Caracterização hierárquica do comportamento dos usuários de sistemas par-a-par na Internet de banda larga

Humberto T. Marques-Neto[1], Emanuel V. do Valle[2], Luis Henrique Castilho[1], Jussara M. Almeida[2], Virgilio A. F. Almeida[2]

[1] Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte - Brasil

[2]Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - Brasil

humberto@pucminas.br, luis.castilho@sga.pucminas.br,
{vianna,jussara,virgilio}@dcc.ufmg.br

Abstract. *Broadband Internet access has been growing in the last years. In order to better manage their resources, it is important for ISPs (Internet Service Providers) to understand the workload generated by users. This paper presents a methodology to characterize broadband Internet user behavior, applied to recent real traffic logs. The characterization of 1.88 million sessions was done in three separated hierarchical levels. The results show a disproportional usage of ISP resources, since less than 3% of all sessions, those who make extensive use of peer-to-peer (P2P), are responsible for almost 58% of the incoming traffic and 74% of the outgoing traffic. It has also been shown that these same sessions are about 12 times longer than sessions without P2P.*

Resumo. *Entender as características da carga de trabalho é uma tarefa fundamental para o provedor de acesso à Internet de banda larga melhorar o gerenciamento da sua infra-estrutura. Este artigo apresenta uma metodologia para caracterização do comportamento do usuário da Internet de banda larga, bem como analisa e discute os resultados de sua aplicação com dados reais de tráfego recente. A caracterização de 1,88 milhão de sessões foi realizada em três níveis hierárquicos. Os resultados mostram que menos de 3% das sessões, as que mais usam sistemas par-a-par (P2P), são responsáveis por cerca de 58% do tráfego de chegada e 74% do tráfego de saída. Além disso, identificou-se que essas sessões são 12 vezes mais longas que sessões sem P2P.*

1. Introdução

O crescimento futuro da Internet banda larga, disponível tanto através de redes a cabo quanto através de redes DSL (*Digital Subscriber Line*), dependerá em quão efetivos serão os provedores de acesso, aqui denominados ISPs (*Internet Service Providers*), na tarefa de gerenciarem os seus recursos. Usuários de Internet de banda larga querem *downloads* em alta velocidade, grande disponibilidade de recursos do ISP e liberdade para fazer qualquer tipo de requisição ou para executar qualquer aplicação, sem filtros ou restrições. Por outro lado, os provedores de acesso precisam evitar a sobrecarga de suas redes e recuperar gastos e investimentos a partir da otimização de uso de seus recursos.

Para otimizar o uso de seus recursos, os ISPs podem controlar o tráfego de seus usuários, regulando as requisições a certos serviços, como, por exemplo, a transferência de arquivos com sistemas par-a-par (P2P). Tanto a mídia especializada [Cerf 2008, Goth 2008] quanto a convencional (por exemplo, New York Times) vêm dando uma atenção considerável para o problema referente ao aumento do uso de sistemas P2P, enfrentado pelos provedores de acesso. A caracterização do comportamento dos usuários pode contribuir para um melhor entendimento da interação dos usuários desses serviços com os ISPs, o que poderia ajudar os provedores a gerenciarem melhor a capacidade de seus recursos provendo uma melhor qualidade do serviço prestado.

Este artigo apresenta uma caracterização hierárquica do comportamento dos usuários de sistemas par-a-par em um ISP de banda larga, uma companhia de TV a cabo brasileira que também provê serviços de banda larga. O comportamento do usuário é definido como uma função do modo como os usuários chegam ao ISP, o tempo que eles ficam on-line, o volume de bytes transferidos e o que eles fazem enquanto estão conectados. Uma metodologia de caracterização foi proposta e aplicada com um conjunto significativo de dados reais coletados na infra-estrutura do provedor, cujos conteúdos permitem a organização da carga de trabalho em sessões de usuários. Tais sessões são classificadas com base na presença ou não de transações de protocolos P2P, mais especificamente eDonkey/eMule e BitTorrent. As análises foram realizadas tanto com o conjunto geral de sessões quanto com sub-conjuntos de sessões, permitindo realizar uma caracterização hierárquica em três níveis: (1) todas as sessões, (2) sessões não-P2P vs. sessões P2P e (3) sessões *light*-P2P vs. *heavy*-P2P.

Os resultados da caracterização mostram os padrões de uso diário das sessões dos usuários que não utilizam sistemas P2P e apontam uma distribuição injusta de banda entre estes usuários e aqueles que utilizam aplicações P2P. Menos de 3% das sessões, ou melhor, as sessões *heavy*-P2P, são responsáveis por cerca de 58% de todo o tráfego de chegada ao provedor e quase 74% do tráfego geral de saída. Além disso, identificou-se que as sessões com muitas requisições de P2P são 12 vezes mais longas do que sessões que não contêm esse tipo de aplicação. Os aspectos analisados da carga de trabalho apresentam resultados semelhantes aos encontrados na literatura quando as análises são realizadas com o conjunto geral de sessões. Porém, analisando a carga de trabalho com maior granularidade, foram encontradas diferenças nas distribuições estatísticas que caracterizam seus diferentes aspectos.

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

2. Trabalhos Relacionados

Além de consolidar e ampliar a utilização de aplicações como o correio eletrônico e a navegação na rede (*browsing*), o acesso à Internet através de redes de banda larga também promove o crescimento do uso de outras aplicações. Videoconferência, TV interativa, jogos, sistemas P2P, aplicações para transmissão e recepção de vídeo pela rede e aplicações que permitem a comunicação em tempo real entre os usuários que estão *on-line* (*Instant Messengers* e *VoIP*) são exemplos de aplicações que passam a ser mais utilizadas

em decorrência das características dessa tecnologia de acesso à Internet [MIT 2005]. Geralmente, a Internet de banda larga está disponível em redes a cabo, de propriedade de empresas de TV por assinatura, e também em redes DSL (*Digital Subscriber Line*), construídas sobre a infra-estrutura da rede de telefonia fixa das companhias de telecomunicações [Dischinger et al. 2007].

Alguns estudos, como [Fukuda et al. 2005] e [Lakshminarayanan et al. 2004], mostram um relacionamento entre a popularização dos sistemas P2P e o aumento da taxa de penetração da Internet de banda larga, particularmente no Japão e nos Estados Unidos. Como muitas dessas aplicações impõem uma carga de trabalho caracterizada por sessões de longa duração com um tráfego de dados intenso e contínuo, os provedores de banda larga precisam otimizar o uso dos seus recursos para cumprir os acordos de níveis de serviço (SLA – *Service Level Agreement*) estabelecidos com seus clientes. Todavia, a construção desse SLA depende do conhecimento que o ISP tem das características do comportamento de seus clientes.

