# Caracterização de tráfego SMTP na rede de origem

**Luis Henrique D. Castilho**[1], **Pedro Henrique B. Las Casas**[1], **Mateus D. Dutra**[1], **Saulo M. R. Ricci**[3], **Humberto T. Marques-Neto**[1], **Artur Ziviani**[2], **Dorgival Guedes**[3], **Jussara M. Almeida**[3], **Virgílio A. F. Almeida**[3]

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

[2]Coordenação de Sistemas e Redes
Laboratório Nacional de Computação Científica (LNCC)
25.651-075 - Petrópolis - Brasil

[3]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
31.270-010 - Belo Horizonte - Brasil

{luis.castilho,pedro.casas,mateus.dutra}@sga.pucminas.br

humberto@pucminas.br, ziviani@lncc.br

{saulomrr,dorgival,jussara,virgilio}@dcc.ufmg.br

***Abstract.*** *The large traffic due to unwanted e-mail (spam) that crosses the Internet may consume network resources that could be put to better uses otherwise. Understanding the characteristics of SMTP (Simple Mail Transfer Protocol) traffic in the Internet provider's network is a fundamental task to enable the development of mechanisms to block unwanted messages at their origin. This paper characterizes the SMTP traffic in an Internet provider from approximately 5,500 broadband users during a 28 day period. Results show that metrics such as the rate a user's SMTP transactions and the number of distinct e-mail servers contacted may be used to stop, at their source and without examining the content of the messages, the transmission of unwanted e-mail.*

***Resumo.*** *O grande volume de tráfego de e-mails indesejados (spam) que circulam na Internet consome recursos que poderiam ser melhor utilizados. Entender as características do tráfego SMTP (Simple Mail Transfer Protocol), sob o ponto de vista do provedor de acesso, é tarefa fundamental para propor mecanismos que permitam o bloqueio de mensagens indesejadas na origem. Este artigo apresenta uma caracterização de tráfego SMTP gerado por cerca de 5.500 usuários de um provedor de Internet de banda larga em um período de 28 dias. Os resultados mostram que métricas, tais como o número de transações SMTP realizadas por um usuário por unidade de tempo e o número de servidores de e-mail distintos contactados por ele, podem ser utilizadas para identificar, ainda na rede de origem e sem a necessidade de inspeção do conteúdo da mensagem, tráfego gerado por usuários que fogem ao padrão de comportamento esperado de usuários comuns, permitindo a identificação de possíveis envios de mensagens indesejadas.*

## 1. Introdução

Relatórios recentes [IronPort 2008, MessageLabs 2009] indicam que cerca de 90% das mensagens eletrônicas (*e-mails*) que circulam pela Internet são mensagens indesejadas, muitas vezes caracterizadas como *spam*. Para reduzir o impacto dessas mensagens indesejadas sobre o serviço de correio eletrônico, provedores de correio eletrônico utilizam filtros de e-mails que as classificam, descartam ou colocam em quarentena, evitando que elas sobrecarreguem as caixas postais dos destinatários. Entretanto, esses filtros não evitam o desperdício de recursos da rede, pois as mensagens recebidas geram tráfego nos *links* e consomem CPU para serem encaminhadas e processadas.

Com o propósito de atenuar o uso desnecessário de recursos da Internet com essas mensagens, técnicas de pré-filtragem poderiam ser utilizadas [Schatzmann et al. 2009, Ramachandran et al. 2007, Hao et al. 2009]. Tais técnicas são aplicadas antes da chegada do e-mail ao servidor destino, por exemplo, na rede de origem da mensagem. A pré-filtragem de mensagens indesejadas poderia ser aplicada, por exemplo, por provedores de acesso à Internet de banda larga. A partir da análise do tráfego SMTP originado em suas redes, esses provedores poderiam bloquear mensagens eletrônicas que, conforme características de comportamento de seus remetentes, seriam descartadas pelo servidor de destino. Além de economizar recursos computacionais utilizados desnecessariamente, a pré-filtragem de e-mails pode contribuir para melhoria da reputação de provedores de serviços de Internet, diminuindo a frequência de sua inserção em listas de bloqueio baseadas em endereços IP como, por exemplo, as mantidas pelo Spamhaus[1].

Para isso, uma caracterização de tráfego SMTP (*Simple Mail Transfer Protocol*) de clientes residenciais de um provedor de Internet de banda larga seria o primeiro passo para entender a carga de trabalho gerada por esse tipo de usuário e, assim, permitir o desenvolvimento de mecanismos que permitam o bloqueio de mensagens indesejadas na sua origem, com base na análise do comportamento de seus remetentes, sem a necessidade de inspeção do conteúdo da mensagem. Este trabalho propõe uma metodologia de caracterização que foi aplicada a um conjunto de dados reais contendo informações agregadas e anonimizadas dos usuários de um provedor de Internet de banda larga residencial. O conjunto de dados analisado é formado por cerca de 6,4 milhões de transações SMTP, realizadas por aproximadamente 5,5 mil usuários distintos em um período de 28 dias.

Os resultados da caracterização mostram que métricas utilizadas no estudo, tais como o número de transações SMTP realizadas por um usuário por unidade de tempo e o número de servidores de e-mail distintos contactados por ele podem ser utilizadas para identificar, ainda na rede de origem e sem a inspeção do conteúdo da mensagem, tráfego que foge ao padrão de comportamento esperado de usuários comuns, permitindo a identificação de possíveis envios de mensagens indesejadas. Além disso, a metodologia proposta permite a organização dos usuários SMTP em grupos com características específicas que evidenciam a diferença de comportamento entre uso normal e uso abusivo do serviço de envio de e-mails.

Este trabalho está organizado em cinco seções. Os trabalhos relacionados são discutidos na seção 2. Na seção 3, a metodologia da caracterização é descrita. A seção 4 apresenta e discute os resultados mais relevantes e a conclusão é apresentada na seção 5.

