, T. (2007). *Virtual honeypots: from botnet tracking to intrusion detection*. Addison-Wesley Professional, first edição.

Pu, C. (2006). Observed trends in spam construction techniques: A case study of spam evolution. Em *In Third Conference on Email and Anti-Spam (CEAS) (2006*.

Ramachandran, A. & Feamster, N. (2006). Understanding the Network-Level Behavior of Spammers. *SIGCOMM Computer Communication Review*, 36(4):291--302.

Rao, J. M. & Reiley, D. H. (2012). The economics of spam. Em *Journal of Economic Perspectives*, Forthcoming Summer.

Richard Clayton (2006). spamHINTS: Happily It's Not The Same. Online. http://www.spamhints.org/.

Sanchez, F.; Duan, Z. & Dong, Y. (2011). Blocking spam by separating end-user machines from legitimate mail server machines. Em *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pp. 116--124, New York, NY, USA. ACM.

Schatzmann, D.; Burkhart, M. & Spyropoulos, T. (2009). Inferring Spammers in the Network Core. Em *Proc. 10th Int'l Conf. on Passive and Active Network Measurement*.

Silva, G. S.; Steding-Jessen, K.; Hoepers, C.; Chaves, M. H. P.; Meira Jr., W. & Guedes, D. (2011). Fatores que afetam o comportamento de spammers na rede. Em *SBSEG 2011*, Brasília, Brasil.

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

Sperotto, A.; Vliek, G.; Sadre, R. & Pras, A. (2009). Detecting Spam at the Network Level. Em *15th Open European Summer School and IFIP TC6.6 Workshop on The Internet of the Future*, Barcelona, Spain.

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

Stone-Gross, B.; Holz, T.; Stringhini, G. & Vigna, G. (2011). The underground economy of spam: a botmaster's perspective of coordinating large-scale spam campaigns. Em *Proceedings of the 4th USENIX conference on Large-scale exploits and emergent threats*, LEET'11, pp. 4--4, Berkeley, CA, USA. USENIX Association.

Symantec (2011). Internet Security Threat Report, Volume 17. `http://www.symantec.com/threatreport`. Acessado em: 19 de março de 2013.

Taveira, D. & Duarte, O. (2008). A Monitor Tool for Anti-Spam Mechanisms and Spammers Behavior. Em *IEEE Network Operations and Management Symposium Workshops*, Salvador, Bahia.

Totti, L. C.; Moreira, R. E. A.; Fazzion, E.; Fonseca, O.; Meira Jr., W.; Guedes, D.; Hoepers, C.; Steding-Jessen, K. & Chaves, M. H. P. (2012). Caracterização Temporal de Estratégias de Disseminação de Spam. Em *Anais do SBRC 2012*, Ouro Preto, Brasil.

Venkataraman, S.; Sen, S.; Spatscheck, O.; Haffner, P. & Song, D. (2007). Exploiting network structure for proactive spam mitigation. Em *Proc. 16th USENIX Security Symposium*.

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

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

Zhuang, L.; Dunagan, J.; Simon, D. R.; Wang, H. J. & Tygar, J. D. (2008). Characterizing botnets from email spam records. Em *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, LEET'08, pp. 2:1--2:9, Berkeley, CA, USA. USENIX Association.

Apêndice A

Visão Geral

Tabela A.1. Visão geral dos dados coletados em todos os *honeypots*.

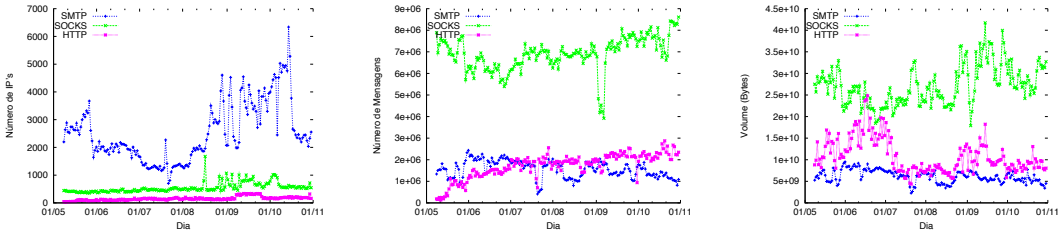
	SMTP(%)	**SOCKS(%)**	**HTTP(%)**	**Total**
Mensagens (milhões)	274,52 (15,32%)	1214,74 (67,78%)	303,02 (16,90%)	1.792,29
Endereços IP	161.398 (89,53%)	18.801 (10,43%)	3.692 (2,05%)	180.262
Prefixos de rede	13.574 (93,22%)	1.343 (9,22%)	168 (1,25%)	14.561
Sistemas Autônomos (AS)	2.355 (95,73%)	330 (13,41%)	45 (1,83%)	2.460
***Country Codes* (CC)**	138 (97,87%)	65 (46,09%)	13 (9,21%)	141
Volume de tráfego (TB)	0,98 (13,97%)	4,35 (61,59%)	1,72 (24,44%)	7,06
Msgs/endereço (milhares/IP)	1,70	64,61	82,07	9,94
Volume/endereço (MB/IP)	6,40	242,61	490,16	41,08
Volume/msg (KB/msg)	3,85	3,84	6,11	4,23

Tabela A.2. 10 Country Codes que mais enviaram mensagens em todos os *honeypots*.

	Mensagens	**Endereços IP**	**Prefixos's**	**ASes**
US	1.087.783.265 (60,69%)	2.365 (1,31%)	1.185 (8,14%)	388 (15,77%)
PH	230.230.162 (12,85%)	171 (0,09%)	45 (0,31%)	9 (0,37%)
CN	91.543.044 (5,11%)	76.197 (42,27%)	5.456 (37,47%)	76 (3,09%)
TW	83.561.691 (4,66%)	64.842 (35,97%)	239 (1,64%)	29 (1,18%)
BR	66.234.708 (3,70%)	4.203 (2,33%)	1.155 (7,93%)	151 (6,14%)
JP	39.139.575 (2,18%)	217 (0,12%)	82 (0,56%)	33 (1,34%)
RU	20.622.747 (1,15%)	4.196 (2,32%)	1.004 (6,89%)	401 (16,30%)
KR	13.195.943 (0,74%)	346 (0,19%)	210 (1,44%)	50 (2,03%)
IN	11.947.247 (0,67%)	17.667 (9,80%)	1.562 (10,73%)	77 (3,13%)
IT	8.115.800 (0,45%)	220 (0,12%)	109 (0,75%)	26 (1,06%)

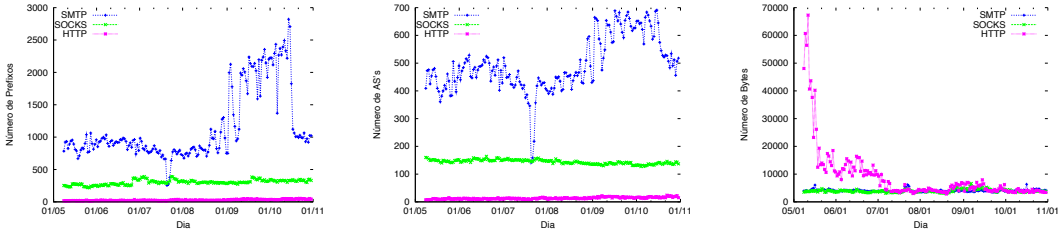
Tabela A.3. 10 Sistemas Autônomos que mais enviaram mensagens em todos os *honeypots*.

	Mensagens	Endereços IP	Prefixos	Volume (GB)	CC	Tipo
10297	896.990.985 (50,05%)	159 (0,09%)	4 (0,03%)	2.714,84 (37,54%)	US	Host/Cloud
9299	147.271.014 (8,22%)	74 (0,04%)	19 (0,13%)	340,70 (4,71%)	PH	Misto
29802	143.452.043 (8,00%)	23 (0,01%)	2 (0,01%)	397,46 (5,50%)	US	Host
6648	78.075.302 (4,36%)	50 (0,03%)	10 (0,06%)	178,45 (2,45%)	PH	DSL/Bus.
3462	70.511.218 (3,93%)	61.598 (34,17%)	108 (0,74%)	275,70 (3,81%)	TW	DSL
4134	56.936.786 (3,18%)	53.972 (29,94%)	4.256 (29,23%)	1.858,41 (25,70%)	CN	DSL
2497	15.674.950 (0,87%)	26 (0,01%)	8 (0,05%)	73,80 (1,02%)	JP	Cloud
4713	12.916.038 (0,72%)	25 (0,01%)	12 (0,08%)	28,98 (0,40%)	JP	Cloud
27699	11.238.638 (0,63%)	1.191 (0,66%)	49 (0,33%)	39,60 (0,55%)	BR	DSL/Bus.
23650	10.707.705 (0,60%)	38 (0,02%)	27 (0,19%)	113,93 (1,58%)	CN	?



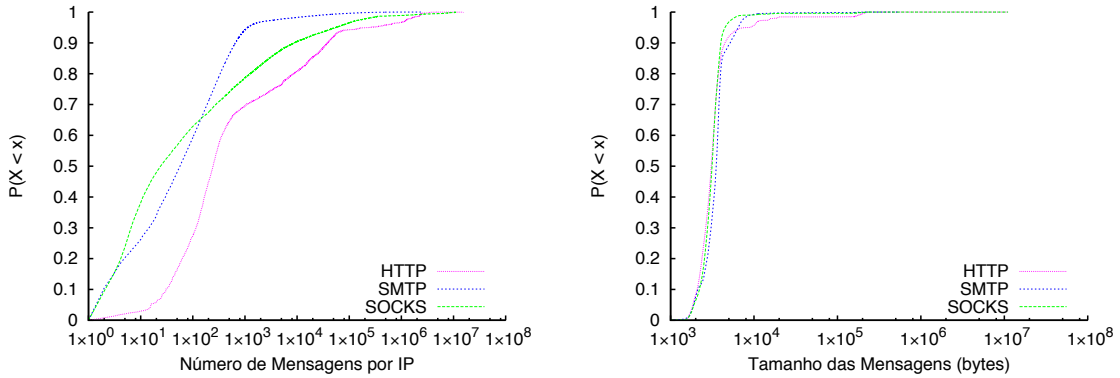
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura A.1. Séries temporais dos endereços, mensagens e volume por protocolo para os dados agregados de todos os *honeypots*.



(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura A.2. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo para os dados agregados de todos os *honeypots*.



(a) Número de mensagens por endereço de origem (b) Tamanho das mensagens

Figura A.3. Distribuições acumuladas para mensagens por protocolo para os dados agregados de todos os *honeypots*.

Apêndice B

Visão por *honeypot*

B.1 *Honeypot* AT-01

Tabela B.1. Visão geral dos dados coletados no *honeypot* AT-01.

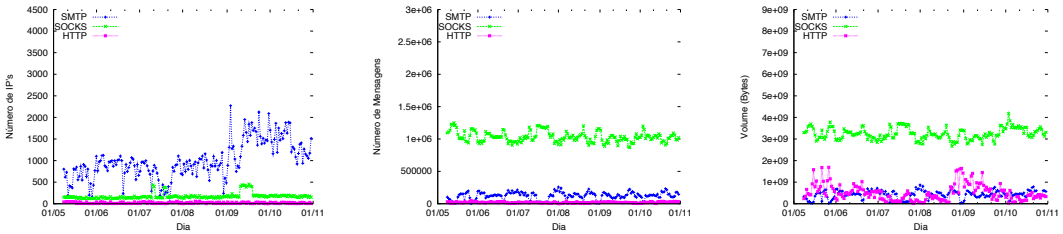
	SMTP	**SOCKS**	**HTTP**	**Total**
Mensagens (milhões)	21,64 (10,46%)	181,87 (87,90%)	3,40 (1,64%)	206,91
Endereços IP	55.329 (96,22%)	2.196 (3,82%)	289 (0,49%)	57.499
Prefixos de rede	7.659 (95,59%)	473 (5,90%)	34 (0,42%)	8.012
Sistemas Autônomos (ASes)	1.849 (96,40%)	184 (9,59%)	11 (0,57%)	1.918
***Country Codes* (CC)**	128 (98,46%)	49 (37,69%)	6 (4,62%)	130
Volume de bytes (GB)	69,57 (10,17%)	532,31 (77,83%)	82,09 (12,00%)	683,98

Tabela B.2. 10 Country Codes que mais enviaram mensagens no *honeypot* AT-01.

	Mensagens	**IP's**	**Prefixos**	**ASes**	**Bytes (GB)**
US	149.782.214	1.252	614	277	452,67
PH	29.544.970	146	28	7	67,93
CN	5.142.403	9.684	2.511	64	87,28
JP	3.967.030	175	64	29	15,87
BR	2.688.847	1.454	679	110	8,44
TW	2.392.758	34.590	129	22	9,36
RU	1.791.447	1.652	207	125	5,52
KR	772.401	207	125	29	2,36
IN	721.514	4.352	928	73	2,29
HK	697.077	73	63	27	2,68

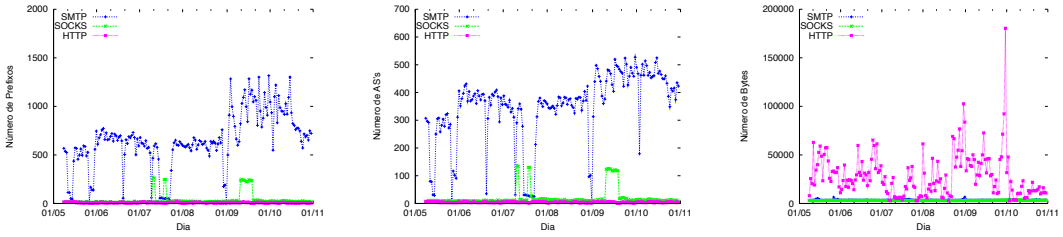
Tabela B.3. 10 Sistemas Autônomos que mais enviaram mensagens no *honeypot* AT-01.

	Mensagens	IP's	Prefixos	Country Code	Bytes (GB)
10297	128.483.909	151	4	US	391,15
9299	19.457.751	66	14	PH	44,57
29802	17.814.727	22	2	US	50,96
6648	9.410.775	49	3	PH	21,63
4134	2.987.827	5.781	1.817	CN	76,79
3462	2.239.445	34.504	71	TW	8,89
2497	1.539.966	25	7	JP	7,22
4713	1.156.424	15	9	JP	2,43
4725	1.052.679	23	7	JP	5,37
21788	886.290	42	26	US	2,61



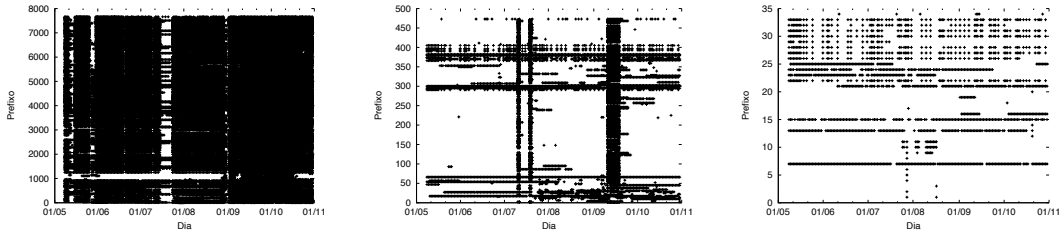
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura B.1. Séries temporais dos endereços, mensagens e volume por protocolo do *honeypot* AT-01.



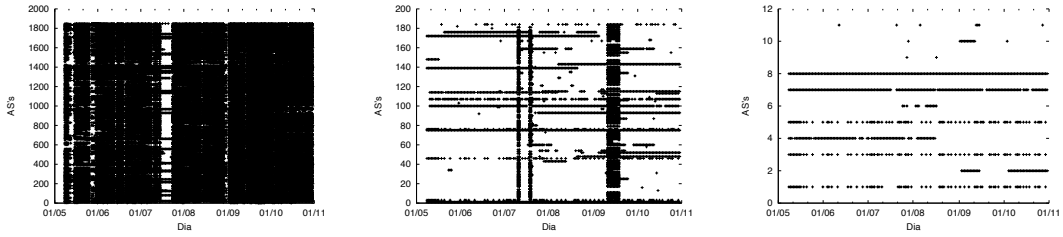
(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura B.2. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo do *honeypot* AT-01.



