Caracterizando Propriedades Essenciais do Tráfego de Redes através de Técnicas de Amostragem Estratificada

Carlos Kamienski[1,2], Tatiene Souza[1], Stenio Fernandes[1], Guthemberg Silvestre[1], Djamel Sadok[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Caixa Postal 7851 – Cidade Universitária – 50.732-970 – Recife – PE

[2]Centro Federal de Educação Tecnológica da Paraíba (CEFET-PB)
Rua 1º de Maio, 720 – 58015-180 – João Pessoa, PB

tatiene@cox.de.ufpe.br, {cak, sflf, gss2, jamel}@cin.ufpe.br

Abstract

Monitoring backbone traffic is a mandatory and vital task to the network management. Such task may be undertaken by either packet or flow measurements. Although modern routers offer monitoring tools for dealing with flow statistics (e.g. NetFlow), several problems still persist. The main difficulty for a proper measurement is the lack of scalability with respect to the link capacity. This work explores a statistical technique called stratified sampling as a powerful tool for deriving properties of the traffic at the flow level. Our results show that the selected samples are significant enough in order to allow further statistical analysis.

Resumo

O monitoramento de tráfego é uma atividade essencial para o gerenciamento de redes e pode ser realizado através da observação dos pacotes ou fluxos. Apesar dos roteadores atuais possuirem ferramentas de monitoramento (e.g., NetFlow), diversos problemas ainda persistem. O principal obstáculo é a falta de escalabilidade relativa à capacidade dos enlaces e estratégias de amostragem têm sido propostas para otimizar o processo de monitoramento. Este trabalho explora a técnica estatística de amostragem estratificada como ferramenta para descrição do comportamento do tráfego (no nível de fluxos). Nossos resultados revelam que as amostras selecionadas são representativas e contêm informações relevantes e suficientes para posterior análise estatística.

1. Introdução

O monitoramento de tráfego em redes de computadores tem se tornado uma atividade essencial para o gerenciamento ativo e passivo em redes de diversos tamanhos, mas principalmente em *backbones* de Provedores de Serviços Internet (ISP). Grandes esforços têm sido feitos na comunidade científica com o objetivo de compreender profundamente como as características de tráfego de diversas aplicações (tradicionais, como Web, maliciosas como *worms* e vírus, ou emergentes como Peer-to-Peer) afetam o comportamento na infraestrutura de rede. Desta forma, estratégias de medições são

essenciais para identificação de comportamentos anômalos (e.g., repentino alto volume de tráfego gerado), bem como para tarefas sazonais como aumento da capacidade de certos enlaces ou para o estabelecimento de cobrança baseada em utilização.

O trabalho de monitoramento pode ser realizado através da observação dos pacotes ou fluxos atravessando os diversos elementos da infraestrutura de rede do ISP. Apesar desta atividade ter se tornado usual com a ajuda de ferrramentas de monitoramento existentes nos roteadores modernos (e.g., CISCO NetFlow [4], JUNIPER JFlow [8]), diversos problemas ainda persistem. O principal obstáculo atual da abordagem de monitoramento de tráfego baseado em medição (de pacotes ou fluxos) é a falta de escalabilidade relativa à capacidade dos enlaces. Em outras palavras, o monitoramento de enlaces com capacidades muito grandes tem como consequência a geração de enormes volumes de dados [14]. À medida que a capacidade dos enlaces e o número de fluxos aumentam, manter contadores para cada fluxo atravessando os roteadores torna-se caro computacionalmente ou economicamente [2].

Desta forma, diversas estratégias de amostragem têm recentemente sido propostas como forma de otimizar o processo de seleção de pacotes (para contabilização de fluxos) [11]-[19] ou seleção de fluxos (para análise estatística do tráfego original) [6]. O processo de amostragem simples (uniforme) não apresenta resultados adequados porque os fluxos IP geralmente tem distribuições de cauda pesada para seus pacotes e bytes [13]. Algumas técnicas existentes de amostragem são dependentes dos tamanhos dos fluxos em que somente contabilizam fluxos relativamente maiores.

Este trabalho explora técnicas estatísticas de amostragem para a análise de tráfego. O método utilizado é conhecido como amostragem estratificada ótima. A principal contribuição deste trabalho está na aplicação de uma nova estratégia de amostragem como poderosa ferramenta para descrição do comportamento do tráfego (no nível de fluxos). Uma consequência direta da técnica proposta é uma substancial redução no número de fluxos necessários (i.e., o tamanho da amostra selecionada) para o cálculo de medidas descritivas dos dados não amostrados originais. Com isto, há um impacto direto nas estratégias de monitoramente, visto que os pontos de medição podem manter e/ou exportar apenas uma pequena fração do número de registros de fluxos. Os resultados da avaliação mostram que através da amostragem estratificada tem-se resultados satisfatórios e promissores.

Na seqüência deste artigo a seção 2 apresenta alguns trabalhos relacionados com técnicas de amostragem em monitoramento de tráfego; a seção 3 descreve a técnica de amostragem utilizada neste trabalho para análise de tráfego; a seção 4 exibe os principais resultados obtidos; e a seção 5 apresenta as conclusões e os possíveis trabalhos futuros.

2. Trabalhos Relacionados

Existe um grande número de trabalhos recentes relacionados ao problema de amostragem de pacotes e fluxos e sua consequente recuperação das estatísticas originais de tráfego a partir de dados amostrados. Alguns deles concentram-se no problema de amostragem desempenhadas principalmente pelos roteadores, enquanto outros preocupam-se mais com a análise e utilização de recursos num ambiente de coleta e medição [11]-[19]. Métodos de amostragem têm recentemente sido sugeridos com o

objetivo de reduzir o volume dos dados em uma infraestrutura de medição de tráfego. O objetivo principal é estimar algumas propriedades do tráfego original (p.ex., a distribuição dos tamanhos dos pacotes) a partir dos pacotes amostrados. Para obter um bom desempenho com um esforço computacional pequeno, uma variedade de estratégias de amostragem vêm sendo estudadas. O grupo de trabalho Packet Sampling (PSAMP WG) da IETF tem discutido algumas técnicas de filtragem e amostragem na seleção de pacotes IP [19] dentro do arcabouço do protocolo PSAMP [16]. Duffield et al [11] também propõem a substituição da amostragem uniforme pela amostragem dependente de tamanho, no qual um objeto de tamanho X é selecionado de acordo com uma probabilidade dependente de tamanho, p. Em [14] os autores apresentam uma abordagem para inferir a distribuição original da duração dos fluxos de tráfego na Internet, baseado nas estatísticas de fluxo formadas a partir de um conjunto de pacotes amostrados.

