Universidade do Minho
Escola de Engenharia

David Esteves Magalhães Martins

Impacto da utilização de técnicas de amostragem na caracterização de fluxos de tráfego

Julho, 2013

David Esteves Magalhães Martins **Impacto da utilização de técnicas de amostragem na caracterização de fluxos de tráfego**

UMinho | 2013

Universidade do Minho
Escola de Engenharia

David Esteves Magalhães Martins

Impacto da utilização de técnicas de amostragem na caracterização de fluxos de tráfego

Dissertação de Mestrado
Mestrado Integrado em Engenharia de Comunicações

Trabalho realizado após orientação da
Professora Doutora Solange Rito Lima

Julho, 2013

Agradecimentos

Esta dissertação marca a conclusão de um fase bastante importante da minha vida, contudo, a realização desta não seria possível sem a ajuda e apoio de algumas pessoas. Por isso, gostaria de agradecer a todos os que contribuíram para realização da mesma.

Em primeiro lugar gostaria de agradecer à minha orientadora, Professora Doutora Solange Rito Lima, pelo apoio e orientação científica, assim como pela disponibilidade que sempre apresentou para me ajudar no desenvolvimento deste trabalho.

Quero agradecer também ao doutorando João Marco Silva, pela constante disponibilidade e ajuda prestada em aspectos importantes deste trabalho.

Por fim, um agradecimento especial aos meus pais, Marcelino e Lúcia, pela paciência e apoio incondicional durante todo o período referente à minha formação, e também a Misia Plusza por me ter apoiado e encorajado numa fase importante da conclusão deste ciclo.

Abstract

The constant development of the Internet and underlying transmission technologies, together with the increasing popularity of provided services, such as multimedia applications and applications using P2P technologies, are contributing to the continuous growth of the network traffic in volume and diversity.

To be able to handle such amount of data while assuring the quality and operation of provided services, traffic-measuring tools are required to implement mechanisms that scale and have a minimum interference on the normal network behaviour.

One of the most common solutions for this purpose involves the implementation of measurement techniques based on traffic sampling. These techniques aim to provide accurate estimations of traffic behaviour and characteristics by processing fractions of the original network traffic.

Another fundamental area of traffic monitoring concerns the traffic classification and characterization, as it supports important tasks such as resource allocation, planning and management, security and quality of service. Attending to this, added up to the mentioned growth in traffic volumes, it is likely that traffic classification and characterization will be increasingly supported by traffic sampling mechanisms.

This work aims to study the impact of traffic sampling mechanisms on the accuracy of traffic flows characterization. This was carried out through the application of classical and adaptive traffic sampling techniques to real traffic traces, which were captured in different real scenarios and opened to public access. The resulting sampled data is then organized under the form of flow records, which are then classified according to their transport and application protocols.

The performance of the distinct traffic sampling techniques in enabling a correct characterization of traffic flows was then assessed taking into account multiple metrics applied to the flow records of sampled traffic, compared to the metrics of the full original traffic.

Resumo

O constante desenvolvimento das tecnologias relacionadas com a Internet e transmissão de dados, juntamente com a crescente popularidade de serviços fornecidos, como aplicações multimédia e aplicações que utilizam tecnologias P2P, contribuem para um contínuo crescimento quer da diversidade do tráfego quer do volume de dados que circulam nas redes.

Para permitir lidar com tais quantidades de dados e ao mesmo tempo garantir a qualidade e o funcionamento dos serviços prestados, as ferramentas de medição de tráfego necessitam de implementar mecanismos escaláveis e de interferir o mínimo possível no comportamento normal da rede.

Uma das soluções mais comuns envolve a implementação de técnicas de medição baseadas em amostragem de tráfego. Estas técnicas têm como objectivo proporcionar estimativas precisas do comportamento do tráfego e suas características através do processamento de fracções do tráfego real.

Outra área fundamental da monitorização de tráfego refere-se à sua classificação e caracterização. Esta suporta tarefas importantes, tais como a alocação de recursos, o planeamento e gestão, a segurança e a qualidade de serviço. Dada a importância desta área adicionado ao mencionado facto do crescimento dos volumes de tráfego, os mecanismos de classificação e caracterização conjugados com mecanismos de amostragem de tráfego constituem um possível cenário cuja adopção futura é tida não só como útil mas também como necessária.

Através da realização deste trabalho, pretende-se contribuir para o estudo do impacto dos mecanismos de amostragem de tráfego na acurácia da caracterização de fluxos de tráfego de rede. Para isso, foram aplicadas técnicas clássicas de amostragem e amostragem adaptativa a colectas de tráfego reais, capturados em diferentes cenários e disponíveis para acesso ao público. Os conjuntos de dados resultantes foram então organizados sob a forma de registos de fluxos e classificados de acordo com o protocolo de transporte e protocolo aplicacional.

O desempenho das diferentes técnicas de amostragem em sustentarem a correcta caracterização de fluxos de tráfego foi avaliado tendo em conta um conjunto de métricas aplicadas aos registos de fluxos do tráfego amostrado, comparando-as com as métricas do tráfego total original.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AS	*Autonomous System*
CAIDA	*The Cooperative Association for Internet Data Analysis*
CPU	*Central Processing Unit*
DNS	*Domain Name System*
DPI	*Deep Packet Inspection*
FTP	*File Transfer Protocol*
HTTP	*Hypertext Transfer Protocol*
HTTPS	*Hypertext Transfer Protocol Secure*
IANA	*Internet Assigned Numbers Authority*
IAT	*Inter Arrival Time*
ICMP	*Internet Control Message Protocol*
IETF	*Internet Engineering Task Force*
IGMP	*Internet Group Management Protocol*
ISP	*Internet Service Provider*
IP	*Internet Protocol*
IPFIX	*IP Flow Information Export*
MDNS	*Multicast Domain Name System*
ML	*Machine Learning*
P2P	*Peer-to-peer*
PCAP	*Packet Capture*
QoS	Qualidade de Serviço (*Quality of Service*)
SCTP	*Stream Control Transmission Protocol*
TCP	*Transmission Control Protocol*
UDP	*User Datagram Protocol*
VoIP	*Voice over IP*

Capítulo 1

Introdução

Actualmente, a Internet é um dos meios de comunicação mais importantes da sociedade devido a sua ubiquidade, rapidez e fiabilidade na comunicação de informação. A sua natureza versátil e flexível incentiva a que cada vez mais esta seja uma ferramenta indispensável na comunicação, incluindo ambos fins pessoais e comerciais. Cada vez mais dispositivos electrónicos de variadas gamas possuem forma de se conectarem à rede, desde os tradicionais computadores pessoais e servidores, a todo o tipo de dispositivos *wireless*, sistemas embebidos, sensores, utensílios domésticos, etc.

Estes são importantes motivos que estimulam o constante crescimento não só da actual vasta gama de serviços e protocolos utilizados nas redes, mas também do desenvolvimento da tecnologia que suporta o transporte de informação. Actualmente, meios de transmissão em fibra óptica e tecnologias *wireless* em conjunto com *routers* e *switches* cada vez mais inteligentes, possibilitam às redes operar em velocidades de transmissão na ordem do Gigabit/s, permitindo a circulação de cada vez maiores volumes de tráfego.

Este crescente desenvolvimento e utilização das tecnologias relacionadas com transmissão de informação contribui para o aumento da complexidade da Internet. Assim, a necessidade de compreender o que se passa nas redes e a forma como estas são utilizadas constitui um requisito crucial para manter e melhorar o seu desempenho de forma a dar suporte a serviços como VoIP, *e-commerce*, *e-banking*, P2P, *streaming* de vídeo e áudio, bem como identificar possíveis ameaças que comprometam o seu funcionamento.

1.1 Classificação e caracterização

Uma das áreas que procura ajudar a satisfazer este requisito é a da classificação e caracterização de tráfego. Através de técnicas eficazes para identificação da origem de fluxos de tráfego consoante as diferentes aplicações e protocolos existentes na rede, a área da classificação de tráfego dá um importante contributo a importantes tarefas como gestão de redes, alocação de recursos, planeamento e *design*, controlo da qualidade de serviço (QoS), segurança e detecção de intrusões.

Os métodos tipicamente utilizados para classificação de tráfego podem ser distinguidos consoante duas abordagens [1]: classificação baseada no conteúdo dos pacotes que constituem o tráfego e classificação baseada em características estatísticas do tráfego.

O método mais comummente aplicado segundo a abordagem baseada no conteúdo dos pacotes, consiste na análise da informação acerca das portas origem/destino ao nível da camada de transporte [2] . Outro método, mais avançado baseia-se na análise do *payload* dos pacotes através de técnicas de *Deep Packet Inspection* (DPI), de forma a identificar protocolos através da identificação de padrões característicos das aplicações ou protocolos.

A classificação, quando efectuada através da abordagem baseada na informação das portas, tem por base um processo bastante simples e com baixos recursos computacionais exigidos. Apesar disto, esta possui várias limitações que restringem bastante os cenários nos quais este tipo de técnica pode ser aplicada. No capítulo 2, estas limitações são expostas mais detalhadamente.

Já segundo a abordagem baseada em técnicas DPI a identificação dos protocolos e aplicações assenta na procura de padrões característicos no *payload* dos pacotes. Apesar da popularidade resultante da elevada eficácia desta abordagem e de ultrapassar as limitações da abordagem baseada na informação das portas, esta apresenta também algumas limitações relacionadas com questões de escalabilidade, perda de eficácia na presença de técnicas de compressão e cifragem de tráfego, e instabilidade dos padrões característicos dos protocolos. As limitações associadas às técnicas de classificação de tráfego baseadas no conteúdo dos pacotes que compõe o tráfego, motivaram o estudo de novas técnicas que se baseiam em propriedades estatísticas do tráfego. Técnicas modernas especificamente desenhadas para suportar este tipo de abordagens são designadas *Machine Learning* (ML). O seu funcionamento base é suportado pela ideia de que através de propriedades tais como a duração dos fluxos, ordem, tamanho e distribuição de pacotes, tempo entre chegada de pacotes num fluxo, etc., é possível determinar quais os protocolos que originam o tráfego [3].

Um factor já referido que tem influência significativa no desempenho das tarefas relacionadas com a monitorização do tráfego, é a falta de escalabilidade no processamento da informação com o constante aumento da capacidade dos *links* e dos volumes de dados. Dispositivos responsáveis pela monitorização, frequentemente *routers* e *switches*, são assim alvo de cada vez maiores exigências que levam ao aumento dos custos associados a este processo. Assim, para evitar que tarefas básicas da monitorização de tráfego, como o seu processamento, análise e armazenamento, criem *bottlenecks* na rede, uma solução típica passa por evitar a captura e consequente análise da totalidade do tráfego. Para alcançar esse objectivo são utilizadas técnicas de medição de tráfego baseadas em amostragem.

1.2 Amostragem

Através da implementação das técnicas de amostragem, pretende-se obter informação a partir da captura parcial do tráfego, de forma que a partir desta seja possível inferir dados representativos da totalidade do tráfego, ao mesmo tempo que se reduz o impacto no normal funcionamento da rede [4].