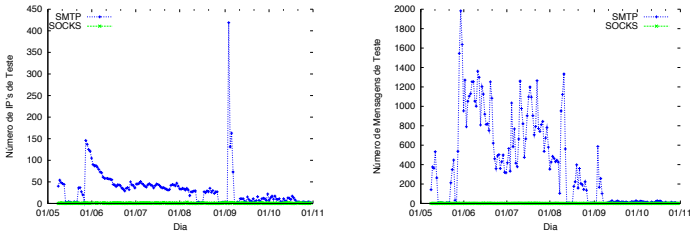
(a) Prefixos SMTP (b) Prefixos SOCKS (c) Prefixos HTTP

Figura B.3. Distribuição dos prefixos ao longo do tempo do *honeypot* AT-01.



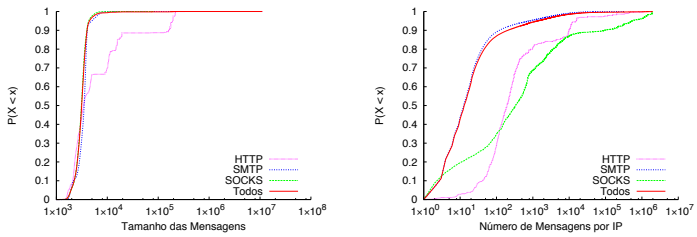
(a) ASes SMTP (b) ASes SOCKS (c) ASes HTTP

Figura B.4. Distribuição dos ASes ao longo do tempo do *honeypot* AT-01.



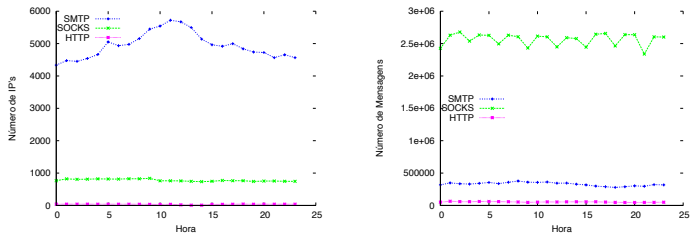
(a) IP's que enviam mensagens de teste (b) Mensagens de teste

Figura B.5. Características da mensagens de teste do *honeypot* AT-01.



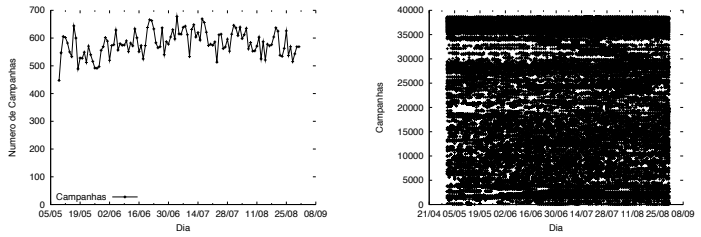
(a) CDF do Tamanho das Mensagens (b) CDF das Mensagens por IP

Figura B.6. Características das CDF's do *honeypot* AT-01.



(a) Número de Endereços IP por hora do dia (b) Mensagens por hora do dia

Figura B.7. Características do *honeypot* AT-01 por hora do dia.



(a) Campanhas por Dia (b) Distribuição das campanhas ativas

Figura B.8. Características das campanhas do *honeypot* AT-01.

B.2 *Honeypot* AU-01

Tabela B.4. Visão geral dos dados coletados no *honeypot* AU-01.

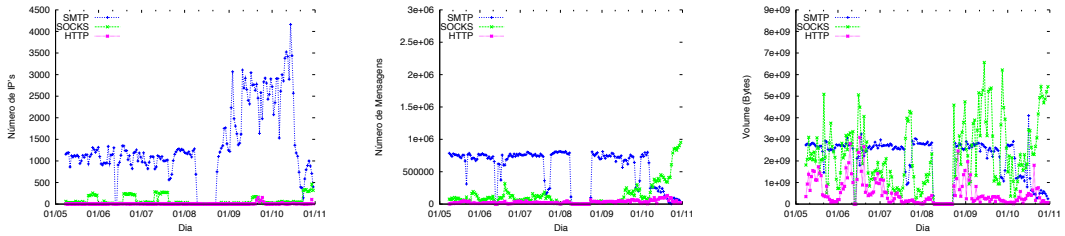
	SMTP	SOCKS	HTTP	Total
Número de Mensagens	97.726.348 (76,87%)	24.163.928 (19,00%)	5.247.579 (4,13%)	127.137.855
Número de IP's	97.676 (97,14%)	2.694(2,68%)	720 (0,72%)	100.547
Número de Prefixos	10.410 (95,12%)	706 (6,45%)	38 (0,35%)	10.944
Número de ASes	2.017 (96,19%)	235 (11,20%)	10 (0,48%)	2.097
Número de CC's	134 (97,81%)	56 (40,88%)	6 (4,38%)	137
Volume de bytes (GB)	342,12 (43,71%)	356,53 (45,55%)	84,06 (10,74%)	782,72

Tabela B.5. 10 Country Codes que mais enviaram mensagens no *honeypot* AU-01.

	Mensagens	IP's	Prefixos	ASes	Bytes (GB)
CN	27.641.751	30.753	4.207	68	436,95
TW	15.264.056	43.259	150	24	62,46
BR	14.505.963	1.737	842	131	49,37
US	9.176.014	1.644	843	331	32,53
RU	7.087.531	3.331	795	341	23,52
KR	4.155.052	267	154	48	13,30
IN	3.832.866	13.306	1.237	74	12,90
HK	2.719.358	87	71	27	10,26
UA	2.524.909	429	224	104	8,51
GB	2.419.285	162	115	54	7,70

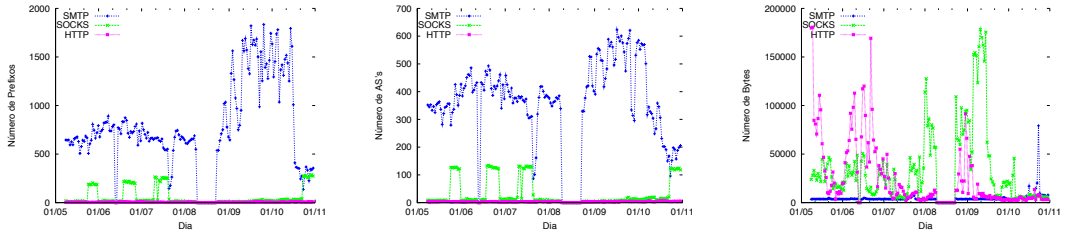
Tabela B.6. 10 Sistemas Autônomos que mais enviaram mensagens no *honeypot* AU-01.

	Mensagens	IP's	Prefixos	Country Code	Bytes (GB)
4134	16.069.992	16.467	3.290	CN	377,34
3462	10.049.868	37.346	71	TW	45,95
4837	3.511.306	7.019	222	CN	12,33
23650	2.744.672	22	17	CN	29,27
28573	2.646.042	318	186	BR	8,64
18881	2.288.589	321	142	BR	7,94
27699	2.094.162	120	40	BR	7,22
38186	1.930.712	2	2	HK	7,73
24164	1.789.730	6	5	TW	5,18
9924	1.717.984	933	7	TW	6,01



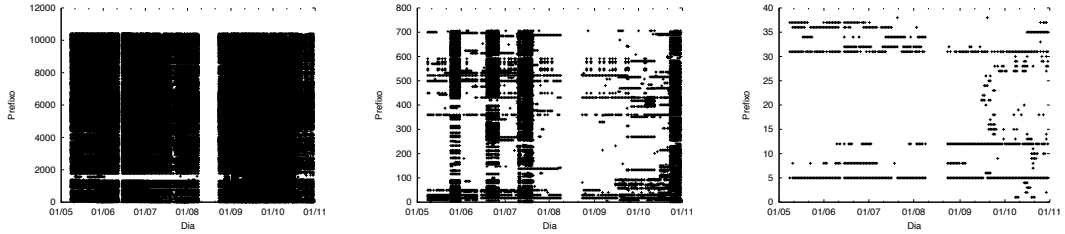
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura B.9. Séries temporais dos endereços, mensagens e volume por protocolo do *honeypot* AU-01,



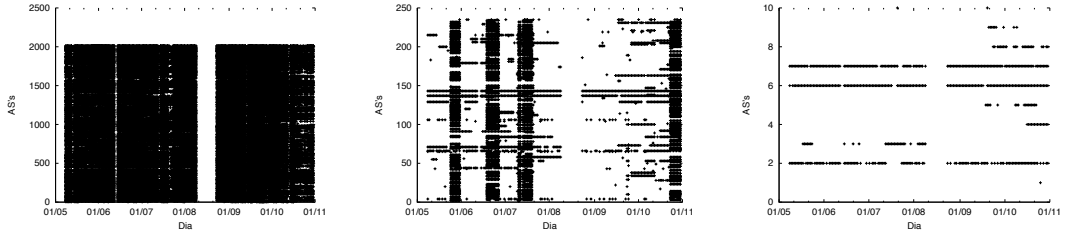
(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura B.10. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo do *honeypot* AU-01.



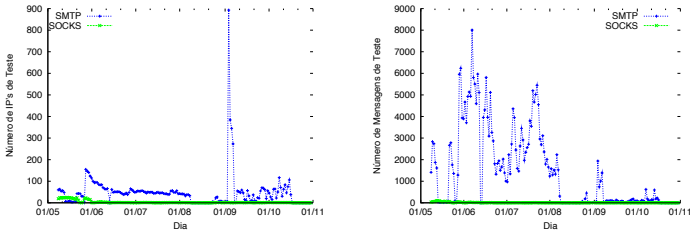
(a) Prefixos SMTP (b) Prefixos SOCKS (c) Prefixos HTTP

Figura B.11. Distribuição dos prefixos ao longo do tempo do *honeypot* AU-01.



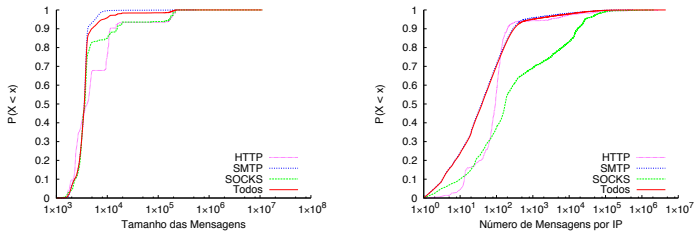
(a) ASes SMTP (b) ASes SOCKS (c) ASes HTTP

Figura B.12. Distribuição dos ASes ao longo do tempo do *honeypot* AU-01.



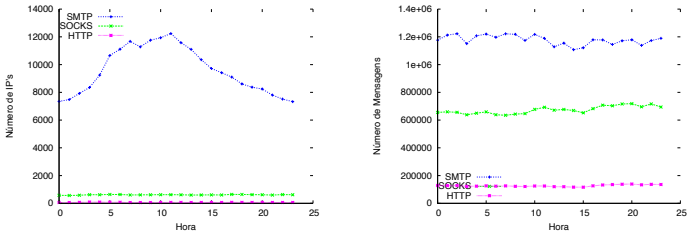
(a) IP's que enviam mensagens de teste (b) Mensagens de teste

Figura B.13. Características da mensagens de teste do *honeypot* AU-01.



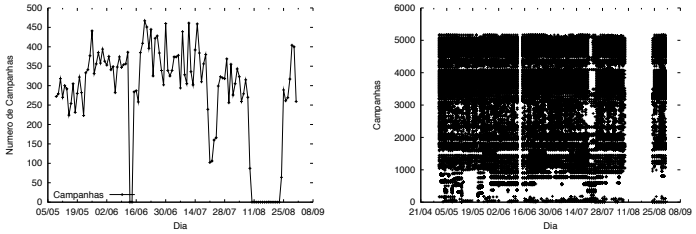
(a) CDF do Tamanho das Mensagens (b) CDF das Mensagens por IP

Figura B.14. Características das CDF's do *honeypot* AU-01.



(a) Número de Endereços IP por hora do dia (b) Mensagens por hora do dia

Figura B.15. Características do *honeypot* AU-01 por hora do dia



(a) Campanhas por Dia (b) Distribuição das campanhas ativas

Figura B.16. Características das campanhas do *honeypot* AU-01.

B.3 *Honeypot* BR-01

Tabela B.7. Visão geral dos dados coletados no *honeypot* BR-01.

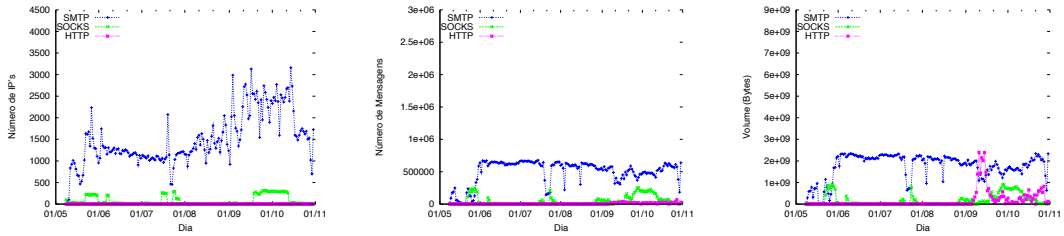
	SMTP	SOCKS	HTTP	Total
Número de Mensagens	87.654.299 (90,14%)	8.519.581 (8,76%)	1.072.890 (1,10%)	97.246.770
Número de IP's	98.292 (98,03%)	2.064 (2,06%)	11 (0,01%)	100.264
Número de Prefixos	9.520 (95,08%)	656 (6,55%)	11 (0,11%)	10.013
Número de ASes	1.988 (96,27%)	212 (10,27%)	4 (0,19%)	2.065
Número de CC's	130 (98,48%)	54 (40,91%)	4 (3,03%)	132
Volume de bytes (GB)	287,48 (84,24%)	28,33 (8,30%)	25,46 (7,46%)	341,27

Tabela B.8. 10 Country Codes que mais enviaram mensagens no *honeypot* BR-01.

	Mensagens	IP's	Prefixos	ASes	Bytes (GB)
BR	14.295.555	1.788	828	124	45,80
CN	1.275.3521	32.652	3.641	67	64,88
US	8.856.338	1.454	747	304	28,00
TW	7.232.834	45.994	148	22	31,44
RU	6.709.622	2.856	783	346	21,04
KR	3.449.639	235	140	44	10,74
IN	3.123.243	9.621	1.109	72	9,95
GB	2.563.238	173	116	59	7,94
UA	2.439.263	421	222	101	7,55
NL	2.224.630	121	57	30	6,97

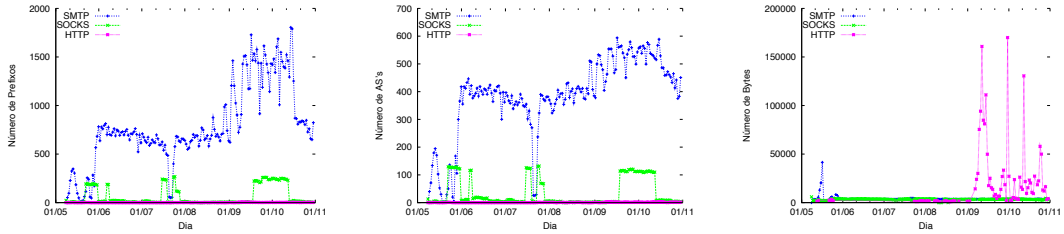
Tabela B.9. 10 Sistemas Autônomos que mais enviaram mensagens no *honeypot* BR-01.