Honh e Veicht [12] apresentaram alguns resultados teóricos para o problema de recuperação de estatísticas de tráfego (superior aos momentos de primeira ordem) a partir do tráfego de redes amostrado. O método denominado de Amostragem Invertida foi aplicado na filtragem de pacotes e na filtragem de fluxos. O trabalho modelou os dados em três níveis: no nível de pacote (densidade espectral do processo de chegada de pacotes), no nível de fluxo (a distribuição do número de pacotes por fluxos) e no nível "interno" ao fluxo (a taxa média de chegada de pacotes pertencentes a um único fluxo).

Estan e Varghese [2] propõem dois algoritmos escaláveis para identificação de fluxos grandes, chamados de "sample and hold" e "multistage filters". Baseado no fato de que poucos fluxos de longa duração dominam o volume de tráfego atual na Internet, eles abordam o problema de identificar tais fluxos, sem manter registro de milhões de fluxos pequenos e de curta duração.

Recentemente, Estan et al [3] propuseram soluções para alguns problemas existentes na arquitetura do NetFlow [4]. Um dos problemas abordados é o fato de que o número de registros mantidos e exportados pelo NetFlow tem uma forte dependência com o volume agregado de tráfego. Ou seja, um instantâneo número grande de fluxos pode saturar o processamento de classificação de fluxos no roteador e/ou congestionar o caminho até seu coletor de registros. Neste último caso, a perda de informações de fluxos entre o roteador e o coletor é extremamente danosa para uma análise estatística precisa [11]. Outro problema abordado em [3], refere-se a falta de uma taxa de amostragem dinâmica no NetFlow. Os autores apontam que uma taxa de amostragem fixa é uma tarefa árdua, pois quando o volume agregado de tráfego é pequeno a taxa de amostragem deveria ser alta para obter-se uma melhor precisão no perfil de tráfego. Por outro lado, se o volume de tráfego cresce substancialmente, esta taxa deveria ser substancialmente diminuída para proteger a infra-estrutura de medição. A principal solução proposta pelos autores é um algoritmo dinâmico para a taxa de amostragem (chamada de Adaptive NetFlow – ANF), no qual garante a geração de um número fixo de registro de fluxos, independente do volume de tráfego agregado passando pelo roteador. Consideramos que nossa proposta de utilização da amostragem estratificada no tratamento de registro de fluxos, pode contribuir fortemente na redução do tráfego gerado entre roteadores e coletores (e posteriormente contribuir na precisão da análise estatística). A flexibilidade da proposta de amostragem estratificada permite sua utilização em combinação com outras propostas, tais como a ANF.

3. Amostragem Estratificada

Nesta seção, descrevemos a técnica de Amostragem Estratificada aplicada na análise de tráfego e sua utilização para redução no volume de dados amostrado.

Na amostragem estratificada [5], uma população de N unidades é primeiramente dividida em subpopulações de $N_1, N_2, \ldots, N_L$ unidades, respectivamente. Essas subpopulações não se superpõem e, juntas abrangem a totalidade da população de tal modo, que $N_1 + N_2 + \ldots + N_L = N$. As subpopulações são denominadas *estratos.* Para que se obtenham todos os benefícios da estratificação, os valores de N_h devem ser conhecidos. Depois de determinados os estratos, seleciona-se uma amostra em cada um deles, sendo as seleções feitas separadamente nos diferentes estratos. As grandezas das amostras dentro dos estratos são denominados $n_1, n_2, \ldots, n_L$, respectivamente.

Quando se selecionam amostras acidentais simples em cada estrato, o processo inteiro é denominado *amostragem acidental estratificada.* A estratificação é uma técnica comum que pode proporcionar um aumento de precisão nas estimativas das características da totalidade da população [5]. Em geral, é possível dividir uma população heterogênea em subpopulações que isoladamente sejam homogêneas. Se todos os estratos são homogêneos, no sentido de que o valor das medidas variem pouco de uma unidade para outra, pode-se obter uma estimativa precisa do valor médio de um estrato qualquer mediante uma pequena amostra desse estrato. Por fim, essas estimativas podem ser combinadas para constituírem uma estimativa precisa do conjunto da população.

A amostragem estratificada pode ser classificada em uniforme, proporcional ou de Bowley e ótima. Na amostragem estratificada uniforme os estratos têm o mesmo tamanho, enquanto que na amostragem proporcional o número de elementos em cada estrato é proporcional ao tamanho do estrato. Por fim, a amostragem estratificada ótima considera além do tamanho do estrato e variabilidade dentro do estrato.

Neste artigo, utiliza-se a amostragem estratificada sem reposição com distribuição ótima. Ou seja, os resultados obtidos consideram o tamanho e a variabilidade do estrato. Suponha que se pretenda utilizar a repartição ótima para um determinado n. A grandeza da amostra, n'_h, no estrato h deve ser

$$n \geq \frac{k^2 \sigma_I^2 N - k^2 \sigma_\sigma (N-1)}{\varepsilon^2 (N-1) + k^2 \sigma_I^2} \tag{1}$$

onde,

$$\sigma_\sigma^2 = \frac{\sum N_h \sigma_h^2}{\sum N_h} - \left(\frac{\sum N_h \sigma_h}{\sum N_h} \right)^2 \qquad \sigma_I^2 = \frac{\sum N_h \sigma_h^2}{\sum N_h}$$

k : quantil $(1-\alpha)$ da distribuição normal padrão, ε : erro de precisão.