Uma das tecnologias mais utilizadas hoje em dia que se baseia neste principio é o *Sampled NetFlow* desenvolvido pela *Cisco*. À medida que os pacotes vão passando pelo ponto de observação, normalmente um encaminhador, estes vão sendo seleccionados ou não, dependendo do funcionamento do método de selecção previamente definido. O *NetFlow* oferece diferentes métodos de selecção: amostragem determinística que consiste na selecção de um pacote a cada N pacotes observados; uma variação desta, mas baseada no tempo, em que um pacote é seleccionado a cada N milissegundos; e a amostragem aleatória que consiste na selecção, segundo uma certa probabilidade, de um pacote num intervalo de N pacotes [5].

Técnicas alternativas têm vindo a ser desenvolvidas, que têm em conta o estado da rede para dinamicamente fazerem o ajuste do processo de amostragem, com o objectivo de diminuir a interferência deste processo no funcionamento da rede nas alturas mais críticas. Este tipo de técnicas são designadas técnicas de amostragem adaptativa. Tipicamente as implementações destas técnicas baseiam o seu carácter adaptativo no ajuste da frequência de amostragem. Isto ocorre através do constante processamento de antecipações do estado futuro da rede com base nas amostras anteriormente colectadas [6][7].

Em [8] é apresentado um estudo, que visa optimizar o processo de selecção das técnicas de medição adaptativas através da implementação de uma técnica Multi-adaptativa. Esta consiste em acrescentar ao carácter adaptativo o ajuste do tamanho

das amostras para além do intervalo de tempo entre as mesmas. Os resultados apresentados mostram que através desta abordagem é possível reduzir significativamente o *overhead* associado a este processo, comparativamente com outras técnicas clássicas de amostragem, mantendo a mesma acurácia na estimativa de parâmetros do tráfego.

1.3 Fluxos

A informação proveniente do processo de captura pode ser analisada e armazenada segundo diferentes granularidades: ao nível do pacote e ao nível do fluxo. A primeira é bastante útil quando se pretende examinar com grande detalhe algum evento ocorrido na rede, que seja, por exemplo, causa de algum problema ou falha de segurança. Isto porque os *traces* de pacotes contêm toda a informação referente a todos os pacotes capturados. Por outro lado, manter esta abordagem não é viável quando se trata de redes de alto débito, uma vez que o volume de dados a processar quer para análise quer para armazenamento pode ser muito grande, mesmo quando na presença de técnicas de amostragem.

A monitorização realizada com granularidade ao nível do fluxo, por sua vez, beneficia de um formato de dados que se coaduna mais com o cenário das redes de alto débito. Apesar do detalhe de informação não ser tão grande como na abordagem ao nível do pacote, os dados são dispostos num formato mais leve e flexível [9]. O funcionamento consiste na manutenção em memória cache do dispositivo medidor, de registos com informação sumarizada proveniente de conjuntos de pacotes com características em comum. A esses registos pode ser adicionada informação extra, tal como a duração temporal do fluxo, o número de pacotes, e o volume de dados [5].

A implementação de métodos de monitorização de tráfego baseados em fluxos é já bastante comum, devido ao volume de dados mais reduzido e à disponibilidade de muitas ferramentas de análise sobre estatísticas de fluxos. O próprio *NetFlow* mencionado anteriormente, é um sistema que se enquadra numa arquitectura de medição de tráfego sob a forma de fluxos. O facto de este ser um *standard* proprietário da *Cisco Systems*, motivou a criação de um *standard* aberto - *IP Flow Information eXport* - IPFIX [10] pelo IETF, baseado na versão 9 do *NetFlow*. O IPFIX descreve diferentes aspectos da monitorização de fluxos, desde modelos de informação até protocolos para transmissão da informação relativa aos fluxos de tráfego IP, de forma a providenciar interoperabilidade entre os dispositivos colectores e dispositivos exportadores de diferentes fabricantes.

1.4 Motivação e objectivos

A manutenção da escalabilidade e baixos custos associados a tarefas como controlo de QoS, alocação de recursos, segurança, entre outras, é um requisito cada vez mais importante na área de engenharia de redes de comunicações. Os principais desafios que se opõem a este requisito são o aumento do débito dos *links* e volumes de dados, bem como a constante transformação do tráfego da Internet, fruto das diversas aplicações que o originam.

A adopção de técnicas de medição baseadas em amostragem de tráfego é imprescindível, principalmente nas redes de alto débito, uma vez que constitui uma solução que proporciona uma redução eficiente do *overhead* associado às tarefas da monitorização, através do processamento parcial do tráfego.

Em comparação com técnicas de amostragem clássicas quanto à eficácia em termos da relação redução de *overhead* / acurácia de estimativas, uma técnica de amostragem Multi-adaptativa apresentou resultados prometedores [8], o que motiva o estudo da aplicação desta técnica noutras áreas da monitorização de tráfego, como a caracterização de fluxos de tráfego.

A caracterização de fluxos de tráfego, tendo como objectivos a análise e compreensão das suas propriedades, tais como composição e dinâmica, é uma área que envolve a utilização de técnicas de classificação. Por sua vez, a classificação eficaz do tráfego vem enfrentando crescentes desafios: a crescente quantidade de técnicas que visam camuflar ou encobrir o tráfego originado por determinados protocolos e aplicações, aliado ao já mencionado aumento da capacidade dos *links* de comunicação e heterogeneidade do tráfego. Desta forma, a conjugação de técnicas de classificação e caracterização de tráfego com técnicas de medição baseadas em amostragem, é apresentada não só como uma tendência natural, mas também como uma necessidade futura.

Neste sentido, pretende-se efectuar um estudo do impacto causado pelas técnicas de amostragem, para efeitos de classificação e consequente caracterização do tráfego, acrescendo a isto o facto de este ser ainda um assunto pouco explorado pela comunidade científica. Para tirar partido de vantagens como flexibilidade e pouco peso computacional associado, a organização do tráfego sob forma de registos de fluxos é uma tendência crescente, como provam a utilização em larga escala de técnologias como o *NetFlow*, e ainda o desenvolvimento do *standard* IPFIX.

Com base nestes motivos, o objectivo geral pretendido através da realização desta dissertação é avaliar e comparar a eficácia de diferentes técnicas de amostragem de tráfego, quanto à sua capacidade de manter a representatividade do tráfego origi-

nal em termos das suas características estatísticas. Isto efectuado após disposição do tráfego sob a forma de registos de fluxos e classificado segundo protocolos de transporte e protocolos aplicacionais.

Para alcançar este objectivo, são realizadas as seguintes tarefas:

- estudo de técnicas de amostragem clássicas e Multi-adaptativa e seus princípios de funcionamento;
- identificação e estudo de métodos para classificação de tráfego e seus princípios de funcionamento;
- desenvolvimento de um ambiente de testes, implementando técnicas de amostragem e consequente classificação do tráfego na forma de fluxos;
- identificação de métricas relativas a tráfego agregado sob forma de fluxos, úteis para a área da caracterização de tráfego;
- comparação do desempenho das técnicas de amostragem utilizadas na correcta caracterização dos fluxos de tráfego existentes, confrontando os resultados da amostragem com os resultados envolvendo o tráfego total original.

1.5 Organização da dissertação

Neste capítulo foi feita uma síntese dos temas relacionados com as tendências, desafios e actuais soluções para aspectos da área da monitorização de tráfego, como a medição baseada em amostragem, classificação e caracterização do tráfego e organização da informação. A partir destes, foram também definidos os objectivos e tarefas a realizar através da elaboração desta dissertação.

O segundo capítulo pretende dar uma visão mais aprofundada das abordagens actualmente utilizadas para fins de medição através de amostragem, classificação e caracterização do tráfego, bem como seus princípios de funcionamento e limitações, sendo referenciados alguns trabalhos relacionados nestas áreas.

No terceiro capítulo, é feita a descrição da forma como foi implementado o ambiente de testes, para permitir estudar o impacto da amostragem na obtenção de uma caracterização precisa do tráfego de rede, quando organizado sob a forma de fluxos.

No quarto capítulo é feita uma análise dos resultados obtidos a partir da implementação prática do ambiente de testes descrito no terceiro capítulo.

Por fim, no quinto capítulo conclui-se a dissertação através de uma reflexão baseada nos resultados obtidos, tendo em consideração os objectivos delineados inicialmente para a elaboração da dissertação.

Capítulo 2

Trabalho relacionado

A crescente velocidade dos *links* e do volume de tráfego na Internet exigem que as técnicas de medição sejam capazes de efectuar análises sobre dezenas de milhões de pacotes por segundo. Para suavizar este processo de forma a manter a sua escalabilidade, uma solução passa por conjugar as técnicas de classificação e caracterização com técnicas de medição baseadas em amostragem de tráfego de forma a permitir a extracção de características do tráfego, sem ter que o processar por inteiro.

Apesar do uso das técnicas de medição baseadas em amostragem oferecerem uma forma para redução da informação a ser processada, para a classificação de tráfego manter a sua eficácia, é importante também que o tráfego resultante da captura seja o mais representativo possível do tráfego real em termos das suas características.

Este capítulo apresenta uma descrição do funcionamento das abordagens básicas utilizadas para medição de tráfego, depois são apresentadas e descritas abordagens distintas utilizadas para classificação de tráfego, e técnicas para medição de tráfego baseada em amostragem. Finalmente são descritos alguns trabalhos que conjugam os conceitos de ambas as áreas.

2.1 Medição de tráfego

As técnicas de medição de tráfego em redes são tipicamente distinguidas através de duas abordagens: técnicas de medição activa e técnicas de medição passiva. As técnicas que conjugam ambas as abordagens são designadas técnicas híbridas.

2.1.1 Medição activa

A medição activa realiza-se através da implementação de técnicas que recorrem a tráfego sintético para efeitos de medição. O seu funcionamento consiste na injecção deste tráfego na rede, cujas características são previamente conhecidas, por forma a obter informação acerca de métricas como atraso, *jitter*, perda e largura de banda, ou informação estrutural sobre links e tabelas de encaminhamento.

Tipicamente as ferramentas utilizadas para este tipo de medição são de simples implementação e o seu funcionamento não implica grandes exigências computacionais. Alguns exemplos deste tipo de ferramentas incluem *ping*, *traceroute* e técnicas de tomografia de rede [11].

A medição activa é a abordagem mais adequada para medições fim-a-fim, por exemplo, a perda de pacotes pode ser obtida através de falha de pacotes com números sequenciais no sistema destino, e o atraso dos pacotes pode ser determinado comparando *time-stamps* de pacotes inseridos pelos terminais origem e destino. É também adequada para obtenção de informação sobre a topologia da rede e localizar nós ou links cuja falha ou congestionamento possa ter impacto no funcionamento da rede.

2.1.2 Medição Passiva

A medição passiva de tráfego, contrariamente à medição activa, não gera tráfego adicional na rede a observar. O seu funcionamento consiste apenas na utilização de técnicas que permitam analisar o tráfego real que passa em determinados pontos de observação. Esta abordagem tem associada a si o desafio provocado pelo constante aumento do volume de dado que circula nas redes.

2.1.3 Comparação Activa / Passiva

Apesar das técnicas de medição activa não serem tão afectadas pelo aumento do volume de dados, têm também a seu favor o facto de que as características do tráfego de prova possam ser conhecidas em detalhe e escolhidas para o propósito mais adequado.

No entanto, estas possuem algumas limitações comparando com as técnicas de medição passiva, nomeadamente, o eventual impacto na estabilidade da rede provocado pelo tráfego de prova injectado, fazendo com que o comportamento que se observa não seja exactamente o que a rede exibiria sem a injecção desse mesmo tráfego.

