# Uma metodologia de classificação para os fluxos de comunicação *

**Luís Felipe M. de Moraes e Guilherme S. Vilela**

[1]Laboratório de Redes de Alta Velocidade – RAVEL
Programa de Engenharia de Sistemas e Computação – COPPE/UFRJ
Caixa Postal: 68.511 – 21941-972 – Rio de Janeiro, RJ

moraes,vilela@ravel.ufrj.br

***Abstract.*** *Traffic characterization is an important tool for the planning and management of computer networks. However, there is still no agreement about what metrics and methodologies should be used. This article proposes and presents a new classification methodology for the network flows in N possible classes. To validate the methodology, a case study with Rede Rio [1] was performed, and important results were obtained. This paper studies the traffic correlation of these measurements, indicating the impacts and possible causes for the observed behavior. An analysis of which applications are responsible for the majority of the network traffic and comparisons with other works are also performed.*

***Resumo.*** *A caracterização de tráfego é hoje um importante instrumento para o planejamento e gerenciamento das redes de computadores. Entretanto, ainda não existe um consenso sobre métricas e metodologias a serem adotadas. O presente trabalho propõe uma nova metodologia para a classificação dos fluxos de comunicação em N classes. Para validar esta metodologia, foi realizado um estudo de caso na Rede Rio. Foi identificado, as distribuições estatísticas do tráfego em função do seu tamanho, duração e taxa. Neste artigo é feito um estudo sobre a correlação destas medidas, indicando os impactos e as possíveis causas do comportamento observado. É apresentada também uma análise sobre quais aplicações são responsáveis pela maior parte do tráfego da rede, assim como comparações com outros trabalhos.*

## 1. Introdução

No decorrer dos últimos anos, o crescimento das redes locais, metropolitanas e de longa distância e o surgimento de novas aplicações e serviços, fizeram com que o tráfego de rede aumentasse em quantidade e diversidade. Em virtude desse aumento, as redes se tornaram mais complexas, fazendo com que a engenharia de tráfego passasse a desempenhar um importante papel para o seu planejamento e gerenciamento.

Uma importante técnica utilizada para a caracterização do tráfego é a medição dos fluxos de comunicação (conjunto de pacotes que possuem características semelhantes, tais como número da porta e endereço IP de origem e destino). Uma grande variedade de mecanismos é utilizada para gerenciar o tráfego dos fluxos, tais como balanceamento

*Este trabalho conta com suporte da FAPERJ (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro).

de carga e políticas de roteamento. Estudos recentes mostram que uma porção muito pequena dos fluxos é responsável pela grande maioria do tráfego total (em bytes) nas redes [2, 3, 4, 5, 6]. É importante compreender as propriedades deste tráfego de fluxos para propósitos de modelagem e monitoração de tráfego. Estudando esses fluxos pode-se compreender uma grande porção do tráfego total, possibilitando a tarifação baseada no uso, detecção de ataques e outras ações de interesse.

Apesar de ser uma preocupação recente, o estudo da caracterização dos fluxos vem crescendo em importância, com o surgimento de diversos estudos sobre seus tamanhos, durações e taxas [2, 7, 8]. No entanto, quase não há estudos sobre como essas características estão relacionadas entre si. De uma forma geral, não é claro como os diferentes tipos de fluxos estão relacionados. Por exemplo, qual é a relação entre fluxos de grande tamanho e os fluxos de longa duração? Os fluxos de longa duração ocorrem devido a transferências de arquivos grandes? As correlações envolvendo características deste tipo de fluxo não são ainda bem conhecidas e as aplicações deste conhecimento podem ser diversas.

Este trabalho tem como finalidade ajudar a responder essas questões. Para isso, uma metodologia de análise e classificação dos fluxos foi desenvolvida, dividindo os fluxos em N diferentes classes. A idéia de se separar o tratamento de diferentes classes de tráfego é utilizada em vários ambientes tais como DiffServ, IntServ, redes ATM, balanceamento de carga e políticas de roteamento. Existem várias metodologias propostas para análise e classificação dos fluxos. No entanto, algumas importantes limitações existem, como por exemplo, a impossibilidade de se classificar mais de duas classes de fluxos e de se analisar um grande volume de tráfego. Além disso, poucas metodologias estudam a correlação de tráfego.

No intuito de validar a metodologia proposta, utilizou-se a rede de pesquisa do Estado do Rio de Janeiro, Rede Rio [1], para a coleta de dados. Através de ferramentas desenvolvidas, os dados foram analisados de forma que fosse possível conhecer a correlação entre três características dos fluxos de comunicação: tamanho, duração e taxa. Além de estudar a forma com que os fluxos estão correlacionados, o trabalho também apresenta a distribuição cumulativa dos fluxos e quais são as aplicações responsáveis pela maior percentagem do tráfego de rede. O trabalho também faz comparações com outro modelo existente.

Os resultados obtidos neste trabalho representam um retrato do comportamento de uma importante rede de pesquisa, mostrando características do tráfego e possíveis respostas envolvendo seu comportamento.

### 1.1. Trabalhos relacionados

A forma mais comum de se classificar os fluxos é dividí-los em duas classes em função de seus tamanhos. A classe formada por fluxos de tamanho muito grande é chamada elefante, e aquela formada por fluxos pequenos é chamada rato.

Ao se falar em fluxo muito grande ou fluxo muito pequeno, uma pergunta importante deve ser respondida: o que deve ser considerado para se classificar um fluxo como muito grande, ou muito pequeno? [2] propõe um valor fixo como sendo o limite entre as duas classes, ou seja, se o fluxo for maior que $x$KB, ele é considerado elefante. Caso contrário, é considerado rato. No trabalho em questão, fluxos menores que 20KB foram

considerados como rato. O artigo também propõe uma classificação para fluxos de longa duração (tartaruga) e pequena duração (libélula) e uma categoria intermediária, estabelecendo como limiar 2s para fluxos muito pequenos, até 15 minutos para fluxos de duração intermediária e maior que 15 minutos para fluxos tartaruga.

Uma forma alternativa para classificação [9] estabelece uma fração $l$ de toda capacidade do enlace, sendo que os fluxos grandes serão os maiores responsáveis pela fração $l$ da utilização do enlace. Outra forma simples de classificação estabelece que os fluxos pesados serão os $N$ maiores. Ambas as formas possuem o problema de estabelecer os valores de $l$ e de $N$, pois não há um método para se obter estes valores.