Neyman [5] estabeleceu um critério de distribuição dos elementos da amostra pelos diferentes estratos a partir da condição de ser mínima a variância resultante. De acordo com esse critério o número n_h de elementos do estrato h, em uma amostragem de n elementos será dado pela expressão:

$$n_h = n \frac{N_h \sigma_h}{\sum N_h \sigma_h}$$

onde, n: tamanho da amostra; N_h: total de unidades; σ_h: desvio padrão dentro dos estratos.O propósito é determinar o tamanho n da amostra que se deve extrair para estimar uma característica qualquer desse universo como, por exemplo, o tamanho médio dos fluxos de tráfego ou sua duração média.

4. Metodologia de Coleta de Tráfego e Análise

Neste artigo foram utilizados quatro conjuntos de dados, cada um contendo traces (rastros) de fluxos de tráfego. Adota-se a definição padrão de fluxo, que é o conjunto de pacotes com os mesmos valores dos campos endereço IP de origem e destino, porta (TCP ou UDP) de origem e destino e protocolo. Os traces foram obtidos no Ponto de Presença de Pernambuco (PoP-PE) da Rede Nacional de Pesquisa (RNP), nos dias 15 a 19 de setembro de 2004. A geração dos arquivos de traces contendo os fluxos foi realizada por um capturador de tráfego baseado em pacotes, desenvolvido especialmente para o Grupo de Trabalho em Computação Colaborativa (GT-P2P), da RNP [7]. Foi coletado todo o tráfego de entrada e saída do PoP-PE, que possuía um enlace de 34 Mbps no período em considerado neste estudo. A Tabela 1 apresenta um resumo das principais características dos traces utilizados.

Tabela 1 – Características dos traces utilizados na análise.

Nome	Data	Horário	Volume (GB)	Número de fluxos
Trace 1	15/09/2004	8h às 12h	28,129	7.013.744
Trace 2	17/09/2004	8h às 12h	37,958	7.497.991
Trace 3	18/09/2004	14h às 18h	23,875	4.086.476
Trace 4	19/09/2004	14h às 18h	62,511	6.571.586

A metodologia utilizada para a aplicação da técnica da amostragem estratificada para fluxos de tráfego foi baseada em seis fases, descritas a seguir:

1. Definição das variáveis: foram analisadas as variáveis volume de tráfego do fluxo em bytes e duração do fluxo em segundos. Elas são utilizadas na maioria dos estudos de análise de fluxos de tráfego porque em conjunto conseguem caracterizar adequadamente o perfil do tráfego;
2. Categorização das variáveis: implica na definição de quantos e quais estratos serão utilizados para categorização de cada variável. Para o volume, foram utilizados oito estratos: menor do que 100B, de 100B a 1KB, de 1KB a 10KB, de 10KB a 100KB, de 100KB a 1MB, de 1MB até 10 MB, de 10 MB até 100MB e acima de 100MB. Para o tempo foram utilizados seis estratos: menos que 1s, de 1 a 10s, de 10 a 100s, de 100 a 1.000s, de 1.000 a 10.000s e acima de 10.000s.
3. Determinação do tamanho da amostra: foram utilizados cinco tamanhos diferentes de amostras, a fim de possibilitar análises das quais se podem tirar sugestões de tamanhos de amostras em situações reais. Os tamanhos de amostra utilizados foram:

- $\mathbf{n_1}$: corresponde a 0.01% do tamanho da população;
- $\mathbf{n_2}$: corresponde a 0.1% do tamanho da população;
- $\mathbf{n_3}$: corresponde a 1% do tamanho da população;
- $\mathbf{n_4}$: corresponde a 10% do tamanho da população;
- $\mathbf{n_5}$: tamanho ótimo obtido através do método de Neyman

Para a determinação de n_5 (tamanho ótimo), foi utilizada a equação (1), utilizando os valores de parâmetros $k = 1.96$ e $\varepsilon = 0.5$, que implicam em um nível de confiabilidade de 95% e uma precisão amostral de nível médio (ε varia entre 0 e 1). Vale ressaltar que as escolhas de k e ε não seguem nenhum modelo, ou seja, são definidos de acordo com critérios subjetivos.

4. Determinação do número de elementos em cada estrato: isso deve ser feito para cada conjunto de dados, para cada tamanho de amostra e para cada uma das duas variáveis observadas.
5. Aplicação do método de simulação estocástica de Monte Carlo, que consiste em escolher uma amostra para cada estrato de cada variável. Esse procedimento foi repetido 1000 vezes (número de réplicas) e utilizou-se o software estatístico R [20]. Em cada replicação, foi coletado o valor de cada métrica e ao final foram calculados a média e o desvio padrão e o intervalo de confiança assintótico ao nível de 99 %.
6. Cálculo e comparação de algumas medidas descritivas (métricas) da amostra e da população. As métricas utilizadas foram a média (do tamanho e da duração dos fluxos), o desvio padrão e a soma. Estas medidas são importantes para que seja possível armazenar uma amostra do tráfego e posteriormente utilizá-la no lugar da população. Por exemplo, uma provável utilização para a soma do volume de tráfego poderia ser num sistema de cobrança. Os valores da média e do desvio padrão amostral foram obtidos através de uma ponderação que considera o tamanho de cada estrato (ou seja, uma média ponderada).