Também as conclusões que se obtêm com a implementação da metodologia activa, podem ser tendenciosas, uma vez que o tráfego de prova utilizado pode não formar uma boa representação do tráfego real, e ser tratado de forma diferente pelos nós da rede [12].

Devido a estas diferenças de princípios de funcionamento entre as metodologias passiva e activa, cada uma das abordagens poderá ser mais útil em domínios de aplicação distintos. Em geral as técnicas de medição activa são utilizadas para responder a questões sobre o estado actual da rede, enquanto que as técnicas de medição passiva se destinam às questões sobre o que se passa na rede e de que forma esta é utilizada.

Ambas as abordagens podem também complementar-se, e é usual que soluções de monitorização em redes utilizem a duas, de forma a combinar os resultados de cada para obtenção de uma visão mais detalhada do comportamento da actividade na rede [13].

No entanto, especificamente para tarefas como caracterização e classificação de tráfego, torna-se mais adequada a utilização de mecanismos baseados em técnicas de medição de tráfego passivas. A própria definição da classificação de tráfego presente em CAIDA [14], diz-nos que esta descreve métodos que classificam tráfego baseado em características observadas passivamente no tráfego de acordo com determinados objectivos.

2.2 Classificação de tráfego

A classificação do tráfego vem ganhando importância à medida que crescem técnicas que visam camuflar o tráfego de certas aplicações bem como a utilização de portas padrão bem conhecidas, de forma a evitar a sua detecção por parte de regras de *firewalls* e outros sistemas de detecção de intrusões.

Como mencionado anteriormente, na área da classificação de tráfego podem ser distinguidas duas abordagens: classificação baseada no conteúdo dos pacotes e classificação baseada em características estatísticas [1].

2.2.1 Baseada no conteúdo

Os mecanismos de classificação de tráfego que se baseiam no conteúdo dos pacotes têm como funcionamento base a análise de determinados parâmetros considerados necessários para identificar os pacotes como sendo de uma determinada aplicação ou protocolo, e posterior mapeamento com uma base de dados de referência cujos

parâmetros estão previamente identificados e associados a aplicações ou protocolos.

Estes mecanismos podem ser ainda divididos em dois tipos consoante a complexidade necessária para extracção dos parâmetros do tráfego, nomeadamente, em abordagem genérica e abordagem avançada.

Portas

Segundo a abordagem genérica, a informação dos cabeçalhos dos pacotes IP é analisada quanto aos seus valores das portas do cabeçalho de transporte, e é feito um mapeamento directo destes valores com as aplicações que as utilizam. Esta informação consta de registos existentes na *Internet Assigned Numbers Authority* (IANA) [15], que associam determinadas aplicações ou protocolos a determinados valores de portas dos cabeçalhos *Transmission Control Protocol / User Datagram Protocol* (TCP/UDP). Usualmente estas são designadas portas bem conhecidas, por exemplo, a porta com valor 80, caso conste no cabeçalho dos pacotes de determinado fluxo analisado, este tráfego é identificado como pertencendo ao protocolo *Hypertext Transfer Protocol* (HTTP).

Este é um dos métodos mais populares devido à sua simplicidade de implementação e baixos recursos computacionais exigidos [16]. Prova disso é a variedade de ferramentas existentes que a utilizam, como *coralRef* [2], *tstat* [17] e *snort* [18].

No entanto, a utilização exclusiva desta informação, não é recomendável para cenários em que seja crucial que os níveis de acurácia na identificação sejam altos. Isto deve-se à crescente heterogeneidade do tráfego e também à crescente implementação de técnicas que visam esconder a real origem do tráfego. Um estudo apresentado em [19] revela que apenas 30 a 70 por cento do tráfego pode ser identificado através deste tipo de abordagem.

Ainda assim devido às vantagens relacionadas com simplicidade e baixo custos associados, a classificação de tráfego baseada neste método é utilizada para efeitos de validação em várias pesquisas [3][20].

Deep Packet Inspection

As abordagens de classificação denominadas avançadas que se baseiam no conteúdo dos pacotes, tipicamente utilizam técnicas de *Deep Packet Inspection* - DPI. As técnicas de DPI destinam-se a efectuar análises a partir da totalidade dos pacotes capturados, ou seja, não apenas o cabeçalho mas também o *payload* é tido em consideração.

O funcionamento das técnicas de classificação baseadas em DPI consiste na extracção de assinaturas do conteúdo dos pacotes. As assinaturas são padrões formados por características dos pacotes que tencionam identificar univocamente cada aplicação ou protocolo. Normalmente são identificados *bytes*, caracteres ou *strings* ou até padrões numéricos como o tamanho do *payload* ou número de pacotes do tipo '*response*' num determinado fluxo.

Após a extracção das assinaturas, o sistema de classificação efectua a comparação destas com as assinaturas previamente associadas a aplicações e protocolos que constam da base de dados de referência, de forma a efectuar o mapeamento e identificação do tráfego.

A classificação efectuada com base em padrões característicos no *payload* dos pacotes apresenta níveis de acurácia elevados, o que aumentou a sua popularidade e motivou a criação de diversas ferramentas que as utilizam, como as *L7-filter* [21], *Wireshark* [22] e *Snort* [18].

No entanto esta abordagem possui várias limitações: implica que o tráfego seja não cifrado, está dependente da estabilidade das assinaturas, isto é, se a assinatura de uma determinada aplicação se alterar, a técnica classificadora terá que se adaptar constantemente às mudanças para continuar a ser capaz de identificar o tráfego com a mesma precisão. Existem também razões legais que inviabilizam a aplicação desta abordagem, por exemplo, em determinados países devido a questões relacionadas com a privacidade dos utilizadores, não é permitido analisar o conteúdo dos pacotes [23].

Apesar das limitações descritas, a principal limitação que advém da implementação destas técnicas são os elevados recursos computacionais exigidos, o que dificulta a utilização deste tipo de técnicas nas redes modernas de alto débito. Devido a isso, grande parte dos estudos sobre este tipo de classificação concentra-se em questões relacionadas com o desempenho, desde o estudo de novas abordagens para reduzir o volume de dados a analisar [24] ou melhorar a eficiência dos algoritmos usados para identificação de assinaturas [24][25], até à optimização do desempenho do sistema de classificação e sua arquitectura de *hardware* de forma a aumentar a eficiência destes processos e minimizar o seu custo [26][27][28].

Para contornar estas limitações, em [29] é proposta uma técnica para classificação de tráfego através da identificação de assinaturas que se baseiam em heurísticas ou padrões de conexão de *hosts*. Esta técnica tem em conta informação fornecida apenas pelos colectores de fluxos (dispositivos de medição de tráfego), sem efectuar qualquer análise ao *payload*, e aos valores das portas na camada de transporte. Este trabalho é complementado em [30], para classificação de um maior número de aplicações,

através da conjugação de informação dos *hosts* capturada a diferentes níveis: nível social, nível de rede e nível aplicacional. Segundo este estudo é possível obter uma acurácia para classificação na ordem dos 95% para cerca de 80% a 90% do tráfego analisado.

2.2.2 Baseada em características estatísticas

Este é o tipo de abordagem mais recente e surge como alternativa às técnicas que baseiam o seu funcionamento sobre técnicas de DPI, possuindo um nível de acurácia semelhante a estas. O funcionamento consiste na identificação de classes de tráfego com base em critérios extraídos de propriedades estatísticas dos fluxos, tais como a sua duração, tamanhos dos pacotes e tempo entre chegadas de pacotes [3].

Para a classificação baseada nesses parâmetros são normalmente utilizados algoritmos *Machine Learning* - ML [31], de forma a tornar possível a análise de grandes volumes de dados. Estes algoritmos são utilizados no processo de mapeamento dos fluxos, consoante as suas características em diferentes classes de tráfego.

Estes algoritmos necessitam apenas de informação ao nível do fluxo para efectuarem a classificação, e permitem contornar limitações das técnicas baseadas em DPI. O tráfego da aplicação Skype, por exemplo, devido a ser uma aplicação baseada num design proprietário e usar mecanismos de cifragem, torna a sua identificação e classificação através de características estatísticas e uso de técnicas ML a única solução possível [32].

Existem dois tipos de abordagens comuns na implementação de algoritmos ML: *supervised* e *unsupervised*. Os algoritmos ML *supervised* implicam uma fase de treino ou aprendizagem anterior à fase de classificação. Na primeira fase é criado o modelo de classificação, ou seja, um conjunto de regras, baseado em instâncias de tráfego previamente conhecidas e classificadas. A partir desse modelo, o classificador ML na fase de classificação é responsável por fazer a associação ao identificar a que classe pertencem as novas instâncias de tráfego através da análise às suas características. O sucesso da classificação está altamente dependente da acurácia das instâncias de tráfego utilizadas para a fase de treino. Em [33] é apresentado um estudo utilizando vários algoritmos ML *supervised* para comparar a sua eficácia na classificação de tráfego em tempo real.

Já os algoritmos ML *unsupervised* consistem no agrupamento em *clusters* de instâncias de tráfego com características semelhantes, sem ter por base qualquer modelo ou referência inicial. Estes são utilizados mais em casos em que o tráfego a analisar possui características desconhecidas ou pouco comuns. Em [34] é sugerida

uma técnica de classificação ML *unsupervised* cujos resultados da classificação superam os obtidos através da técnica ML *supervised* apresentada em [35] como tendo alta precisão na classificação de tráfego.

2.3 Técnicas de amostragem

Para a colecta de informação a partir do tráfego observado passivamente na rede, a implementação de técnicas de amostragem vem ganhando cada vez mais importância no domínio das redes de alto débito. Estas técnicas destinam-se a permitir a redução do *overhead* e dos custos associados ao processo de medição provocados pelo constante aumento do volume de dados que circulam nas redes. Dispositivos de processamento e armazenamento de dados, largura de banda e ferramentas de análise, compõem o domínio de recursos cuja optimização forma o motivo principal para a adopção destes mecanismos.

2.3.1 Técnicas convencionais

Tipicamente, as técnicas de amostragem podem ser caracterizadas segundo o método de selecção que utilizam. Técnicas convencionais têm por base princípios de funcionamento com baixa complexidade, utilizando regras fixas para determinar quando activar e parar a captura de amostras. Essas regras utilizadas são tipicamente sistemáticas ou aleatórias [36].

O processo de amostragem através das técnicas sistemáticas consiste na selecção de pacotes de acordo com uma função determinística. Esta função define o início e duração da selecção. Por sua vez, o funcionamento das técnicas de amostragem aleatórias consiste na utilização de uma função aleatória para gerar os intervalos de amostragem.

Ainda segundo o mecanismo que desencadeia a selecção das amostras, as técnicas de amostragem podem funcionar segundo diferentes abordagens:

Sistemática *count-based*

Segundo esta técnica, o início e fim dos intervalos de amostragem são definidos de acordo com o posicionamento de chegada dos pacotes, ou seja, cada pacote é seleccionado caso os seus números de ordem de chegada relativamente à última amostra seleccionada, corresponda ao valor definido no mecanismo de selecção.

Sistemática *time-based*

Neste caso, o funcionamento é similar à técnica *count-based* com a diferença que os intervalos de amostragem são definidos pelo tempo de chegada dos pacotes. Os pacotes são então seleccionados caso cheguem ao ponto de observação entre os intervalos de tempo definidos para início e fim de selecção.