	Mensagens	IP's	Prefixos	Country Code	Bytes (GB)
3462	6.330.496	25.661	58	TW	28,67
4134	6.303.436	13.519	2.589	CN	43,59
28573	2.618.919	202	131	BR	8,21
4837	2.346.755	4.411	198	CN	8,17
18881	2.153.289	395	132	BR	6,80
27699	1.939.883	101	35	BR	6,25
4230	1.483.980	84	32	BR	4,68
16276	1.339.778	102	14	CA	4.21
10429	1.057.256	21	15	BR	3,44
8167	989.739	59	47	BR	3,35



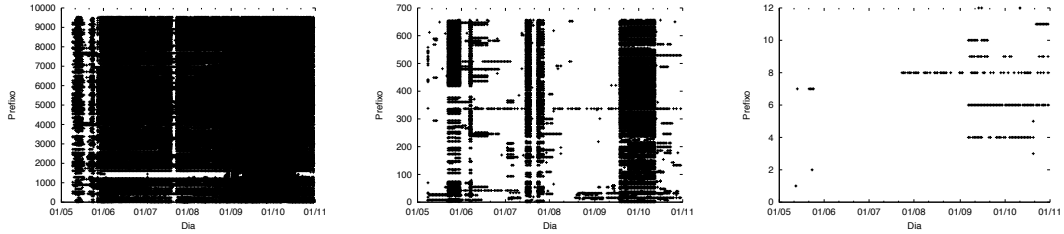
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura B.17. Séries temporais dos endereços, mensagens e volume por protocolo do *honeypot* BR-01.



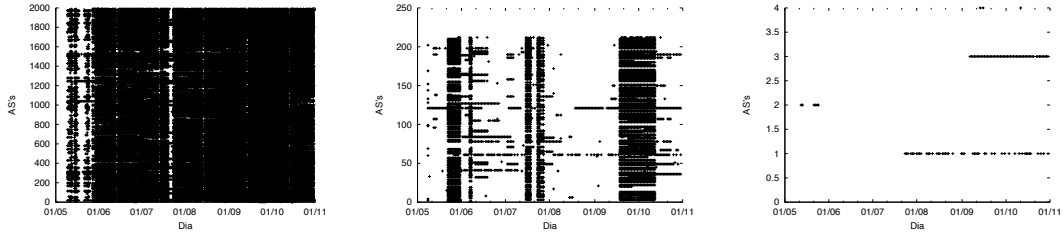
(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura B.18. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo do *honeypot* BR-01.



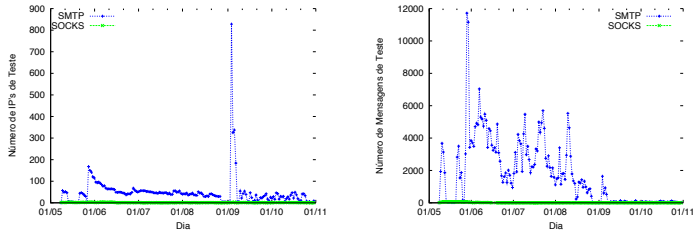
(a) Prefixos SMTP (b) Prefixos SOCKS (c) Prefixos HTTP

Figura B.19. Distribuição dos prefixos ao longo do tempo do *honeypot* BR-01.



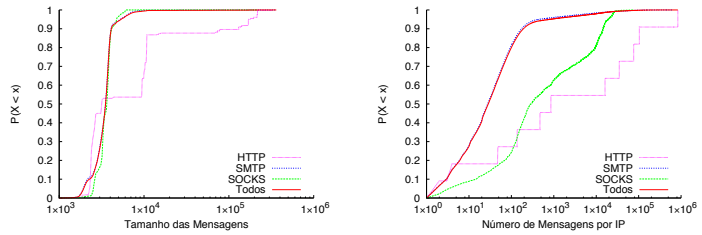
(a) ASes SMTP (b) ASes SOCKS (c) ASes HTTP

Figura B.20. Distribuição dos ASes ao longo do tempo do *honeypot* BR-01.



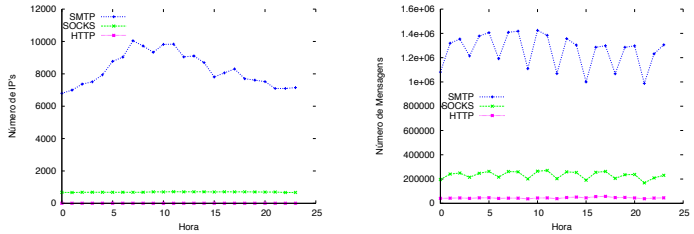
(a) IP's que enviam mensagens de teste (b) Mensagens de teste

Figura B.21. Características da mensagens de teste do *honeypot* BR-01.



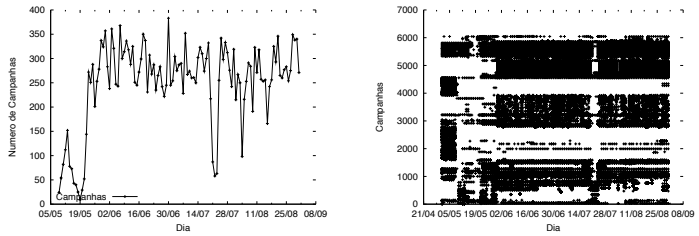
(a) CDF do Tamanho das Mensagens (b) CDF das Mensagens por IP

Figura B.22. Características das CDF's do *honeypot* BR-01.



(a) Número de Endereços IP por hora do dia (b) Mensagens por hora do dia

Figura B.23. Características do *honeypot* BR-01 por hora do dia



(a) Campanhas por Dia (b) Distribuição das campanhas ativas

Figura B.24. Características das campanhas do *honeypot* BR-01.

B.4 *Honeypot* BR-02

Tabela B.10. Visão geral dos dados coletados no *honeypot* BR-02.

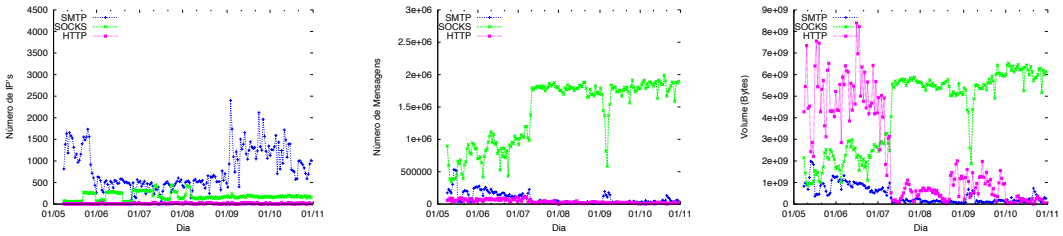
	SMTP	SOCKS	HTTP	Total
Número de Mensagens	15.489.192 (5,71%)	249.133.615 (91,80%)	6.764.800 (2,49%)	271.387.607
Número de IP's	77.049 (96,61%)	2.818 (3,53%)	130 (0,17%)	79.746
Número de Prefixos	6.556 (92,46%)	654 (9,22%)	35 (0,49%)	7.091
Número de ASes	1.376 (93,10%)	225 (15,22%)	10 (6,76%)	1.478
Número de CC's	117 (%)	56 (%)	4 (%)	121
Volume de bytes (GB)	68,92 (6,12%)	710,54 (63,00%)	348,25 (30,88%)	1.127,72

Tabela B.11. 10 Country Codes que mais enviaram mensagens no *honeypot* BR-02.

	Mensagens	IP's	Prefixos	ASes	Bytes (GB)
US	191.531.092	733	406	206	563,81
PH	42.566.314	144	24	6	97,17
CN	9.142.727	22.356	2.735	57	357,35
JP	6.995.214	143	46	19	25,38
TW	5.606.070	43.888	101	18	24,67
BR	3.790.653	982	536	100	14,10
HK	997.952	57	47	23	3,98
KR	837.444	154	104	39	2,89
IN	799.942	6.256	979	64	3,15
TH	776.514	45	38	20	2,77

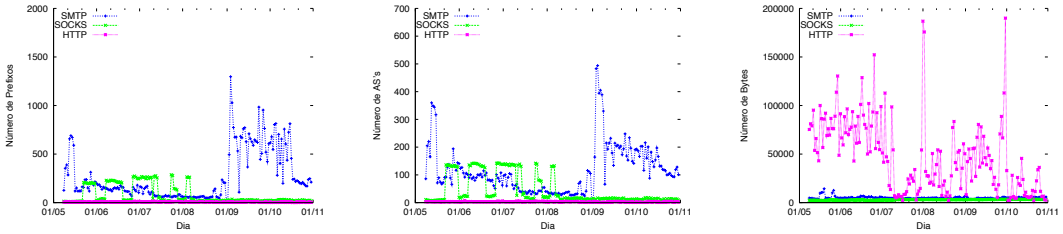
Tabela B.12. 10 Sistemas Autônomos que mais enviaram mensagens no *honeypot* BR-02.

	Mensagens	IP's	Prefixos	Country Code	Bytes (GB)
10297	152.742.213	154	3	US	454,73
29802	33.671.020	22	1	US	90,11
9299	29.388.957	14	1	PH	66,40
6648	12.141.274	48	2	PH	27,97
4134	5.969.012	16.024	2.075	CN	335,43
3462	5.558.379	43.821	62	TW	24,51
4713	3.055.023	20	10	JP	6,51
2497	2.141.739	24	8	JP	10,03
21788	2.126.417	21	12	US	6,31
4725	1.483.526	16	5		



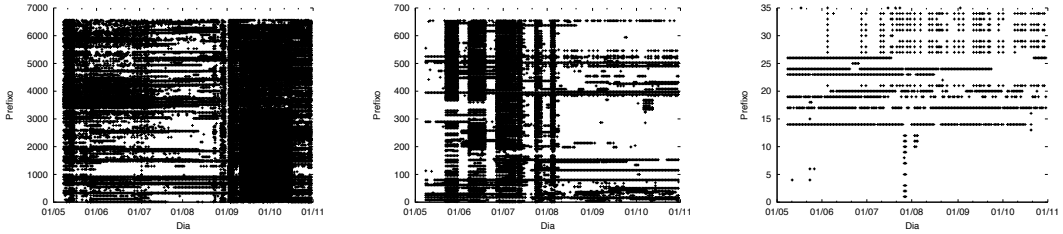
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura B.25. Séries temporais dos endereços, mensagens e volume por protocolo do *honeypot* BR-02.



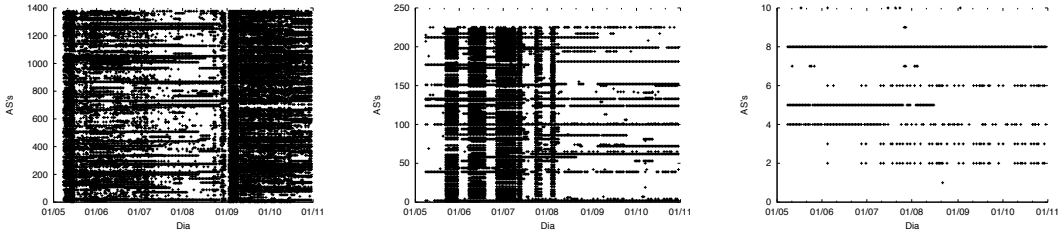
(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura B.26. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo do *honeypot* BR-02.



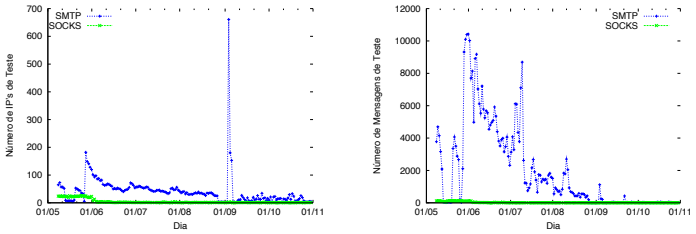
(a) Prefixos SMTP (b) Prefixos SOCKS (c) Prefixos HTTP

Figura B.27. Distribuição dos prefixos ao longo do tempo do *honeypot* BR-02.



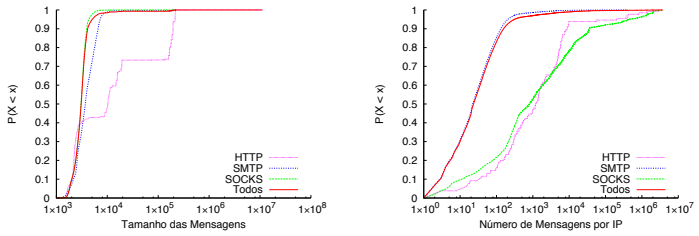
(a) ASes SMTP (b) ASes SOCKS (c) ASes HTTP

Figura B.28. Distribuição dos ASes ao longo do tempo do *honeypot* BR-02.



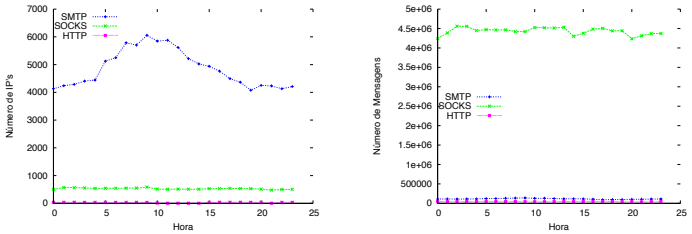
(a) IP's que enviam mensagens de teste (b) Mensagens de teste

Figura B.29. Características da mensagens de teste do *honeypot* BR-02.



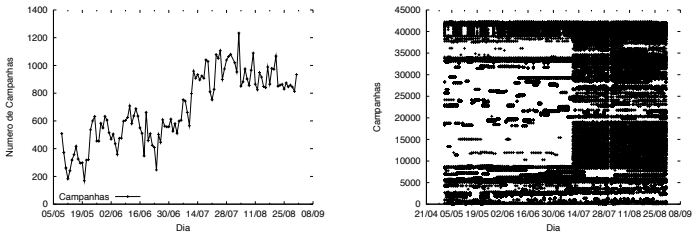
(a) CDF do Tamanho das Mensagens (b) CDF das Mensagens por IP

Figura B.30. Características das CDF's do *honeypot* BR-02.



(a) Número de Endereços IP por hora do dia (b) Mensagens por hora do dia

Figura B.31. Características do *honeypot* BR-02 por hora do dia



(a) Campanhas por Dia (b) Distribuição das campanhas ativas

Figura B.32. Características das campanhas do *honeypot* BR-02.

B.5 *Honeypot* EC-01

Tabela B.13. Visão Geral dos dados coletados no *honeypot* EC-01.

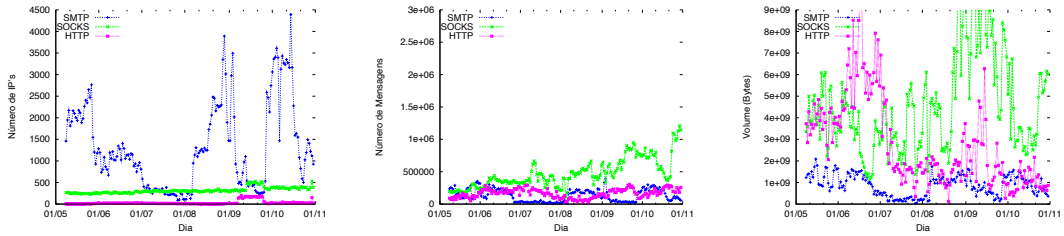
	SMTP	**SOCKS**	**HTTP**	**Total**
Mensagens	23.440.496(17,24%)	83.510.524 (61,42%)	29.020.814 (21,34%)	135.971.834
Endereços IP's	105.795 (93,32%)	6.907 (6,09%)	2.597 (2,29%)	113.360
Prefixos de rede	8.399 (90,67%)	1.050 (11,33%)	78 (8,42%)	9.263
Sistemas Autônomos (AS)	1.551 (92,82%)	287 (17,17%)	22 (13,17%)	1.671
***Country Codes* (CC)**	124 (97,64%)	62 (48,82%)	7 (5,51%)	127
Volume de tráfego (GB)	143,01 (10,56%)	757,63 (55,96%)	453,26 (34,48%)	1.353,91

Tabela B.14. 10 Country Codes que mais enviaram mensagens no *honeypot* EC-01.