5. Resultados da Análise

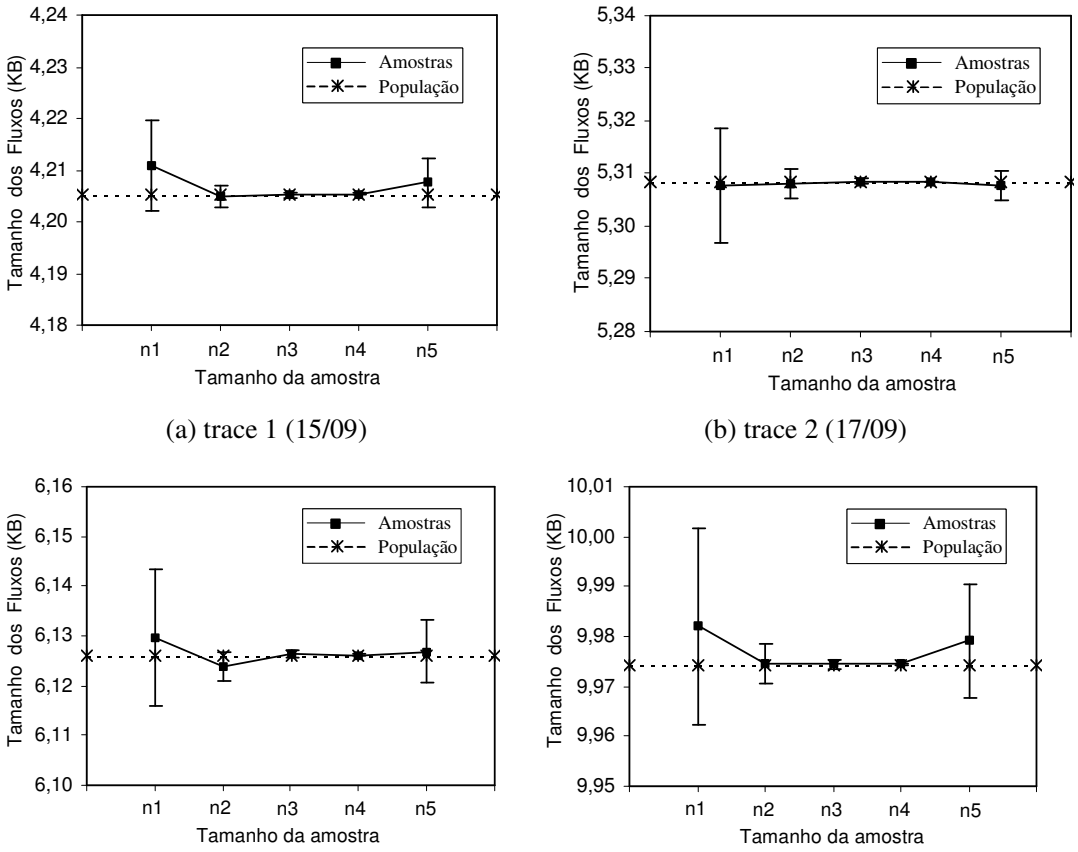
As seções seguintes descrevem os resultados obtidos na análise dos traces de fluxos usando a amostragem estratificada, conforme descrito na seção 4. Em todos os gráficos das sub-seções 5.1 e 5.2 os valores representam a média da métrica analisada e as barras verticais representam os intervalos de confiança assintóticos ao nível de 99 % de acordo com os valores das 1000 replicações realizadas. Foram também calculados os valores para os cinco tamanhos amostrais (n_1, n_2, n_3, n_4, n_5), que estão representados nos gráficos. É importante enfatizar que o tamanho amostral n_5 refere-se ao tamanho ótimo obtido através do método de Neyman. Para todos os conjuntos de dados analisados e variáveis (volume e duração), o valor de n_5 ficou entre os valores de n_1 e n_2, respectivamente 0,01 e 0,1 % do tamanho populacional.

5.1. Volume de Tráfego

A Figura 1 apresenta os gráficos da média da variável volume para os quatro conjunto de dados estudados (trace 1, trace 2, trace 3, trace 4). Através da análise gráfica, nota-se que à medida que o tamanho amostral é aumentado (n_1 a n_4), mais preciso é o intervalo

de confiança, isto é, menor é a sua amplitude (diferença entre o limite superior e o limite inferior) como pode ser visto na Figura 1. A linha pontilhada em cada gráfico é o valor da média populacional, que pertence a todos os intervalos de confiança das amostras para os quatro conjuntos de dados considerados. Isto significa que os intervalos construídos a partir das amostras são representativos o que permite utilizar as amostras no lugar da população para esta variável. Por exemplo, na Figura 1a a média populacional é 4,21 que corresponde a linha pontilhada pertence aos intervalos de confiança para todos os tamanhos amostrais.

A importância destes resultados reside em ressaltar o compromisso existente entre tamanho de amostra e precisão (e variabilidade) nos resultados. Por exemplo, para obter uma alta precisão, deve-se escolher amostras maiores, porque em geral para amostras com tamanho a partir de 1% (n_4) da população, os resultados das amostras são muito próximos aos da população. Por outro lado, quando uma precisão menor puder ser suportada, é possível utilizar tamanhos amostrais de até 0,01% (n_1) ou então o tamanho ótimo de Neyman.



(a) trace 1 (15/09)

(b) trace 2 (17/09)

(c) trace 3 (18/09)

(d) trace 4 (19/09)

Figura 1 – Média do tamanho dos fluxos para os quatro traces analisados.

Levando em consideração que os resultados para os quatro traces da Figura 1 apresentam características semelhantes, por restrições de espaço a partir de agora os resultados serão apresentados apenas para o conjunto 'trace 3', do dia 18/09/2004 (a

escolha do dia foi aleatória). Neste trace o tamanho da população (N) é igual a 4.086.476 fluxos. Conseqüentemente, os valores dos tamanhos amostrais são: n_1 = 409; n_2 =4.085; n_3 = 40.864 e n_4=408.645. O tamanho n_5 deve ser calculado separadamente para cada variável, uma vez considera a variabilidade e o tamanho do estrato, de modo que n_5=905 para a variável tempo e n_5=1.313 para a variável volume.