Aleatória *n-Out-of-N* ou *stratified*

Na aleatória *n-out-of-N*, n pacotes são então seleccionados de um intervalo de N pacotes. Esta técnica combina o intervalo fixo entre amostras (ora temporal ora baseado na posição dos pacotes) da técnica de amostragem sistemática, com a amostragem aleatória ao seleccionar n pacotes para amostra, dentro de cada intervalo.

Aleatória probabilística ou simples

No método amostragem probabilística, a decisão de seleccionar um pacote para a amostra é feita de acordo com uma probabilidade predefinida.

Cada pacote possui a mesma probabilidade P de ser seleccionado. Este mecanismo constitui a amostragem probabilística do tipo uniforme, no entanto, segundo outra abordagem a probabilidade de selecção pode não ser a mesma para cada pacote, sendo esta designada por amostragem probabilística não uniforme. Estas abordagens são apresentadas em mais detalhe em [36].

Ainda segundo [36] a introdução de um função aleatória para o processo de amostragem oferece melhoramentos em relação às técnicas sistemáticas na medida em que o nível de acurácia que se pode obter da representatividade do tráfego total é maior. Isto é válido principalmente para casos em que o tráfego apresenta um comportamento uniforme.

O valor da frequência de amostragem para estas técnicas é normalmente definido com base no volume médio da carga expectável, na distribuição do tráfego ou apenas segundo um valor a partir do qual resulte um nível de *overhead* admissível.

2.3.2 Amostragem adaptativa

Para além das técnicas convencionais, as técnicas de amostragem adaptativa vêm ganhando interesse devido aos resultados obtidos na diminuição da interferência que o processo de amostragem causa no funcionamento da rede. Através da sua capacidade de ajustar a frequência de amostragem dependendo do estado em que a rede se encontra, esta abordagem permite uma maior redução do *overhead* quando

comparado com as técnicas convencionais. Tipicamente o seu carácter adaptativo funciona reagindo a constantes estimativas efectuadas de parâmetros específicos, tal como atraso, *jitter* [37] ou perda de pacotes [38].

Em [6] são comparados os desempenhos de técnicas de amostragem adaptativas com as técnicas convencionais. Segundo este estudo as técnicas adaptativas, através da sua flexibilidade em adaptarem-se a alterações no comportamento da rede, permitem em alguns casos obter uma redução do número de amostras colectadas para metade, mantendo os níveis de acurácia de estimativas semelhantes.

Para implementação de técnicas de medição adaptativas, métodos tipicamente utilizados baseiam-se em lógica *fuzzy* [39] e predição linear [40][41].

Segundo o método lógica *fuzzy*, o funcionamento das técnicas consiste na aplicação de frequências de amostragem de acordo com comportamentos exibidos pela rede, tendo por base situações anteriores semelhantes.

Já segundo o método Predição Linear, é utilizado um parâmetro de referência calculado a partir das amostras de tráfego. São efectuadas previsões acerca do seu valor futuro com base nas amostras colectadas. Caso as previsões se confirmem dentro de um mínimo exigido, a frequência de amostragem é reduzida. Caso contrário, se as previsões não apresentam um determinado nível de acurácia, significa que o comportamento do tráfego se alterou significativamente, o que faz com que a frequência de amostragem aumente, ou seja, diminui o intervalo entre amostras. Este parâmetro de referência pode ser a representação de valores como o débito médio, a taxa média de perda de pacotes ou outro parâmetro mensurável através dos pacotes da amostra [8].

Multi-adaptativa

Através do método predição linear, apenas os intervalos entre amostras variam, mantendo-se os tamanhos das amostras constantes durante todo o processo de medição. Esta é uma particularidade explorada em [42], em que através do desenvolvimento de uma técnica baseada no principio da predição linear, se propõe que a selecção de amostras se efectue através de uma abordagem Multi-adaptativa. Isto é alcançado tornando adaptável não apenas o intervalo entre as amostras mas também o tamanho das mesmas. Para isso durante o processo de selecção é considerado o nível de actividade na rede, ou seja, caso o nível de actividade na rede aumente, a frequência de amostragem tende a aumentar também, mas para evitar sobrecarregar o dispositivo de medição nesta fase crítica, reduz-se o tamanho das amostras a colectar. Caso contrário, se o nível de actividade na rede diminuir, a frequência de amostragem também diminui, mas aumenta o tamanho das amostras.

Em comparação com as técnicas sistemáticas *time-based* e adaptativa baseada em predição linear, o controlo destes dois parâmetros por parte da técnica Multi-adaptativa demonstra reduzir significativamente o *overhead* associado ao processo de medição, sem no entanto comprometer a acurácia das estimativas [42].

2.4 Classificação de tráfego com amostragem

Apesar da tendência para as técnicas de classificação implementarem mecanismos para minimizar exigências computacionais associadas a este processo, minimizar custos associados aos crescentes volumes de tráfego é algo crucial para manter a sua operacionalidade nas redes de alto débito.

A conjugação destas técnicas com técnicas de amostragem constituem uma possível solução para este desafio. No entanto, à redução de *overhead* proporcionada pela amostragem, está associada a perda de parte significativa da informação obtida, o que pode causar impacto na precisão da obtenção de características do tráfego e consequente classificação. Esse impacto é analisado em [43] através da aplicação de um algoritmo de classificação *Machine Learning supervised* (C4.5), sobre conjuntos de dados com diferentes características e resultantes da aplicação de diferentes frequências de amostragem. No mesmo trabalho conclui-se, entre outras descobertas, que no caso em que os dados utilizados para a fase de treino da técnica de ML sejam obtidos a partir da mesma frequência de amostragem que é utilizada para a validação, os resultados da classificação obtida são são bastante bons mesmo utilizando taxas de amostragem baixas.

Em [44] é efectuado um estudo acerca da precisão da classificação de tráfego utilizando o mesmo algoritmo ML (C4.5), mas com uma inovação na fase de treino. Para esta fase é desenvolvida uma nova técnica baseada na ferramenta L7-Filter [21] com o objectivo de a tornar automática. Para construção da base de dados de referência, são utilizadas apenas características presentes em fluxos resultantes da amostragem através do sistema *Netflow*. É demonstrado que da alteração efectuada ao processo de treino, resulta um aumento significativo da acurácia da classificação sob utilização de amostragem.

Em [32] é efectuado um estudo acerca do efeito causado pela amostragem na acurácia da classificação de tráfego específico da aplicação Skype através da ferramenta *opensource Skypeness* [45]. Aqui é demonstrado que após utilização de técnicas de amostragem convencionais, mesmo com taxas de amostragem altas, as características base utilizadas na identificação deste tipo de tráfego, tal como o tempo entre a chegada de pacotes, são distorcidas, resultando assim na diminuição da eficácia da

ferramenta em estudo. No entanto, através de uma modificação à ferramenta de detecção, multiplicando os tempos de chegada entre pacotes pela taxa de amostragem utilizada, os resultados tornam-se semelhantes aos casos em que não há amostragem.

2.5 Resumo

Neste capítulo foram inicialmente apresentadas as abordagens básicas para medição de tráfego - Activa e Passiva. Posteriormente foram apresentadas abordagens distintas de classificação de tráfego, bem como alguns trabalhos relacionados nesta área. O mesmo foi feito em relação às técnicas de amostragem, expondo técnicas de amostragem convencionais simples e técnicas baseadas em amostragem adaptativa.

Finalmente foram descritos de forma breve alguns trabalhos que conjugam classificação e amostragem de tráfego de forma a mitigar os desafios relacionados com os crescentes volumes de tráfego.

O próximo capítulo descreve a forma como foi implementado um ambiente de testes para estudo do impacto resultante dos processos de amostragem na caracterização dos fluxos de tráfego. É também efectuada a identificação de métricas para análise dos conjuntos de tráfego resultantes da amostragem e discutida a forma como estes serão analisados e comparados.

Capítulo 3

Ambiente de testes

Para possibilitar a extracção de características do tráfego e posteriormente efectuar análises sobre o mesmo, foi criado um ambiente de testes cujo funcionamento pretende representar os dispositivos de captura e medição de tráfego. Para tal foi desenvolvida uma aplicação com a função de processar instâncias de tráfego com granularidade ao nível dos pacotes IP e, dependendo do método de selecção de amostras, capturar e armazenar informação referente a essas mesmas instâncias.

Para isso dota-se a aplicação da possibilidade de operar com diferentes técnicas de amostragem, e também da capacidade de armazenamento do tráfego sob forma de registos de fluxos, para posteriormente se efectuarem as análises a partir destes.

Este capítulo apresenta a descrição da metodologia utilizada para validação deste estudo, desde a descrição das colectas de tráfego até às características utilizadas para análise e comparação de resultados. São também descritas as técnicas de medição baseadas em amostragem utilizadas, e a forma como a agregação da informação ao nível do fluxo é efectuada.

3.1 Conceitos introdutórios

As técnicas de medição baseadas em amostragem têm usualmente associadas a si um conjunto de termos para descreverem parâmetros base cujas definições podem variar de acordo com as fontes. Assim, de acordo com os conceitos utilizados neste trabalho, são definidos os termos amostras, tamanho de amostra, intervalo entre amostras e parâmetro de referência. Estes conceitos são ilustrados na Figura 3.1

Amostras

Conjuntos de pacotes seleccionados a partir do tráfego total que passa em determinado ponto de observação na rede. Dados referentes às amostras são armazenados

para posteriormente serem analisados.

Tamanho das amostras

Dependendo do tipo de técnica de amostragem utilizada, o tamanho das amostras pode ser baseado no número de pacotes, sendo cada amostra composta por 1 ou mais pacotes, ou baseada no tempo, sendo cada amostra definida pelo intervalo de tempo durante o qual os pacotes observados são seleccionados.

Intervalo entre amostras

Da mesma forma que o tamanho das amostras, o intervalo entre amostras pode ser definido pelo número de pacotes que não são seleccionados para amostra, ou definido pelo período de tempo durante o qual nenhuma selecção de pacotes ocorre.

Parâmetro de referência

Para o caso das técnicas de medição baseadas em amostragem adaptativa, este é o parâmetro calculado a partir das amostras de tráfego e que é tido em conta para alterar o regime de amostragem adoptado. O valor utilizado em [8] é referente ao débito médio da ligação.

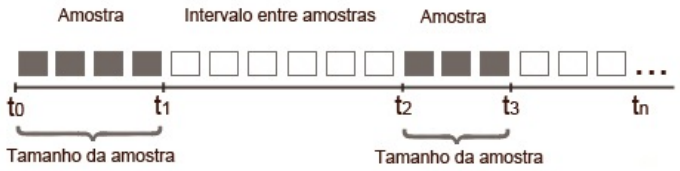

Figura 3.1: *Conceitos de amostragem*

3.2 Metodologia de testes

De uma forma resumida, o funcionamento do ambiente de testes desenvolvido consiste nas seguintes tarefas:

- Inicialmente é seleccionado um *trace* de tráfego com informação ao nível do pacote IP em formato PCAP (packet capture), sobre o qual o algoritmo de selecção de tráfego e classificação irá incidir.

- Posteriormente é escolhida a técnica de amostragem a utilizar e dependendo desta são definidos os seus parâmetros iniciais. De seguida dá-se início ao processo de medição e aquando da selecção dos pacotes, a aplicação efectua

consultas à tabela de registos de fluxos para verificar se existe já algum registo activo sobre os dados do fluxo a que corresponde cada pacote. Caso exista, as variáveis de contagem são actualizadas, caso não exista, é criado um novo registo e inicializadas as variáveis desse registo.