	Mensagens	**IP's**	**Prefixos**	**ASes**	**Bytes (GB)**
TW	32.189.332	39.603	118	18	118,96
BR	27.401.989	3.369	795	110	101,80
CN	26.808.399	53.165	3.884	58	932,86
US	10.989.366	918	663	261	47,62
IT	4.631.745	163	82	18	14,81
KR	3.097.625	168	113	38	10,71
RU	2.528.879	2.034	569	255	11,56
HK	2.456.565	55	42	21	9,66
IN	2.280.162	8.829	1.068	63	8,28
MY	1.961.158	28	26	10	6,91

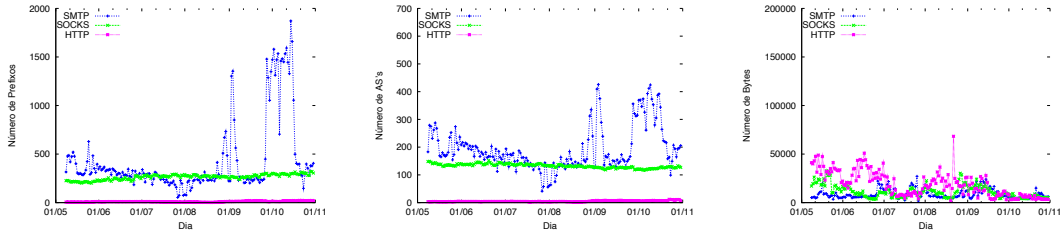
Tabela B.15. 10 Sistemas Autônomos que mais enviaram mensagens no *honeypot* EC-01.

	Mensagens	**IP's**	**Prefixos**	**Bytes (GB)**	
3462	26.212.302	14.337	47	TW	100,61
4134	19.668.435	16.236	2.718	CN	870,82
27699	5.799.741	637	32	BR	21,32
8167	4.649.230	57	41	BR	15,89
23650	4.354.828	15	9	CN	47,95
3269	3.866.336	22	17	IT	12,06
18881	3.472.309	533	138	BR	13,70
4230	2.911.968	65	19	BR	10,47
28573	2.863.030	108	79	BR	10,63
24164	2.107.404	7	5	TW	5,92



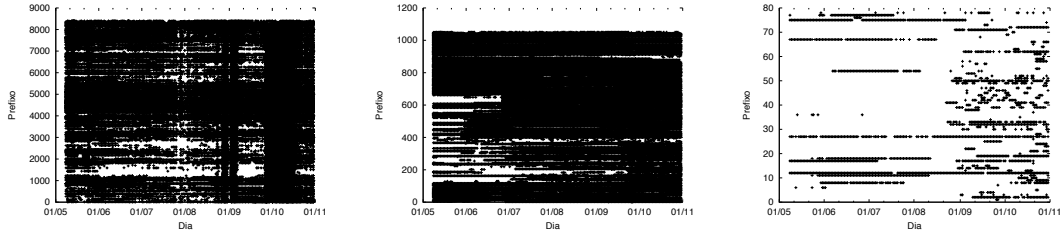
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura B.33. Séries temporais dos endereços, mensagens e volume por protocolo do *honeypot* EC-01.



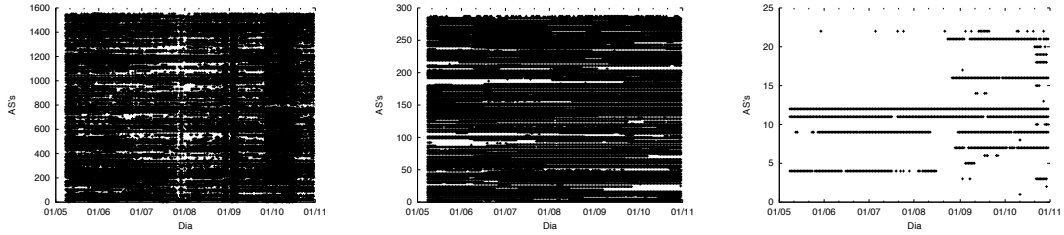
(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura B.34. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo do *honeypot* EC-01.



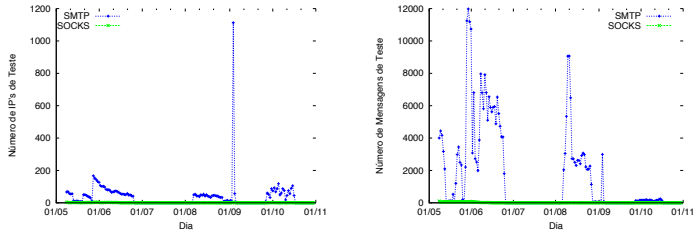
(a) Prefixos SMTP (b) Prefixos SOCKS (c) Prefixos HTTP

Figura B.35. Distribuição dos prefixos ao longo do tempo do *honeypot* EC-01.



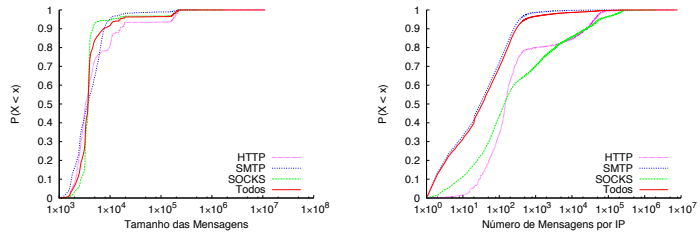
(a) ASes SMTP (b) ASes SOCKS (c) ASes HTTP

Figura B.36. Distribuição dos ASes ao longo do tempo do *honeypot* EC-01.



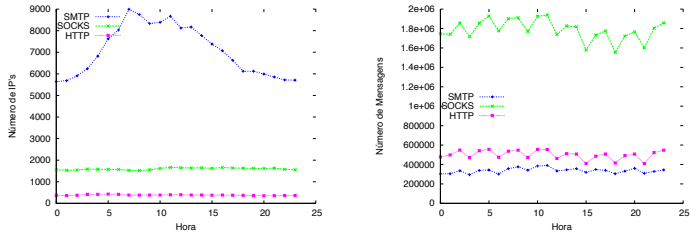
(a) IP's que enviam mensagens de teste (b) Mensagens de teste

Figura B.37. Características da mensagens de teste do *honeypot* EC-01.



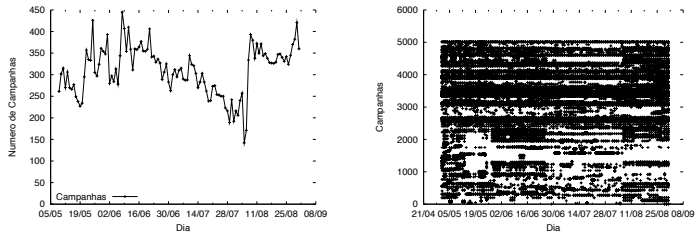
(a) CDF do Tamanho das Mensagens (b) CDF das Mensagens por IP

Figura B.38. Características das CDF's do *honeypot* EC-01.



(a) Número de Endereços IP por hora do dia (b) Mensagens por hora do dia

Figura B.39. Características do *honeypot* EC-01 por hora do dia.



(a) Campanhas por Dia (b) Distribuição das campanhas ativas

Figura B.40. Características das campanhas do *honeypot* EC-01.

B.6 *Honeypot* NL-01

Tabela B.16. Visão Geral dos dados coletados no *honeypot* NL-01.

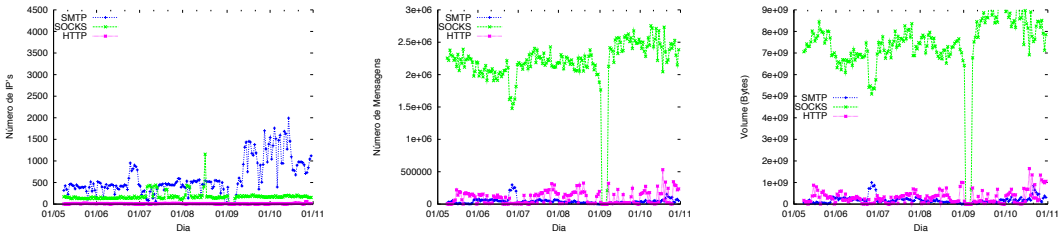
	SMTP	SOCKS	HTTP	Total
Número de Mensagens	7.584.298 (1,87%)	381.823.478 (93,91%)	17.177.007 (4,22%)	406.584.783
Número de IP's	61.651 (94,96%)	3.409 (5,25%)	77 (0,12%)	64.923
Número de Prefixos	5.814 (92,89%)	533 (8,52%)	20 (0,32%)	6.259
Número de ASes	1.188 (92,67%)	195 (15,21%)	8 (0,62%)	1.282
Número de CC's	118 (99,16%)	52 (43,70%)	5 (4,20%)	119
Volume de bytes (GB)	30,48 (2,36%)	1.209,31 (93,69%)	51,00 (3,95%)	1.290,80

Tabela B.17. 10 Country Codes que mais enviaram mensagens no *honeypot* NL-01.

	Mensagens	IP's	Prefixos	ASes	Bytes (GB)
US	344.245.561	564	288	168	1030,23
PH	38.937.219	136	20	7	89,86
JP	7.854.430	141	46	18	29,37
TW	5.527.192	41.693	83	13	21,73
CN	4.778.186	12.147	2.614	50	101,13
BR	752.947	817	461	76	2,54
IN	335.491	4.838	869	60	1,17
TH	322.880	38	31	17	1,13
RU	261.489	1.504	476	233	0,86
AE	252.503	9	9	2	0,91

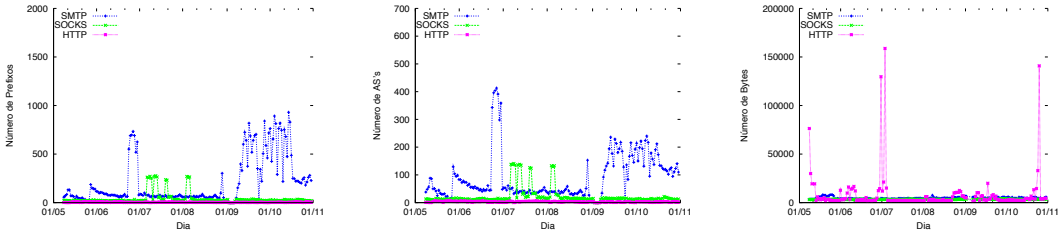
Tabela B.18. 10 Sistemas Autônomos que mais enviaram mensagens no *honeypot* NL-01.

	Mensagens	IP's	Prefixos	Country Code	Bytes (GB)
10297	293.902.576	149	3	US	889,79
29802	44.672.082	21	1	US	123,62
9299	23.991.704	56	10	PH	55,32
6648	14.184.963	49	3	PH	32,58
3462	5.474.017	41.638	55	TW	21,58
21788	4.633.320	11	7	US	13,52
4134	3.117.042	7.395	2.023	CN	87,99
4713	3.014.601	17	8	JP	6,21
2497	2.726.833	22	7	JP	12,60
4725	1.828.149	21	6	JP	9,35



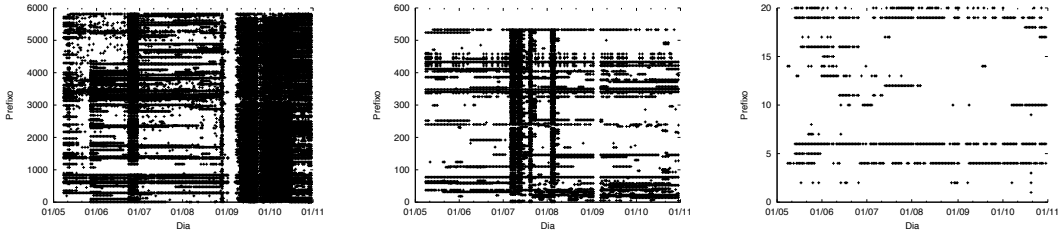
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura B.41. Séries temporais dos endereços, mensagens e volume por protocolo do *honeypot* NL-01.



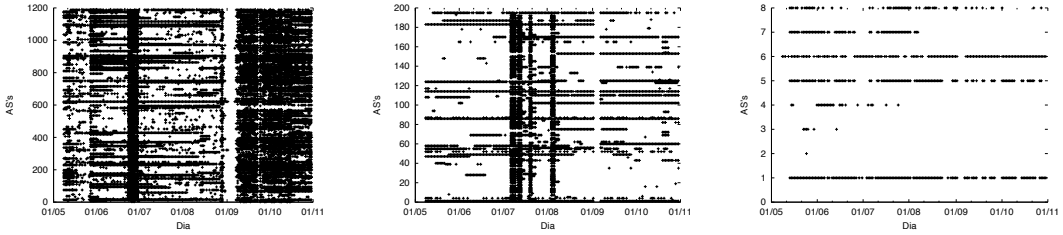
(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura B.42. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo do *honeypot* NL-01.



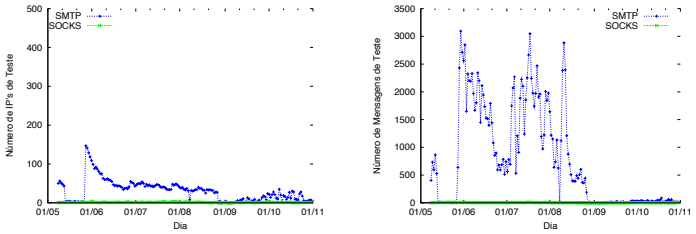
(a) Prefixos SMTP (b) Prefixos SOCKS (c) Prefixos HTTP

Figura B.43. Distribuição dos prefixos ao longo do tempo do *honeypot* NL-01.



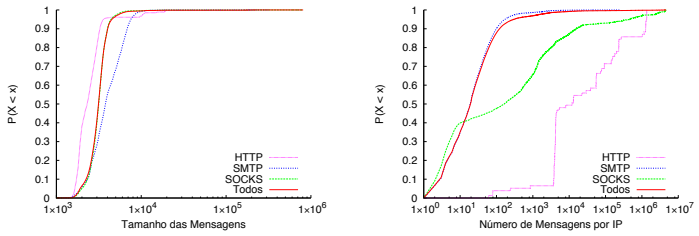
(a) ASes SMTP (b) ASes SOCKS (c) ASes HTTP

Figura B.44. Distribuição dos ASes ao longo do tempo do *honeypot* NL-01.



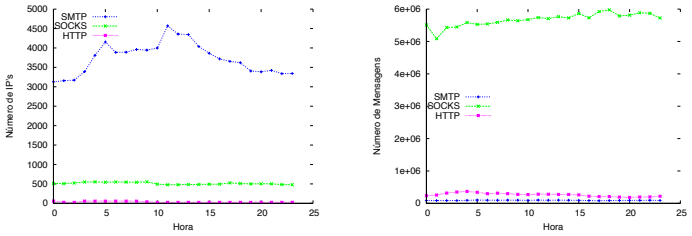
(a) IP's que enviam mensagens de teste (b) Mensagens de teste

Figura B.45. Características da mensagens de teste do *honeypot* NL-01.



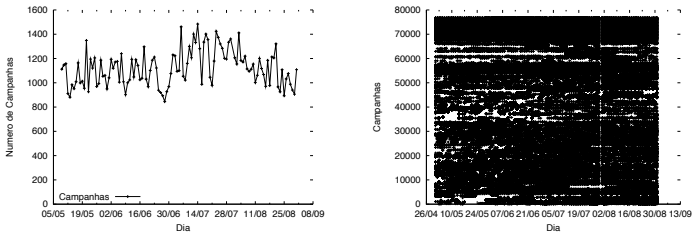
(a) CDF do Tamanho das Mensagens (b) CDF das Mensagens por IP

Figura B.46. Características das CDF's do *honeypot* NL-01.