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

## 2. Referencial Teórico

Entre os trabalhos de caracterização de comportamento de usuários de e-mail destaca-se o de [Barabasi 2005]. O estudo aponta que, enquanto muitas ações humanas são aleatoriamente distribuídas ao longo do tempo, sendo bem aproximadas por processos de Poisson, o envio de e-mails é marcado pelo envio de rajadas de mensagens seguido de longos períodos de inatividade. Esse comportamento é consequência de um processo de tomada de decisão baseado em prioridades, o que leva o tempo de chegada dos eventos a ser melhor modelado por distribuições de cauda pesada. Os resultados do estudo são utilizados como base nas análises realizadas neste trabalho.

Na área de caracterização de cargas de trabalho de e-mail, o estudo de [Gomes et al. 2004] busca características que diferenciem *spam* de mensagens legítimas. São analisados o processo de chegada de mensagens, o tamanho das mensagens e a popularidade e a localidade temporal de endereços remetentes. Numa extensão deste trabalho, [Gomes et al. 2005] analisam também uma carga de trabalho de tráfego de e-mail, dessa vez a fim de levantar propriedades de grafos traçados entre remetentes e destinatários. Ambos os estudos encontram diferenças entre *spam* e e-mails legítimos nos aspectos escolhidos para análise. Porém, ambos utilizam dados da camada de aplicação no destino e não da camada de rede, como proposto aqui. O trabalho apresentado em [Gomes et al. 2009], também baseado em dados coletados da camada de aplicação, utiliza o conceito de entropia da comunicação para analisar o comportamento de *spammers* ao longo do tempo. Os resultados apontam que tráfego legítimo apresenta menor entropia que tráfego oportunista, gerado pelo envio de *spam*. Conjectura-se que as diferenças encontradas em todos os três artigos se devem ao comportamento distinto de um usuário legítimo de e-mail, envolvido em relações sociais com os destinatários, e de *spammers*, enviando e-mails indiscriminadamente para seus alvos.

Ainda na área de caracterização, o estudo de [Ramachandran e Feamster 2006] tenta determinar características de tráfego, dessa vez da camada de rede, que sejam comuns a *spammers*. O trabalho analisa características como a persistência de endereços IP e rotas e características específicas de *botnets*[2]. Outro trabalho que busca caracterizar o comportamento de *spammers* na rede é o trabalho de [Calais et al. 2009], que observa os padrões de comunicação presentes em uma campanha de *spam*. Apesar desses dois trabalhos focarem em aspectos de tráfego, eles se baseiam na observação em um ponto interior da rede, enquanto neste trabalho discutimos os padrões de tráfego observados internamente a um provedor de acesso, próximo aos clientes que geram esse tráfego.

Entre os estudos que apresentam técnicas de pré-filtragem de e-mail, o trabalho de [Ramachandran et al. 2007] apresenta um sistema chamado *SpamTracker*. O sistema utiliza uma técnica de *behavioral blacklisting* (bloqueio por comportamento) que classifica o *host* que envia a mensagem de e-mail baseado em seu comportamento e não em sua identidade, como seu IP, por exemplo. Porém, enquanto em [Ramachandran et al. 2007] a identidade e o comportamento de um *host* são determinados pelos domínios para onde o mesmo envia mensagens, o que pode agrupar tráfego de equipamentos de rede distintos numa mesma identidade, neste trabalho analisa-se o comportamento de usuários, identificados unicamente como detalhado na seção 3.

[2] *Botnets* são grupos de computadores infectados por *malware*, chamados neste caso de *bots*, controlados remotamente e utilizados muitas vezes para o envio de *spam* ou para atacar outras redes de computadores.

Por fim, o estudo de [Hao et al. 2009], também apresenta um sistema de pré-filtragem de e-mail e faz um levantamento de características da camada de rede e da camada de aplicação que podem ser utilizadas em pré-filtragem de *spam*. As características analisadas são divididas em características que podem ser obtidas a partir de um único pacote, características que podem ser obtidas a partir de uma única mensagem de e-mail e características agregadas, coletadas ao longo do tempo. O estudo propõe um sistema de reputação baseado nessas características. O trabalho aqui apresentado se relaciona diretamente a este estudo, utilizando inclusive uma métrica proposta pelos autores, explicada em detalhes também na seção 3. Entretanto, a análise desta métrica é consideravelmente distinta nos dois trabalhos. Além disso, os estudos possuem objetivos distintos. Enquanto aquele estudo visava propor um sistema de pré-classificação de mensagens de e-mail como *spam* ou mensagens legítimas, o estudo apresentado aqui visa diferenciar comportamentos dos usuários durante o envio de e-mails.

## 3. Metodologia de Caracterização

A caracterização de tráfego SMTP foi feita sobre duas fontes de dados: (a) o *log* do tráfego de um provedor de Internet banda larga cobrindo o período de 01 a 28 de Março de 2009 (28 dias), e (b) o *log* do serviço de DHCP prestado pelo provedor aos seus assinantes nesse mesmo período. O *log* de tráfego foi coletado por equipamentos da plataforma *Cisco Service Control Engine* (SCE) [Cisco 2008], que contém amostras das transações realizadas através da infra-estrutura do provedor. Uma transação é uma conexão TCP ou um fluxo de dados UDP que é coletado e analisado do ponto de vista das camadas de rede e de aplicação. Nesse processo o equipamento extrai as principais informações sobre a comunicação e armazena os dados com os endereços IP de origem e destino e com o serviço/protocolo sendo utilizado. Os principais campos de uma transação são: data/hora inicial, duração, serviço/protocolo, volume de bytes recebidos, volume de bytes enviados, e endereços IP de origem e destino.

O *log* do serviço de DHCP foi utilizado para identificar os usuários do provedor através do *MAC address* do equipamento utilizado para acessar a Internet. As duas fontes de dados foram integradas com base no endereço IP do assinante e pela data e hora, campos presentes em ambos os *logs*. Vale mencionar que os campos de endereço foram todos anonimizados anteriormente à nossa análise.