- No final é retornada a tabela preenchida com os registos de fluxos de forma a possibilitar a extracção de informação para que sobre esta se possa efectuar a análise das características do tráfego de rede organizado em fluxos.

De seguida, os principais aspectos das tarefas mencionadas serão descritos mais detalhadamente.

3.3 Técnicas de amostragem

Para estudo do impacto da amostragem na caracterização e classificação de tráfego, foram implementadas diferentes técnicas de amostragem. Estas são de seguida apresentadas bem como seus princípios de funcionamento.

3.3.1 Técnica de amostragem Multi-adaptativa

Esta é uma técnica proposta em [8], baseada nas técnicas apresentadas em [6] e [37] fazendo uso dos seus princípios de funcionamento mas adicionando à capacidade adaptativa os tamanhos das amostras para além dos intervalos entre as mesmas.

A capacidade adaptativa da técnica é baseada em predição linear, devido à simplicidade de implementação e baixas exigências ao nível de consumo de recursos.

O seu funcionamento baseia-se na medição do nível de actividade da rede. Quando é detectado um aumento de actividade, é diminuído o valor do intervalo entre amostras de forma a aumentar a frequência de amostragem. Este procedimento é necessário para que o novo padrão de comportamento seja identificado. O funcionamento é análogo para o caso inverso, ou seja, quando é detectada uma diminuição do nível de actividade na rede, o intervalo entre amostras também diminui.

Paralelamente a estas adaptações, devido ao maior consumo de recursos a nível de processamento e armazenamento causados pelo aumento da frequência de amostragem, o tamanho das amostras é também alterado. Desta forma, detecções de aumentos na actividade da rede produzem também diminuições no tamanho das amostras, e vice versa.

De forma mais detalhada, a técnica Multi-adaptativa funciona da seguinte forma:

inicialmente são definidos pelo utilizador os parâmetros tamanho inicial da amostra, o intervalo inicial entre amostras e a ordem, ou seja, o número de amostras a considerar no cálculo da previsão do parâmetro de referência. De seguida dá-se início ao processo de medição e são captadas as primeiras N amostras. Com base nestas primeiras amostras é então calculado o valor do parâmetro de referência, e seguidamente é calculado o valor previsto para o mesmo.

Após captura de nova amostra, é calculada a partir desta o novo parâmetro de referência. Este novo parâmetro de referência é então comparado ao valor previsto previamente calculado, e baseando-se na taxa de diferença entre os dois é feito o ajuste dos valores dos parâmetros intervalo entre amostras e tamanho das amostras. Este processo repete-se então indefinidamente até que o processo de amostragem termine.

Para o ajuste dos valores de intervalo entre amostras e tamanho das amostras, é levado em conta a taxa de diferença entre o parâmetro de referência calculado a partir da amostra actual e o parâmetro de referência previsto. Este valor, denominado m, é portanto o factor indicador do nível de actividade na rede.

A tomada de decisão é baseada na comparação do valor m com limites m(máximo) e m(mínimo) definidos previamente. Estas tomadas de decisão baseadas no valor e m são apresentadas na Tabela 3.1.

Tabela 3.1: *Parâmetro m. Rapidamente significa uma variação de 25%, lentamente significa uma variação de 10%*

Factor 'm'	Significado	Intervalo entre amostras	Tamanho das amostras
m < m(min)	maior actividade que a prevista	diminui (rapidamente)	diminui (rapidamente)
m > m(max)	menor actividade que a prevista	aumenta (actual+1s)	aumenta (lentamente)
m(min) <= m <= m(max)	previsão correcta	mantém-se	mantém-se
m indefinido	rede estável	aumenta (2*actual)	aumenta (lentamente)

Em [6], os valores de m(máximo) e m(mínimo) são definidos de forma experimental, respectivamente como 1.1 e 0.9 devido ao bom desempenho apresentado para diferentes tipos de tráfego, e por isso são estes os valores utilizados na técnica Multi-adaptativa.

A ordem N utilizada, ou seja, o número de amostras a ter em conta no cálculo preditivo, é 3, por ser o valor cujo desempenho é mais preciso para diferentes tipos de tráfego, tendo em conta o *overhead*, e a acurácia de estimativas, de acordo com o trabalho em [8].

3.3.2 Técnicas de amostragem clássicas

Para a análise das características do tráfego e comparação dos resultados com os obtidos através da técnica Multi-adaptativa, foram escolhidas técnicas de amostragem de um conjunto de técnicas anteriormente denominadas convencionais. Designadamente as técnicas sistemática *count-based* e aleatória *n-Out-of-N* foram seleccionadas.

As razões que motivaram esta escolha estão relacionadas com a sua simplicidade de funcionamento e implementação e também com o facto de fazerem parte de ferramentas largamente adoptadas por operadores de redes e ISPs (*Internet Service Provider*), como, por exemplo, a ferramenta de medição *NetFlow* da *Cisco*. O *Sampled NetFlow* é uma variante desta ferramenta adaptada para as redes de alto débito, uma vez que implementa os mecanismos de amostragem mencionados de forma a permitir operar com grandes volumes de tráfego.

Apesar do crescente estudo e desenvolvimento de técnicas de medição avançadas baseadas em amostragem, como as desenvolvidas em [46][47], grande parte destas obtêm bons resultados apenas sob condições muito específicas ou direccionados para domínios específicos de aplicação. Tipicamente para a implementação de técnicas de amostragem em dispositivos que são comercializados, técnicas simples e robustas bem especificadas e documentadas em *standards* são preferidas a esquemas sofisticados mas com domínios de aplicação mais restritos [48].

Sistemática e Aleatória

Ambas as técnicas implementadas são *count-based*, com tamanho de amostra fixo de 1 pacote, mantendo-se estes parâmetros constantes durante todo o processo de medição.

No caso da técnica Sistemática, o intervalo entre amostras depende apenas da taxa de amostragem definida (taxa de amostragem - 1) e é fixo durante todo o processo. Já para o caso da técnica Aleatória o intervalo entre amostras é variável, ou seja, entre cada amostra o intervalo entre amostras é definido entre 0 e 2*[*taxa de amostragem*]-2. Estes conceitos e funcionamento de ambas as técnicas são ilustrados na Figura 3.2.

Na amostragem Sistemática, os pacotes são seleccionados para amostra de forma determinística, com frequência 1 de *N*. No exemplo da Figura 3.2, pode ver-se que em cada intervalo de 5 pacotes, o 1º pacote é sempre o seleccionado.

Na amostragem Aleatória, os pacotes são agrupados em intervalos de N pacotes, a partir dos quais apenas um é escolhido para amostra de forma aleatória. No

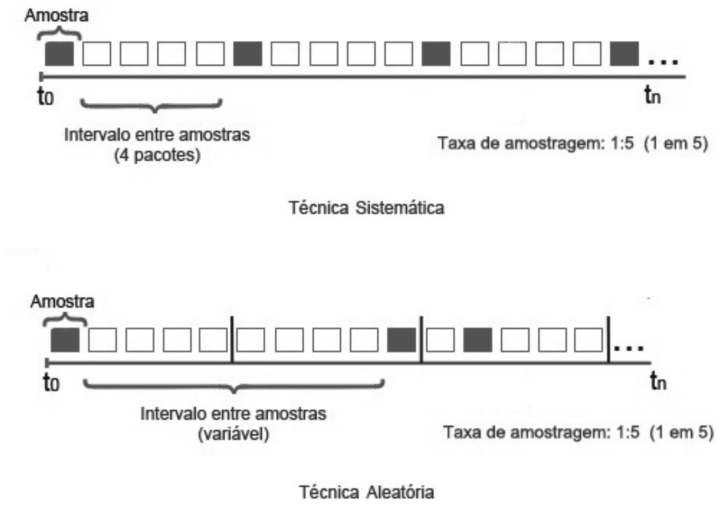

Figura 3.2: *Funcionamento das técnicas Sistemática e Aleatória*

caso da Figura 3.2, para cada conjunto de 5 pacotes, 1 pacote é seleccionado, mas contrariamente à amostragem sistemática em vez da selecção recair sempre sobre um pacote na mesma posição, o algoritmo gera um número aleatório entre 1 e 5, e o pacote que corresponde à posição do número gerado é seleccionado.

Para o caso em que a aplicação é executada com a técnica de amostragem aleatória, os resultados referentes ao número de fluxos final variam cada vez que o algoritmo é executado. Assim, para este caso específico, o mecanismo foi executado 100 vezes e o conjunto de dados utilizado para análise foi o correspondente àquele cujo número de fluxos se aproxima mais do número médio de fluxos obtido tendo em conta todas as execuções.

Taxas de amostragem

Segundo [49], apesar de poderem ser definidas entre 1 e 65535, as frequências de amostragem mais comummente utilizadas são 1:100 e 1:1000. Ou seja, em cada 100 ou 1000 pacotes que passam no ponto de medição, 1 é seleccionado para amostra. Nos dispositivos Cisco da série 7500, a percentagem média de diminuição da carga no CPU é de 75% para o caso da taxa 1:100, e 82% para 1:1000. Quando os pacotes são amostrado a uma taxa de 1:100, aproximadamente 80% dos fluxos irão constar da *NetFlow cache*, ao mesmo tempo que se obtém uma redução de dados na ordem dos 99% em comparação com a abordagem sem amostragem. Esta redução de fluxos permite menos entradas na cache, diminuindo a utilização de CPU e ao mesmo tempo reduzindo os volumes de dados a exportar e consequente utilização de largura de banda.

A informação mencionada sobre a utilização e desempenho que resulta da aplicação das taxas de amostragem 1:100 e 1:1000, e também a sua utilização por parte de

vários estudos, como em [43][44], formam as razões tomadas em conta para a escolha destes valores para aplicação de cada uma das técnicas, sistemática e aleatória nos testes efectuados no Capítulo 4.

3.4 Classificação de tráfego em fluxos

A classificação ou organização do tráfego em fluxos, é o passo que se segue à selecção de amostras. Os fluxos consistem em conjuntos de pacotes com propriedades em comum que são observados e seleccionados em um determinado período de tempo. Estes são organizados da seguinte forma: quando um pacote é seleccionado, o dispositivo que efectua a medição verifica se esse pacote já possui um fluxo activo, ou seja, se já há informação a ser armazenada para pacotes com as mesmas características de fluxo. Essa informação tipicamente inclui contadores de pacotes e *bytes* que vão sendo actualizadas consoante são seleccionados mais pacotes.

Tipicamente os fluxos são definidos por um conjunto de valores, denominado *flow-key*, que determinam a forma como o fluxo é identificado. Valores comummente utilizados, como nos trabalho [30][50], são 5 campos presentes nos pacotes IP: endereços IP origem e destino, portas TCP/UDP origem e destino e protocolo de transporte. Esta é também a *flow-key* escolhida para identificação de fluxos na aplicação desenvolvida. A Figura 3.3 ilustra de forma sucinta o processo de selecção e classificação em fluxos.