(a) Número de Endereços IP por hora do dia (b) Mensagens por hora do dia

Figura B.47. Características do *honeypot* NL-01 por hora do dia.



(a) Campanhas por Dia (b) Distribuição das campanhas ativas

Figura B.48. Características das campanhas do *honeypot* NL-01.

B.7 *Honeypot* TW-01

Tabela B.19. Visão Geral dos dados coletados no *honeypot* TW-01.

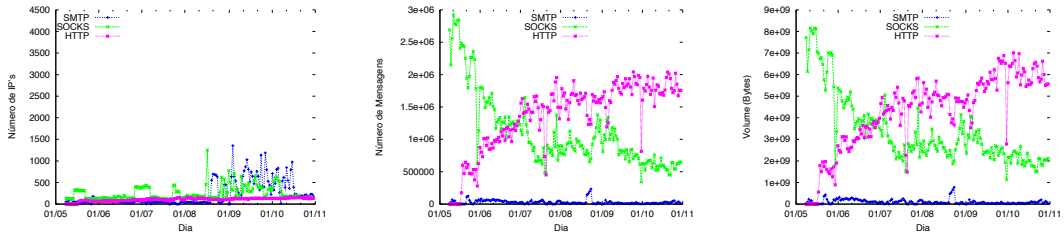
	SMTP	SOCKS	HTTP	Total
Número de Mensagens	5.032.636 (1,14%)	195.857.407 (44,48%)	239.438.886 (54,38%)	440.328.929
Número de IP's	19.447 (61,14%)	12.286 (38,62%)	1.008 (3,17%)	31.809
Número de Prefixos	5.771 (89,71%)	771 (11,98%)	83 (12,90%)	6.433
Número de ASes	1.087 (89,61%)	226 (18,63%)	25 (2,06%)	1.213
Número de CC's	110 (96,49%)	57 (50,00%)	11 (9,64%)	114
Volume de bytes (GB)	17,11 (1,35%)	552,60 (43,48%)	701,28 (55,17%)	1271,00

Tabela B.20. 10 Country Codes que mais enviaram mensagens no *honeypot* TW-01.

	Mensagens	IP's	Prefixos	ASes	Bytes (GB)
US	300.733.921	525	308	167	903,57
PH	102.978.546	138	26	5	238,86
JP	15.727.381	154	46	19	63,86
TW	13.879.569	11.298	117	14	41,63
CN	1.109.165	9.992	2.425	54	3,63
BR	707.121	768	503	70	2,34
TH	459.881	26	34	17	1,62
IN	309.855	4.918	996	60	1,05
RU	240.682	1.412	477	221	0,79
GE	236.081	2	2	2	0,82

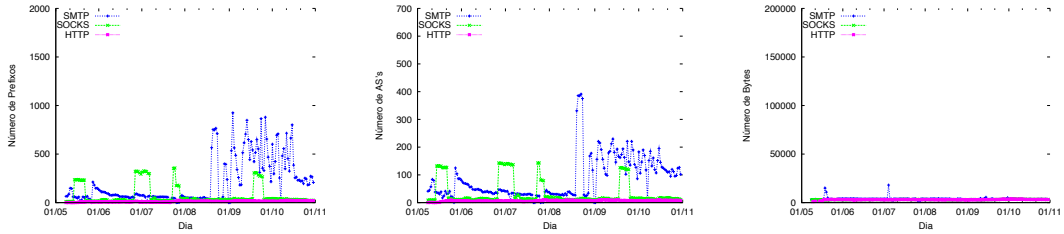
Tabela B.21. 10 Sistemas Autônomos que mais enviaram mensagens no *honeypot* TW-01.

	Mensagens	IP's	Prefixos	Country Code	Bytes (GB)
10297	259.501.630	150	3	US	788,38
9299	65.031.126	58	9	PH	152,23
29802	39.085.845	21	1	US	108,57
6648	36.748.579	49	9	PH	83,40
3462	13.290.857	11.235	67	TW	39,99
2497	8.432.475	24	6	JP	39,98
4713	3.935.612	17	8	JP	7,80
4725	2.729.666	22	6	JP	13,35
21788	1.563.944	9	9	US	4,61
9658	1.164.823	19	4	PH	3,12



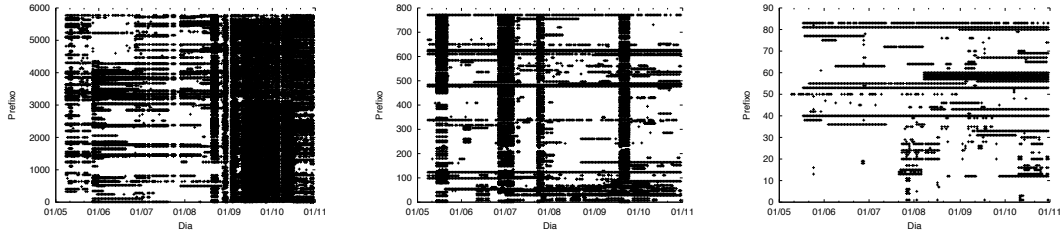
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura B.49. Séries temporais dos endereços, mensagens e volume por protocolo do *honeypot* TW-01.



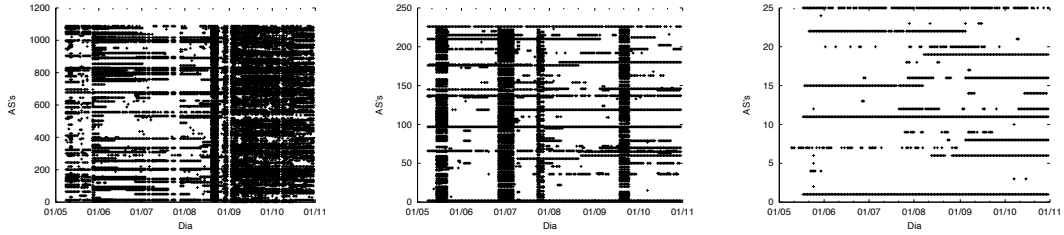
(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura B.50. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo do *honeypot* TW-01.



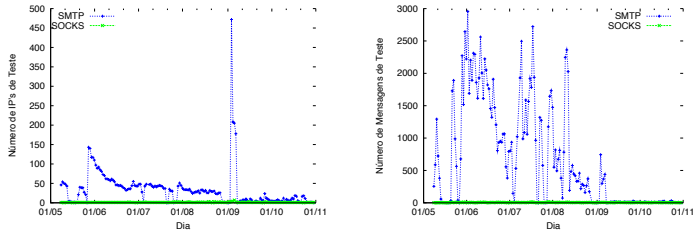
(a) Prefixos SMTP (b) Prefixos SOCKS (c) Prefixos HTTP

Figura B.51. Distribuição dos prefixos ao longo do tempo do *honeypot* TW-01.



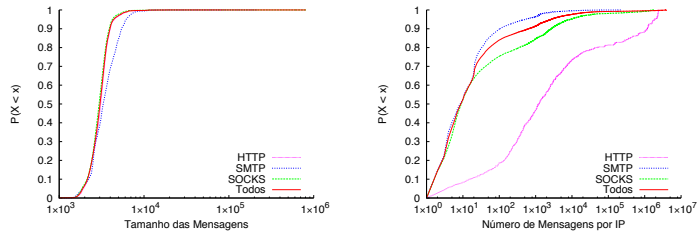
(a) ASes SMTP (b) ASes SOCKS (c) ASes HTTP

Figura B.52. Distribuição dos ASes ao longo do tempo do *honeypot* TW-01.



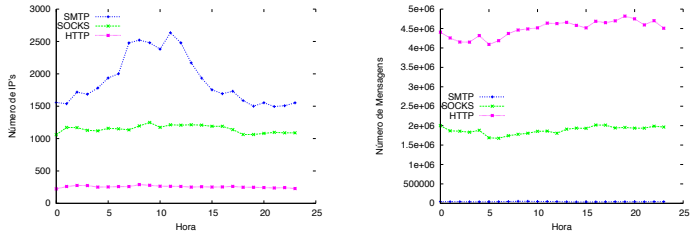
(a) IP's que enviam mensagens de teste (b) Mensagens de teste

Figura B.53. Características da mensagens de teste do *honeypot* TW-01.



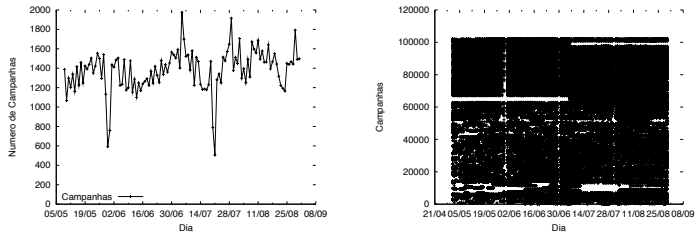
(a) CDF do Tamanho das Mensagens (b) CDF das Mensagens por IP

Figura B.54. Características das CDF's do *honeypot* TW-01.



(a) Número de Endereços IP por hora do dia (b) Mensagens por hora do dia

Figura B.55. Características do *honeypot* TW-01 por hora do dia.



(a) Campanhas por Dia (b) Distribuição das campanhas ativas

Figura B.56. Características das campanhas do *honeypot* TW-01.

B.8 *Honeypot* UY-01

Tabela B.22. Visão Geral dos dados coletados no *honeypot* UY-01.

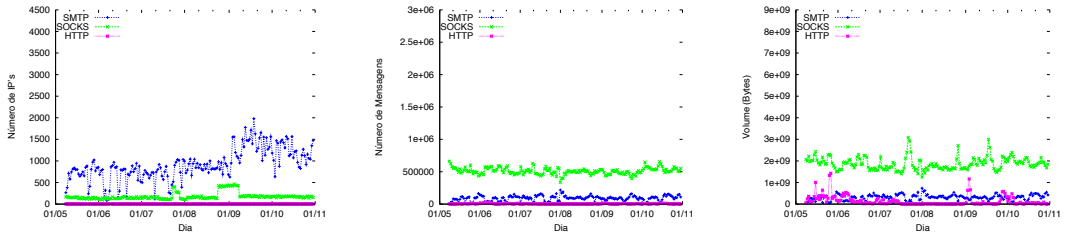
	SMTP	SOCKS	HTTP	Total
Número de Mensagens	15.954.283 (14,95%)	89.867.474 (84,20%)	902.390 (0,85%)	106.724.147
Número de IP's	49.877 (96,46%)	1.888 (3,65%)	9 (0,01%)	51.709
Número de Prefixos	7.615 (95,81%)	452 (56,87%)	9 (0,11%)	7.948
Número de ASes	1.838 (96,53%)	175 (9,19%)	4 (0,21%)	1.904
Número de CC's	127 (99,21%)	48 (37,50%)	3 (2,34%)	128
Volume de bytes (GB)	51,25 (13,48%)	307,19 (80,78%)	21,84 (5,74%)	380,28

Tabela B.23. 10 Country Codes que mais enviaram mensagens no *honeypot* UY-01.

	Mensagens	IP's	Prefixos	ASes	Bytes (GB)
US	72.468.759	1.221	604	275	220,81
PH	14.821.424	141	26	8	34,27
CN	4.166.892	10.114	2500	60	73,58
BR	2.091.633	1.240	669	111	6,59
JP	2.002.095	160	64	29	8,47
TW	1.469.880	28.436	127	22	5,84
RU	1.274.270	1.735	677	324	3,93
KR	570.942	205	126	40	1,75
IN	544.174	4.461	923	68	1,72
UA	440.663	318	193	97	1,37

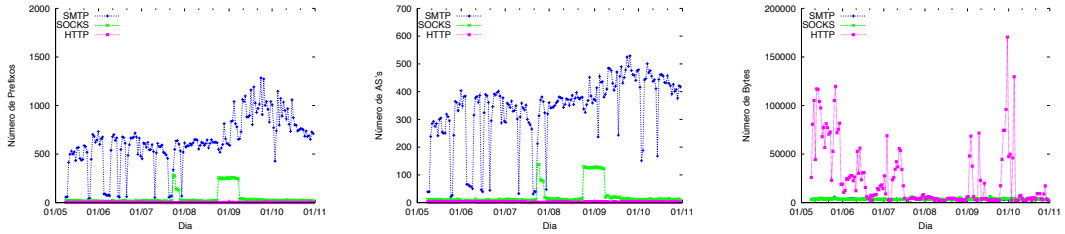
Tabela B.24. 10 Sistemas Autônomos que mais enviaram mensagens no *honeypot* UY-01.

	Mensagens	IP's	Prefixos	Country Code	Bytes (GB)
10297	62.238.706	141	3	US	190,38
9299	8.907.984	7	5	PH	20,53
29802	8.205.928	21	1	US	24,15
6648	5.579.927	48	2	PH	12,83
4134	2.456.537	5.681	1.684	CN	65,26
3462	1.355.854	14.932	54	TW	5,48
2497	832.130	23	7	JP	3,89
4725	601.879	14	6	JP	3,08
4837	485.547	2.037	161	CN	1,52
4713	435.794	2	2	JP	0,98



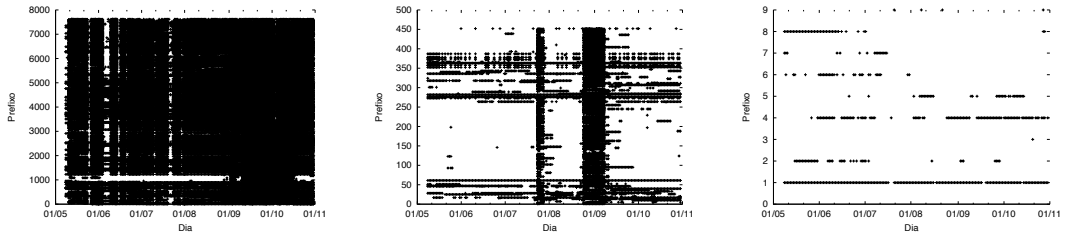
(a) Endereços IP ao longo do tempo (b) Mensagens ao longo do tempo (c) Volume (bytes) ao longo do tempo

Figura B.57. Séries temporais dos endereços, mensagens e volume por protocolo do *honeypot* UY-01.



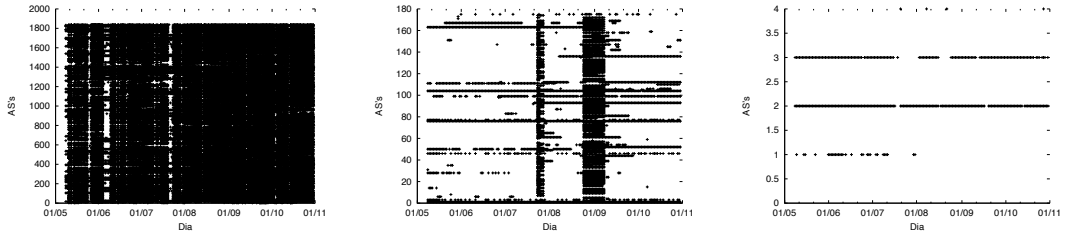
(a) Prefixos ao longo do tempo (b) ASes ao longo do tempo (c) Tamanho das Mensagens por Protocolo

Figura B.58. Séries temporais dos prefixos, ASes e tamanho da mensagem por protocolo do *honeypot* UY-01.



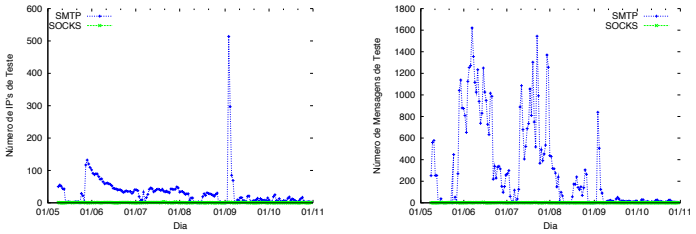
(a) Prefixos SMTP (b) Prefixos SOCKS (c) Prefixos HTTP

Figura B.59. Distribuição dos prefixos ao longo do tempo do *honeypot* UY-01.