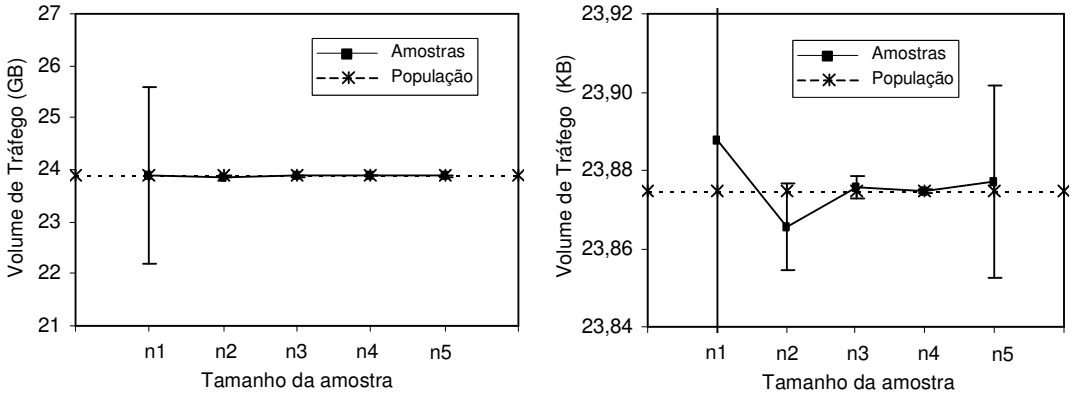
A Tabela 2 mostra o tamanho populacional (N) dentro de cada estrato e os tamanhos amostrais (n_1, n_2, n_3, n_4, n_5) da variável volume (bytes). Por exemplo, dentro do estrato 5 que são os valores entre 100KB e 1M existem 6221 informações. Em n_3 é preciso uma amostra de apenas 2696 para representar esta população, enquanto que em n_4 é necessário uma amostra de tamanho 6221, ou seja, neste caso o tamanho amostral é igual ao tamanho populacional.

Tabela 2 - Tamanho populacional (N) e amostral (n_1, n_2, n_3, n_4, n_5) por estrato da variável volume (em bytes) do trace 3 (dia 18/09).

Estrato	N	n_1	n_2	n_3	n_4	n_5
1: <100B	2.108.964	10	801	14.937	169.125	55
2: 100B-1KB	1.527.082	14	650	11.513	129.429	63
3: 1KB-10KB	231.631	12	198	2.738	29.551	42
4: 10KB-100KB	210.636	57	635	7.040	72.379	184
5: 100KB-1MB	6.221	27	267	2.696	6.221	85
6: 1MB-10MB	1.314	91	907	1.314	1.314	291
7: 10MB-100MB	607	178	607	607	607	572
8: >=100MB	21	21	21	21	21	21
Total	**4.086.476**	**410**	**4.086**	**40.866**	**408.647**	**1.313**

A Figura 2 apresenta a soma da variável volume em escala normal (a) e em escala reduzida (b). Ou seja, a Figura 2b faz um zoom na Figura 2a para apresentar em mais detalhes os resultados para n_2, n_3, n_4 e n_5 uma vez que o intervalo de confiança para n_1 teve uma amplitude significativamente maior do que os outros tamanhos de amostra. Pode-se observar que a soma populacional (23,88) pertence a todos os intervalos construídos a partir das diferentes amostras selecionadas.

O fato da soma populacional poder ser representada pela soma amostral pode ser utilizado para fazer cobrança de utilização com base em pequenos percentuais de dados armazenados. Por exemplo, no caso de uma amostra de tamanho 0.01% do tamanho populacional, como n_1 na Figura 2a, um provedor poderia cobrar o usuário pelo limite inferior do intervalo de confiança (no caso 22,2 GB) com uma confiabilidade estatística de 99%.

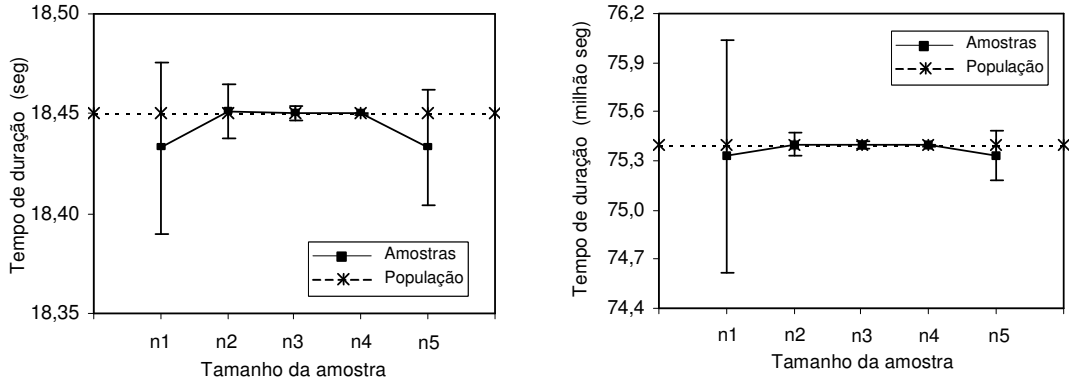


(a) soma (b) soma (escala reduzida)

Figura 2 – Soma do tamanho dos fluxos para o trace 3 (18/09); (a) escala normal; (b) escala reduzida (zoom).

5.2. Duração

A Figura 3 apresenta os gráficos da média e da soma da variável tempo para o trace 3. Através da análise gráfica, nota-se que à medida que se aumenta o tamanho amostral, mais preciso é o intervalo de confiança. Além disso, pode-se observar claramente que a média populacional (18,45) que corresponde à linha pontilhada em (a) e a soma populacional (75,4) que corresponde à linha pontilhada em (b) pertencem aos seus respectivos intervalos de confiança para os diferentes tamanhos amostrais considerados. É possível observar também que para as duas métricas, tamanhos amostrais pequenos (n_1 e n_5) têm uma tendência de subestimar o real valor dos parâmetros da população.



(a) média (em segundos) (b) soma (em milhões de segundos)

Figura 3 – Média e soma do tempo de duração dos fluxos para o trace 3 (18/09).

A Tabela 3 apresenta o tamanho populacional (N) dentro de cada estrato e os tamanhos amostrais (n_1, n_2, n_3, n_4, n_5) da variável tempo (segundos). Por exemplo, dentro do estrato 3, que são os valores entre 10 e 100 segundos, existem 391.609 informações. Para n_1 é necessária uma amostra de tamanho 80, enquanto que para n_5 a amostra é de tamanho 177 (para representar a população do estrato 3). Através da

fórmula (1) (equação de Neyman utilizada para obter o tamanho ótimo), é necessária uma amostra total de tamanho igual a 905.