Existem diversos estudos na literatura sobre Internet que apresentam caracterizações de cargas de trabalho. Alguns analisam cargas de trabalho tradicionais, compostas por acessos a documentos, imagens e domínios presentes na *Web* [Arlitt 2000, Barford et al. 1999], enquanto outros caracterizam cargas de trabalho de serviços mais específicos, tais como, distribuição de mídia sob-demanda e ao vivo [Costa et al. 2004, Veloso et al. 2006], sistemas P2P [Gummadi et al. 2003, Hamada et al. 2004, Sen and Wang 2004], *Web Proxy* [Arlitt et al. 1999] e, mais recentemente, IPTV [Cha et al. 2008].

Entretanto, estudos recentes com uma caracterização do tráfego geral da Internet banda larga ainda são escassos na literatura. O trabalho de [Dischinger et al. 2007] analisa algumas características do serviço oferecido por provedores de banda larga na América do Norte e na Europa. Apesar dos autores apresentarem medições de propriedades, tais como, capacidade da conexão, tempo de *round-trip* (RTT) e *jitter* dos pacotes, taxa de perda de pacotes, tamanho da fila e políticas de descarte de pacotes de 1.894 usuários residencias de banda larga, o estudo não caracteriza as sessões desses usuários por não disporem de dados de tráfego reais coletados da infra-estrutura de um ISP.

A partir da caracterização do comportamento dos usuários da Internet de banda larga passa a ser possível propor mecanismos mais justos de controle de tráfego que promovam o bem-estar coletivo no contexto do provedor de acesso e melhorar métricas para avaliação da qualidade do serviço percebido pelo usuário, tais como, desempenho, disponibilidade de acesso, segurança e custo.

3. Metodologia de caracterização

Esta seção apresenta uma metodologia de caracterização do comportamento de usuários da Internet de banda larga sob o ponto de vista de um provedor de acesso. Entender as características do comportamento desses usuários é uma tarefa que pode melhorar a qualidade de serviço do ambiente criado pela Internet de banda larga e, além disso, contribuir para o desenvolvimento e evolução das aplicações utilizadas nesse ambiente. O objetivo principal da metodologia é delinear um processo sistemático para analisar as atividades dos usuários enquanto estão conectados na infra-estrutura do ISP, quantificando e qualificando a carga de trabalho gerada por eles.

A metodologia proposta contempla a análise de sete aspectos chave da carga de trabalho de um ISP de banda larga utilizados para a identificação de características do comportamento dos usuários e também do tráfego gerado. São eles: (i) processo de chegada das sessões dos usuários à infra-estrutura do ISP, (ii) processo de saída das sessões dos usuários do ISP, (iii) duração das sessões, (iv) bytes recebidos durante as sessões dos usuários, (v) bytes enviados nas sessões, (vi) os principais serviços e (vii) atividades de comércio eletrônico utilizadas na Internet de banda larga. Os processos de chegada e de saída de sessões, assim como a duração dessas sessões, provêm informações sobre o aspecto temporal da carga de trabalho gerada pelos usuários. Já o volume de tráfego, a popularidade dos serviços e a classificação das requisições de comércio eletrônico provêm e qualificam a carga de trabalho gerada pelos usuários.

Para realizar a caracterização foram utilizadas duas fontes de dados: (a) o log[1] de tráfego de um ISP de Internet banda larga, referentes ao mês de junho de 2008, 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], e contém amostras agrupadas do uso da infra-estrutura do ISP. Este log é formado basicamente por amostras dos fluxos das *transações* geradas por aplicações/protocolos. Os principais campos de cada transação são: data/hora inicial, duração, serviço[2], protocolo, volume de bytes recebidos e enviados e os endereços IP envolvidos. O segundo log, do serviço de DHCP, foi utilizado para identificar os usuários do ISP através do MAC Address do seu equipamento utilizado para acessar a Internet. Este log é a transcrição da comunicação[3] entre esses equipamentos e o servidor de DHCP.

As transações foram agrupadas em *sessões*. Uma sessão é definida como um conjunto de transações de um mesmo usuário do ISP que possuem um período de inatividade inferior a uma hora. Em seguida, as duas fontes de dados foram integradas pelo endereço IP e pelo *timestamp*, campos presentes em ambos os logs. Após a junção das duas fontes de dados foi possível identificar o usuário de cada sessão através de seu MAC Address. O agrupamento de cerca de 71 milhões de transações do log de tráfego gerou aproximadamente 2,8 milhões de sessões. Cada sessão é caracterizada pelos seguintes dados: data e hora de início, duração, serviços/protocolos utilizados, volume de bytes transferidos e MAC Address do usuário responsável por aquele tráfego.

Após a geração das sessões, foram removidas aquelas que não poderiam ser utilizadas na caracterização do comportamento dos usuários do ISP de banda larga. Em resumo, foram removidas: sessões com IPs não encontrados no log do serviço de DHCP; sessões de transações que não puderam ser associadas a um MAC Address; sessões de assinantes não residenciais, pois, além de possuir características específicas, representam menos de 1% do total de sessões; sessões com duração igual a zero, provavelmente devido a problemas na coleta dos dados; e, por último, sessões *outliers*, ou seja, sessões com um número de bytes transferidos desproporcional em relação ao conjunto total de sessões. Nessa última remoção o ponto de corte foi determinado pela média do número de bytes transferidos em uma sessão, acrescido/reduzido por duas vezes o desvio padrão. Após as remoções restaram cerca de 1,88 milhão de sessões.

[1]Arquivo com histórico de um conjunto de transações computacionais.
[2]HTTP, SMTP, POP3, VoIP, BitTorrent, etc.
[3]Fornecimento, renovação e expiração de *leasings* de IPs.

Considerando o alto número de transações P2P (eDonkey/eMule ou BitTorrent), aproximadamente 26% do total, e a importância desse tipo de tráfego para o planejamento e gerenciamento da infra-estrutura do ISP na caracterização proposta, as sessões foram primeiramente classificadas em P2P ou não-P2P, se contêm, ou não, pelo menos uma transação P2P identificada e classificada pelo SCE. Além disso, devido à alta variabilidade do volume de bytes recebidos em sessões P2P, estas foram classificadas em *light* e *heavy*-P2P, com o intuito principal de separar as sessões que fazem uso casual de P2P daquelas que fazem uso intenso desses protocolos. Como o tráfego de bytes recebidos em 85% de todas as sessões P2P não excede 100 MB, resolveu-se classificar esse grupo como sessões *light*-P2P. Assim, as sessões *heavy*-P2P são aquelas que possuem ao menos uma transação de eDonkey/eMule ou de BitTorrent e, por sua vez, transferiram mais de 100MB.