A forma de classificação proposta em [2] possui alguns problemas. Os valores escolhidos foram arbitrários, ou seja, foram escolhidos sem utilizar algum método, o que pode levar ao questionamento dos valores escolhidos (por que 15 minutos e não 14?, por exemplo). Outra colocação importante é o fato do tráfego de rede poder apresentar comportamentos diferentes de uma rede para outra, ou seja, o parâmetro escolhido como limiar em uma determinada rede pode não ser adequado em outra. Além disto, a classificação dos fluxos em apenas 2 conjuntos também deve ser evitada, uma vez que as informações sobre fluxos de tamanhos intermediários são perdidas.

Uma tentativa de resolver dois dos problemas citados acima foi proposta nos trabalhos [6, 8]. A classificação é feita baseada na média e no desvio padrão, e propõe uma classificação para as taxas dos fluxos, além das variáveis tamanho e duração, sendo os fluxos de alta taxa chamados de chitá. Desta forma os fluxos grandes seriam aqueles que tivessem seu valor superior a média somada a três vezes o desvio padrão.

Esta classificação remove o problema da arbitrariedade e da diversidade das redes. No entanto, ela ainda classifica os fluxos em apenas duas classes, ou seja, os fluxos intermediários não podem ser identificados adequadamente. Outro problema importante se deve ao fato que o tráfego de rede muitas vezes apresenta grandes variações, resultando em um comportamento de cauda pesada [10, 11]. Nestes casos o uso do desvio padrão não é adeguado pois seu valor pode tender para o infinito.

Em uma tentativa de resolver todos os problemas citados acima [12] propõe uma classificação baseada em fundamentos estatísticos, utilizando testes de hipóteses. A metodologia utilizada analisa o comportamento de um fluxo ao longo do tempo e realiza testes para verificar em qual classe de $n$ possíveis a sua distribuição melhor se encaixa. Apesar do método resolver os problemas mencionados anteriormente, ele possui uma grave limitação, pois não consegue analisar grandes quantidades de fluxos (ex. todos os fluxos de um Backbone). A necessidade de analisar cada fluxo ao longo de toda sua duração faz com que o processamento e o espaço de armazenamento das informações requeridos sejam proibitivos.

Trabalhos recentes de análise de fluxos vêm percebendo comportamentos de diversidade e disparidade, ou seja, a relação entre o número de fluxos e seus tamanhos não é trivial e muitas vezes apresenta uma distribuição de cauda pesada. Pode-se citar como exemplo o fato da arquitetura da Internet ser bastante diversa e dificilmente apresentar pontos com falha. No entanto, cerca de 80-90% das rotas dos pacotes que nela trafegam utilizam cerca de apenas 20 provedores. De forma similar [7] observou que apenas cerca de 80 ASs (*Autonomous Systems*), entre milhares observados, contribuiam para 95% do

tráfego nos enlaces.

Trabalhos analisando o comportamento do tamanho, duração e taxa dos fluxos também mostram que há uma grande diversidade e disparidade entre o número e volume que diferentes classes de fluxos possuem. A diversidade pode ser definida como a presença de um grande número de fluxos distintos. A disparidade é a concentração de tamanho (volume) em um pequeno número de fluxos.

Em [2] foi observado que cerca de 98% dos fluxos possuem duração igual ou inferior a 15 minutos. No entanto, os 2% restantes são responsáveis por cerca de 60% dos bytes trafegados. Zhang *et al* [13] mostra que o tráfego medido é dominado por fluxos de pequeno tamanho, enquanto os fluxos grandes são os maiores responsáveis pelo tráfego em bytes. [8] mostra que 36% dos fluxos de alta taxa correspondem a 2% do tráfego em Bytes. Resultados semelhantes foram encontrados em [6], onde os fluxos considerados elefantes representavam 0,071% e seu volume em bytes 62%. O trabalho mostra também que os fluxos tartaruga (0,35%) foram responsáveis por 63% do tráfego em bytes. Nem a diversidade, nem a disparidade devem ser consideradas boas ou ruins. Por outro lado, o impacto imposto por estes tipos de comportamento deve ser analisado.

Alguns trabalhos recentes vêm analisando o comportamento dos fluxos ao longo do tempo. Estes trabalhos verificaram que os fluxos podem mudar de taxa no decorrer de sua existência e assim apresentar grande volume durante certo intervalo de tempo e pequeno volume em outro intervalo. O trabalho [14] mostra que após 25 minutos 40% do tráfego de grande volume (elefante) deixa de apresentar este comportamento.

O artigo está organizado da seguinte forma: a metodologia proposta é explicada na segunda seção. Na terceira são feitas diversas análises sobre os dados coletados, apresentando os resultados encontrados. Finalmente, na quarta seção, é feita uma conclusão.

## 2. Metodologia Proposta

### 2.1. Classificação dos fluxos

Pretende-se classificar os fluxos em relação a três variáveis: tamanho, duração e taxa. O procedimento utilizado é o mesmo para cada uma das variáveis e possibilita a classificação dos fluxos em $N$ classes, onde $N$ deve ser escolhido de forma que melhor atenda as necessidades de cada rede. É importante ressaltar que não faz parte do escopo deste trabalho indicar qual deve ser o número de classes escolhido para cada rede. A resposta para esta questão não abrange simplesmente o fato de que o maior numero é o melhor. O nível de precisão necessário dependerá do uso que se pretende fazer com a classificação.

Para a aplicação do algoritmo de classificação e caracterização, é necessário que os fluxos sejam coletados e armazenados em um banco de dados contendo as informações para cada fluxo. As informações são: tamanho em bytes, duração em segundos, taxa em bytes/s, porta de origem e porta de destino.

O banco de dados é utilizado para gerar as distribuições dos fluxos e suas freqüências relativas segundo seus tamanhos, durações e taxas. Baseando-se nestas duas funções, é possível iniciar o algoritmo de classificação dos fluxos em $N$ classes. Uma vez escolhido o valor de $N$ deve-se descobrir o valor do fator de corte, que será chamado de $C$. O fator de corte é definido como:

$$C = \frac{1}{N} \sum_{i=1}^{i=X_f} x_i * p_x, \tag{1}$$

onde $p_x$ indica a probabilidade da variável aleatória $X$ possuir o valor $x_i$, lembrando que neste caso existem três variáveis aleatórias: tamanho, duração e taxa e $X_f$ é o número de amostras. Portanto, todos estes passos devem ser seguidos para cada uma das variáveis.