Tabela 3 - Tamanho populacional (N) e amostral (n_1, n_2, n_3, n_4, n_5) por estrato da variável tempo de duração (em segundos) do trace 3 (dia 18/09).

Estrato	N	n_1	n_2	n_3	n_4	n_5
1: <1	1.839.265	3	46	1.983	60.716	8
2: 1|-10	1.696.252	44	448	5.884	96.551	97
3: 10|-100	391.609	80	800	8.333	92.031	177
4: 100|-1000	154.751	199	1.993	20.065	154.751	441
5: 1000|-10000	4.400	60	599	4.400	4.400	132
6: >=10000	199	23	199	199	199	50
Total	**4.086.476**	**409**	**4.085**	**40.864**	**408.645**	**905**

5.3. Análise por Estrato

Através dos resultados apresentados nesta seção pode-se ter uma visão geral do que ocorre dentro de cada estrato, uma vez que a técnica de amostragem estratificada é a base deste artigo. Nesta subseção, todas os resultados se referem ao tamanho ótimo de amostra, calculado pelo método de Neyman (ou seja, n_5).

A Tabela 4 apresenta a média, o desvio padrão e a soma da população e da amostra, e seus respectivos vieses relativos (percentuais) para a variável volume (em bytes). Os principais resultados encontram-se resumidos a seguir. Primeiro, em relação à média podemos observar que em todos os casos a média da amostra superestima a média da população, exceto no estrato 4 e 5 onde os vieses relativos são -0.10 % e –0.17 %, respectivamente. Segundo, os valores da média geral populacional e amostral (6,13) são iguais para este tamanho de amostra (n_5), que corresponde a 0.032% do tamanho da população. Terceiro, através da estimativa do desvio padrão total (485,63 KB) percebe-se que este valor subestima o desvio padrão populacional (493,68 KB), que produz um viés relativo de -1.63%. Quarto, é possível afirmar que a diferença entre a soma populacional e amostral é relativamente pequena. Por exemplo, no estrato 2 a soma populacional é 466,91 e a soma amostral é 467,76 o que resulta uma diferença de apenas 0.04.

A Tabela 5 apresenta a média, desvio padrão e a soma da população e da amostra, e seus respectivos vieses relativos percentual para a variável tempo (em segundos). Os principais resultados encontram-se resumidos a seguir. Primeiro, em relação à média podemos observar que em todos os casos a média da amostra subestima a média da população, exceto para o estrato 2 e 3 onde os vieses são 0,22 % e 0,09 %, respectivamente. Segundo, os valores da média geral populacional (18,45) e da média geral amostral (18,43) estão bastante próximos o que produz um viés relativo pequeno (-0,11 %). Terceiro, a maior diferença entre a média amostral e a média populacional é apresentada no estrato 5 onde o viés é -0,40 %. Quarto, em relação à segunda medida descritiva calculada, desvio padrão, nota-se que a estimativa do desvio padrão da amostra subestima o desvio padrão populacional em todos os casos, exceto para o estrato 2. Quinto, através da estimativa do desvio padrão total (164,92) percebe-se que

este valor é bastante distante do desvio padrão populacional (192,16) o seu viés relativo é de -14,18 %. Sexto, em relação à soma dos valores da amostra é relativamente próxima da soma dos valores da população. Por exemplo, a soma da população do estrato 3 é 16.540,16 enquanto que para a amostra este valor é 16.556,04. A diferença relativa entre esses valores é apenas 0,10 %, onde o tamanho da amostra é 905 que corresponde a 0.022% do tamanho da população (n_5). Sétimo, a soma geral populacional é 75.396,61 e a amostral é 75.328,03 e a diferença relativa percentual entre esses valores é –0,09 %, pode-se notar que a estimativa da soma amostral subestima a soma populacional.

Tabela 4 – Média, desvio padrão e soma da variável volume de tráfego para a população e para o tamanho ótimo de amostra (n_5) por estrato do trace 3 (dia 18/09).

Estr.	Média (KB)			Desvio padrão (KB)			Soma (MB)		
	População	Amostra	Viés (%)	População	Amostra	Viés (%)	População	Amostra	Viés (%)
1	0,0739	0,0740	0,14	0,0209	0,0209	0,00	152,29	152,39	0,06
2	0,3131	0,3137	0,19	0.2155	0.2163	0,37	466,91	467,76	0,04
3	2,8957	2,8976	0,07	2,0410	2,0522	0,55	655,01	655,45	0,07
4	20,35	20,33	-0,10	11,25	11,25	0,02	4.185,13	4.182,51	-0,06
5	243,14	242,73	-0,17	181,76	180,84	-0,51	1.477,11	1.474,62	-0,01
6	4.128,53	4.132,11	0,09	2.941,59	2.943,19	0,05	5.297,75	5.302,33	0,09
7	14.672,24	14.675,20	0,02	12.506,17	12.520,32	0,11	8.697,32	8.699,06	0,02
8	171.453,64	171.453,64	0,00	77.101,68	77.101,68	0,00	3.516,14	3.516,14	0,00
Total	6,13	6,13	0,00	493,68	485,63	-1,63	24.447,65	24.450,27	0,10