Com base nessa classificação, é possível realizar uma caracterização hierárquica através de três níveis: (1) todas as sessões, (2) sessões não-P2P vs. sessões P2P e (3) sessões *light*-P2P vs. *heavy*-P2P. A partir das cargas de trabalho específicas de cada nível hierárquico, foi realizada a identificação dos processos de chegada e de saída de sessões da infra-estrutura do provedor, a identificação de suas respectivas durações, a contabilização do volume de dados trafegado e a identificação dos principais serviços de Internet e das atividades de comércio eletrônico utilizados nas sessões de usuários. Ressalta-se que a quantidade de dados transferidos em cada sessão de usuário é caracterizada de acordo com a sua "direção" sob o ponto de vista do usuário: dados de chegada (*incoming* bytes) e dados de saída (*outgoing* bytes). O passo seguinte é a determinação da distribuição estatística de cada aspecto do comportamento do usuário analisando a que mais se aproxima dos dados coletados utilizando tanto o método *least-square fit* [Trivedi 2002] quanto a análise dos gráficos das distribuições. Para avaliar as variações no tráfego ao longo do dia, as análises estatísticas das sessões de cada nível hierárquico foram realizadas para cada hora, de dias da semana e finais de semana.

A avaliação do padrão de acesso dos serviços foi realizada hierarquicamente para cada classe e em cada hora de um dia típico, revelando os serviços mais significativos da Internet de banda larga. O padrão de requisições HTTP dos usuários também foi analisado, com base nos domínios acessados em serviços HTTP, HTTPS e *streaming* sobre HTTP. O mapeamento dos acessos foi feito com base na identificação de termos-chave ou *tags* (youtube, forum, banner, etc.) presentes nas URLs e na sua posterior vinculação a uma das categorias de comércio eletrônico proposto por [Rappa 2004]. As categorias propostas neste estudo são: intermediários, publicidade, informacionais, comerciais, manufatura (direto), afiliados, comunidades, assinatura e sob-demanda.

4. Resultados

Esta seção apresenta e discute os resultados mais relevantes encontrados na caracterização hierárquica do comportamento do usuário de Internet de banda larga. A seção 4.1 apresenta uma visão geral da carga de trabalho. A identificação do processo de chegada e de saída das sessões, a identificação de suas respectivas durações, bem como a identificação da quantidade de bytes enviados e recebidos em cada sessão de cada nível hierárquico são caracterizadas na seção 4.2. A popularidade dos serviços e as atividades de comércio eletrônico dos usuários de Internet de banda larga são apresentadas nas seções 4.3 e 4.4, respectivamente.

4.1. Visão geral da carga de trabalho

Uma visão geral da carga de trabalho dos dois primeiros níveis hierárquicos desta caracterizarão é provida na Tabela 1 e do terceiro nível hierárquico é apresentado na Tabela 2. Os logs reais do ISP de Internet de banda larga a cabo que foram utilizados na caracterização são de um período de 28 dias (de 08/06/2008 a 05/07/2008), durante o qual cerca de 1.880.000 sessões foram identificadas. A partir da Tabela 1 observa-se que mais de 90% das sessões foram classificadas como não-P2P e transferiram apenas 40% de todos os bytes recebidos e 16% dos bytes enviados. Com base na Tabela 2 pode-se destacar que menos de 3% de todas as sessões (*heavy-P2P*) são responsáveis por aproximadamente 58% de todos os bytes recebidos e 74% dos bytes enviados. Além disso as sessões *heavy*-P2P foram cerca de 12 vezes maiores que as sessões não-P2P.

Tabela 1. Visão geral da carga de trabalho em uma semana típica.

	Todos	**não-P2P**	**P2P**
Total de sessões	1.879.315	1.703.919	175.396
Total de bytes enviados (%P2P)(TB)	59,15 (80,23%)	9,37 (0,00)	49,78 (95,33%)
Total de bytes recebidos (%P2P)(TB)	116,44 (45,45%)	47,06 (0,00)	69,38 (76,28%)
Média (CV) da duração das sessões (hora)	1,79 (3,17)	1,29 (2,66)	6,65 (2,15)
Média (CV) de serviços distintos / sessão	1,96 (0,59)	1,94 (0,59)	2,55 (0,61)
Média (CV) de amostras / sessão	3.053,68 (15,79)	1.129,56 (5,19)	21.745,92 (7,15)
Média (CV) de bytes enviados / sessão (MB)	33,00 (26,58)	5,77 (122,08)	297,57 (6,15)
Média (CV) de bytes recebidos / sessão (MB)	64,97 (14,68)	28,96 (16,64)	414,76 (6,54)

Tabela 2. Visão geral da carga de trabalho em uma semana típica.

	P2P	***light*-P2P**	***heavy*-P2P**
Total de sessões	175.396	127.963	47.433
Total de bytes enviados (%P2P)(TB)	49,78 (95,33%)	6,00 (93,87%)	43,78 (95,53%)
Total de bytes recebidos (%P2P)(TB)	69,38 (76,28%)	2,08 (28,38%)	67,30 (77,76%)
Média (CV) da duração das sessões (hora)	6,65 (2,15)	3,50 (1,42)	15,16 (1,60)
Média (CV) de serviços distintos / sessão	2,55 (0,61)	2,43 (0,60)	3,32 (0,60)
Média (CV) de amostras / sessão	21.745,92 (7,15)	8.238,02 (2,13)	58.187,04 (5,06)
Média (CV) de bytes enviados / sessão (MB)	297,57 (6,15)	49,13 (5,58)	967,81 (3,51)
Média (CV) de bytes recebidos / sessão (MB)	414,76 (6,54)	17,03 (1,40)	1.487,76 (3,40)

A partir da junção do log de tráfego com o log DHCP, foi possível identificar as sessões dos usuários, identificadas pelo MAC Address. A Figura 1 apresenta a porcentagem de usuários que realizou uma combinação de diferentes tipos de sessão ao longo do mês analisado. Um mesmo usuário pode realizar, por exemplo, sessões *heavy*-P2P e também sessões não-P2P. O valor 20,68% na interseção central da figura representa a porcentagem de usuários que criaram todos os tipos de sessão.