Figura 3.3: *Processo de selecção de pacotes e classificação em fluxos*

Posteriormente à execução da aplicação e obtido o conjunto de dados organizados ao nível do fluxo, seguem-se as fases de identificação do tráfego e análise das suas características. O objectivo é tratar a informação obtida de forma a que se possa avaliar comparativamente com o conjunto de tráfego original, e analisar o impacto sofrido pela actuação do processo de amostragem, sobre características de interesse para identificação do tráfego.

A análise e comparação das características é efectuada ao nível dos protocolos de transporte TCP ou UDP que originaram os fluxos, e também com mior granularidade, consoante o protocolo aplicacional identificado. Este tipo de análise é

tipicamente útil para compreensão das tendências aplicacionais e seu impacto na rede ao longo do tempo [51] [52]. Por isso é necessário associar cada fluxo ao protocolo de transporte e protocolo aplicacional que o gerou.

Estas associações foram efectuadas tendo por base os valores do campo 'protocolo' do cabeçalho IPv4, como identificado na Figura 3.4. Outros valores para além dos correspondentes aos protocolos TCP (valor 6) e UDP (valor 17), como ICMP (valor 1), IGMP (valor 2) SCTP (valor 132), etc., também se encontram em alguns fluxos, mas devido a representarem um volume pouco significativo comparativamente aos primeiros, todos estes foram considerados num grupo designado 'Outros'.

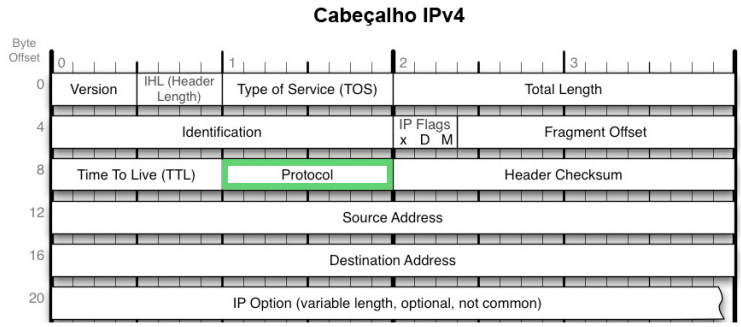

Figura 3.4: *Cabeçalho IP e campo considerado para classificação em fluxos*

Relativamente à identificação dos fluxos consoante os protocolos aplicacionais, esta foi efectuada com base na informação dos valores das portas dos cabeçalhos TCP ou UDP que originaram os fluxos, como indicado na Figura 3.5.

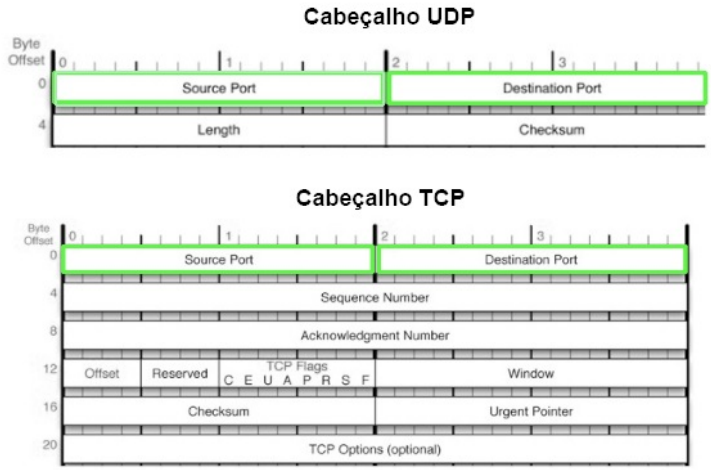

Figura 3.5: *Cabeçalhos UDP e TCP, e campos considerados para classificação em fluxos*

Como mencionado anteriormente no capítulo 2, através de uma classificação com base apenas na informação dos valores das portas, o conhecimento que se pode obter acerca do serviço (aplicação/protocolo) em utilização, nunca é completamente fiável, uma vez que não existe qualquer restrição na associação de valores de portas a determinados serviços comuns a todos os sistemas operativos.

No entanto, este tipo de classificação de fluxos foi escolhido por ser o mais simples e menos exigente, e também por se considerar que a acurácia da classificação dos fluxos não é essencial no âmbito deste estudo. Considera-se sim relevante a manutenção da coerência no método utilizado para a classificação, uma vez que o objecto principal de estudo é o impacto na caracterização dos fluxos resultantes da aplicação de diferentes padrões de amostragem.

O valor correspondente ao número das portas pode estar definido no intervalo entre 1 e 65535. Em aplicações cliente-servidor tipicamente o serviço é fornecido através de uma porta oficial designada "bem conhecida"ou qualquer outra porta "não oficial". As aplicações foram então identificadas pelos seu valores das portas bem conhecidas de acordo com [15] e as restantes aplicações e protocolos que utilizam outra gama de portas que não se encontram nessa lista foram identificadas com base em [53], onde está dísponivel para consulta uma base da dados de portas oficiais e não oficiais, obtidas através da combinação de portas fornecidas pela IANA e portas associadas a aplicações após investigação e/ou submissão por parte de utilizadores.

Assim, para a análise efectuada neste trabalho, fluxos utilizando as portas 80, 8080, 3128 e 6588 são classificados como pertencendo a tráfego HTTP, portas 20 e 21 - FTP, porta 52 - DNS, etc. Já os fluxos que sejam identificados como originados por determinados jogos como *Quake*, *Call of duty*, *Battlefield*, etc., foram todos identificados dentro de um conjunto denominado 'jogos'. Da mesma forma, aplicações P2P como *eDonkey*, *BitTorrent*, *Napster*, etc., ficaram agrupadas sob o conjunto denominado P2P.

3.5 *Traces* utilizados

Para estudo e validação deste trabalho foram utilizados dois *traces* com diferentes tipos de tráfego, colectado em dois tipos distintos de ambientes. Um foi gerado durante a conferência *Sigcomm* em 2008 pelos seus participantes, é composto apenas por tráfego *wireless*, colectado na parte Ethernet da rede [54]. O outro *trace* utilizado, foi obtido na CAIDA [55], é um *trace* colectado numa conexão OC48 que tem uma capacidade de transmissão máxima de 2.448 Gbits/s. A taxa média de transmissão de dados foi cerca de 350 Mbits/s e a captura foi efectuada durante duas horas.

Devido ao grande volume de dados do *trace* relativo à captura na ligação OC48, este volume foi reduzido ao equivalente a uma captura de cerca de 120 segundos, de forma a permitir manter em níveis toleráveis o processamento e armazenamento de informação no ambiente de testes desenvolvido.

Já relativamente ao *trace* oriundo da captura na conferência *Sigcomm*, a informa-

ção relativamente ao tráfego não IP foi filtrada, como por exemplo, pacotes enviados pelos pontos de acesso para efeitos de sincronização (*beacon interval*). A capacidade de transmissão de dados máxima da conexão é de 54 Mbit/s. O tempo de captura do conjunto de dados utilizado é de aproximadamente 25600 segundos. A Tabela 3.2, descreve em maior detalhe cada um dos *traces* utilizados:

Tabela 3.2: *Descrição dos traces utilizados*

Trace	Tipo de ligação	Pacotes	Fluxos	Bytes	Pacotes/fluxo	Bytes/fluxo	Duração (s)
Sigcomm	wireless - 802.11	1.519.680	127705	1037536564	11,9	8124,48	24588,44
OC48	Fibra optica - OC48	7.723.415	496640	3915378446	15.55	7883,74	119,7

3.6 Caracterização e comparação

A caracterização de tráfego, neste caso relativa ao tráfego disposto sob forma de fluxos, pode ser obtida a partir de fluxos agregados ou particularidades de pacotes ou fluxos individuais.

O volume dos conjuntos de dados resultantes da aplicação das técnicas de amostragem em estudo, diferem, quer devido a especificidades relativas ao funcionamento das próprias técnicas quer devido às diferentes frequências de amostragem utilizadas. Assim, o *overhead*, avaliado a nível de redução de dados resultante da aplicação de cada técnica de amostragem, é considerado de forma a evitar comparações tendenciosas.

As características estudadas, relativas aos fluxos agregados, referem-se a:

- Distribuições do número de pacotes por fluxo, para permitir obter uma noção da quantidade de fluxos que possuem determinado número de pacotes e inferir sobre a acurácia das estimativas. Perspectiva-se que quanto maior o número de pacotes por fluxo maior a acurácia das estimativas que se podem obter e vice-versa;

- Distribuições do tamanho dos fluxos (bytes), para permitir efectuar uma análise acerca da representatividade dos conjuntos resultantes da aplicação das técnicas de amostragem face ao tráfego original;

- Proporções do volume de dados (fluxos, pacotes e bytes), para permitir avaliar a acurácia obtida dos conjuntos resultantes após aplicação de amostragem, em termos das quantidades obtidas ao nível dos protocolos de transporte e protocolos aplicacionais.

Relativamente às características extraídas de fluxos isolados, são analisados fluxos individuais através do critério média do tempo entre chegadas de pacotes - Mean Inter arrival times (*mean* IAT), que está directamente relacionado com a duração temporal dos fluxos. Os fluxos seleccionados para análise são os mais significativos em termos de volume de dados (*bytes*), e referentes a cada um dos protocolos aplicacionais classificados.

Em síntese, as características extraídas dos fluxos utilizadas para análise são:

- Fluxos individuais
  - Nº total de pacotes
  - Volume de dados (*bytes*)
  - Duração (segundos)
  - Tempo entre a chegada de pacotes (média) - IAT
- Fluxos agregados
  - Nº total de pacotes
  - Volume de dados (*bytes*)
  - Média de pacotes por fluxo
  - Média *bytes* por fluxo

Critérios comparativos

Após identificadas as características em estudo, cada um dos conjuntos de fluxos resultantes do processo de medição com aplicação de diferentes técnicas de amostragem, são comparados tendo por base os seguintes critérios:

- nível de redução de dados;
- nível de representatividade do tráfego original:
  - fluxos agregados:
    - acurácia na caracterização dos fluxos em termos dos protocolos de transporte (TCP/UDP e Outros);
    - acurácia na caracterização dos fluxos em termos dos protocolos aplicacionais (http, dns, p2p, etc. e Outros);
  - fluxos individuais:
    - acurácia na identificação das propriedades temporais dos fluxos referentes a cada protocolo aplicacional identificado;

Para quantificar o quanto se distanciam os resultados das estimativas dos conjuntos de dados resultantes das técnicas de medição com amostragem face ao tráfego total original, é utilizada a métrica erro relativo. O erro relativo é a diferença (em módulo) entre o valor estimado (após medição com amostragem) e o valor real (referente ao tráfego original), dividido por este último (e multiplicado por 100 caso se apresentem os valores em percentagem), como mostra a fórmula:

$$Erro = \frac{|valor_{estimado} - valor_{real}|}{valor_{real}} \tag{3.1}$$

3.7 Resumo

Neste capítulo foi apresentada a descrição do funcionamento do ambiente de testes criado para análise do impacto dos processos de amostragem nas características úteis para identificação de fluxos. Foram definidos os conceitos gerais utilizados nos processos de amostragem, descritas as técnicas de amostragem utilizadas e seu funcionamento e parâmetros de inicialização, até à posterior classificação da informação sob a forma de fluxos, em que é apresentada a definição de fluxo utilizada, e a forma como cada fluxo é classificado de acordo com o protocolo que o origina. Posteriormente foram identificadas as características dos fluxos individuais e agregados, que são utilizadas para a análise comparativa.