Um vez descoberto o valor do fator de corte $C$, é possível efetuar a divisão das classes. Um fluxo que pertence a primeira classe é aquele que está contido entre $x_0 \leq X \leq x_{n_1}$, onde $x_0$ é o menor valor encontrado para aquela variável aleatória, e $x_{n_1}$ será o último valor de $X$ que ainda satisfaz a inequação (2),

$$\sum_{i=0}^{i=n_1} x_i * p_x \leq C. \tag{2}$$

As demais classes seguirão o mesmo raciocínio, ou seja, a segunda classe será composta por aqueles fluxos onde $x_{n_1} < X \leq x_{n_2}$, onde $x_{n_2}$ será o último valor de $X$ que ainda satisfaz a inequação 2 (substituindo-se $i = 0$ por $i = n_1$ e $i = n_1$ por $i = n_2$). O algoritmo se repete até que as $N$ classes sejam obtidas. Desta forma, é possível conhecer o quanto cada classe é reponsável pela média. Neste modelo proposto, a n-ésima classe será considerada como elefante caso o cálculo seja feito com a variável tamanho, tartaruga no caso da duração e chitá para a variável taxa.

### 2.2. Correlação

Um importante fator a ser estudado é a correlação entre as variáveis dos fluxos. Tal estudo torna possível o conhecimento de como elas estão relacionadas. Pode-se estudar mecanismos de tarifação de redes, distinguir um tráfego padrão de um malicioso e obter importantes informações para os projetistas de rede e de equipamentos [8].

Através dos dados armazenados no banco de dados, é possível calcular o coeficiente de correlação das variáveis tamanho, taxa e duração e desta forma pode-se saber como estas variáveis estão inter-relacionadas.

Para relacionar as diferentes categorias de fluxo (elefante, tartaruga e chitá), duas análises devem ser feitas: análise da percentagem dos fluxos que pertencem a duas categorias diferentes (ex: percentagem de fluxos elefante e tartaruga) e análise da percentagem dos fluxos que pertencem a uma determinada categoria, dado que já pertencem a outra. A seguir é descrito o algoritmo utilizado para estas análises.

Para realizar o cálculo da percentagem dos fluxos que pertencem a duas categorias, todos os fluxos contidos no banco de dados devem ser consultados. É verificado então, se um determinado fluxo apresenta valores compatíveis com as classes que se deseja comparar. Caso apresente, é somado o valor de 1 para o contador dos fluxos. Ao final do processo, o valor do contador é dividido pelo número total de fluxos. De forma semelhante, na análise da percentagem dos fluxos que pertencem a uma categoria (dado que já pertencem a outra), o banco de dados deve ser consultado. Deve-se verificar se a condição desejada foi respeitada. Caso positivo, deve-se verificar, então, se a outra

**Tabela 1. Descrição dos dados coletados**

| | Rede Rio |
|---|---|
| Data de início | 26/09/2005 |
| Data de término | 11/11/2005 |
| Número de fluxos | 1.860.549.382 |
| Número de bytes | 22.004.957.909.714 |

variável também atendeu a condição. Se estas duas condições forem satisfeitas é somado 1 ao contador de fluxos. Ao final do processo, o contador é dividido pelo número total de fluxos que respeitaram a primeira condição.

### 2.3. Comportamento das Aplicações

Para a análise dos serviços mais utilizados, foi criada uma tabela com três campos: número da porta, número de bytes e número de fluxos. É importante identificar as principais aplicações que trafegam na rede para que se possa conhecer melhor o tipo de tráfego. A tabela foi construída analisando o número da porta de origem e destino de cada fluxo medido. Após a verificação da porta, é somado ao campo "número de bytes"da tabela, o valor do tamanho do fluxo, e somado ao campo "número de fluxos" uma unidade. Após a análise de todos os fluxos medidos, a tabela é ordenada pelo campo "número de bytes".

## 3. Resultados

Nesta seção é feito um estudo de caso na Rede Rio [1] utilizando a metodologia descrita. Serão feitas diversas análises dos fluxos medidos, mostrando qual é a correlação entre as diferentes categorias de fluxos, como é a distribuição cumulativa das diferentes classes dos fluxos. Serão feitas também comparações com o modelo de classificação que utiliza a média e o desvio padrão [6].

### 3.1. Dados medidos

Os dados foram obtidos das interfaces ligadas à Embratel e à RNP do roteador de borda da Rede Rio[1]. O tráfego coletado foi aquele que entrava e saía do roteador, ou seja, todo tráfego com destino à Rede Rio e todo o tráfego com origem na Rede Rio. Os dados correspondem a sete semanas de coletas (o maior período de coleta em relação as referências consultadas), realizadas no horário de maior utilização da rede, das 10Hs às 16Hs, excluindo-se os finais de semana e feriados.

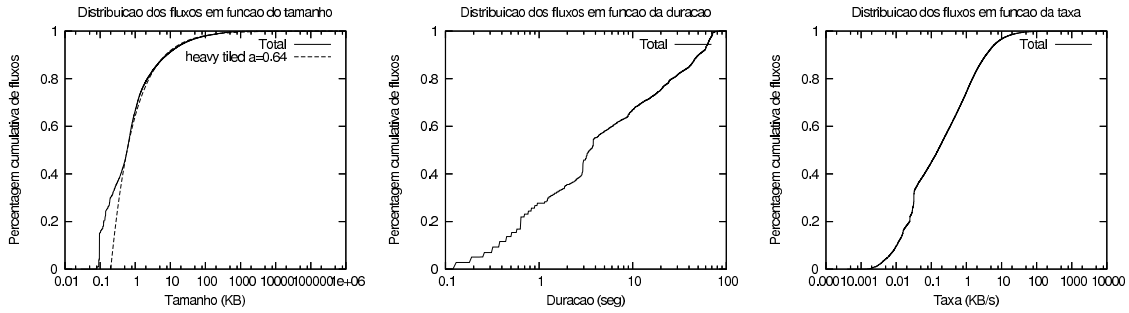
Na Tabela 1 são mostrados quais dados foram obtidos e o período de obtenção dos mesmos.

É interessante enfatizar o tempo de medição bem como para o volume de tráfego medido. Em relação a estes aspectos, os dados obtidos apresentam pelo duas vantagens em relação a trabalhos anteriores. A primeira vantagem é o grande volume de tráfego coletado, que permite ter uma maior confiabilidade nos resultados, uma vez que quanto maior o tamanho da amostra, maior a probabilidade dos resultados obtidos representarem o comportamento real da rede. Enquanto medições feitas em [15] foram da ordem de 30 milhões de pacotes, neste trabalho foram analisados mais de 30 bilhões. O trabalho [8] realiza suas medições com cerca de 3 milhões de fluxos. Já no presente trabalho, são

analisados cerca de 1.8 bilhão. Os artigos [15, 8] fizeram apenas 2 horas de medição e em [2] o tráfego é medido em um período de 10 horas.