A Figura 2(a) mostra o número de sessões ativas e criadas em um dia típico, nesse caso uma quarta-feira. Nota-se nessa figura que poucos usuários do ISP de banda larga criam suas sessões na virada do dia. A ociosidade dos recursos do ISP nas primeiras horas do dia confirmam o padrão diário de utilização da Internet apresentado na literatura [Floyd and Paxson 2001]. Analisando a Figura 2(b), e também a Tabela 1, é possível notar que, para o agrupamento geral de sessões, a razão média entre o número de bytes recebidos e o número de bytes enviados por sessão fica entre um e dois. Contudo, esta razão não segue a mesma proporção nas sessões não-P2P e tampouco em sessões P2P. Isso mostra a representatividade das sessões P2P no tráfego de bytes enviados, devido ao papel de servidores que os usuários de sistemas P2P desempenham.

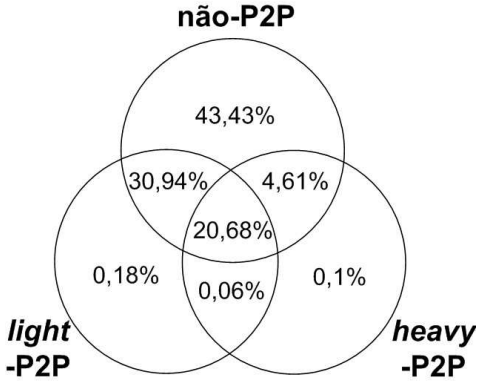
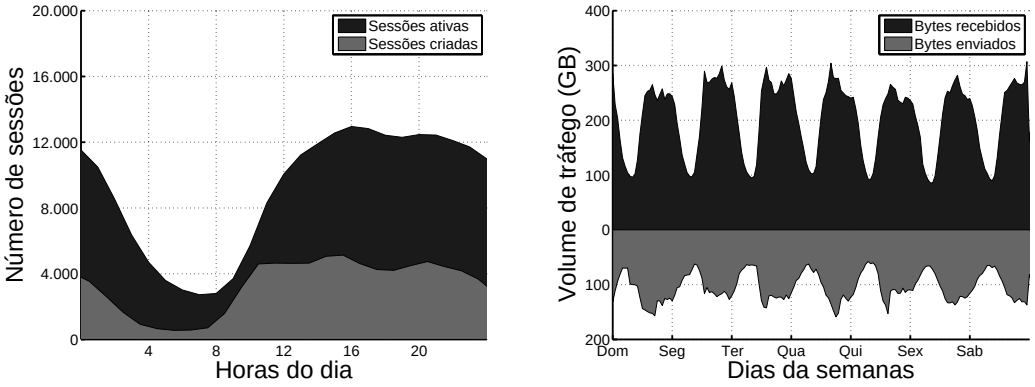


Figura 1. Percentual de usuários por tipo de sessão.



(a) Sessões ativas e criadas em um dia. (b) Volume de tráfego em uma semana.

Figura 2. Criação e atividade das sessões e volume de tráfego.

4.2. Características da carga de trabalho

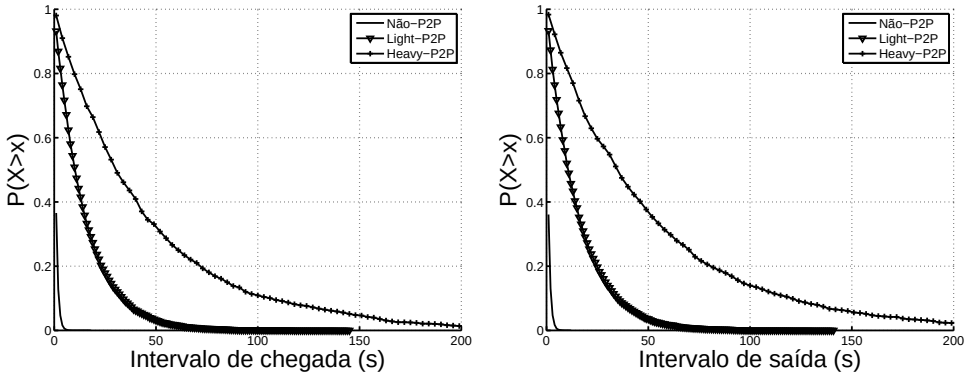
Esta seção analisa cinco aspectos utilizados na caracterização hierárquica do comportamento dos usuários. Os processos de chegada e de saída são analisados na seção 4.2.1, a análise da duração das sessões é descrita na seção 4.2.2 e a seção 4.2.3 examina o volume de bytes recebidos e enviados em uma sessão. A análise foi feita para os três níveis hierárquicos definidos, em cada hora, tanto dos dias de semana quanto dos finais de semana, exibindo nos gráficos o horário de maior demanda (19hs). Várias distribuições estatísticas foram calculadas para a identificação daquela que mais se aproxima dos dados reais utilizando a técnica *least-square-fit* [Trivedi 2002] e, posteriormente, realizou-se uma inspeção visual dos gráficos gerados.

4.2.1. Processo de chegada e de saída de sessões

Esta seção caracteriza os processos de chegada e de saída de sessões dos usuários durante cada hora em dias de semana e finais de semana. As sessões que iniciaram e finalizaram dentro das 24 horas de um determinado dia foram agrupadas para se caracterizar o comportamento do usuário em diferentes períodos do dia.

Tabela 3. Sumário das distribuições IAT e IDT das sessões.