Em termos gerais, os logs continham 68,2 milhões de transações associadas a 48,7 mil usuários. Entretanto, foram desconsideradas transações que estavam ativas no início ou no fim do período de coleta (17,45% do total, devido a transações de longa duração) e transações de assinantes não-residenciais (6,09% do total), uma vez que optamos por focar nossa análise em padrões comportamentais de usuários residenciais apenas. Dado o foco deste trabalho, restringimos nossa análise a transações SMTP apenas. Note que especificamente para SMTP, uma única transação (uma conexão) pode ser usada para entregar diversos e-mails a um mesmo servidor. Portanto, este estudo se diferencia de trabalhos anteriores [Gomes et al. 2004] por focar no tráfego em nível de transações e não de mensagens individuais.

Finalmente, optamos por filtrar transações que representavam possíveis erros de coleta, tais como transações com duração nula (0,14%) ou com zero bytes enviados (0,1%) ou recebidos (menos de 0,01%). Também foram removidas transações SMTP

que enviaram menos de 160 bytes (1,89%) ou que receberam menos de 80 bytes (0,02%). Esses limiares foram definidos por corresponderem ao número mínimo de bytes necessário para se estabelecer e encerrar uma conexão TCP (considerando-se 40 bytes para os cabeçalhos IP e TCP nos pacotes do *three-way handshake* e de finalização). Note-se que essas transações consideradas inválidas podem ser indicativas, por exemplo, de uma busca por servidores SMTP ativos realizada por usuários maliciosos ou *bots*. Pretendemos considerar os padrões que emergem dessas transações em trabalhos futuros.

Dentre as transações restantes após a filtragem, foram analisadas as transações SMTP: 6,4 milhões de transações, realizadas por 5,5 mil usuários distintos. Os padrões de comportamento desses usuários, no que tange às transações SMTP por eles realizadas, foram então caracterizados, de acordo com cinco métricas principais: número de transações SMTP realizadas no período de coleta, número de servidores SMTP distintos acessados, tamanho das transações (em bytes enviados), distância geodésica entre os endereços IP de origem e de destino das transações e tempo entre chegadas de transações sucessivas de um usuário no provedor (*inter-arrival time* – IAT).

O número de transações por usuário é útil para distinguir usuários que fazem pouco uso de SMTP daqueles que o utilizam com grande intensidade. Entretanto, o uso isolado dessa métrica para detecção de *spammers* pode levar a altas taxas de falsos-positivos e de falsos-negativos. Além disso, nós consideramos o número de servidores SMTP distintos acessados como uma métrica de interesse. Consideramos que, enquanto o uso de poucos servidores é o esperado para usuários legítimos, o acesso a um número muito grande pode indicar a operação de *open proxies*[3] ou de *open mail relays*[4] sendo explorados para o envio de *spam* [Taveira e Duarte 2008, Calais et al. 2008] por usuários maliciosos ou *bots*.

Considerou-se também o tamanho médio das transações SMTP realizadas por um usuário, uma vez que trabalhos anteriores [Gomes et al. 2004, Taveira e Duarte 2008] indicaram que mensagens de *spam* tendem a ser menores que mensagens legítimas. Mesmo não dispondo do tamanho de cada mensagem individual em uma transação SMTP, espera-se que a análise dos tamanhos das transações SMTP, conjuntamente com as demais métricas, possa explicitar padrões de comportamento diferentes entre os usuários.

Já a escolha da distância geodésica[5] como métrica para análise se baseia na hipótese de que conexões maliciosas, para envio de *spam*, tendem a ocorrer entre IPs mais distantes que conexões legítimas [Hao et al. 2009]. Segundo os autores, conexões legítimas ocorrem como parte de relações sociais existentes, tais como e-mails enviados a colegas de trabalho pelo servidor SMTP da empresa, tendendo assim a percorrerem menores distâncias. Conexões maliciosas por outro lado não são parte de relações sociais, tendendo a ocorrer entre clientes e servidores mais distantes. Para obter a distância geodésica entre dois IPs foi utilizada a base de dados GeoLite Country [MaxMind 2009], que possui em muitos casos apenas a precisão de país, ou seja, a maior parte dos endereços encontrados em um país recebem um mesmo par latitude-longitude.

[3]Servidores HTTP ou SOCKS que servem de intermediários para conexões a outras máquinas.

[4]Servidores SMTP que permitem o envio de e-mails a partir de outras máquinas sem qualquer tipo de autenticação, ou validação.

[5]Menor distância entre dois pontos ao longo da superfície da Terra.

**Tabela 1. Métricas para o conjunto de usuários que realizaram transações SMTP.**

| | Todos os usuários | | |
|---|---|---|---|
| | **Mínimo** | **Média (CV)** | **Máximo** |
| Número de transações SMTP | 1 | 1.160 (6,33) | 306.099 |
| Número de servidores SMTP distintos acessados | 1 | 449 (4,55) | 53.268 |
| Tamanho das transações SMTP (KB) | 0,16 | 500 (7,34) | 170.339 |
| Distância geodésica entre os IPs envolvidos (km) | 0 | 3.965 (0,97) | 17.499 |
| Tempo entre chegadas de transações SMTP (h) | 0,00028 | 41 (1,32) | 318 |

Finalmente, o tempo entre chegadas (IAT) é calculado como o tempo decorrido entre duas transações SMTP consecutivas de um mesmo usuário. Ele indica com que frequência o usuário inicia um envio de mensagens. Pela sua definição, o IAT só pode ser calculado para usuários com pelo menos duas transações SMTP.

## 4. Resultados

Ao analisar os dados coletados, apresentamos primeiramente uma caracterização da carga como um todo, destacando a alta variabilidade dos dados nesse caso. Em seguida apresentamos a aplicação da técnica de agrupamento.

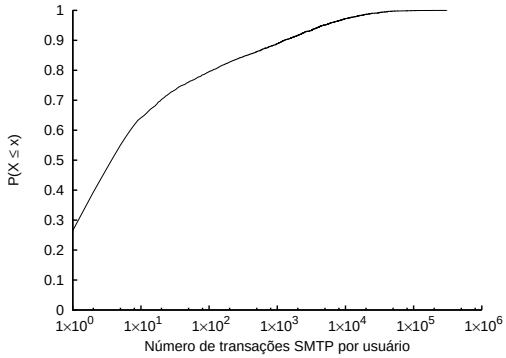
### 4.1. Análise Geral da Carga de Trabalho

Esta seção analisa as cinco métricas escolhidas como chave para compreender quantitativamente e qualitativamente a carga de trabalho de envio de e-mail, aplicadas aqui aos usuários que realizaram atividade SMTP. Todas as métricas foram calculadas para todos os usuários que realizaram pelo menos uma transação SMTP ao longo do período de coleta, exceto pelo cálculo do IAT, que só faz sentido para usuários que realizaram pelo menos duas transações.