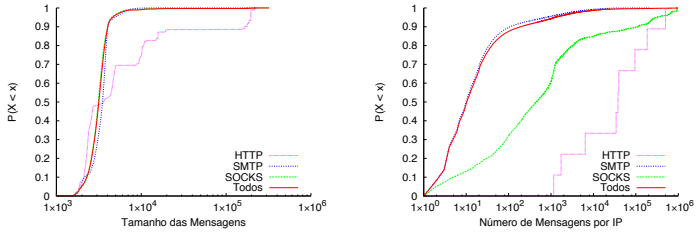
(a) ASes SMTP (b) ASes SOCKS (c) ASes HTTP

Figura B.60. Distribuição dos ASes ao longo do tempo do *honeypot* UY-01.



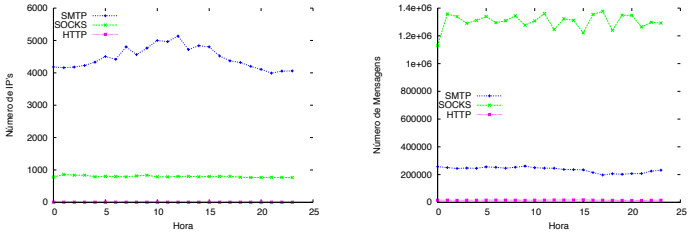
(a) IP's que enviam mensagens de teste (b) Mensagens de teste

Figura B.61. Características da mensagens de teste do *honeypot* UY-01.



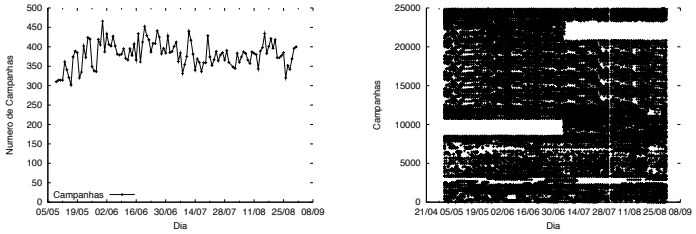
(a) CDF do Tamanho das Mensagens (b) CDF das Mensagens por IP

Figura B.62. Características das CDF's do *honeypot* UY-01.



(a) Número de Endereços IP por hora do dia (b) Mensagens por hora do dia

Figura B.63. Características do *honeypot* UY-01 por hora do dia.



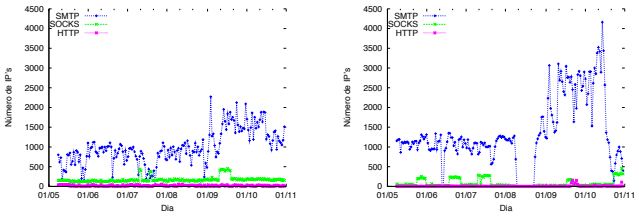
(a) Campanhas por Dia (b) Distribuição das campanhas ativas

Figura B.64. Características das campanhas do *honeypot* UY-01.

Apêndice C

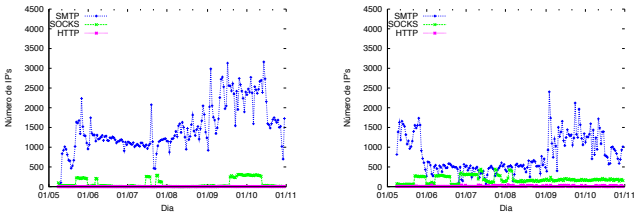
Visão por característica

C.1 Endereços IP por protocolo



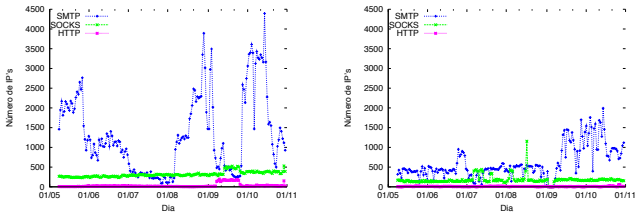
(a) AT-01 (b) AU-01

Figura C.1. Endereços IP por protocolo dos *honeypots* AT-01 e AU-01.



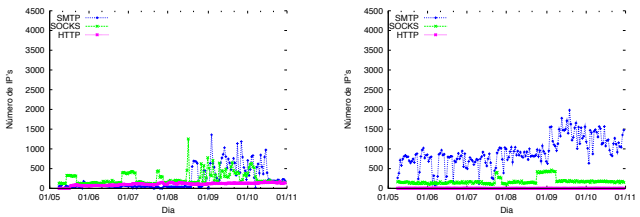
(a) BR-01 (b) BR-02

Figura C.2. Endereços IP por protocolo dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

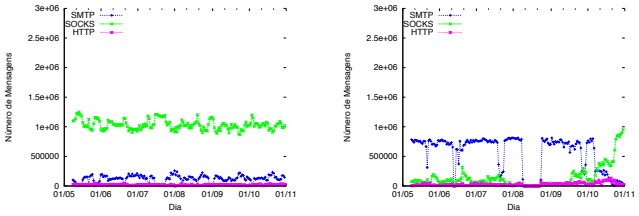
Figura C.3. Endereços IP por protocolo dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

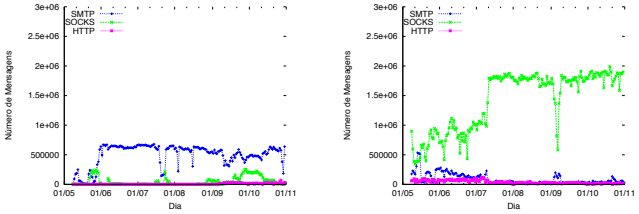
Figura C.4. Endereços IP por protocolo dos *honeypots* TW-01 e UY-01.

C.2 Mensagens por protocolo



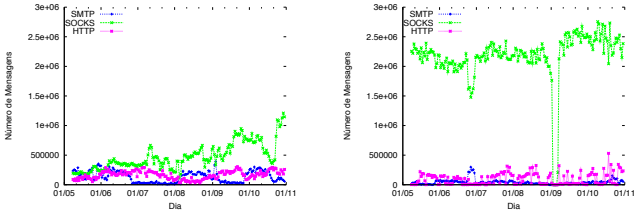
(a) AT-01 (b) AU-01

Figura C.5. Mensagens por protocolo dos *honeypots* AT-01 e AU-01.



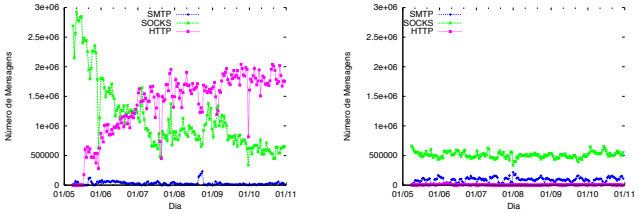
(a) BR-01 (b) BR-02

Figura C.6. Mensagens por protocolo dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

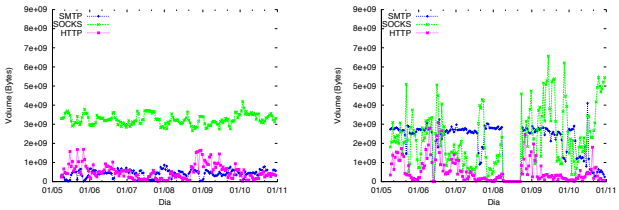
Figura C.7. Mensagens por protocolo dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

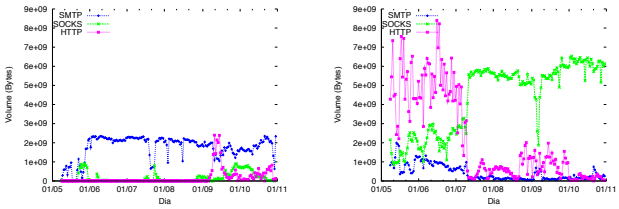
Figura C.8. Mensagens por protocolo dos *honeypots* TW-01 e UY-01.

C.3 Volume (bytes) por protocolo



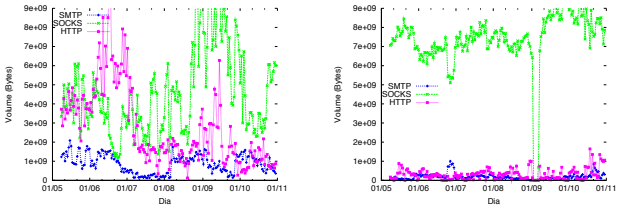
(a) AT-01 (b) AU-01

Figura C.9. Volume (bytes) por protocolo dos *honeypots* AT-01 e AU-01.



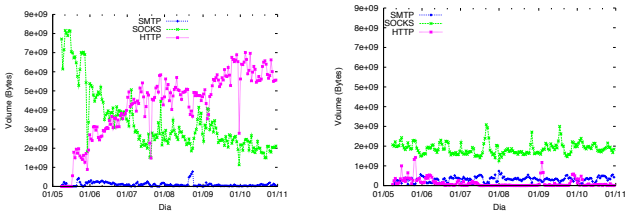
(a) BR-01 (b) BR-02

Figura C.10. Volume (bytes) por protocolo dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

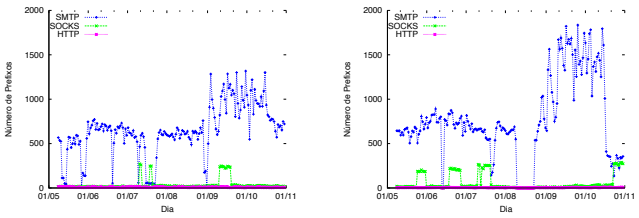
Figura C.11. Volume (bytes) por protocolo dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

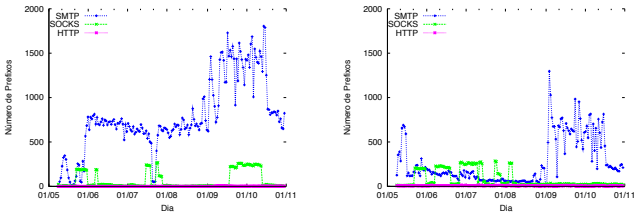
Figura C.12. Volume (bytes) por protocolo dos *honeypots* TW-01 e UY-01.

C.4 Prefixos de rede por protocolo



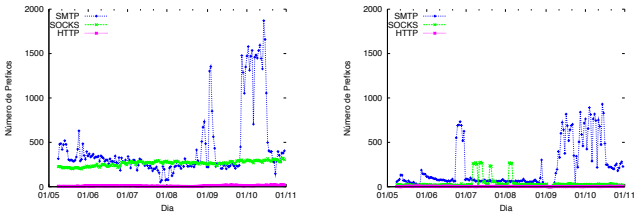
(a) AT-01 (b) AU-01

Figura C.13. Prefixos de rede por protocolo dos *honeypots* AT-01 e AU-01.



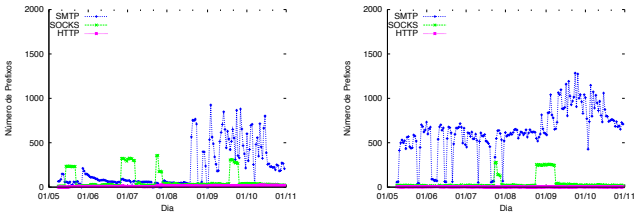
(a) BR-01 (b) BR-02

Figura C.14. Prefixos de rede por protocolo dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

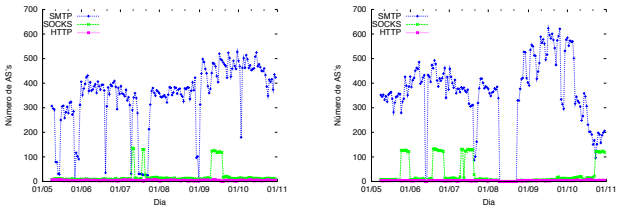
Figura C.15. Prefixos de rede por protocolo dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

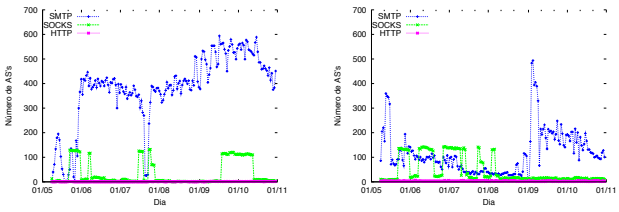
Figura C.16. Prefixos de rede por protocolo dos *honeypots* TW-01 e UY-01.

C.5 Sistemas Autônomos por protocolo



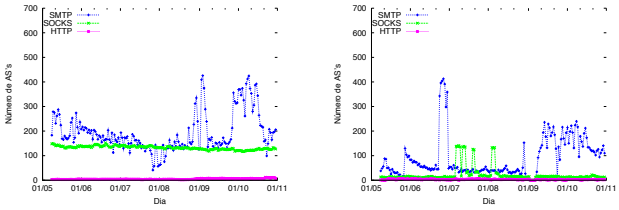
(a) AT-01 (b) AU-01

Figura C.17. Sistemas Autônomos por protocolo dos *honeypots* AT-01 e AU-01.



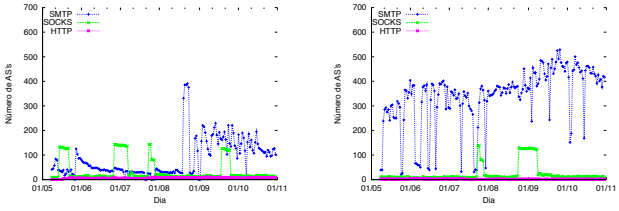
(a) BR-01 (b) BR-02

Figura C.18. Sistemas Autônomos por protocolo dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

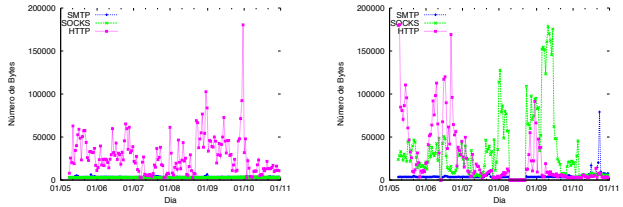
Figura C.19. Sistemas Autônomos por protocolo dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

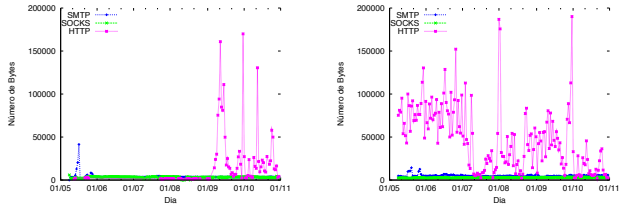
Figura C.20. Sistemas Autônomos por protocolo dos *honeypots* TW-01 e UY-01.

C.6 Tamanho das mensagens por protocolo



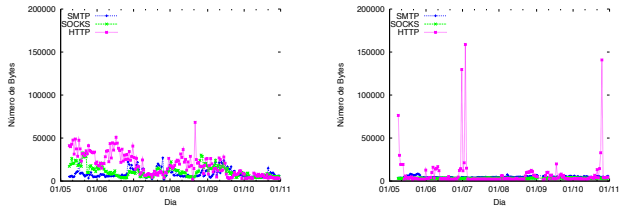
(a) AT-01 (b) AU-01

Figura C.21. Tamanho das mensagens por protocolo dos *honeypots* AT-01 e AU-01.