Os resultados obtidos e respectiva análise serão apresentados no capítulo 4.

Capítulo 4

Análise de resultados

Neste capítulo é apresentada a análise e discussão dos resultados segundo as métricas obtidas dos vários conjuntos de registos de fluxos, que resultam da medição do tráfego com aplicação das técnicas de amostragem descritas na secção 3.3 (Multi-adaptativa, Sistemática 1:100 e 1:1000 e Aleatória 1:100 e 1:1000)

Ambos os *traces* utilizados para análise (OC48 e *Sigcomm*) são bastante heterogéneos, contendo grande variedade de protocolos e aplicações. De notar que no *trace* *Sigcomm* não foi identificado nenhum fluxo originado por qualquer aplicação P2P ou relativamente a aplicações identificadas como Jogos, provavelmente devido ao seu uso estar bloqueado por estar fora do contexto da conferência. Já relativamente ao *trace* OC48, são identificados vários fluxos contendo estes serviços.

Para a análise ao desempenho das técnicas de amostragem, os seus resultados são sempre comparados com o conjunto referente ao tráfego original total, sem aplicação de qualquer processo de amostragem, e depois comparados os resultados das técnicas de amostragem entre si. Os conjuntos resultantes de tráfego serão então analisados segundo os critérios definidos na secção 3.6 (redução de dados e representatividade do tráfego total quanto às caracteristicas de fluxos individuais e fluxos agragados).

4.1 Redução de dados processados

Primeiramente, para análise das características do tráfego na sua forma original, este é processado de forma a obter um conjunto agregado sob forma de fluxos sem aplicação de qualquer técnica de amostragem.

As Tabelas 4.1 e 4.2 mostram a quantidade de fluxos, pacotes e *bytes*, correspondentes a cada *trace* de tráfego em estudo, antes e após aplicação de cada uma das técnicas de amostragem.

Tabela 4.1: *Dados do trace OC48 antes e após amostragem*

Conjunto	Fluxos	Pacotes	Bytes
Tráfego Total	496640	7723415	3915378446
Multi-Adaptativa	103730	430670	219509853
Sistemática 1:100	39652	77234	38965771
Aleatória 1:100	39476	77234	38929675
Sistemática 1:1000	6089	7723	3910472
Aleatória 1:1000	6069	7723	3931802

Tabela 4.2: *Dados do trace Sigcomm antes e após amostragem*

Conjunto	Fluxos	Pacotes	Bytes
Tráfego Total	127705	1519680	1037536564
Multi-Adaptativa	15774	80731	57224042
Sistemática 1:100	7838	15196	10349353
Aleatória 1:100	7783	15196	10378390
Sistemática 1:1000	1113	1519	1033280
Aleatória 1:1000	1119	1519	1008603

Os gráficos das Figura 4.1 e 4.2, mostram a percentagem de dados (fluxos, pacotes e *bytes*) após a aplicação de cada uma das técnicas de amostragem, relativamente ao tráfego total.

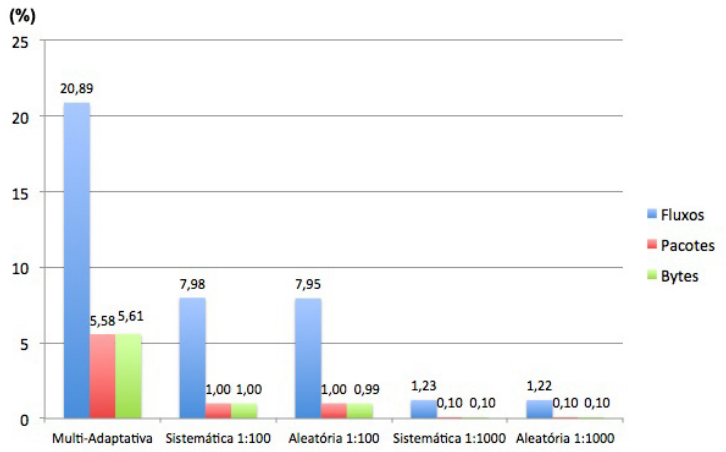

Figura 4.1: *Percentagem de dados após amostragem - OC48*

Observando os gráficos, pode verificar-se que em termos de número de fluxos, para cada um dos *traces* OC48 e *Sigcomm*, obtém-se aproximadamente 21% e 12.5% respectivamente, através da aplicação da técnica Multi-adaptativa. Estes volumes de fluxos são originados a partir da selecção de cerca de 5.6% e 5.3% dos pacotes observados, respectivamente. Os valores correspondentes à percentagem de volume em *bytes* são sempre muito próximos dos valores correspondentes aos volume de pacotes.

Relativamente às técnicas de amostragem sistemática e aleatória com taxa de amostragem 1:100, não existem diferenças significativas entre as duas. A percen-

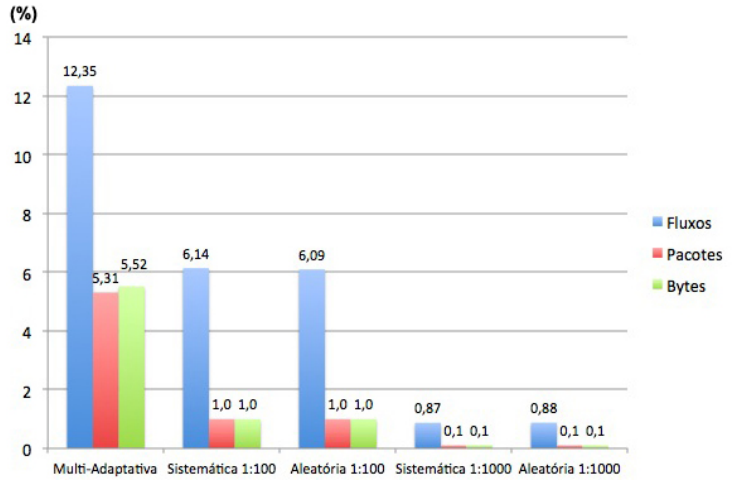

Figura 4.2: *Percentagem de dados após amostragem - Sigcomm*

tagem de fluxos em relação à totalidade do tráfego é de aproximadamente 8% e 6%, originados por cerca de 1% do total de pacotes observados. Quanto às mesmas técnicas mas com taxas de amostragem na ordem de 1:1000, em termos proporcionais verifica-se um ligeiro aumento na quantidade de fluxos identificados, ou seja, a quantidade de fluxos em relação à quantidade de pacotes é maior para as técnicas de menor frequência (1:1000) comparativamente com as técnicas de maior frequência (1:100), conforme se pode ver nos gráficos das Figuras 4.1 e 4.2. O valor percentual de fluxos face ao tráfego total é, no entanto, muito baixo para o caso das técnicas (1:1000).

Pode afirmar-se que, a quantidade de informação relativamente aos fluxos, é tanto maior quanto maior for a quantidade de pacotes seleccionados a partir dos mecanismos de amostragem. Assim, para realização de uma avaliação mais imparcial às técnicas de amostragem, considera-se a relação pacotes colectados/fluxos resultantes.

Com base nesta relação, observa-se que da aplicação das técnicas Sistemática ou Aleatória, a percentagem de fluxos que se obtém é de 6 a 10 vezes superior à percentagem de pacotes que os originam. Já relativamente à técnica Multi-adaptativa a percentagem de fluxos vai de aproximadamente 2 a 3 vezes o valor da percentagem de pacotes. Ou seja, a partir das técnicas Sistemática e Aleatória, obtém-se maior quantidade de fluxos a partir de um menor volume de pacotes, em comparação com a técnica Multi-adaptativa.

4.2 Representatividade do tráfego

Para facilitar a representação e análise dos tamanhos dos fluxos, foram definidos intervalos de tamanho em pacotes e *bytes* nos quais são dispostos os volume des fluxos correspondentes. Os intervalos definidos permitem identificar os volumes de fluxos em percentagem e também informar sobre a sua relevância em função do volume total de fluxos resultante.

4.2.1 Distribuição de fluxos após amostragem

Analisando as Tabelas 4.3 e 4.4, referentes à distribuição de pacotes, verifica-se que para as técnicas Sistemática e Aleatória, 80 a 90% dos fluxos resultantes, contêm apenas 1 pacote, enquanto que segundo a técnica Multi-adaptativa, este valor anda à volta dos 50% para ambos os *traces*. De forma genérica, as técnicas de amostragem tendem a sobrestimar o número de fluxos de pequena dimensão, fruto do próprio processo de amostragem. A técnica Multi-Adaptativa, ao seleccionar amostras contendo pacotes sequenciais (ao contrário das técnicas Sistemática e Aleatória) tem influência no decréscimo do número de fluxos com 1 só pacote.

Tabela 4.3: *Percentagem de fluxos por volume de pacotes - OC48*

Intervalo (Pacotes)	Tráfego Total	Multi Adaptativa	Sistemática 100	Aleatória 1-100	Sistemática 1000	Aleatória 1-1000
1	27,416	53,013	76,690	76,743	86,188	85,961
2 - 4	26,541	31,765	17,114	16,876	12,399	12,523
4 - 8	24,572	7,415	3,538	3,648	1,117	1,203
8 - 16	10,404	3,933	1,604	1,664	0,230	0,280
16 - 32	5,503	2,262	0,731	0,730	0,049	0,016
32 - 64	2,710	0,983	0,265	0,281	0,016	0,016
64 - 128	1,325	0,380	0,048	0,051	-	-
128 - 256	0,676	0,195	0,005	0,003	-	-
256 - 512	0,476	0,040	0,005	0,005	-	-
512 - 1024	0,227	0,012	-	-	-	-
1024 - 2048	0,087	0,002	-	-	-	-
2048 - 4096	0,048	0,001	-	-	-	-
4096 - 8192	0,012	-	-	-	-	-
8192 - 16384	0,003	-	-	-	-	-
16384 - 32768	0,0002	-	-	-	-	-
acima de 1024	0,0002	-	-	-	-	-

Tabela 4.4: *Percentagem de fluxos por volume de pacotes - Sigcomm*

Intervalo (Pacotes)	Tráfego Total	Multi Adaptativa	Sistemática 100	Aleatória 1-100	Sistemática 1000	Aleatória 1-1000
1	40,846	55,680	82,279	81,601	90,386	87,578
1 a 4	29,987	25,707	13,281	14,005	6,918	9,473
4 a 8	13,069	9,249	2,360	2,261	1,078	1,966
8 a 16	7,371	4,723	1,072	1,105	1,258	0,804
16 a 32	4,635	2,536	0,536	0,565	0,359	0,179
32 a 64	2,210	1,262	0,255	0,244	-	-
64 a 128	0,955	0,463	0,153	0,141	-	-
128 a 256	0,486	0,209	0,051	0,064	-	-
256 a 512	0,214	0,108	0,013	0,013	-	-
512 a 1024	0,124	0,057	-	-	-	-
1024 a 2048	0,054	0,006	-	-	-	-
2048 a 4096	0,027	-	-	-	-	-
4096 a 8192	0,013	-	-	-	-	-
8192 a 16384	0,005	-	-	-	-	-
acima de 1024	0,004	-	-	-	-	-

As Tabelas 4.5 e 4.6 e correspondentes gráficos representados nas Figuras 4.3 e

4.4, referentes às distribuições de frequências, indicam a quantidade de fluxos por tamanho em *bytes*.