### 3.2. Tráfego

As Figuras de 1(a) a 1(c) mostram as distribuições cumulativas dos fluxos de acordo com o tamanho, duração e taxa respectivamente. É importante conhecer tais distribuições, para fins de modelagem de tráfego e planejamento de novas redes. A Figura 1(a), mostra que cerca de 91% dos fluxos possuem tamanho de até 10KBytes. A partir deste resultado é possível concluir que o tráfego é predominantemente composto por fluxos de tamanhos pequenos. As implicações deste fato podem ser diversas. Pode-se citar por exemplo, que as análises e simulações do comportamento do protocolo TCP são focadas em seu comportamento em regime permanente, com saturação da rede (ex: a transmissão de longos arquivos), assumindo que o grande volume de fluxos elefante não seria afetado pela presença de pequenos fluxos (ratos). No entanto, enquanto os fluxos elefante sofrem o controle de congestionamento do protocolo TCP, os fluxos de pequeno tamanho não são controlados pelo algoritmo do protocolo, uma vez que eles são enviados e recebidos antes que o TCP tenha a oportunidade de aplicar o controle. Desta forma, um grande volume de pequenos fluxos pode gerar perdas de pacotes, aumentando o congestionamento da rede. A distribuição encontrada teve comportamento similar em [8] e [15]. É interessante notar que que a distribuição do tamanho dos fluxos se aproxima muito de uma distribuição de cauda pesada [16] com valor para $\alpha = 0.64$.




(a) Distribuição do tamanho dos fluxos (b) Distribuição da duração dos fluxos (c) Distribuição da taxa dos fluxos

**Figura 1. Distribuição cumulativa do tamanho, duração e taxa dos fluxos**

A Figura 1(b) mostra que mais de 85% dos fluxos duram até 25 segundos e cerca de 50% duram até 3 segundos. Pode-se verificar que, em sua maioria, os fluxos são de pequena duração, assim como observado em [2, 8, 15].

Na Figura 1(c) observa-se que mais de 85% dos fluxos possuem taxa igual ou inferior a 2 Kbytes/s, indicando que a grande maioria dos fluxos possuem taxas de transmissão relativamente pequenas. Em [8] a distribuição da taxa se mostrou superior. Cerca de 85% das taxas eram iguais ou inferiores a 10Kbytes/s. Zhang *et al* [15] verificou uma distribuição semelhante, com a grande maioria dos fluxos apresentando taxas inferiores a 10Kbits/s.

Na Tabela 2 é possível verificar os serviços responsáveis pela maior fração do tráfego. O tráfego Web representa cerca de 44% do tráfego da Rede Rio, representando 43% dos fluxos. Pode-se observar também uma grande percentagem de tráfego *Peer to*

*Peer* que representa cerca de 15% do tráfego total em bytes, confirmando o crescimento que as aplicações P2P vêm tendo na Internet. É interessante ressaltar que este valor é certamente inferior ao valor real, uma vez que nas medições feitas foram consideradas apenas as portas definidas por cada tipo de aplicação, enquanto as aplicações P2P permitem que o usuário modifique as portas utilizadas no intuito de burlar os *firewalls*.

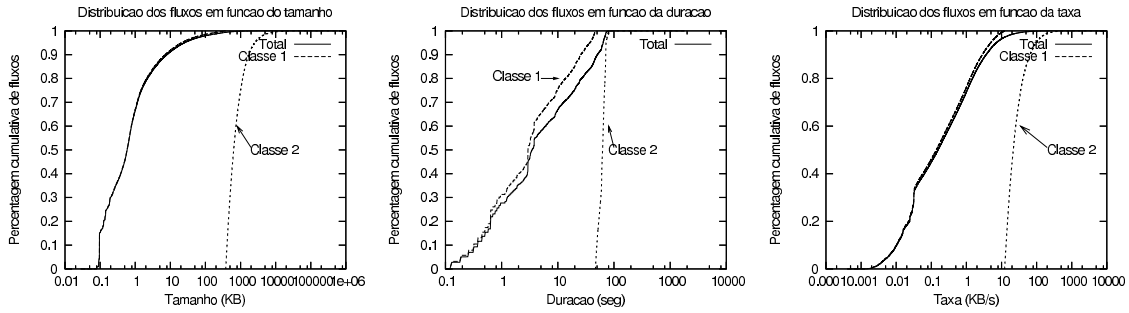
**Tabela 2. Serviços utilizados**

| | Percentagem de bytes | Percentagem de Fluxos |
|---|---|---|
| Web | 44.93 | 43.02 |
| E-Donkey | 11.51 | 11.94 |
| SMTP | 4.86 | 3.20 |
| BitTorrent | 3.18 | 1.07 |
| HTTPS | 2.31 | 3.81 |
| FTP | 0.57 | 0.04 |
| SSH | 0.50 | 0.94 |
| DNS | 0.24 | 4.35 |
| Outros | 31.90 | 31.63 |

### 3.2.1. Classificação com duas classes

Nesta seção serão apresentados os resultados utilizando a metodologia proposta para duas classes. Os fluxos foram considerados de pequeno tamanho (ratos ou classe 1) se tivessem tamanho igual ou inferior a 370 KBytes, e os de grande tamanho (elefantes) aqueles tivessem tamanho maior que 370 KBytes. Os fluxos foram considerados de longa duração (tartaruga) se tivessem duração superior a 46.6 segundos. Em relação a taxa, foram classificados como alta taxa (chitá) aqueles com taxa superior a 12.3 KBps.

É possível perceber um comportamento semelhante nas Figuras 2(a), 2(b) e 2(c), onde a classe 1 tem uma distribuição quase idêntica a distribuição total (classe 1 + classe 2). Devido a utilização de apenas duas classes, não é possível conhecer o comportamento dos fluxos intermediários.