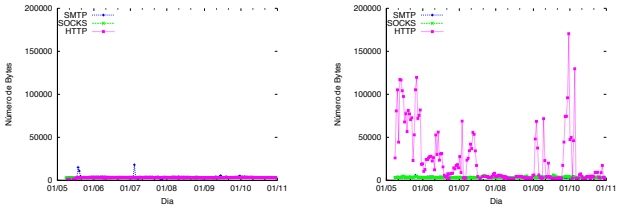
(a) BR-01 (b) BR-02

Figura C.22. Tamanho das mensagens por protocolo dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

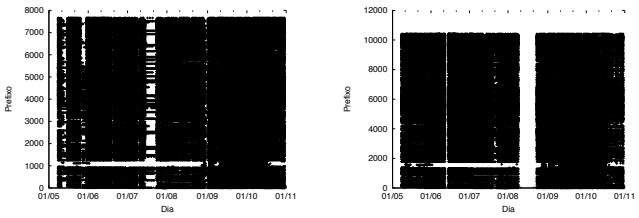
Figura C.23. Tamanho das mensagens por protocolo dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

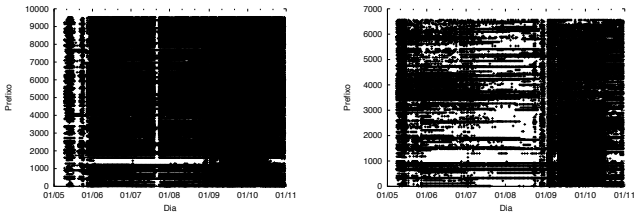
Figura C.24. Tamanho das mensagens por protocolo dos *honeypots* TW-01 e UY-01.

C.7 Distribuição de prefixos SMTP



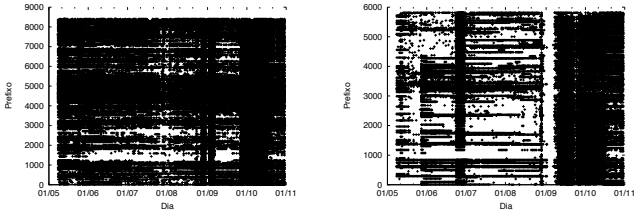
(a) AT-01 (b) AU-01

Figura C.25. Distribuição de prefixos SMTP dos *honeypots* AT-01 e AU-01.



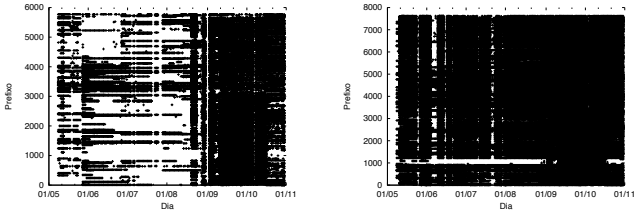
(a) BR-01 (b) BR-02

Figura C.26. Distribuição de prefixos SMTP dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

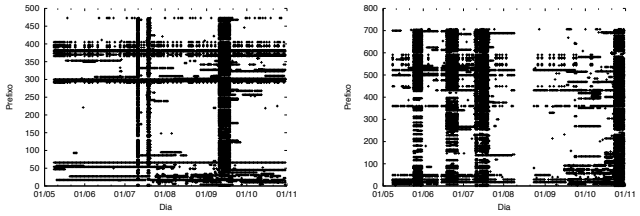
Figura C.27. Distribuição de prefixos SMTP dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

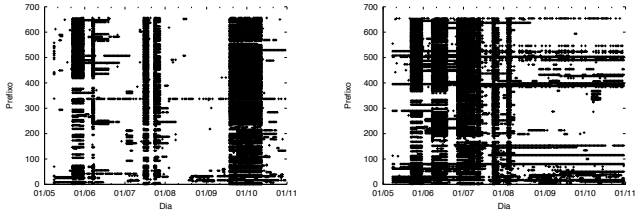
Figura C.28. Distribuição de prefixos SMTP dos *honeypots* TW-01 e UY-01.

C.8 Distribuição de prefixos SOCKS



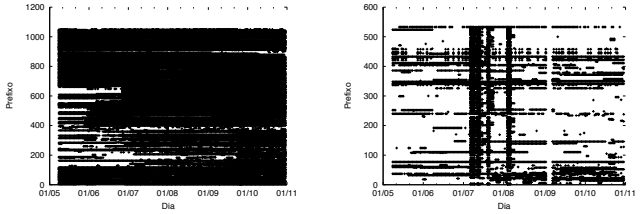
(a) AT-01 (b) AU-01

Figura C.29. Distribuição de prefixos SOCKS dos *honeypots* AT-01 e AU-01.



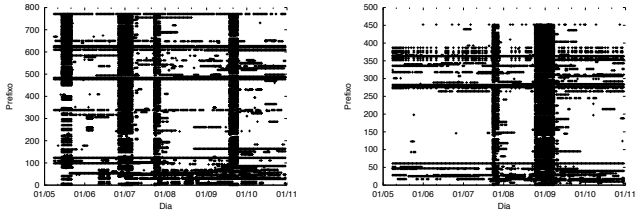
(a) BR-01 (b) BR-02

Figura C.30. Distribuição de prefixos SOCKS dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

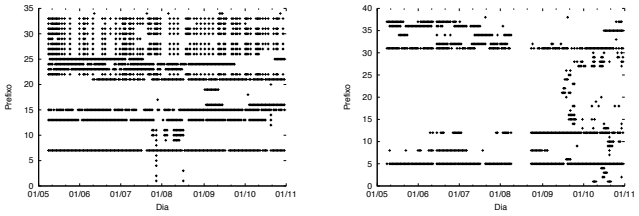
Figura C.31. Distribuição de prefixos SOCKS dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

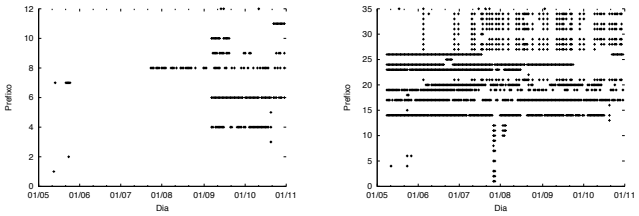
Figura C.32. Distribuição de prefixos SOCKS dos *honeypots* TW-01 e UY-01.

C.9 Distribuição de prefixos HTTP



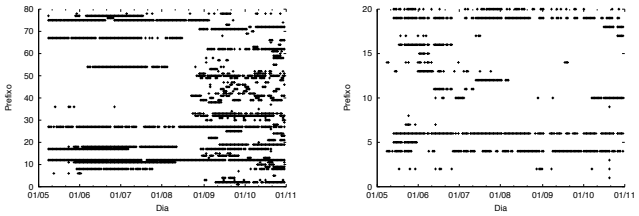
(a) AT-01 (b) AU-01

Figura C.33. Distribuição de prefixos HTTP dos *honeypots* AT-01 e AU-01.



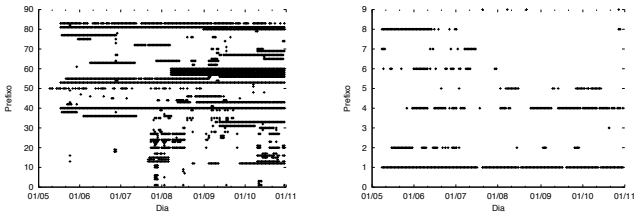
(a) BR-01 (b) BR-02

Figura C.34. Distribuição de prefixos HTTP dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

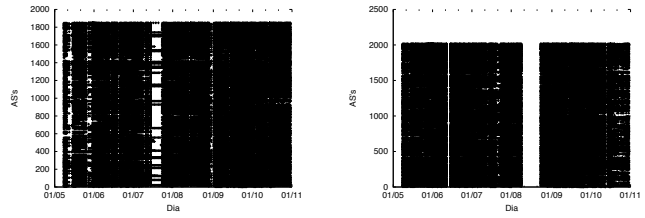
Figura C.35. Distribuição de prefixos HTTP dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

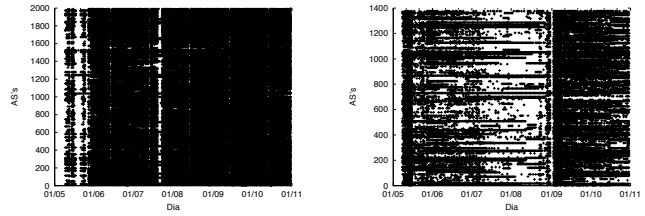
Figura C.36. Distribuição de prefixos HTTP dos *honeypots* TW-01 e UY-01.

C.10 Distribuição de ASes SMTP



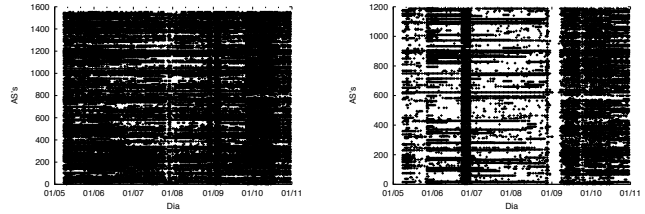
(a) AT-01 (b) AU-01

Figura C.37. Distribuição de ASes SMTP dos *honeypots* AT-01 e AU-01.



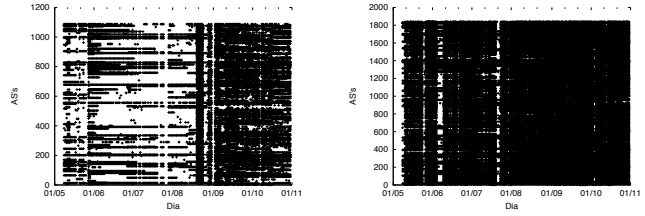
(a) BR-01 (b) BR-02

Figura C.38. Distribuição de ASes SMTP dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

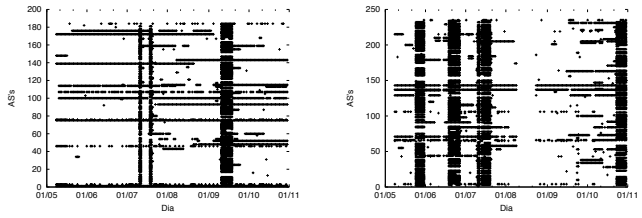
Figura C.39. Distribuição de ASes SMTP dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

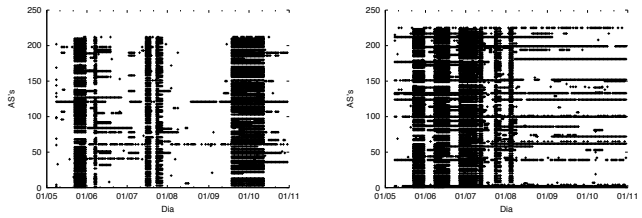
Figura C.40. Distribuição de ASes SMTP dos *honeypots* TW-01 e UY-01.

C.11 Distribuição de ASes SOCKS



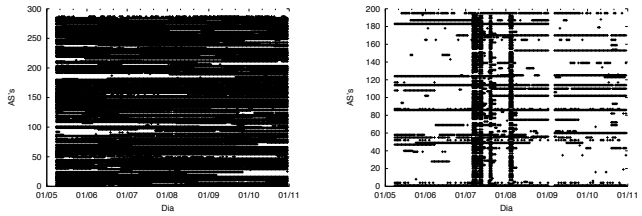
(a) AT-01 (b) AU-01

Figura C.41. Distribuição de ASes SOCKS dos *honeypots* AT-01 e AU-01.



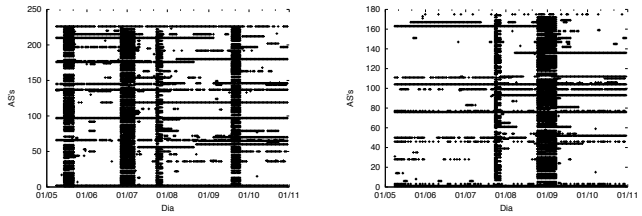
(a) BR-01 (b) BR-02

Figura C.42. Distribuição de ASes SOCKS dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

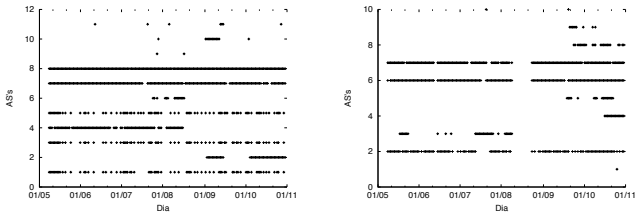
Figura C.43. Distribuição de ASes SOCKS dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

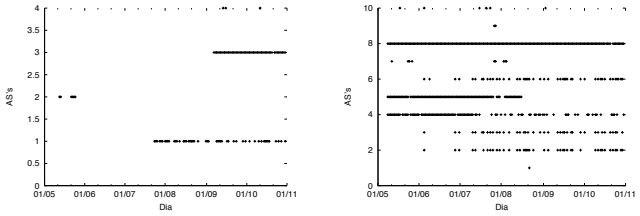
Figura C.44. Distribuição de ASes SOCKS dos *honeypots* TW-01 e UY-01.

C.12 Distribuição de ASes HTTP



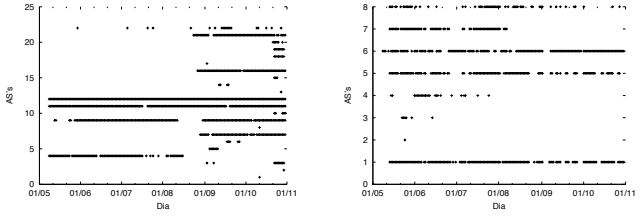
(a) AT-01 (b) AU-01

Figura C.45. Distribuição de ASes HTTP dos *honeypots* AT-01 e AU-01.



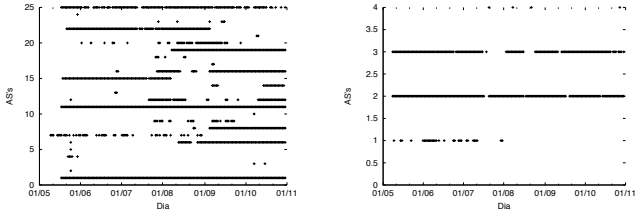
(a) BR-01 (b) BR-02

Figura C.46. Distribuição de ASes HTTP dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

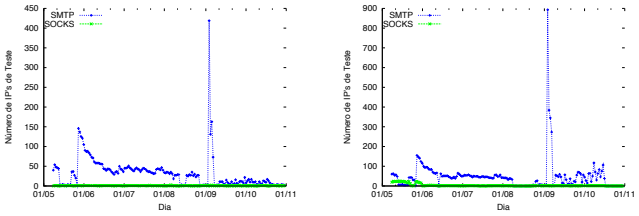
Figura C.47. Distribuição de ASes HTTP dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

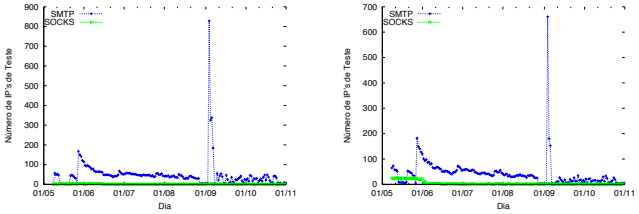
Figura C.48. Distribuição de ASes HTTP dos *honeypots* TW-01 e UY-01.

C.13 IP's das mensagens de teste



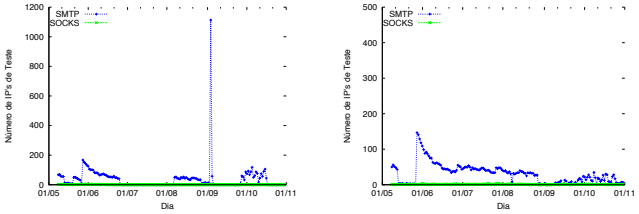
(a) AT-01 (b) AU-01

Figura C.49. IP's das mensagens de teste dos *honeypots* AT-01 e AU-01.