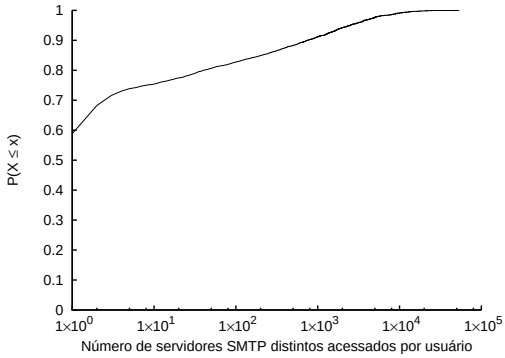
Analisando-se a Figura 1(a), que apresenta a função de distribuição acumulada ou CDF (*Cumulative Distribution Function*) da métrica número de transações SMTP, observa-se que 65% dos usuários realizaram 10 transações SMTP ou menos no período de coleta. Pouco mais de 10% dos usuários realizam mais de 1.000 transações SMTP e menos de 5% realizam mais de 10.000. Os usuários que fazem uso intenso de SMTP são responsáveis pela alta média de 1,1 mil transações SMTP por usuário, apresentada na Tabela 1. O CV de 6,33 dessa média e a enorme diferença entre o número mínimo (1) e máximo (306.099) de transações SMTP indica a alta variabilidade dessa métrica no conjunto de dados. Percebe-se que os usuários possuem comportamentos bem distintos no uso de SMTP, o que reforça a divisão dos usuários em grupos, como proposto.

Analisando-se a distribuição do número de servidores SMTP distintos acessados, apresentada na Figura 1(b), vemos que cerca de 70% dos usuários utilizaram 5 ou menos servidores SMTP distintos em suas transações, pouco menos de 10% utilizaram mais de 1.000 servidores SMTP distintos e, destes, apenas cerca de 1% utilizaram mais de 10.000 servidores distintos durante o período da amostra. Observa-se na Tabela 1 os altos valores da média (449 servidores acessados) e do CV (4,55) e a grande diferença entre o número mínimo (1) e máximo (53.268) de servidores.

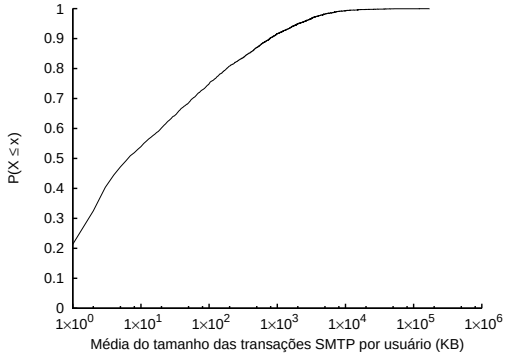
A CDF do tamanho das transações SMTP, apresentada na Figura 1(c), mostra que cerca de 55% dos usuários realizam transações SMTP menores que aproximadamente 10 KB e que menos de 10% realizam transações SMTP maiores que aproximadamente 1 MB.



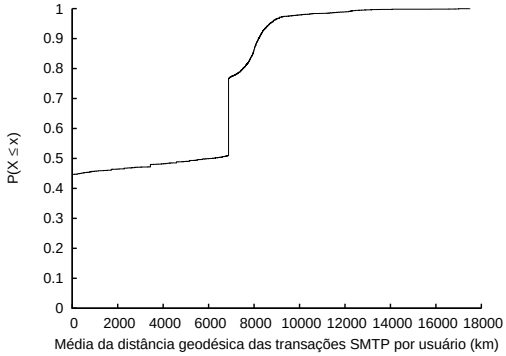

(a) Número de transações SMTP.



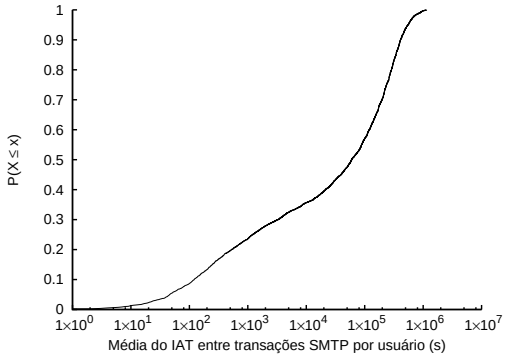

(b) Número de servidores SMTP acessados.




(c) Tamanho das transações SMTP.




(d) Distância geodésica das transações SMTP.




(e) IAT das transações SMTP.

**Figura 1. CDFs para o conjunto de usuários que realizaram transações SMTP.**

Menos de 1% dos usuários realizam transações SMTP maiores que 10 MB. A Tabela 1 mostra novamente a alta variabilidade dos dados por usuário: média de 500 KB, CV de 7,34 e diferença de 166MB entre o menor e o maior tamanho médio por usuário.

Observa-se na Tabela 1 que a média da distância geodésica por usuário é de 3.965 km, com um CV de 0,97. A Figura 1(d), que apresenta a função de distribuição dessa métrica, mostra que cerca de 45% dos usuários possuem média de distância geodésica igual a 0 km, o que significa que esses usuários realizam transações SMTP apenas em servidores brasileiros. A Figura mostra ainda que cerca de 25% dos usuários possuem média de distância geodésica entre 6.000 km e cerca de 7.000 km. Esse degrau no gráfico é devido à baixa resolução da base de dados utilizada, como mencionado anteriormente.