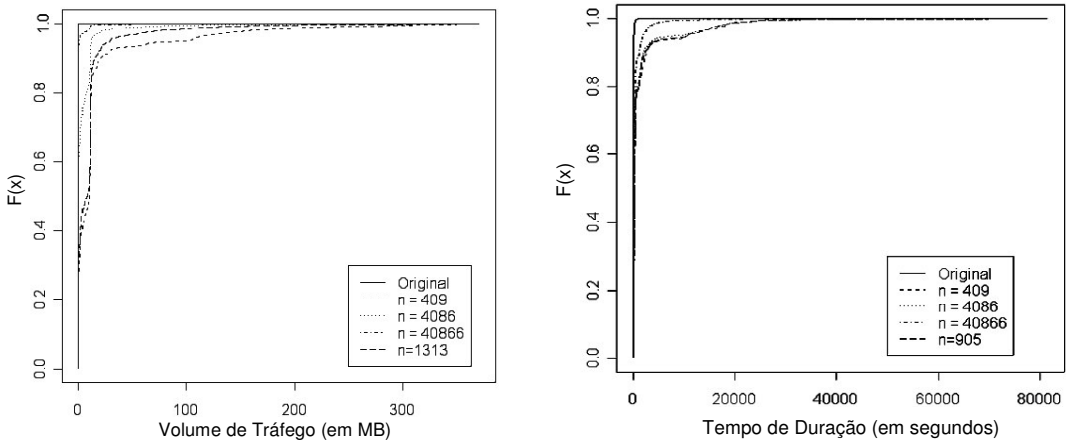
Tabela 5 – Média, desvio padrão e soma da variável tempo de duração para a população e para o tamanho ótimo de amostra (n5) por estrato do trace 3 (dia 18/09).

Estrato	Média			Desvio padrão			Soma (1000 segundos)		
	População	Amostra	Viés (%)	População	Amostra	Viés (%)	População	Amostra	Viés (%)
1	0,10	0,10	-0,00	0,21	0,20	-0,05	188,63	185,21	-1,18
2	4,64	4,65	0,22	2,83	2,83	0,00	7.866,53	7.889,85	0,30
3	42,24	42,28	0,09	22,34	22,33	-0,04	16.540,16	16.556,04	0,10
4	238,83	238,51	-0,12	141,11	141,07	-0,02	36.959,33	36.909,11	-0,14
5	2.257,06	2.247,93	-0,40	1.491,11	1.476,73	-0,96	9.931,05	9.890,87	-0,40
6	19.652,76	19.582,64	-0,36	12.385,19	12.246,31	-1,12	3.910,90	3.896,95	-0,36
Total	18.45	18.43	-0.11	192,16	164,92	-14.18	75.396.61	75.328.03	-0.09

5.4. Distribuição de Probabilidade

A Figura 4 apresenta a função de distribuição empírica acumulada (ECDF) para a população e para os diferentes tamanhos amostrais, consideradas as variáveis volume (Figura 4a) e tempo (Figura 4b). A visualização gráfica proporcionada pela figura

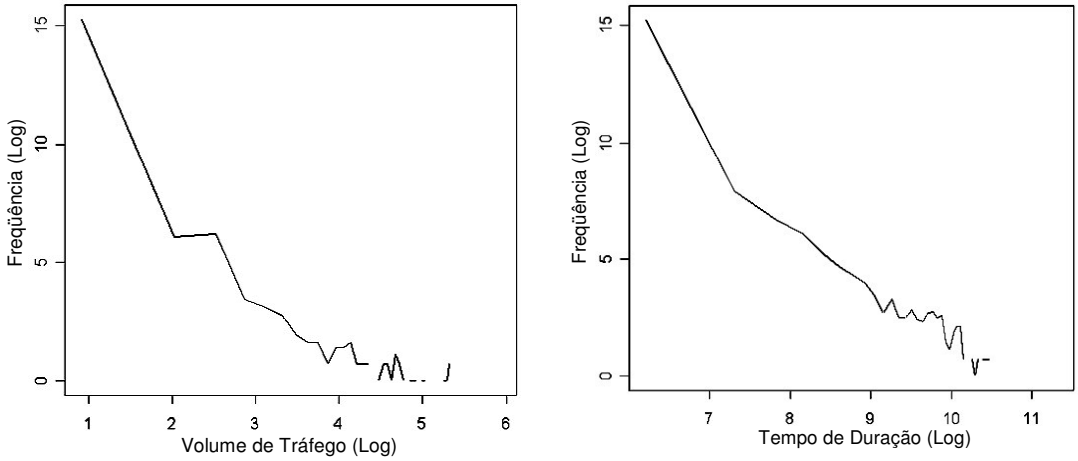
corrobora com as conclusões apresentadas anteriormente. Em outras palavras, pode-se observar que para os dois casos apresentados a ECDF das amostras acompanha a ECDF da população. A suavização das curvas (para volume e tempo) é maior para tamanhos menores de amostras. Isso se deve ao fato de que, conforme mostrado na Tabela 2 e Tabela 3, para tamanhos pequenos de amostras um menor número de elementos dos estratos inicias é tomado, devido à baixa variabilidade da sua população. À medida que o tamanho amostral aumenta, a amostra segue mais fielmente a curva da população, porque mais elementos dos estratos iniciais são incluídos.



(a) (b)

Figura 4 – Função de Distribuição Empírica Acumulada (ECDF) empírica das variáveis volume e tempo de duração.

A Figura 5 apresenta os gráficos da distribuição das variáveis volume e tempo, respectivamente, em escala log-log (escala logarítmicas nos dois eixos). Nota-se que estas distribuições aparentam ter as propriedades Zipf (distribuição de lei de potência), ou seja, um grande número de ocorrências de valores muito pequenos e a existência de cauda pesada para ocorrência de valores muito grandes.



(a) (b)

Figura 5 – Distribuição empírica de volume e tempo em escala log-log.

Em outras palavras, uma distribuição de Zipf indica a freqüência de ocorrência de valores muito pequenos e a existência de cauda pesada para ocorrência do *i*-ésimo valor é proporcional a $i^{-\alpha}$, onde α é o parâmetro ou coeficiente da distribuição. Estas distribuições têm uma aparência linear quando plotadas em escalas logarítmas, como mostra a Figura 5. A existência de leis de potência na Internet vem sendo estudada há alguns anos e suas implicações são bem conhecidas para a modelagem de tráfego e o planejamento de redes, por exemplo.