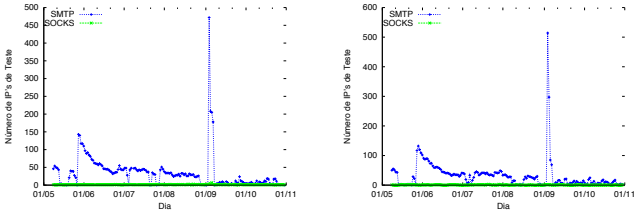
(a) BR-01 (b) BR-02

Figura C.50. IP's das mensagens de teste dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

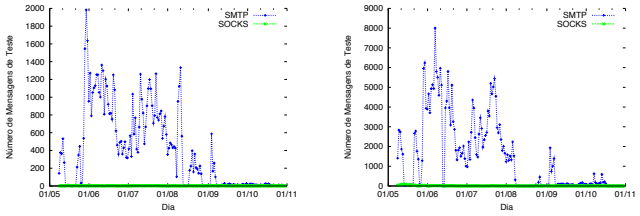
Figura C.51. IP's das mensagens de teste dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

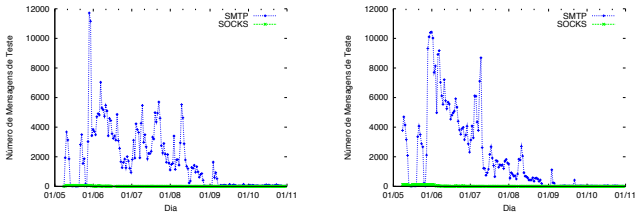
Figura C.52. IP's das mensagens de teste dos *honeypots* TW-01 e UY-01.

C.14 Mensagens de teste por protocolo



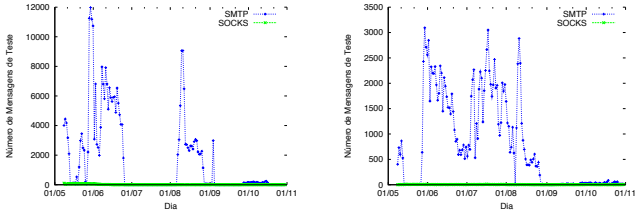
(a) AT-01 (b) AU-01

Figura C.53. Mensagens de teste por protocolo dos *honeypots* AT-01 e AU-01.



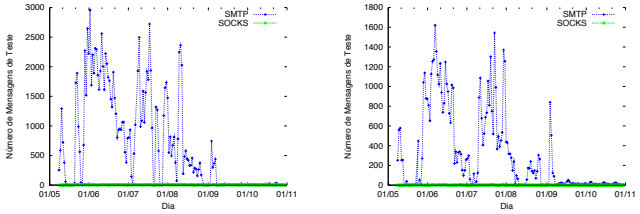
(a) BR-01 (b) BR-02

Figura C.54. Mensagens de teste por protocolo dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

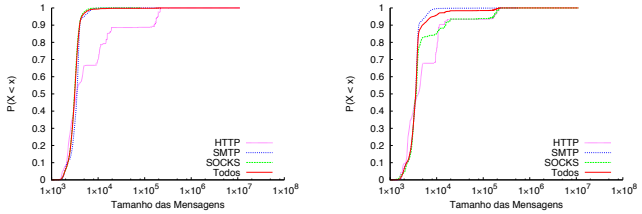
Figura C.55. Mensagens de teste por protocolo dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

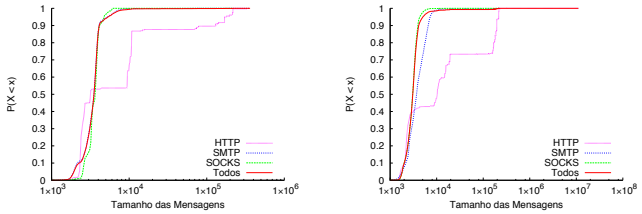
Figura C.56. Mensagens de teste por protocolo dos *honeypots* TW-01 e UY-01.

C.15 CDF do tamanho das mensagens



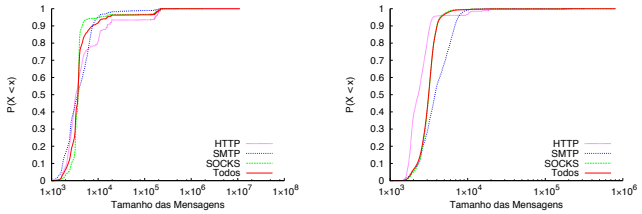
(a) AT-01 (b) AU-01

Figura C.57. CDF do tamanho das mensagens dos *honeypots* AT-01 e AU-01.



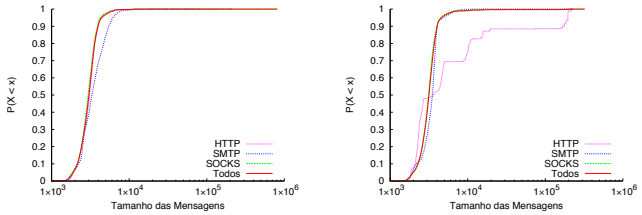
(a) BR-01 (b) BR-02

Figura C.58. CDF do tamanho das mensagens dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

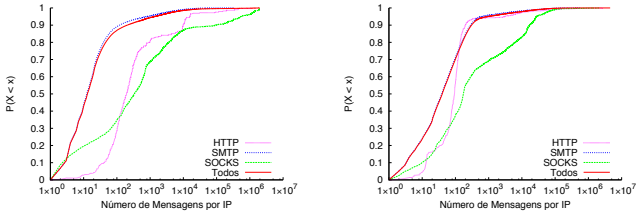
Figura C.59. CDF do tamanho das mensagens dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

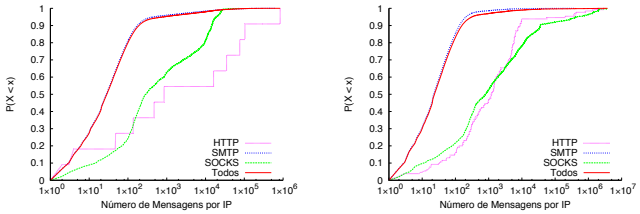
Figura C.60. CDF do tamanho das mensagens dos *honeypots* TW-01 e UY-01.

C.16 CDF das mensagens por IP



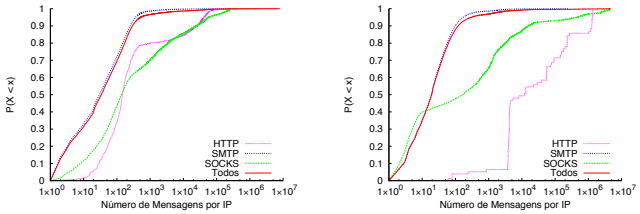
(a) AT-01 (b) AU-01

Figura C.61. CDF das mensagens por IP dos *honeypots* AT-01 e AU-01.



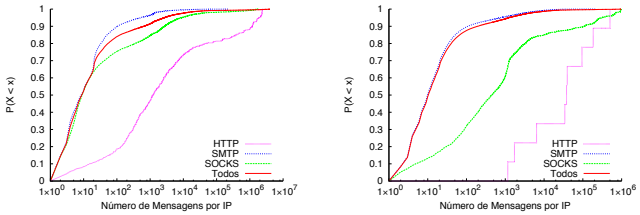
(a) BR-01 (b) BR-02

Figura C.62. CDF das mensagens por IP dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

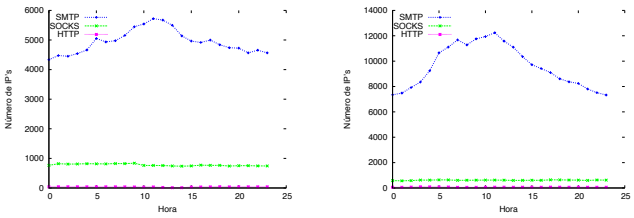
Figura C.63. CDF das mensagens por IP dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

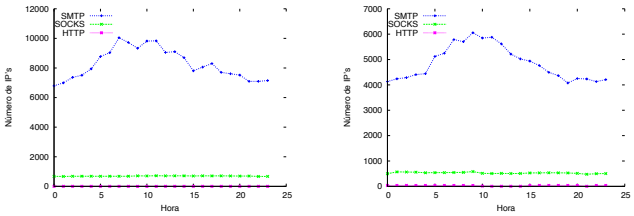
Figura C.64. CDF das mensagens por IP dos *honeypots* TW-01 e UY-01.

C.17 Endereços IP por hora do dia



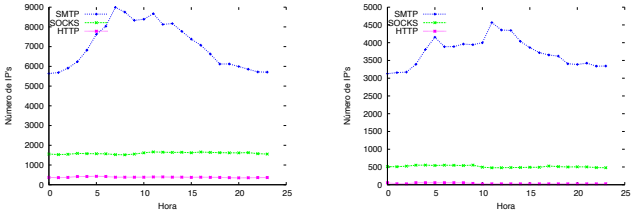
(a) AT-01 (b) AU-01

Figura C.65. Endereços IP por hora do dia dos *honeypots* AT-01 e AU-01.



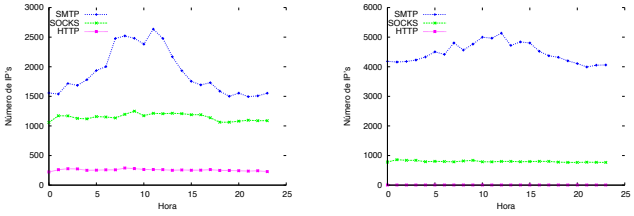
(a) BR-01 (b) BR-02

Figura C.66. Endereços IP por hora do dia dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

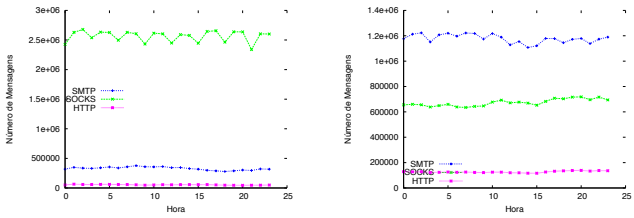
Figura C.67. Endereços IP por hora do dia dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

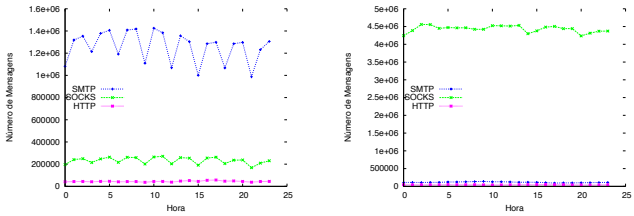
Figura C.68. Endereços IP por hora do dia dos *honeypots* TW-01 e UY-01.

C.18 Mensagens por hora do dia



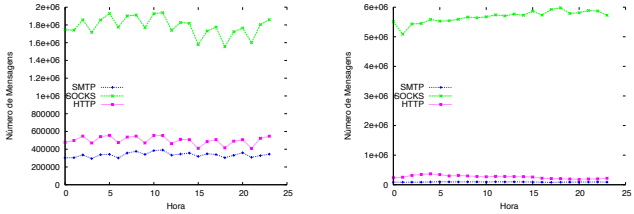
(a) AT-01 (b) AU-01

Figura C.69. Mensagens por hora do dia dos *honeypots* AT-01 e AU-01.



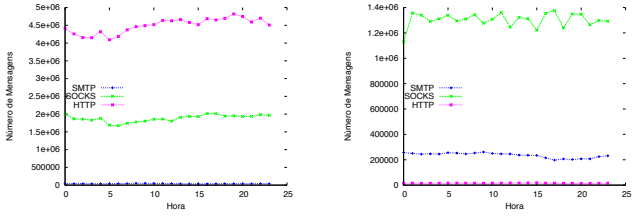
(a) BR-01 (b) BR-02

Figura C.70. Mensagens por hora do dia dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

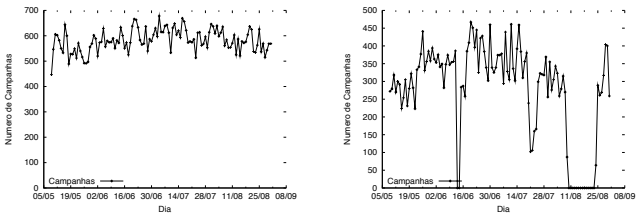
Figura C.71. Mensagens por hora do dia dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

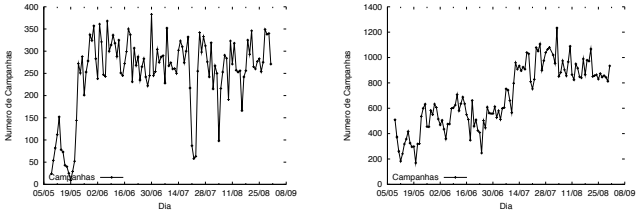
Figura C.72. Mensagens por hora do dia dos *honeypots* TW-01 e UY-01.

C.19 Campanhas por dia



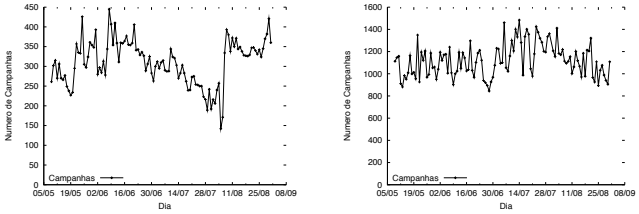
(a) AT-01 (b) AU-01

Figura C.73. Campanhas por dia dos *honeypots* AT-01 e AU-01.



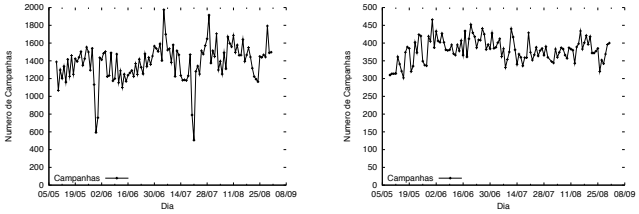
(a) BR-01 (b) BR-02

Figura C.74. Campanhas por dia dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

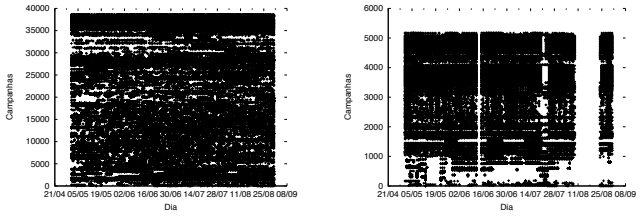
Figura C.75. Campanhas por dia dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

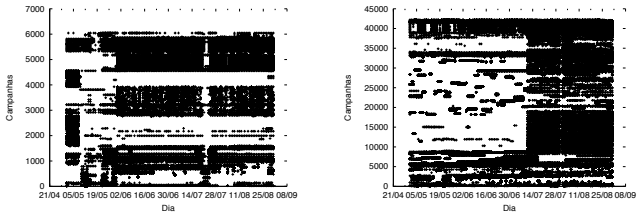
Figura C.76. Campanhas por dia dos *honeypots* TW-01 e UY-01.

C.20 Distribuição das campanhas ativas



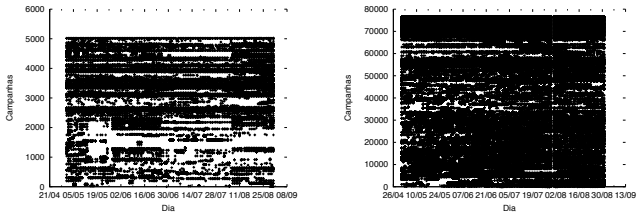
(a) AT-01 (b) AU-01

Figura C.77. Distribuição das campanhas ativas dos *honeypots* AT-01 e AU-01.



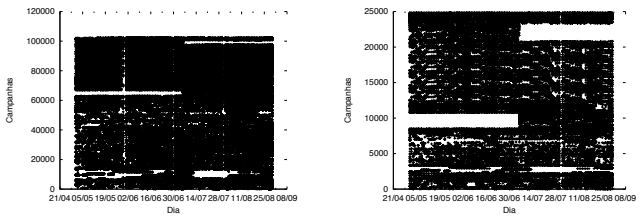
(a) BR-01 (b) BR-02

Figura C.78. Distribuição das campanhas ativas dos *honeypots* BR-01 e BR-02.



(a) EC-01 (b) NL-01

Figura C.79. Distribuição das campanhas ativas dos *honeypots* EC-01 e NL-01.



(a) TW-01 (b) UY-01

Figura C.80. Distribuição das campanhas ativas dos *honeypots* TW-01 e UY-01.