**Tabela 2. Métricas para os grupos de uso baixo, médio e intenso de SMTP.**

| | Uso baixo de SMTP | | Uso médio de SMTP | Uso intenso de SMTP |
|---|---|---|---|---|
| | Subgrupo 1 | Subgrupo 2 | | |
| | Média (CV) | Média (CV) | Média (CV) | Média (CV) |
| Número de transações SMTP | 1,00 (0,00) | 3,57 (0,45) | 911,20 (2,85) | 38.056,82 (1,01) |
| Número de servidores SMTP distintos acessados | 1,00 (0,00) | 1,33 (0,52) | 425,94 (2,43) | 12.462,84 (0,60) |
| Tamanho das transações SMTP (KB) | 641,48 (4,35) | 855,67 (5,40) | 329,59 (11,35) | 1,71 (0,59) |
| Distância geodésica entre os IPs envolvidos (km) | 2.640,67 (1,32) | 2.171,06 (1,43) | 5.061,49 (0,74) | 8.414,93 (0,15) |
| Tempo entre chegadas de transações SMTP (h) | Não se aplica | 120,88 (0,41) | 16,11 (1,25) | 0,02 (0,57) |

Cerca de 1% dos usuários possuem média de distância geodésica maior que 12.000 km. A alta variabilidade dos dados agregados, responsável pela média pouco representativa (3.965 km) e pelo CV alto (0,97), não permite conclusões sobre a validade da métrica, o que justifica novamente a análise dos usuários divididos em grupos distintos.

A CDF do IAT das transações SMTP no provedor de acesso é apresentada na Figura 1(e). Observa-se que cerca de 25% dos usuários possuem IAT médio menor que aproximadamente 15 min (1.000 s), que cerca de 35% possuem IAT médio menor que cerca de 2 horas e meia (10.000 s), que mais de 40% dos usuários possuem IAT médio menor que aproximadamente 1 dia e 3 horas (100.000 s). A Tabela 1 apresenta a alta média do IAT (41 horas) quando todos os usuários que realizaram atividade SMTP são considerados em conjunto. Os usuários que realizaram poucas transações ao longo do período de coleta influenciam diretamente esta média, devido ao grande valor de IAT associado a eles.

### 4.2. Análise do Agrupamento dos Usuários

Esta seção analisa os grupos de usuários (*clusters*) resultantes do processo de agrupamento (*clustering*), realizado com o uso de todas as métricas propostas em conjunto. Na busca de comportamentos distintos entre os grupos indicados pelo método X-means e analisando-se a média e o CV das métricas calculadas para cada grupo na Tabela 2, pode-se perceber que os dois primeiros grupos, respectivamente com 1.452 (26,5%) e 973 (17,8%) usuários, possuem métricas com valores muito próximos, além de bem distintos dos outros dois grupos. A razão para a separação em dois grupos nesse caso foi aparentemente o peso do agrupamento de usuários que tiveram apenas uma transação SMTP no primeiro grupo. Com base nessa análise inicial, decidimos considerar os dois primeiros grupos como subgrupos de um mesmo grupo maior, representando todos os usuários de uso baixo de SMTP. Os dois grupos restantes representam usuários de uso médio (2.958 usuários, 53,9% do total) e intenso (96 usuários, 1,8% do total) de SMTP, respectivamente. A caracterização apresentada nesta seção tem como foco estes três grupos. Assim, é possível analisar o vetor de métricas dos usuários sob uma nova perspectiva, visando dessa vez caracterizar o grupo ao qual o usuário faz parte. Com essa caracterização dos grupos de uso baixo, médio e intenso de SMTP pode-se analisar quão efetivas são as métricas adotadas na diferenciação de comportamentos distintos de envio de e-mail.

#### 4.2.1. Grupo de Uso Baixo de SMTP

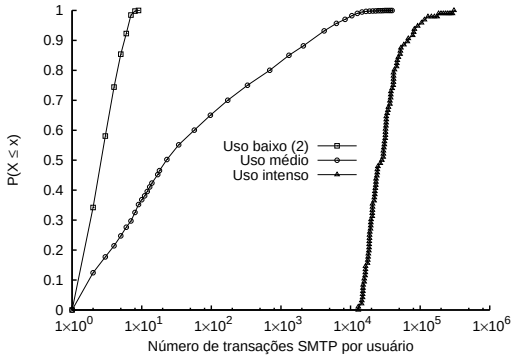
O grupo de uso baixo de SMTP é o grupo mais uniforme, com menor variabilidade, entre os três. Observa-se mais uma vez na Tabela 2 que em média seus usuários realizam poucas transações SMTP (1,0 ou 3,6), acessam poucos servidores SMTP dis-

tintos (1,00 ou 1,33) e executam suas transações com longos intervalos de inatividade entre elas (121 horas de inatividade). A distância geodésica média entre os IPs origem e destino e, principalmente, o tamanho médio das transações SMTP possuem uma variabilidade maior nesse grupo.

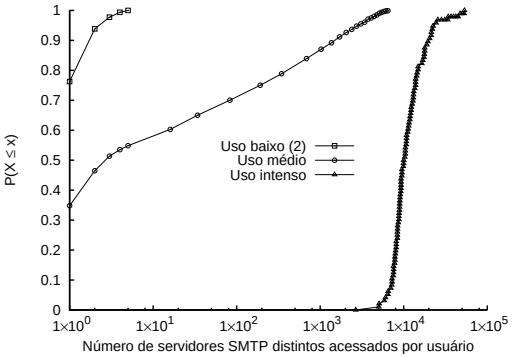
Como mencionado anteriormente, o primeiro subgrupo é formado exclusivamente pelos usuários que realizaram uma única transação SMTP ao longo do período de coleta. Por esse motivo, não há uma curva para este subgrupo nas Figuras 2(a), 2(b) e 2(e). A distância geodésica média desse subgrupo, mostrada na figura 2(d), possui um comportamento simples, já que com apenas uma transação, essa pode ocorrer dentro do Brasil (mais de 60% dos casos) ou fora dele (onde há pequenas variações na distância, mas provavelmente a maioria dos acessos é direcionada a servidores nos EUA[6]). Esses acessos podem ser explicados pelo número de usuários que utilizam grandes provedores de serviço de e-mail como Gmail, Yahoo!, Hotmail, entre outros, que possuem servidores localizados principalmente nos EUA. Menos de 1% dos usuários possuem distância geodésica média maior que 7.000 km, com menos de 0,1% destes chegando a 17.500 km. O tamanho médio das transações SMTP desse subgrupo é a métrica com maior variabilidade, apresentando um CV de 4,35. O tamanho médio das transações varia de 1 KB, correspondendo ao envio de mensagens curtas que contêm apenas texto, a mais de 44 MB, que poderia representar o envio de mensagens com anexos maiores, como fotos ou vídeos. Pela Figura 2(c) tem-se que mais de 50% dos usuários desse subgrupo efetuam transações SMTP menores que cerca de 10 KB, que cerca de 90% efetuam transações menores que 1 MB e que menos de 5% dos usuários efetuam transações maiores que aproximadamente 10 MB. Percebe-se que mesmo efetuando uma única transação SMTP ao longo do período de coleta, os usuários desse subgrupo possuem comportamentos bem distintos quanto ao tamanho das transações efetuadas.