		Dia de semana				Fim de semana			
		Média(ms)	**CV**	**Distrib.**	**Parâmetros**	**Média(ms)**	**CV**	**Distrib.**	**Parâmetros**
Todos	IAT	761-7.022	0,98-1,17	Exp. (λ)	1,31e-3–1,42e-4	789-5.552	0,99-1,01	Exp. (λ)	1,27e-3–1,80e-4
	IDT	765-5.907	1,01-1,16	Exp. (λ)	1,31e-3–1,69e-4	849-5.001	1,00-1,01	Exp. (λ)	1,18e-3–2,00e-4
não-P2P	IAT	834-8.471	0,98-1,26	Exp. (λ)	1,20e-3–1,18e-4	876-6.495	0,99-1,01	Exp. (λ)	1,14e-3–1,54e-4
	IDT	852-7.269	1,01-1,20	Exp. (λ)	1,17e-3–1,38e-4	924-5.887	1,00-1,01	Exp. (λ)	1,08e-3–1,70e-4
P2P	IAT	8.707-40.309	0,99-1,03	Weibull (α, β)	1,03e-4–2,70e-5 1,01e0–9,93e-1	7,742-37.666	1,01-0,99	Weibull (α, β)	1,30e-4–2,60e-5 1,00e-0–1,00e0
	IDT	7.480-32.414	1,00-1,04	Weibull (α, β)	1,33e-4–4,60e-5 1,00e0–9,64e-1	8.665-32.756	1,04-1,03	Weibull (α, β)	1,43e-4–4,40e-5 9,77e-1–9,67e-1
***light*-P2P**	IAT	11.720-51.847	1,00-1,02	Weibull (α, β)	8,10e-5–2,30e-5 1,01e0–9,84e-1	11.015-46.550	1,03-0,95	Weibull (α, β)	8,60e-5–1,70e-5 1,01e0–1,02e0
	IDT	10.603-44.576	1,00-1,04	Weibull (α, β)	9,30e-5–3,60e-5 1,00e0–9,57e-1	12.440-42.731	1,01-1,03	Weibull (α, β)	8,60e-5–4,40e-5 9,93e-1–9,42e-1
***heavy*-P2P**	IAT	24.526-210.649	1,30-1,06	Weibull (α, β)	6,80e-5–3,00e-6 9,53e-1–1,03e0	25.456-185.200	1,02-1,02	Weibull (α, β)	4,60e-5–7,00e-6 9,86e-1–9,76e-1
	IDT	25.232-116.405	0,99-0,95	Weibull (α, β)	3,30e-5–7,00e-6 1,02e0–1,02e0	27.478-136.257	1,07-1,01	Weibull (α, β)	6,30e-5–1,30e-5 9,48e-1–9,52e-1

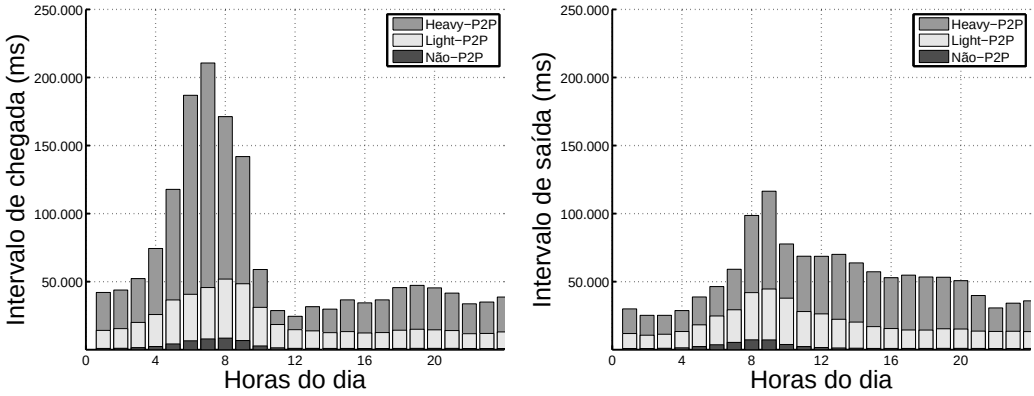


(a) Processo de chegada de sessões às 19hs. (b) Processo de saída de sessões às 19hs.

Figura 3. ICDFs do IAT e IDT das sessões em um dia típico.

Pode-se observar claramente pela Tabela 3 a divisão entre as distribuições do processo de chegada (IAT – *inter-arrival times*) e de saída (IDT – *inter-departure times*) das sessões de usuários não-P2P e P2P, caracterizadas respectivamente pelas distribuições Exponencial e Weibull, tanto para dias de semana quanto para finais de semana. A diferença entre as duas distribuições ocorre devido ao tempo entre chegada/saída de sessões P2P serem mais esparsos, o que aumenta a cauda da curva, caracterizando a distribuição de cauda pesada Weibull. A Tabela 3 sumariza os resultados encontrados provendo o intervalo entre a menor e a maior média das diferentes horas do dia, e o coeficiente de variação (CV) dos processos de chegada e de saída, assim como os valores do parâmetro λ da distribuição Exponencial e dos parâmetros α e β da distribuição Weibull. Já as Figuras 3(a) e 3(b) apresentam as ICDFs[4], respectivamente, dos processos de chegada e de saída das sessões não-P2P, *light*-P2P e *heavy*-P2P.

[4]*Inverse Cumulative Distribution Frequency* (ICDF).



(a) IAT. (b) IDT.

Figura 4. Médias do IAT e IDT em um dia típico.

Os valores de λ da distribuição Exponencial que modelam o IAT e o IDT das sessões dos usuários não-P2P indicam que estas sessões são iniciadas e finalizadas com uma frequência relativamente alta – uma a cada 834 a 8.471 ms (IAT) e uma a cada 852 a 7.269 ms (IDT) – em horas em dias de semana. Por outro lado, sessões P2P não são iniciadas e finalizadas às mesmas taxas: uma a cada 8.707 a 40.409 ms (IAT) e uma a cada 7.480 a 32.414 ms (IDT). Ou seja, sessões de usuários P2P chegam e saem com menos frequência que sessões não-P2P. As Figuras 4(a) e 4(b) apresentam as médias do IAT e IDT para as horas de um dia da semana típico, onde, quanto maior for o tempo entre chegada/saída de sessões, menor será o número de sessões que estão sendo criadas/finalizadas.

4.2.2. Duração das sessões

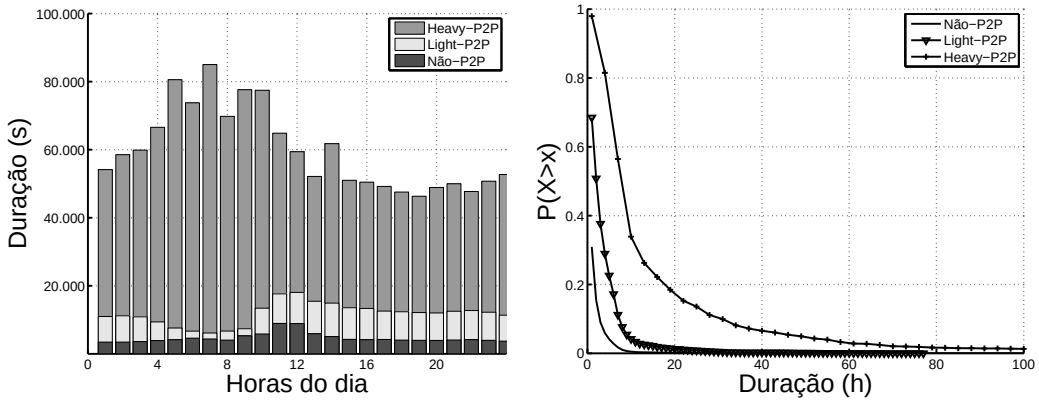
Esta seção analisa a duração das sessões dos usuários de Internet de banda larga. A duração é caracterizada separadamente para grupos de sessões iniciadas em um mesmo dia, tanto em dias de semana quanto em finais de semana.