(a) Distribuição do tamanho dos fluxos (b) Distribuição da duração dos fluxos (c) Distribuição da taxa dos fluxos

**Figura 2. Distribuição cumulativa das diferentes classes dos fluxos com N=2**

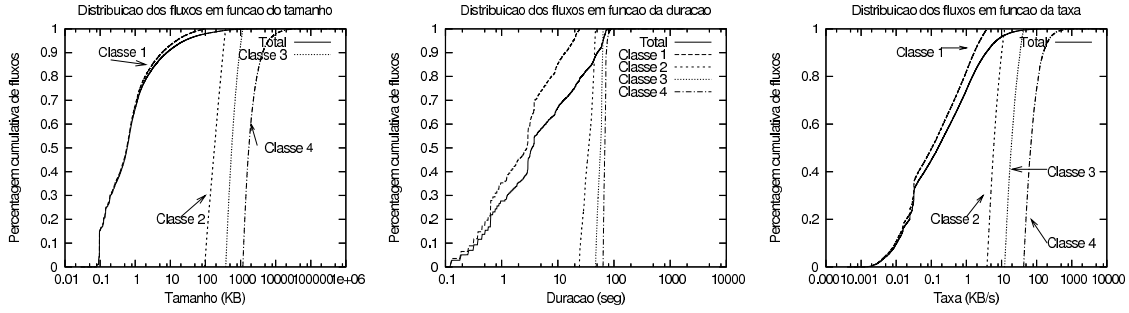
### 3.2.2. Classificação com quatro classes

Nesta seção, os fluxos serão classificados utilizando quatro classes. O principal motivo de se utilizar mais classes se deve ao fato de que a utilização de um número pequeno de classes pode fazer com que haja fluxos de comportamentos substancialmente diferentes pertencentes a uma mesma classe. Um número maior de classes permite realizar distinções importantes. Executando a metodologia para quatro classes algumas diferenças puderam ser identificadas.

A Figura 3(a) mostra a distribuição do tamanho das diferentes classes. Nota-se que, devido ao comportamento de cauda pesada, a classe 1 ainda tem uma distribuição muito semelhante a total. No entanto, já é possível analisar o comportamento de classes intermediárias (2 e 3) que antes estavam mascaradas. As classes 1 e 2 foram divididas em duas. Utilizando quatro classes os fluxos elefantes foram considerados aqueles com tamanho superior a 1.1MB. A classe 3 foi considerada entre 370KB e 1.1MB, sendo os fluxos de tamanho entre 93KB e 370KB classificados como pertencentes a classe 2.

A Figura 3(b) mostra a distribuição da duração das quatro classes. Os fluxos da classe 1 foram classificados como sendo aqueles de duração entre 0.10s e 24.19s. Pode-se perceber que a classe 1 não possui uma curva tão simular em relação a distribuição total. Os fluxos com durações entre 24.19s e 46.80s foram considerados pertencentes a classe 2. A classe 3 foi considerada como sendo de fluxos com durações entre 46.80s e 63.87. Finalmente os fluxos de longa duração (tartaruga) foram aqueles com duração superior a 63.87s.

A Figura 3(c) mostra o comportamento da distribuição da taxa dos fluxos. Os fluxos da classe 1 (taxas entre 0.41Bytes/s e 3.86KBytes/s) apresentaram uma distribuição muito similar da distribuição total. Os fluxos da classe 2, foram considerados de taxa entre 3.86KBps e 12.29KBps. A classe 3 foi composta de fluxos com taxas entre 12.29KBps e 42.80KBps. Os fluxos chitá foram aqueles com taxas superiores a 42.80KBps.




(a) Distribuição do tamanho dos fluxos (b) Distribuição da duração dos fluxos (c) Distribuicao da taxa dos fluxos

**Figura 3. Distribuição cumulativa das diferentes classes dos fluxos com N=4**

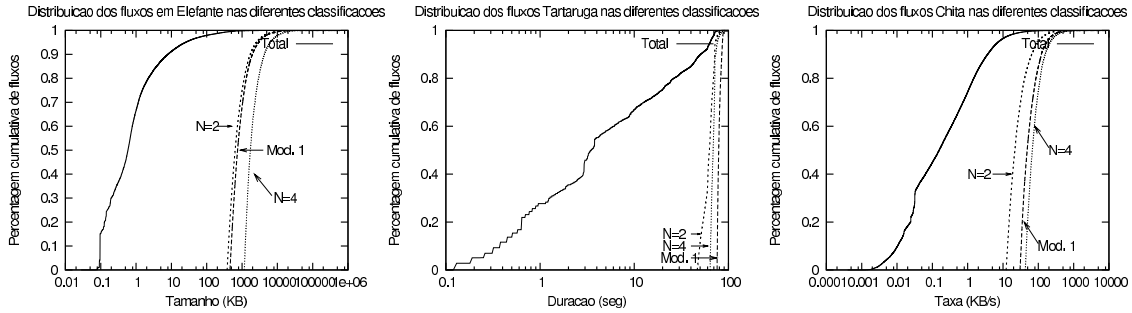
### 3.2.3. Comparações

Nesta seção serão realizadas comparações entre a metodologia proposta e os métodos que utilizam a média e o desvio padrão como forma de classificação. A tabela 3 mostra os valores encontrados.

**Tabela 3. Média e desvio padrão dos Fluxos**

| | Média | Desvio Padrão | Média + 3*desvio padrão |
|---|---|---|---|
| Tamanho (KBytes) | 11.54 | 145.03 | 446.65 |
| Duração (seg) | 14.23 | 20.51 | 75.74 |
| Taxa (KBps) | 1.86 | 9.75 | 30.36 |

Nas Figuras 4(a) a 4(c) é possível observar algumas importantes diferenças entre os dois métodos. Pode-se observar que modelo proposto com $N = 4$ foi o que classificou os fluxos elefantes e chitá com os maiores valores para o tamanho e taxa respectivamente. Como a variável duração foi a que apresentou menor coeficiente de variação, o modelo proposto com $N = 4$ classificou os fluxos tartaruga com um limiar menor que o método que utiliza a média e o desvio padrão. É possível notar também que o método proposto com $N = 2$ foi o que classificou os fluxos como pesados com os menores valores.




(a) Distribuição do tamanho dos fluxos (b) Distribuição da duração dos fluxos (c) Distribuição da taxa dos fluxos

**Figura 4. Comparação das diferentes distribuições cumulativas dos fluxos**

## 3.3. Correlação

Nesta seção serão apresentados diversos resultados obtidos utilizando o modelo proposto para $N = 4$. Usando um número maior de classes algumas distinções importantes podem ser feitas. Utilizando 4 classes é possível obter uma maior granularidade nas classes e assim ter uma idéia melhor de quais fluxos são realmente elefantes. Apesar de alguns estudos focarem nas características dos fluxos, poucos esforços estão sendo feitos para se conhecer como estas características interagem, ou seja, como os fluxos estão relacionados. Por exemplo, os fluxos elefantes são em sua maioria também tartaruga? A resposta para esta questão depende dos enlaces utilizados para transferência de grandes arquivos. Existem várias aplicações para se saber a relação entre os diferentes tipos de fluxos, como por exemplo, métodos de cobrança que podem se basear no tamanho ou na duração dos fluxos trafegados. O conhecimento destas relações também pode oferecer dados importantes para se diferenciar os tráfegos maliciosos dos benignos.