O segundo subgrupo também é formado por usuários que realizam poucas transações SMTP. O número de transações SMTP realizadas e o número de servidores SMTP distintos acessados, que não possuem variabilidade no subgrupo anterior, aqui apresentam baixa variabilidade (CVs de 0,45 e 0,52, respectivamente). Como as médias das duas métricas são baixas, esse CV não é significativo, sendo possível concluir que o subgrupo efetua poucas transações SMTP, utilizando poucos servidores SMTP distintos. Já o intervalo entre chegadas, varia de mais de 1 dia (10.000 s) a mais de 11 dias (100.000 s), como observado na Figura 2(e). A distância geodésica média desse subgrupo possui uma variabilidade semelhante à do subgrupo anterior (CV de 1,43), variando entre 0 e 9.900 km. Pela Figura 2(d) pode-se observar, assim como no subgrupo anterior, que a distância geodésica média de cerca de 65% dos usuários é igual a 0 km e que a de mais de 99% é menor que 6.900 km, o que indica novamente acesso a servidores brasileiros e a servidores nos EUA. A distância geodésica média dos demais, menos de 1% dos usuários, é menor que 10.000 km. O tamanho médio das transações SMTP, novamente como no subgrupo anterior, apresenta um CV alto (5,4) e, analisando a Figura 2(c), varia de 1 KB a aproximadamente 100 MB. Analisando as duas curvas pode-se observar que os usuários do segundo subgrupo possuem médias de tamanho maiores em geral, mas que as duas curvas convergem. De qualquer forma, percebe-se uma variabilidade grande de comportamento quanto ao tamanho das transações efetuadas, o que pode ser devido à utilização

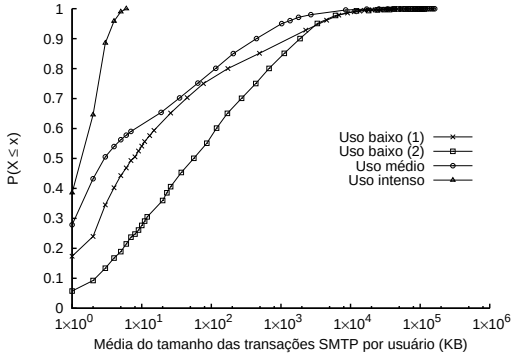
[6]A distância geodésica entre o Brasil e os EUA, pela base de dados utilizada, é de 6.877 km.



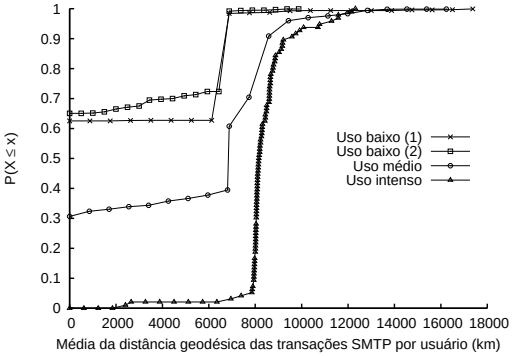

(a) Número de transações SMTP.



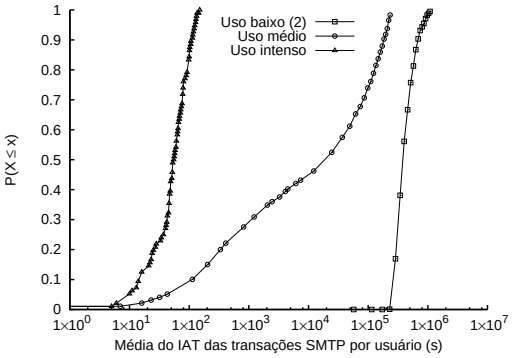

(b) Número de servidores SMTP acessados.




(c) Tamanho das transações SMTP.




(d) Distância geodésica das transações SMTP.




(e) IAT das transações SMTP.

**Figura 2. CDFs das métricas por grupo de uso baixo, médio e intenso de SMTP.**

de muitos anexos grandes, fato normalmente associado a usuários não *spammers*.

### 4.2.2. Grupo de Uso Médio de SMTP

O grupo de uso médio de SMTP é o grupo com maior número de usuários e com maior variabilidade nos dados. Pela Tabela 2 nota-se que os CVs do número de transações SMTP (2,85), do número de servidores SMTP distintos acessados (2,43) e do tamanho médio das transações SMTP (11,35) são os maiores do grupo. Os CVs da distância geodésica média (0,74) e do IAT médio (1,25) são menores, mas ainda assim significativos, como mostrado abaixo.

Pela Figura 2(a), percebe-se que o número de transações SMTP fica abaixo de 10 para cerca de 35% dos usuários desse grupo, o que representa menos de 1 transação SMTP por dia no período de 28 dias analisado. Porém, cerca de 65% dos usuários do grupo efetuam 100 transações SMTP ou menos no mesmo período, ou seja, mais de 3 transações SMTP por dia. Pouco menos de 20% dos usuários efetuam mais de 1.000 transações SMTP nesse período, sendo que menos de 5% destes realizam mais de 10.000 transações SMTP, ou seja, mais de 350 transações SMTP por dia em média. Conclui-se que o grupo não apresenta uniformidade quanto ao número de transações SMTP, parecendo agregar comportamentos possivelmente legítimos e comportamentos possivelmente abusivos.