A distribuição estatística da duração das sessões mostra a influência que as sessões *heavy*-P2P exercem sobre a duração geral de todas as sessões. Apesar da duração das sessões não-P2P e *light*-P2P terem se aproximado mais da distribuição Gamma, as distribuições características dos usuários P2P e geral foram ajustadas a Lognormal. Já a distribuição das sessões P2P no final de semana se ajustou a Weibull, provavelmente devido à curta duração das sessões não-P2P e *light*-P2P. Observa-se que as sessões P2P e também as *heavy*-P2P seguem a distribuição Lognormal, o que é coerente com outros resultados apresentados em [Floyd and Paxson 2001] e [Veloso et al. 2006].

Nota-se que sessões de usuários P2P não são criadas com tanta freqüência, mas são mais longas. As Figuras 5(a) e 5(b), apresentam, respectivamente, a média das durações em cada hora dos dias da semana e as ICDFs das sessões não-P2P, *light*-P2P e *heavy*-P2P, onde pode-se observar que a grande quantidade de sessões criadas no início da noite têm durações crescentes que se estendem ao longo da madrugada.

Tabela 4. Sumário das distribuições da duração das sessões.

	Dia de semana				Fim de semana			
	Média (s)	**CV**	**Distrib.**	**Parâmetros**	**Média (s)**	**CV**	**Distrib.**	**Parâmetros**
Todos	4.872-12.190	3,25-2,48	Logn. (μ)	7,27e0–8,42e0 1,57e0–1,40e0	5.294-16.223	3,13-6,17	Logn. (μ)	7,39e0–7,86e0 1,54e0–1,91e0
não-P2P	3.434-8.933	2,81-2,34	Gamma (α, β)	2,98e-1–2,90e-1 1,15e+4–3,08e+4	3.735-6.723	2,10-8,76	Gamma (α, β)	3,00e-1–2,64e-1 1,24e+4–2,55e+4
P2P	14.785-38.264	4,09-1,65	Logn. (μ)	8,16e0–9,90e0 1,70e0–1,15e0	13.490-70.522	2,55-2,97	Weibull (α,β)	3,44e-2–4,51e-3 4,01e-1–5,18e-1
***light*-P2P**	6.120-18.044	2,26-1,11	Gamma (α, β)	2,75e-1–6,04e-1 2,22e+4–2,99e+4	5.829-16.729	2,15-1,78	Gamma (α, β)	2,96e-1–4,79e-1 1,97e+4–3,49e+4
***heavy*-P2P**	46.305-85.013	1,69-1,91	Logn. (μ)	1,01e+1–1,06e+1 1,16e0–1,24e0	41.627-166.536	1,45-1,96	Logn. (μ)	1,01e+1–1,12e+1 1,06e0–1,26e0



(a) Média da duração em dia típico. (b) Duração das sessões às 19hs.

Figura 5. Médias e ICDFs da duração das sessões em um dia típico.

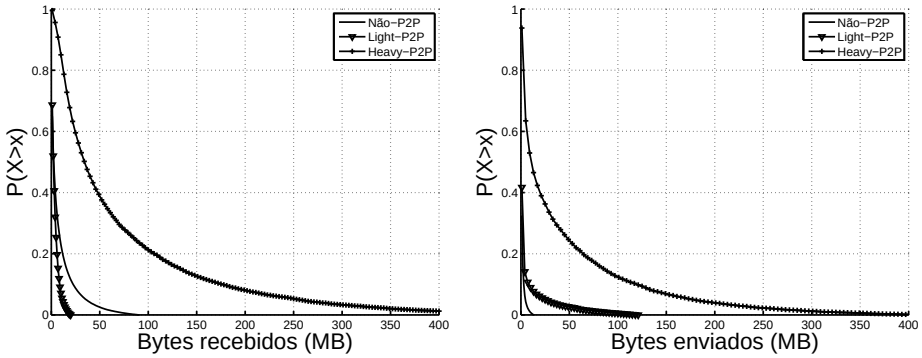
4.2.3. Volume de tráfego recebido e enviado

Esta seção caracteriza o volume total de bytes recebidos (*incoming* bytes) e enviados (*outgoing* bytes) nas sessões estabelecidas pelos usuários do ISP de Internet banda larga. Assim como na seção 4.2.1 e 4.2.2, as análises foram realizadas para grupos de sessões iniciadas em um mesmo dia e foram segregadas por nível hierárquico e por hora do dia, tanto em dias de semana quanto em finais de semana, para se caracterizar o comportamento do usuário em diferentes períodos do dia.

A Tabela 5 apresenta as distribuições do volume de bytes recebidos/enviados de cada nível hierárquico, com seus respectivos parâmetros, tanto nos dias de semana quanto nos dias de final de semana. Pode-se observar que o comportamento do tráfego geral se assemelha mais ao do usuário não-P2P, cujo volume de bytes enviados se aproximou mais da distribuição de Pareto, devido à grande quantidade de sessões com pequeno número de bytes enviados. Já o volume de bytes recebidos se aproximou mais da distribuição de Gamma, devido à presença de sessões de maior tráfego. As Figuras 6(a) e 6(b), apresentam, respectivamente, a ICDF das sessões não-P2P, *light*-P2P e *heavy*-P2P, onde pode-se perceber que enquanto as sessões não-P2P possuem maior número de bytes recebidos, as sessões *light*-P2P apresentam maior volume de bytes enviados.

Tabela 5. Sumário das distribuições dos bytes recebidos/enviados das sessões.