Entender a natureza das taxas dos fluxos na Internet também é importante por diversas razões. Para compreender o quanto o desempenho de uma aplicação irá melhorar aumentando as taxas de transmissão, deve-se primeiramente saber o que está limitando a taxa de transmissão. Fluxos que são limitados por congestionamento da rede devem ter

uma atenção diferente daqueles que são limitados pelo tamanho dos *buffers* das máquinas. Muitos algorimos de roteamento para controlar a utilização do enlace pelos fluxos foram propostos, e o desempenho e a escabilidade destes algoritmos dependem da natureza das taxas dos fluxos [17, 18]. Desta forma, saber mais sobre as taxas pode auxiliar no desenvolvimento destes algoritmos. Além disso, o conhecimento da natureza das taxas dos fluxos pode auxiliar a criação de melhores modelos do tráfego da Internet. Tais modelos podem ser úteis na geração de simulação de modelos e no estudo de problemas na rede.

Zhang *et al* [15] mostrou que existe uma forte correlação entre o tamanho dos fluxos e suas taxas. A hipótese formulada para justificar este comportamento foi que os usuários escolhem o tamanho dos arquivos que vão ser transferidos baseado no tamanho do enlace disponível, isto é, quanto maior o enlace, maior os arquivos transferidos. No entanto o artigo [8] propõe que este comportamento pode ser melhor explicado devido ao comportamento dos protocolos para fluxos pequenos/médios. Tal informação é importante pois pode indicar que algumas mudanças devem ser feitas nos protocolos, como por exemplo aumentar o tamanho da janela inicial do TCP, no intuito de se melhorar o desempenho do protocolo.

Resultados encontrados em [8] indicam que o tráfego *web* é responsável pela grande parte dos fluxos elefantes e que o DNS é o principal responsável pelos tráfegos de longa duração. O artigo mostra também que há uma correlação entre tamanho e duração.

A Tabela 4 mostra como os fluxos estão relacionados. Verifica-se que apenas 0.09% dos fluxos são elefante e tartaruga, e apenas 0.03% são elefante e chitá. Observa-se também uma pequena relação entre os fluxos de alta taxa e alta duração. Ao analisar a coluna de bytes, percebe-se que apesar de uma pequena relação entre os fluxos elefante e tartaruga, eles representam 19.51% do tráfego em bytes. Isto mostra que um quinto do tráfego da rede é composto por fluxos de longa duração e grande tamanho, o que indica um perfil de tráfego contínuo.

**Tabela 4. Percentagem de fluxos pertencentes a duas categorias**

| | Elefante | | Chitá | |
|---|---|---|---|---|
| | % Bytes | % Fluxos | % Bytes | % Fluxos |
| Tartaruga | 19.51 | 0.09 | 8.82 | 0.02 |
| Chitá | 13.05 | 0.03 | - | - |

A Tabela 5 apresenta a probabilidade de um fluxo pertencer a uma categoria, dado que já pertence a outra. É verificado que existe uma grande relação entre os fluxos elefantes e tartarugas. Um fluxo elefante tem probabilidade 77.75% de ser tartaruga. Pode-se observar que apenas 0.32% dos fluxos de longa duração possuem altas taxas de transmissão, confirmando o fato da maioria dos fluxos tartaruga não serem de grande tamanho. Em relação aos fluxos com alta taxa, 6.66% são elefantes e 3.32% são tartaruga.

Ao se analisar a percentagem de bytes da Tabela 5, é possível visualizar uma grande variação. Apesar de apenas 6.66% dos fluxos chitá serem elefante, eles representam cerca de 82% dos bytes, mostrando que uma pequena porção do tráfego chitá é responsável pela maior percentagem de bytes. O mesmo pode ser verificado em relação aos fluxos de longa duração. A única relação em que não há uma grande diferença entre a percentagem de fluxos e bytes é entre os fluxos elefante e tartaruga.

**Tabela 5. Percentagem de fluxos pertencentes a uma categoria dado que pertence a outra**

| | Dado | | | | | |
|---|---|---|---|---|---|---|
| | Elefante | | Tartaruga | | chitá | |
| Valor Esperado | % Fluxos | % Bytes | % Fluxos | % Bytes | % Fluxos | % Bytes |
| Elefante | - | - | 1.77 | 33.47 | 6.66 | 82.96 |
| Tartaruga | 77.75 | 78.05 | - | - | 3.32 | 56.05 |
| Chitá | 28.11 | 52.19 | 0.32 | 15.12 | - | - |

Os estudos [8, 13] mostram que há uma alta correlação entre o tamanho do fluxo e sua taxa. A maioria dos fluxos de alta taxa (93,34%) não são de grande tamanho, o que nos mostra uma característica de tráfego por rajada. No entanto, uma pequena porção de fluxos de alta taxa (fluxos chitá e elefante) responde pela grande maioria dos bytes trafegados. Isto ocorre provavelmente devido a transmissões de grandes arquivos em enlaces de alta velocidade.

Na Tabela 6 pode-se verificar a grande falta de relação entre o número de fluxos e volume de tráfego que cada categoria é responsável. Resultados semelhantes foram verificados em [2] e [7]. A tabela indica que apesar de somente 0.12% dos fluxos serem elefantes, eles representam 25% do tráfego da rede. O mesmo é observado em relação a duração e a taxa, havendo uma grande variação entre a percentagem de fluxos e de bytes. Tal comportamento deve ser analisado, uma vez que as infra-estruturas de rede e os protocolos atuais são desenvolvidos para um tráfego de rajada. A maior utilização do enlace de rede por um longo período de tempo, devido ao aumento da duração e do tamanho dos fluxos, modifica o perfil do tráfego. Em função deste comportamento [3] propõe a utilização de roteadores "sensíveis a carga", que utilizariam rotas especiais para tais tipos de fluxos. Em relação aos fluxos chitá, é observado que os percentuais de fluxos e bytes são próximos, mostrando que os mesmos têm pouco impacto no tráfego de rede.