O número de servidores SMTP distintos acessados, observando-se a Figura 2(b), possui uma variação similar. Mais de 55% dos usuários utilizam 10 servidores SMTP ou menos, o que poderia ser considerado uso legítimo de serviço de envio de e-mail. Porém, pouco menos de 30% dos usuários desse grupo utilizam 100 servidores SMTP ou mais, o que passa a ser um comportamento suspeito. E menos de 15% dos usuários utilizam 1.000 servidores SMTP distintos ou mais, o que pode ser compreendido como comportamento abusivo, já que espera-se que usuários legítimos de e-mail não utilizem tal número de servidores para envio de suas mensagens. Novamente, o grupo não apresenta uniformidade quanto à métrica.

O tamanho das transações SMTP apresenta uma variação similar à do grupo de uso baixo de SMTP, porém, mais acentuada devido à presença de valores extremos. Pela Figura 2(c) tem-se que mais de 60% dos usuários efetuam transações SMTP com tamanho médio inferior a aproximadamente 10 KB e que cerca de 95% efetuam transações SMTP menores que 1 MB. Os demais 5% realizam transações SMTP de tamanho médio entre 1 MB e mais de 100 MB, o que faz com que o CV da média seja alto.

Quanto à média da distância geodésica dos IPs envolvidos nas transações SMTP deste grupo, percebe-se, analisando a Figura 2(d), que cerca de 30% dos usuários possuem média de distância geodésica igual a 0 km, indicando o uso de servidores SMTP brasileiros, e que 60% dos usuários possuem média de distância geodésica menor que 6.900 km, indicando o uso de servidores SMTP localizados nos EUA. Cerca de 90% possuem média de distância geodésica menor que 9.000 km e mais de 95% possuem média menor que 10.000 km.

Por fim, a análise do tempo entre chegadas das transações SMTP deste grupo, com base na Figura 2(e), revela que mais de 45% dos usuários possuem média de IAT menor que aproximadamente 2h e meia (10.000 s) e que cerca de 75% possuem média menor que 1 dia e 3 horas (100.000 s). Diferentemente do grupo de uso baixo de SMTP, a maior parte dos usuários desse grupo já apresentam IAT menor que aproximadamente 1 dia.

### 4.2.3. Grupo de Uso Intenso de SMTP

O grupo de uso intenso de SMTP é formado por apenas 96 usuários, 1,75% do total de usuários que efetuaram transações SMTP no período analisado. A Tabela 2 indica que o grupo possui baixa variabilidade, com os CVs das métricas variando de 0,15 a 1,01. As médias desse grupo são bem distintas das médias dos demais grupos, sendo consideravelmente maiores, como no caso das contagens de transações SMTP e de servidores

SMTP distintos acessados, ou menores, como no caso do tamanho médio das transações SMTP e do IAT médio dessas transações.

O número de transações SMTP efetuadas pelos usuários desse grupo já indica o comportamento potencialmente abusivo dos mesmos. Analisando a Figura 2(a), percebe-se que todos os usuários do grupo efetuam mais de 10.000 transações SMTP no período de 28 dias analisado, equivalente a efetuar mais de 350 transações SMTP diariamente. Cerca de 5% efetuam mais de 100.000 transações SMTP. Destes 5%, dois usuários se destacam, efetuando 184.084 e 306.099 transações SMTP no período de coleta. O alto número de transações SMTP dos usuários deste grupo pode ser indicativo de infecção por *malware*, fazendo com que a máquina do usuário se comporte como um *bot* de envio de *spam*, ou de comportamento abusivo do próprio usuário.

O número de servidores SMTP distintos acessados também indica comportamento abusivo. Pela Figura 2(b), percebe-se que todos os usuários utilizam mais de 1.000 servidores SMTP distintos e que cerca de 50% utilizam 10.000 servidores ou mais. Os dois usuários destacados acima utilizam, respectivamente, 47.435 e 53.268 servidores SMTP distintos, sendo os usuários que possuem maior número de servidores SMTP utilizados. Como dito anteriormente, o uso de diversos servidores SMTP distintos pode indicar abuso de *open proxies* e *open mail relays* para envio de *spam*.

O grupo de uso intenso de SMTP é o único grupo que apresenta baixa variabilidade no tamanho das transações SMTP efetuadas. Considerando o comportamento aparentemente abusivo dos usuários desse grupo, apontado pelas demais métricas analisadas, esse resultado encontra apoio na literatura [Gomes et al. 2004, Taveira e Duarte 2008], que indica que mensagens de *spam* tendem a ser menores em tamanho que mensagens legítimas. O resultado reforça a suspeita de comportamento abusivo por parte dos usuários deste grupo. Pela Figura 2(c) percebe-se que o tamanho médio das transações SMTP deste grupo varia de 1 KB a menos de 10 KB apenas, com um CV de 0,59, como apresentado na Tabela 2.

A média da distância geodésica dos IPs envolvidos nas transações SMTP deste grupo apresenta menor variabilidade e possui a maior média dos três grupos analisados. Cerca de 50% dos usuários possuem média da distância geodésica pouco maior que 8.000 km, enquanto 90% possuem média menor que cerca de 9.500 km. Menos de 5% dos usuários possuem média menor que 12.000 km, sendo esse o maior valor encontrado no grupo. Considerando a suspeita de envio de *spam* por usuários do grupo, estes resultados reforçam a hipótese apresentada em [Hao et al. 2009], de que *spam* tende a percorrer maiores distâncias que e-mails legítimos.

Como consequência do alto número de transações SMTP efetuadas pelos usuários deste grupo, pode-se observar pela Tabela 2 e pela Figura 2(e) os baixos valores de IAT e a baixa variabilidade da métrica no grupo. Cerca de 85% dos usuários possuem IAT médio menor que 1 minuto e meio (100 s) e nenhum usuário do grupo possui IAT médio maior ou igual a 2 minutos e meio. Isso significa que cada usuário deste grupo executa uma nova transação SMTP a cada 1 ou 2 minutos.