		Dia de semana				Fim de semana			
		Média(MB)	**CV**	**Distrib.**	**Parâmetros**	**Média(MB)**	**CV**	**Distrib.**	**Parâmetros**
Todos	IN	11,78-24,37	1,97-2,00	Gamma (α,β)	2,20e-1–1,83e-1 5,34e1–1,33e2	11,70-22,36	2,01-2,07	Gamma (α,β)	2,23e-1–1,81e-1 5,25e1–1,24e2
	OUT	3,44-15,14	3,08-2,32	Pareto (α,k)	6,23e-1–7,77e-1 2,22e-1–2,40e0	3,65-14,64	3,35-2,43	Pareto (α,k)	5,75e-1–7,06e-1 1,62e-1–1,64e0
não-P2P	IN	4,04-8,16	2,06-1,76	Gamma (α,β)	1,66e-1–2,16e-1 2,43e1–3,79e1	3,88-8,26	2,12-1,70	Gamma (α,β)	1,56e-1–2,21e-1 2,49e1–3,73e1
	OUT	0,58-1,11	2,30-1,66	Pareto (α,k)	1,13e0–1,27e0 2,89e-1–6,06e-1	0,50-0,93	2,25-1,56	Gamma (α,β)	1,34e-1–1,73e-1 3,73e0–5,36e0
P2P	IN	33,20-47,16	1,72-1,55	Weibull (α,β)	2,70e-1–2,78e-1 4,61e-1–4,11e-1	35,02-47,07	1,76-1,59	Weibull (α,β)	2,80e-1–2,95e-1 4,48e-1–3,98e-1
	OUT	21,95-34,46	2,06-1,61	Weibull (α,β)	5,41e-1–4,23e-1 3,37e-1–3,44e-1	24,20-36,22	2,02-1,62	Weibull (α,β)	5,49e-1–4,34e-1 3,26e-1–3,34e-1
***light*-P2P**	IN	1,78-3,78	1,25-1,13	Exp. (λ)	5,61e-1–2,64e-1	1,74-3,65	1,33-1,13	Exp. (λ)	5,76e-1–2,74e-1
	OUT	4,00-6,96	3,12-3,08	Pareto (α,k)	6,25e-1–6,20e-1 2,63e-1–4,60e-1	4,01-6,28	3,13-3,17	Pareto (α,k)	6,21e-1–6,06e-1 2,60e-1–3,74e-1
***heavy*-P2P**	IN	65,49-72,37	1,25-1,22	Weibull (α,β)	2,95e-2–2,99e-2 8,60e-1–8,39e-1	66,58-72,77	1,28-1,28	Weibull (α,β)	3,67e-2–3,41e-2 8,11e-1–8,11e-1
	OUT	38,83-48,14	1,63-1,34	Weibull (α,β)	1,61e-1–1,05e-1 5,71e-1–6,32e-1	42,38-51,08	1,53-1,34	Weibull (α,β)	1,62e-1–1,07e-1 5,54e-1–6,20e-1



(a) Volume de bytes recebidos pelas sessões às 19hs.

(b) Volume de bytes enviados pelas sessões às 19hs.

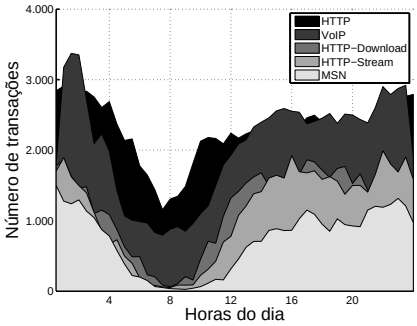
Figura 6. Distribuições dos bytes transferidos nas sessões em um dia típico.

4.3. Popularidade dos serviços

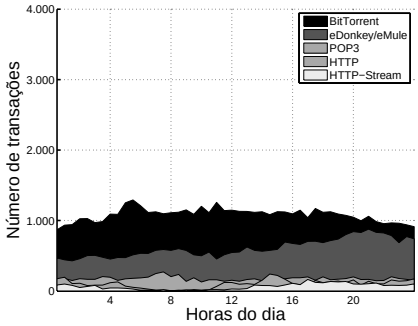
Esta seção analisa os tipos de serviço mais acessados pelos usuários do ISP. A Figura 7 apresenta os cinco serviços mais populares das sessões não-P2P e *heavy*-P2P durante um dia típico, neste caso uma quarta-feira.

A Figura 7(a) apresenta o padrão de requisição das sessões não-P2P aos serviços classificados previamente pelo SCE. O volume de transações ao longo do dia segue o padrão de acesso diurno, também presente na literatura [Cha et al. 2008]. Nota-se a importância dos serviços VoIP e HTTP-Stream e supõe-se que isto ocorra devido à popularização da comunicação via telefonia IP e devido à expansão de transmissão de conteúdos de *sites*, tais como, YouTube, MySpace, Last.fm, entre outros.

Observa-se na Figura 7(b) o comportamento *always-on* dos usuários *heavy*-P2P, demonstrado pela regularidade do número de requisições a sistemas P2P (BitTorrent e eDonkey/eMule) durante todas as horas do dia. Nota-se que o serviço POP3, além do P2P, apresentou alta representatividade nesta classe de usuários.



(a) Popularidade de serviços das sessões não-P2P em um dia típico.

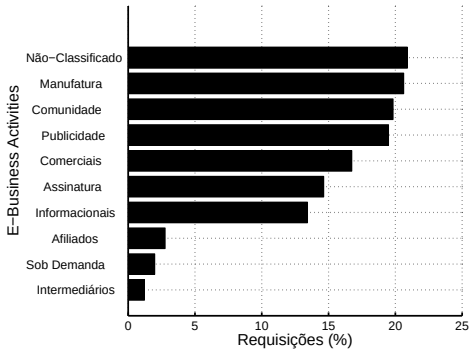


(b) Popularidade de serviços das sessões *heavy*-P2P em um dia típico.

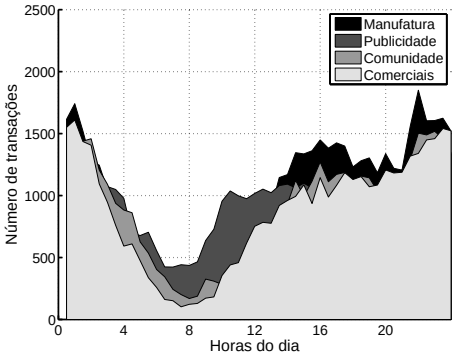
Figura 7. Popularidade de serviços.

4.4. Atividades de comércio eletrônico

Esta seção apresenta uma classificação das requisições do protocolo HTTP do usuário de Internet banda larga com base na taxonomia proposta por [Rappa 2004] que agrupa as atividades de comércio eletrônico em nove categorias. A Figura 8 apresenta o padrão de requisições HTTP dos usuários do ISP. Apesar de apenas 31% dos domínios terem sido classificados nas categorias propostas, estes correspondem a 79% de todos os acessos a *websites* realizados pelos usuários do ISP. Os percentuais de requisições por atividade de comércio eletrônico são apresentados na Figura 8(a).



(a) Requisições por categoria de comércio eletrônico.



(b) Transações das principais categorias de comércio eletrônico em um dia típico.