6. Conclusões

Neste artigo foi apresentado um estudo da viabilidade da utilização da técnica de amostragem estratificada para o tráfego de redes constituído de traces de fluxos, com o objetivo de diminuir a quantidade de dados que precisam ser armazenados e processados. Foram analisados quatro traces, obtidos no PoP-PE da RNP, no mês de setembro de 2004. Os resultados apresentados revelam que as amostras selecionadas para todos os casos considerados (variáveis e métricas) são representativas da população. Isto é, os estratos das amostras carregam informações relevantes da população, a partir das quais é possível inferir, por exemplo, a média ou soma do tamanho e da duração dos fluxos. Foi mostrado que isto é possível inclusive para tamanhos amostrais de apenas 0,01 % ou 0,1 % do tamanho da população.

Os resultados deste artigo podem ser diretamente aplicados pela comunidade de redes de computadores, como provedores de serviços, fabricantes de equipamentos e pesquisadores. Os provedores podem utilizar a amostragem estratificada, por exemplo, para efetuar contabilização e cobrança baseada em utilização (volume). Como uma amostra pequena é suficiente para representar a soma da população, a cobrança pode ser feita através da soma da amostra. Fabricantes podem utilizar para otimizar e diminuir os requisitos de hardware, além de implementar a amostragem estratificada diretamente nos equipamentos. Pesquisadores podem prescindir de todo um conjunto completo de dados, para fazer modelagem e análises envolvendo grandes quantidades de dados. Dessa forma, análises de períodos mais longos se tornam possíveis, uma vez que somente amostras precisam ser armazenadas.

Este trabalho apresenta várias possibilidades para identificar trabalhos futuros. A mais importante diz respeito à utilização da amostragem estratificada para a amostragem em tempo real em roteadores (que seria objeto de implementação por fabricantes). A técnica apresentada neste artigo assume que se possui todo o conjunto de fluxos de dados para o cálculo dos estratos e escolha dos elementos (fluxos) em cada estrato. Do ponto de vista de diminuir a necessidade de armazenamento, uma opção mais interessante é fazer a amostragem diretamente nos roteadores, ou seja, o roteador pode descartar pacotes seletivamente de acordo com uma probabilidade pré-computada para cada estrato, baseada na freqüência relativa do tamanho do estrato amostra em relação ao estrato populacional. Outros trabalhos futuros incluem a comparação entre o método bootstrap [6] e a técnica de amostragem estratificada ótima e por fim, a aplicação da técnica de amostragem estratificada com distribuição ótima considerando o custo de armazenamento para os dados.

Referências

[1] A. Kumar, J. Xu, L. Li, and J. Wang, “Space-code bloom filter for efficient traffic flow measurement" in Proc. ACM SIGCOMM Internet Measurement Conference, 2003.

[2] C. Estan & G. Varghese. “New Directions in Traffic Measurement and Accounting”, Proceedings of the ACM SIGCOMM 2002, August 2002.

[3] C. Estan, K. Keysy, D. Moore & G. Varghese, “Building a Better NetFlow”, ACM SIGCOMM 2004, Aug. 30-Sept. 3, 2004, Portland, Oregon, USA

[4] CISCO NetFlow, http://www.cisco.com/warp/public/732/Tech/nmp/netflow.

[5] Cocharan, William G., *Sampling Techniques*, 3ª ed. New York: John Willey, 1977.

[6] Fernandes, S., Correia, T., Kamienski, C., Sadok, D. & Karmouch, A., “Estimating Properties of Flow Statistics using Bootstrap”, IEEE MASCOTS 2004, Outubro 2004.

[7] Grupo de Trabalho em Computação Colaborativa (GT-P2P) – Rede Nacional de Pesquisa, http://www.rnp.br/pd/gts2004-2005/p2p.html, acessado em dezembro/2004.

[8] JUNIPER JFlow IP Stats, http://www.juniper.net/products/junose/105017.html

[9] K. Papagiannaki, N. Taft, Z.-L. Zhang, C. Diot. "Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models". IEEE INFOCOM 2003, Mar 2003.

[10] K. Papayanaki, N. Taft, C. Diot. "Impact of Flow Dynamics on Traffic Engineering Design Principles". IEEE Infocom 2004. 7-11 March 2004. Hong-Kong.

[11] N. Duffield, C. Lund. “Predicting Resource Usage and Estimation Accuracy in an IP Flow Measurement Collection Infrastructure”. ACM Internet Measurement Conference 2003.

[12] N. Hohn & D. Veitch. “Inverting Sampled Traffic”, ACM Internet Measurement Conference - IMC’03, October 27–29, 2003, Miami Beach, Florida, USA.

[13] N.G. Duffield, C. Lund, M. Thorup, “Charging from sampled network usage”, Proceedings of the ACM SIGCOMM Internet Measurement Workshop, Nov. 2001.

[14] N.G. Duffield, C. Lund, M. Thorup, “Estimating Flow Distributions From Sampled Flow Statistics”, Proceedings of the ACM SIGCOMM 2003, Karlsruhe, Germany, Aug. 2003.

[15] N.G. Duffield, et al., “Properties and Prediction of Flow Statistics from Sampled Packet Streams”, ACM SIGCOMM Internet Measurement Workshop, November 6-8, 2002.

[16] Nick Duffield (Editor), “A Framework for Packet Selection and Reporting”, Internet Draft, October 2004 (Expires: April 2005)

[17] Ratul Mahajan et. al., “Controlling High Bandwidth Aggregates in the Network”, ACM SIGCOMM Computer Communication Review, Volume 32 , Issue 3 (Jul 2002), 2002

[18] Silvestre, G, Kamienski, C., Fernandes, S., & Sadok, D., “Análise Quantitativa e Qualitativa de Tráfego P2P baseada na Carga Útil dos Pacotes”, submetido ao SBRC 2005, Dezembro 2004.

[19] T. Zseby, M. Molina, F. Raspall, N. Duffield, S. Niccolini, “Sampling and Filtering Techniques for IP Packet Selection”, Internet Draft, Expires: April 2005, October 2004.

[20] The R Project for Statistical Computing, http://www.r-project.org/, accessed March 05.