A Tabela 7 mostra a correlação dos fluxos. O coeficiente de correlação encontrado entre duração e o tamanho dos fluxos indica que há uma pequena correlação entre estas duas variáveis. Este comportamento pode ser explicado pelo fato de que um aumento no tamanho do fluxo, considerando-se um taxa de transmissão constante, aumentará a duração. O coeficiente de correlação entre taxa e tamanho dos fluxos foi o que apresentou maior valor. Este resultado indica que, para uma grande parte do tráfego o aumento de tamanho implica em aumento de taxa, isto ocorre provavelmente devido a transferências de grandes arquivos com altas taxas. O coeficiente de correlação entre taxa e duração foi ligeiramente negativo, mostrando que são poucos os fluxos em que o aumento da taxa é seguido pela diminuição da duração.

**Tabela 6. Percentagem do tráfego de cada categoria**

| | Elefante | Tartaruga | Chitá |
|---|---|---|---|
| Fluxos | 0.12% | 5.10% | 0.49% |
| Bytes | 25.00% | 58.29% | 15.72% |

**Tabela 7. Correlação dos fluxos**

| | Coeficiente de Correlação |
|---|---|
| Tamanho e Duração | 0.15 |
| Tamanho e Taxa | 0.29 |
| Taxa e Duração | -0.07 |

As Tabelas 8 e 9 mostram as correlações das categorias dos fluxos. A Tabela 8 indica que não há correlação entre os fluxos de grande tamanho e longa duração e que há uma pequena correlação negativa entre os fluxos de alta taxa e longa duração, indicando que tais tipos de categorias não possuem relação. O coeficiente de correlação chitá e elefante apresenta o maior valor, mostra que em geral, quando os fluxos são elefantes e chitá, o aumento do tamanho do fluxo é acompanhado pelo aumento da taxa.

**Tabela 8. Correlação entre as categorias**

| | Coef. de Correlação |
|---|---|
| Elefante e Tartaruga | 0.00 |
| Tartaruga e Chitá | -0.01 |
| Chitá e Elefante | 0.39 |

**Tabela 9. Correlação entre os fluxos**

| | Coeficiente de Correlação Dado | | |
|---|---|---|---|
| | Elefante | Tartaruga | chitá |
| Tamanho e Taxa | 0.50 | 0.99 | 0.21 |
| Tamanho e Duração | 0.02 | 0.04 | 0.68 |
| Duração e Taxa | -0.27 | 0.03 | -0.01 |

A Tabela 9 fornece informações sobre a correlação dos fluxos dado que eles pertencem a uma determinada categoria. Algumas correlações chamam atenção. Nos fluxos de alta taxa, o tamanho e a duração desses fluxos apresentam alta correlação, indicando que nos fluxos chitá o aumento de tamanho implica em aumento de duração. Este fato também explica a pequena correlação entre duração e taxa nos fluxos chitá. Quando o tamanho é aumentado, aumenta-se também a duração, e a taxa conseqüentemente sofre pouca alteração. Os fluxos de longa duração apresentam um coeficiente de correlação próximo de 1, indicando que nos fluxos tartaruga, o aumento de tamanho significará em aumento da taxa. Pode-se também observar uma correlação negativa entre duração e taxa nos fluxos de grande tamanho.

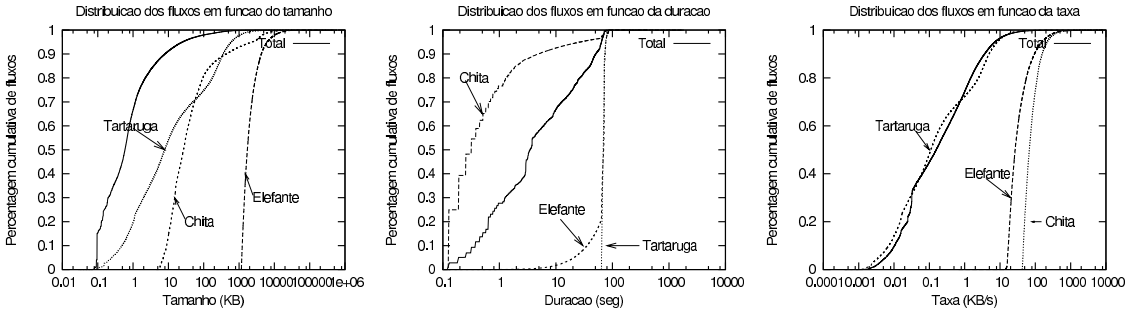
A Figura 5(a) mostra a distribuição dos fluxos de cada categoria em função do tamanho. Na literatura ainda não existe um consenso sobre o que causa fluxos de longa duração. Não se sabe se eles são causados devido ao comportamento dos protocolos/usuário ou devido a transferência de grandes arquivos em enlaces com baixas taxas [8]. Na análise feita é possível observar que cerca de 30% dos fluxos de longa duração tem tamanho de até 2KB, 60% até 14KB e 90% até 345KB, o que mostra que a grande maioria dos fluxos de longa duração são fluxos de pequeno tamanho. Este fato indica que fluxos de longa duração são causados devido à enlaces de baixa capacidade ou ao comportamento de protocolos utilizados. O gráfico mostra também que apenas 6% dos fluxos de alta taxa possuem tamanho superior a 1.1MB.

A Figura 5(b), observa-se que cerca de 22% dos fluxos elefante duram até 63 segundos. Mais de 95% dos fluxos de alta taxa duram menos de 26 segundos.

Finalmente, na Figura 5(c) observa-se que cerca de 71% dos fluxos de grande tamanho têm taxa de até 42KB/s, indicando que a maioria dos fluxos elefantes possuem baixas taxas de transmissão. Os fluxos tartaruga apresentaram as menores taxas, com cerca de 90% dos fluxos possuindo taxas até 4.8KB/s.

A partir dos gráficos analisados, conclui-se que os fluxos elefantes têm grande tamanho, grande duração e pequenas taxas. Os fluxos tartaruga têm pequeno tamanho, grande duração e pequenas taxas. Os fluxos chitá, por outro lado, têm pequeno tamanho,

pequena duração e alta taxa.




(a) Distribuição do tamanho dos fluxos (b) Distribuição da duração dos fluxos (c) Distribuicao da taxa dos fluxos

**Figura 5. Distribuição cumulativa das diferentes categorias dos fluxos**

Na Tabela 10 são mostrados os principais serviços responsáveis por cada categoria de fluxo. Em todas as categorias, a aplicação responsável pela maior parte dos fluxos é a Web. É interessante notar que as aplicações *peer to peer* (E-Donkey e BitTorrent) são responsáveis por 7.12% dos fluxos de grande tamanho e por 17.05% dos fluxos de longa duração, mostrando o importante papel destes tipos de aplicação no tráfego de rede.