## 5. Conclusões

Neste trabalho foi apresentada e aplicada uma metodologia de caracterização hierárquica do uso do protocolo SMTP dos usuários de um provedor de acesso à Internet

de banda larga. A metodologia consiste na análise de um conjunto de métricas da carga de trabalho de forma a diferenciar comportamentos de interesse. No caso, busca-se diferenciar comportamentos legítimos de envio de e-mail de comportamentos suspeitos ou abusivos, analisando apenas dados obtidos da camada de rede de conexões SMTP. A metodologia proposta foi aplicada em um conjunto significativo de dados reais de um provedor de acesso à Internet. A carga de trabalho do provedor foi analisada como um todo e em seguida foi utilizado um algoritmo de agrupamento para dividir os usuários segundo seus padrões de comportamento, quando foram identificados três perfis claramente distintos: usuários uso baixo, médio e intenso de SMTP.

A análise dos grupos de usuários sugere que as métricas escolhidas são eficazes na distinção entre os comportamentos procurados. O grupo de uso baixo de SMTP, que compreende 44,3% dos usuários e menos de 1% do total de transações, possui comportamento aparentemente legítimo no envio de mensagens: realizando transações SMTP com longos períodos de inatividade entre elas (como indica o trabalho de [Barabasi 2005]), utilizando poucos servidores distintos e possuindo tamanhos de transação variáveis. O grupo de uso médio de SMTP, com 53,9% dos usuários e 42,4% do total de transações, possui um comportamento indefinido, parecendo agregar comportamentos legítimos e abusivos num mesmo grupo. Por fim, o grupo de uso intenso de SMTP, com 1,8% dos usuários e 58% das transações, possui comportamento aparentemente abusivo: realizando grandes quantidades de transações, em milhares de servidores SMTP distintos, fazendo de transações pequenas (o que pode ser indicativo de envio de *spam*, segundo os estudos de [Gomes et al. 2004, Taveira e Duarte 2008]) e com suas transações percorrendo grandes distâncias geodésicas (outro indicativo de envio de *spam*, segundo [Hao et al. 2009]). Logo, foi possível, com o uso das métricas escolhidas, apontar comportamentos distintos na carga de trabalho, que podem ser interpretados como uso legítimo de serviço de envio de e-mail e uso suspeito ou abusivo deste serviço.

Melhorias e extensões dessa metodologia podem contribuir para a pré-filtragem de conexões SMTP abusivas, como as utilizadas para o envio de *spam* e de *malware*. Como trabalho futuro propõe-se a busca de métricas que possam diminuir a porcentagem de usuários no grupo heterogêneo citado acima, assim como o desenvolvimento de um algoritmo que use estas métricas para realizar uma pré-identificação e filtragem de *spam*.

## Agradecimentos

Esta pesquisa é parcialmente financiada pelo Instituto Nacional de Ciência e Tecnologia para a Web - INCTWeb (MCT/CNPq 573871/2008-6), pelo Projeto REBU (CTInfo/CNPq 55.0995/2007-2), pela FAPEMIG, pela FAPERJ e pelo Fundo de Incentivo à Pesquisa da PUC-Minas (FIP-2009/3504-S1).

## Referências

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211.

Calais, P. H., Pires, D. E. V., Guedes, D. O., Jr., W. M., Hoepers, C., e Steding-Jessen, K. (2008). A Campaign-based Characterization of Spamming Strategies. Em *Proceedings of the Fifth Conference on Email and Anti-Spam - CEAS 2008*, pág. 1–10, Mountain View, CA, USA. CEAS.

Calais, P. H., Pires, D. E. V., Guedes, D. O., Jr., W. M., Hoepers, C., Steding-Jessen, K., e Chaves, M. (2009). Caracterização de Encadeamento de Conexões para Envio de Spams. Em *Anais do XXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2009)*, pág. 1–14, Recife.

Cisco (2008). Cisco Service Control Application for Broadband Reference Guide. Disponível em: `http://www.cisco.com/en/US/docs/cable/serv_exch/serv_control/broadband_app/rel316/scabbrg/scabbrg.pdf`.

Gomes, L. H., Almeida, R. B., Bettencourt, L. M. A., Almeida, V., e Almeida, J. M. (2005). Comparative Graph Theoretical Characterization of Networks of Spam and Legitimate Email. Em *Proceedings of the Second Conference on Email and Anti-Spam - CEAS 2005*, Stanford, CA, USA. CEAS.

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

Gomes, L. H., Cazita, C., Almeida, J. M., Almeida, V., e Meira, Jr., W. (2004). Characterizing a Spam Traffic. Em *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pág. 356–369, New York, NY, USA. ACM.

Hao, S., Syed, N. A., Feamster, N., Gray, A., e Krasser, S. (2009). Detecting Spammers with SNARE: Spatio-temporal Network-level Automatic Reputation Engine. Em *Usenix Security '09*, Montreal, QC, Canada. USENIX Association.

IronPort (2008). 2008 Internet Security Trends. Disponível em: `http://www.ironport.com/pdf/Trends_Report_IronPort_2008.pdf`.

MaxMind (2009). GeoLite Country Database. Disponível em: `http://www.maxmind.com/app/geoip_country`.

MessageLabs (2009). MessageLabs Intelligence: May 2009. Disponível em: `http://www.messagelabs.com/mlireport/MLIReport_2009_05_May_FINAL.pdf`.

Ramachandran, A. e Feamster, N. (2006). Understanding the network-level behavior of spammers. *SIGCOMM Comput. Commun. Rev.*, 36(4):291–302.

Ramachandran, A., Feamster, N., e Vempala, S. (2007). Filtering Spam with Behavioral Blacklisting. Em *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*, pág. 342–351, New York, NY, USA. ACM.

Schatzmann, D., Burkhart, M., e Spyropoulos, T. (2009). Inferring Spammers in the Network Core. Em *PAM '09: Proceedings of the 10th International Conference on Passive and Active Network Measurement*, pág. 229–238, Berlin, Heidelberg. Springer-Verlag.

Taveira, D. e Duarte, O. (2008). A Monitor Tool for Anti-Spam Mechanisms and Spammers Behavior. Em *2008 IEEE Network Operations and Management Symposium Workshops - NOMS 08*, pág. 101–108, Piscataway, NJ, USA. IEEE Computer Society.