Figura 8. Categorias de comércio eletrônico da Internet de banda larga.

Observa-se a alta representatividade de requisições da categoria manufatura, que inclui *websites* como microsoft.com, hp.com, dell.com, entre outros. Destaca-se também o alto número de acesso à categoria de comunidades, que inclui acessos a *websites* como Orkut, UOL, MySpace, entre outros. Acessos a *websites* de conteúdo livre, produzido pelos próprios usuários, como YouTube, Flickr e Blogspot, também se encontram nessa categoria. A categoria publicidade, com praticamente a mesma quantidade de requisições que as categorias de manufatura e de comunidades, inclui portais como Google, Yahoo!, Terra e UOL, que disponibilizam conteúdo e serviços junto a áreas reservadas para propagandas, sejam elas pré-definidas ou definidas pelo padrão de navegação do usuário,

tais como Google AdWords e Yahoo! Search Marketing. As categorias de portais intermediários, como MercadoLivre e eBay, de assinatura, a exemplo do Terra, Globo e Estadão, e também os *websites* de coleta de dados sobre usuários e seus hábitos de consumo, como Google, Yahoo!, DoubleClick e Right Media, apresentaram um número de requisições semelhante pelos usuários do ISP. A Figura 8(b) mostra como os acessos às principais categorias de comércio eletrônico se distribuem ao longo de um dia típico.

5. Conclusões

Este artigo apresenta uma metodologia de caracterização hierárquica do comportamento de usuários de sistemas par-a-par na Internet de banda larga e sua respectiva aplicação com um conjunto significativo de dados reais de um provedor de acesso a cabo. As fontes de dados utilizadas permitiram a organização da carga de trabalho em sessões de usuários, que foram classificadas com base na presença ou não de transações de protocolos P2P. A caracterização hierárquica foi realizada através de três níveis: (1) todas as sessões, (2) sessões não-P2P vs. sessões P2P e (3) sessões *light*-P2P vs. *heavy*-P2P.

Os resultados encontrados mostram a desproporção do consumo, pois, menos de 3% das sessões são responsáveis por cerca de 58% de todo o tráfego de chegada ao provedor e 74% do tráfego geral de saída. Além disso, identificou-se que sessões com muitas requisições de P2P são 12 vezes mais longas do que sessões que não fazem P2P. Os aspectos analisados da carga de trabalho apresentam resultados semelhantes aos encontrados na literatura quando as análises são realizadas com o conjunto geral de sessões. Porém, analisando a carga de trabalho com maior granularidade, foram encontradas diferenças nas distribuições estatísticas que caracterizam seus diferentes aspectos.

Essa caracterização hierárquica do comportamento de usuários cria condições para os ISPs de banda larga aprimorarem o gerenciamento da sua infra-estrutura tecnológica e o planejamento da prestação do serviço de acesso à Internet, por exemplo, através da simulação baseada nas distribuições estatísticas e também da diferenciação de preço baseada no comportamento do usuário ao longo do tempo.

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 (CT-Info/CNPq 55.0995/2007-2) e pelo Fundo de Incentivo à Pesquisa da PUC-Minas (FIP-2009/3504-S1).

Referências

Arlitt, M. (2000). Characterizing web user sessions. *SIGMETRICS Performance Evaluation*, 28(2):50–63.

Arlitt, M., Friedrich, R., and Jin, T. (1999). Workload characterization of a web proxy in a cable modem environment. *SIGMETRICS Performance Evaluation*, 27(2):25–36.

Barford, P., Bestavros, A., Bradley, A., and Crovella, M. (1999). Changes in web client access patterns: Characteristics and caching implications. *World Wide Web*, 2:15–28.

Cerf, V. (2008). What's a reasonable approach for managing broadband networks?. *Google Public Policy Blog.* (http://googlepublicpolicy.blogspot.com/2008/08/ whats-reasonable-approach-for-managing.html).

Cha, M., Rodriguez, P., Crowcroft, J., Moon, S., and Amatriain, X. (2008). Watching television over an ip network. In *IMC '08: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA. ACM.

CISCO (2008). Cisco service control application for broadband reference guide. (http://www.cisco.com/en/us/docs/cable/serv_exch/serv_control/broadband_app/rel317/scabbrg/scabbrg.html).

Costa, C. P., Cunha, I. S., Borges, A., Ramos, C. V., Rocha, M. M., Almeida, J. M., and Ribeiro-Neto, B. (2004). Analyzing client interactivity in streaming media. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 534–543, New York, NY, USA. ACM.

Dischinger, M., Haeberlen, A., Gummadi, K. P., and Saroiu, S. (2007). Characterizing residential broadband networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 43–56, New York, NY, USA. ACM.

Floyd, S. and Paxson, V. (2001). Difficulties in simulating the internet. *IEEE/ACM Transactions on Networking*, 9(4):392–403.

Fukuda, K., Cho, K., and Esaki, H. (2005). The impact of residential broadband traffic on japanese isp backbones. *ACM SIGCOMM Computer Communications Review*, 35(1):15–21.

Goth, G. (2008). Isp traffic management: Will innovation or regulation ensure fairness? *IEEE Distributed Systems Online*, 9(9).

Gummadi, K. P., Dunn, R. J., Saroiu, S., Gribble, S. D., Levy, H. M., and Zahorjan, J. (2003). Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. *SIGOPS Oper. Syst. Rev.*, 37(5):314–329.

Hamada, T., Chujo, K., Chujo, T., and Yang, X. (2004). Peer-to-peer traffic in metro networks: analysis, modeling and policies. *IEEE/IFIP Network Operations & Management Symposium (NOMS 2004)*.

Lakshminarayanan, K., Padmanabhan, V. N., and Padhye, J. (2004). Bandwidth estimation in broadband access networks. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 314–321, NY, USA. ACM Press.

MIT (2005). The broadband incentive problem. In *MIT Communications Futures Program (CFP) and Cambridge University Communications Research Network.*

Rappa, M. A. (2004). The utility business model and the future of computing services. *IBM Syst. J.*, 43(1):32–42.

Sen, S. and Wang, J. (2004). Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Transactions on Networking*, 12(2):219–232.

Trivedi, K. (2002). *Probability & Statistics with Reliability, Queueing, and Computer Science Applications*. John Wiley & Sons, 2nd edition.

Veloso, E., Almeida, V., Wagner Meira, J., Bestavros, A., and Jin, S. (2006). A hierarchical characterization of a live streaming media workload. *IEEE/ACM Transactions on Networking*, 14(1):133–146.