**Tabela 10. Percentagem das aplicações mais utilizadas de acordos com as categorias dos fluxos**

| Posição | Elefante (% bytes, % fluxos) | Tartaruga (% bytes, % fluxos) | Chitá (% bytes, % fluxos) |
|---|---|---|---|
| 1 | Web (47.44%, 43.56% ) | Web (26.94%, 18.56% ) | Web (58.25%, 87.61% ) |
| 2 | SMTP (8.34%, 8.50%) | Edonkey (16.36%, 12.90%) | SMTP (11.69%, 4.45%) |
| 3 | Edonkey (4.61%, 7.12%) | BitTorrent (4.66%, 4.15%) | LDM (3.47%, 0.16%) |
| 4 | FTP (1.61%, 1.29%) | SMTP (2.75%, 1.90%) | FTP (1.71%, 0.24%) |
| 5 | HTTPS (1.58%, 1.70%) | LDM (1.41%, 0.43%) | HTTPS (1.67%, 4.07%) |

## 4. Conclusões

A engenharia de tráfego vem se tornando fundamental para as redes de computadores. Para se projetar e gerenciar uma rede é imprescindível que se conheça o tráfego de rede. Conhecendo o comportamento do tráfego é possível projetar equipamentos melhores, implementar mecanismos de qualidade de serviço e de tarifação, além de possibilitar uma identificação de comportamentos anômalos na rede. É também importante conhecer os serviços utilizados e o comportamento dos protocolos para uma melhor gerência da rede.

O trabalho propõe uma nova metodologia para classificação dos fluxos, possibilitando a divisão em $N$ classes. Utilizando a metodologia proposta e os dados obtidos

no estudo de caso, foi possível conhecer o comportamento de uma importante rede de pesquisa.

Pode-se observar através dos gráficos de distribuição que a maior parte dos fluxos da rede é formada por pacotes de tamanho pequeno, de alta taxa e pequena duração. Foi observado também uma grande disparidade entre o volume de fluxos e seu tamanho em bytes. Apesar de uma pequena fração dos fluxos apresentarem grande tamanho, eles são responsáveis por cerca de 25% do tráfego da rede.

Nos estudos sobre correlação foi observado que existe uma grande relação entre o tamanho e a taxa dos fluxos, mostrando que se um fluxo possui um grande tamanho, provavelmente também será de alta taxa. Novamente na correlação pode-se observar uma grande disparidade entre a percentagem de fluxos e bytes relacionados.

As análises feitas sobre os serviços mais utilizados mostram que aplicações *peer to peer* são responsáveis por um grande volume de tráfego, além de serem um dos maiores responsáveis pela percentagem dos fluxos de longa duração e grande tamanho. O crescimento de aplicações P2P tem diversas implicações, a mais importante talvez sendo a mudança no perfil do tráfego de rede.

Como trabalhos futuros, pretende-se fazer comparações mais profundas com os outros métodos propostos e estudar o comportamento dos fluxos ao longo do tempo.

## Referências


[1] "Rede rio." `http://www.rederio.br/`, último acesso em 19/12/2005.

[2] K. C. N.Brownlee, "Understanding internet traffic streams: Dragonies and tortoises," *IEEE Communications*, vol. 40, pp. 110–117, Out 2002.

[3] R. J. Shaikh A. and S. K., "Load-sensitive routing of long-lived ip flows," *Proceedings of the ACM SIGCOMM. ACM*, pp. 215–226, 1999.

[4] P. L. Fang W., "Inter-as traffic patterns and their implications," *Proceeding of IEEE GLOBECOM 99*, pp. 1859–1868, 1999.

[5] J. J. Bhattacharyya S., Diot C. and T. N., "Opop-level and access-link-level traffic dynamics in a tier-1 pop," *Proceeding of ACM SIGCOMM Internet Measurement Workshop*, pp. 39–54, 2001.

[6] L. F. M. de Moraes and G. Vilela, "Caracterização de tráfego utilizando fluxos de comunicação," in *Anais do XXIII Simpósio Brasileiro de Redes de Computadores (SBRC2005)*, (Fortaleza, CE, Brasil), maio 2005.

[7] k. c. Andre Broido, Young Hyun, "Their share: diversity and disparity in ip traffic," *PAM Workshop*, 2004.

[8] J. K.Lan, "On the correlation of internet flow characteristics," 2003. `ftp://ftp.isi.edu/isi-pubs/tr-574.pdf`, último acesso em 19/12/2005.

[9] D. C. Papagiannaki K., Taft N., "Impact of flow dynamics of traffic engineering principles," *INFOCOMM*, 2004.

[10] S. J. e. D. M. R. Cáceres, P. B. Danzing, "Characteristics of wide-area tcp/ip conversation," *Computer Communication Review*, vol. 21, 1991.

[11] W. E. Leland and D. V. Wilson, "High time-resolution measurement and analusis of lan traffic:implications for lan interconnection," *Proceedings of IEEE Infocomm*, pp. 1360–1366, 1991.

[12] K. S. N. T. Konstantina Papagiannaki, Augustin Soule and R. Emilion, "Flow classification by histograms or how to go on safari in the internet," *SIGMETRICS/Performance*, Junho 2004.

[13] Z. Y. and Q. L., "Understanding the end-to-end performance impact of red in a heterogeneous environment," *Cornell CS Technical Report TR*, 2000.

[14] K. Papagiannaki, *Provisioning IP Backbone Networks Based on Measurements*. PhD thesis, Department of Computer Science, University College London, Londres, Inglaterra, 2003.

[15] V. P. Yin Zhang, Lee Breslau and S. Shenker, "On the characteristics and origins of internet flow rates," *Proceedings of ACM SIGCOMM*, Ago 2002.

[16] M. E. Crovella and M. S. Taqqu, "Estimating the heavy tail index from scaling properties," *Methodology and Computing in Applied Probability*, vol. 1, 1999.

[17] S. F. R. Mahajan and D. Wetherall, "Controlling high-bandwidth flows at congested router," *Proceedings of 9th International Conference on Network Protocols*, Novembro 2001.

[18] A. M. H. Martin and J. Cleary, "Analysus of internet delay times," *Proceedings of Passive and Active Measurements workshop*, 2000.