Tabela 4.5: *Distribuição dos tamanhos dos fluxos e respectivos volumes - OC48*

Intervalo (Bytes)	Bytes	%	nº Fluxos	%
até 500	41550342	1,06	259053	52,16
500-1000	76679134	1,96	111406	22,43
1000-2000	49124432	1,25	35489	7,15
2000-4000	71373787	1,82	25166	5,07
4000-8000	93744981	2,39	16513	3,32
8000-16000	205989472	5,26	18146	3,65
16000-32000	374619800	9,57	16190	3,26
32000-64000	330411164	8,44	7407	1,49
64000-128000	301864098	7,71	3408	0,69
128000-256000	311659746	7,96	1738	0,35
256000-512000	455223155	11,63	1247	0,25
512000-1024000	316974576	8,10	451	0,09
1024000-2048000	306138637	7,82	213	0,04
2048000-4096000	384364693	9,82	133	0,03
4096000-8192000	357871540	9,14	63	0,01
8192000-16384000	125028261	3,19	12	0,00
16384000-32768000	74918871	1,91	4	0,00
acima de 32768000	37841757	0,97	1	0,00

Tabela 4.6: *Distribuição dos tamanhos dos fluxos e respectivos volumes - OC48*

Intervalo (Bytes)	Bytes	%	nº Fluxos	%
até 500	20989633	2,02	64720	50,68
500-1000	14889694	1,44	21071	16,50
1000-2000	20114440	1,94	13819	10,82
2000-4000	27809524	2,68	9880	7,74
4000-8000	42480008	4,09	7458	5,84
8000-16000	50340542	4,85	4496	3,52
16000-32000	59637633	5,75	2643	2,07
32000-64000	74141597	7,15	1660	1,30
64000-128000	85186858	8,21	958	0,75
128000-256000	86113741	8,30	488	0,38
256000-512000	106812378	10,29	302	0,24
512000-1024000	75682691	7,29	108	0,08
1024000-2048000	70791098	6,82	51	0,04
2048000-4096000	82969605	8,00	27	0,02
4096000-8192000	82667530	7,97	15	0,01
8192000-16384000	96686504	9,32	8	0,01
16384000-32768000	0	0	0	0
acima 32768000	40223088	3,88	1	0,001

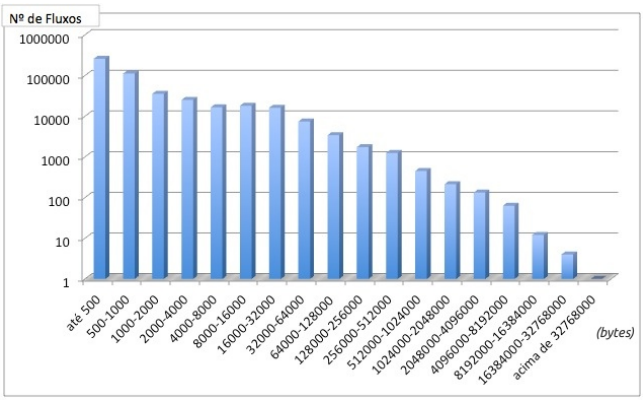

Figura 4.3: *Distribuição de fluxos - OC48*

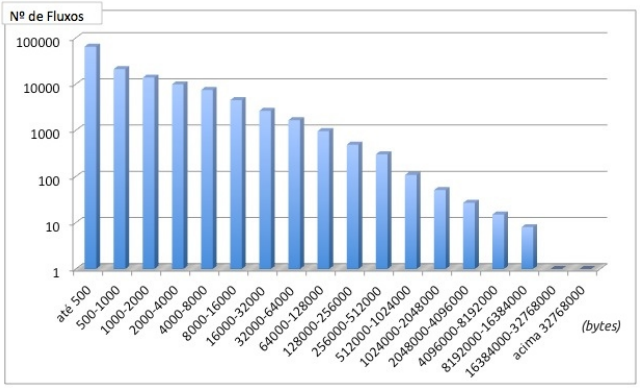

Figura 4.4: *Distribuição de fluxos - Sigcomm*

Em ambos os gráficos da distribuição de frequências, pode verificar-se que em termos de quantidade de fluxos, a maioria (mais de 50%) possui em termos individuais um volume de dados menor ou igual a 500 *bytes*, correspondendo a um total de 1,06% do volume total de tráfego para o caso do tráfego OC48 e cerca de 2,02% no caso do tráfego *Sigcomm*.

Estes resultados, permitem confirmar que, como referido em vários estudos [56][57], uma pequena percentagem dos fluxos contribui para uma grande percentagem do volume de tráfego. O estudo apresentado em [56] mostra que em várias medições efectuadas ao tráfego entre pares AS (*Autonomous System*), cerca de 9% do total de fluxos corresponde a mais de 90% do volume de tráfego em *bytes*.

Após amostragem, apesar de não representado graficamente, verificou-se que as

distribuições de frequências dos conjuntos resultantes mantêm a mesma tendência do tráfego total (como nas Figuras 4.3 e 4.4), ou seja, as maiores quantidades de fluxos correspondem sempre aos fluxos com menor tamanho, apresentando distribuições do tipo *heavy-tailed*, ou seja, a forma da assimptota da distribuição é hiperbólica [58].

Nos gráficos das Figuras 4.5 e 4.6 está representada a distribuição dos fluxos por tamanho e respectivo volume que representam para o tráfego global (*MBytes*). Nestes gráficos pode ver-se que em ambas as distribuições, o intervalo mais relevante em termos de volume é aquele cujos tamanhos de fluxos se compreendem entre 256KB e 512KB. Ambos acima dos 10% do total de volume de dados como se pode ver nas Tabelas 4.5 e 4.6, correspondendo a aproximadamente cerca de 0,25% da quantidade total de fluxos em ambos os casos.

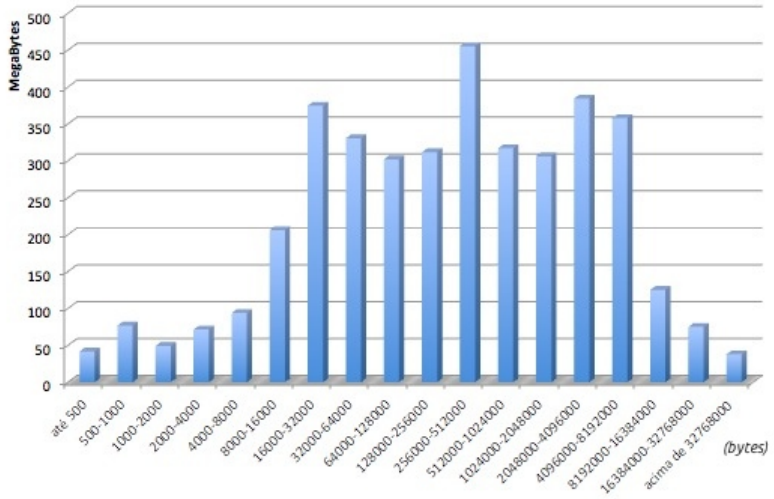

Figura 4.5: *Distribuição de fluxos do tráfego total - OC48*

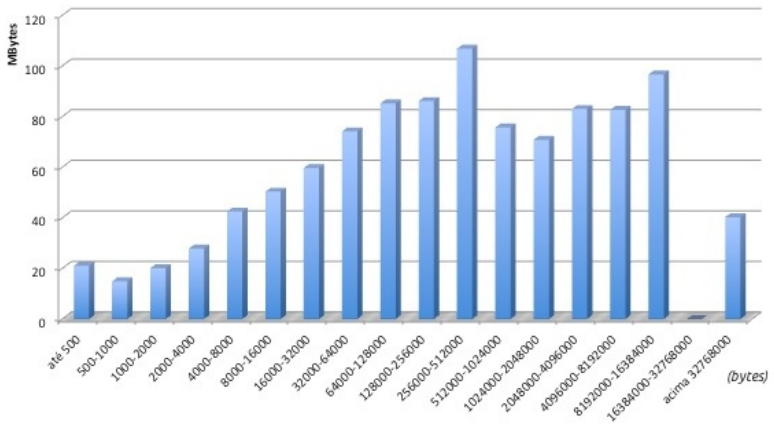

Figura 4.6: *Distribuição de fluxos do tráfego total - Sigcomm*

Relativamente às distribuições dos volumes referentes ao tráfego após amostragem, ao comparar os gráficos destas com os referentes ao tráfego total (Figuras 4.5

e 4.6), verifica-se um nível de semelhança maior entre os gráficos tráfego total e técnica Multi-adaptativa, do que relativamente aos gráficos das restantes técnicas de amostragem, como se pode verificar através dos gráficos das Figuras 4.7 e 4.8.

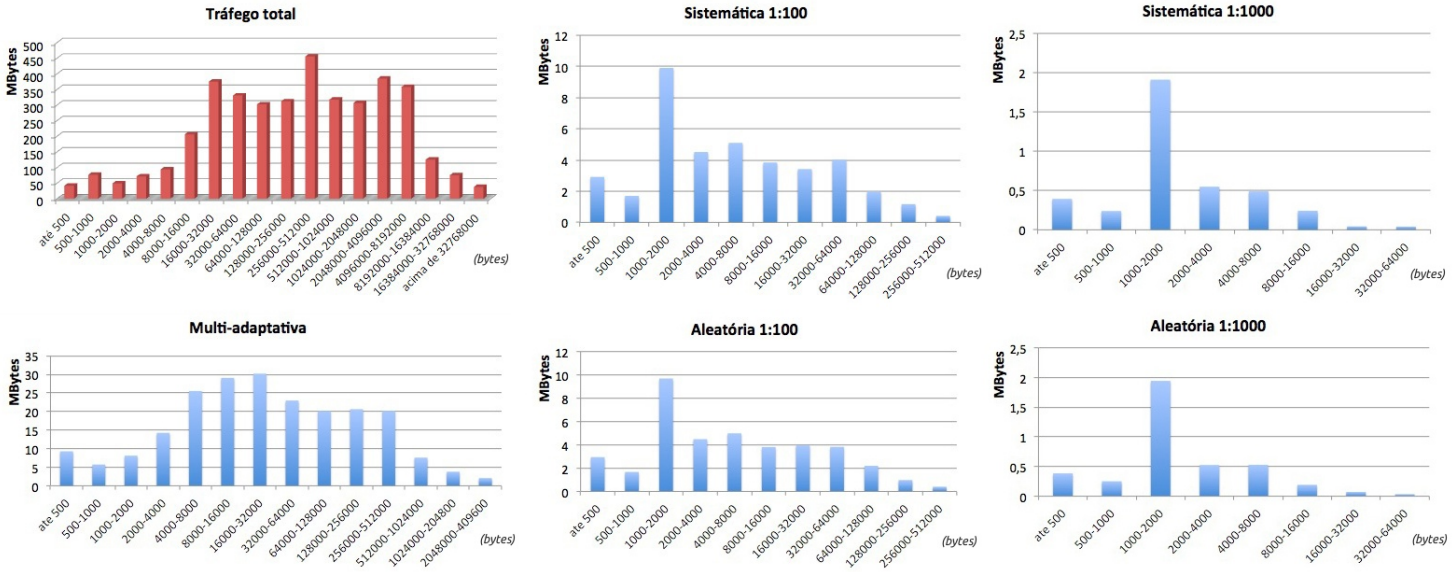

Figura 4.7: *Comparação da distribuição de fluxos após amostragem - OC48*

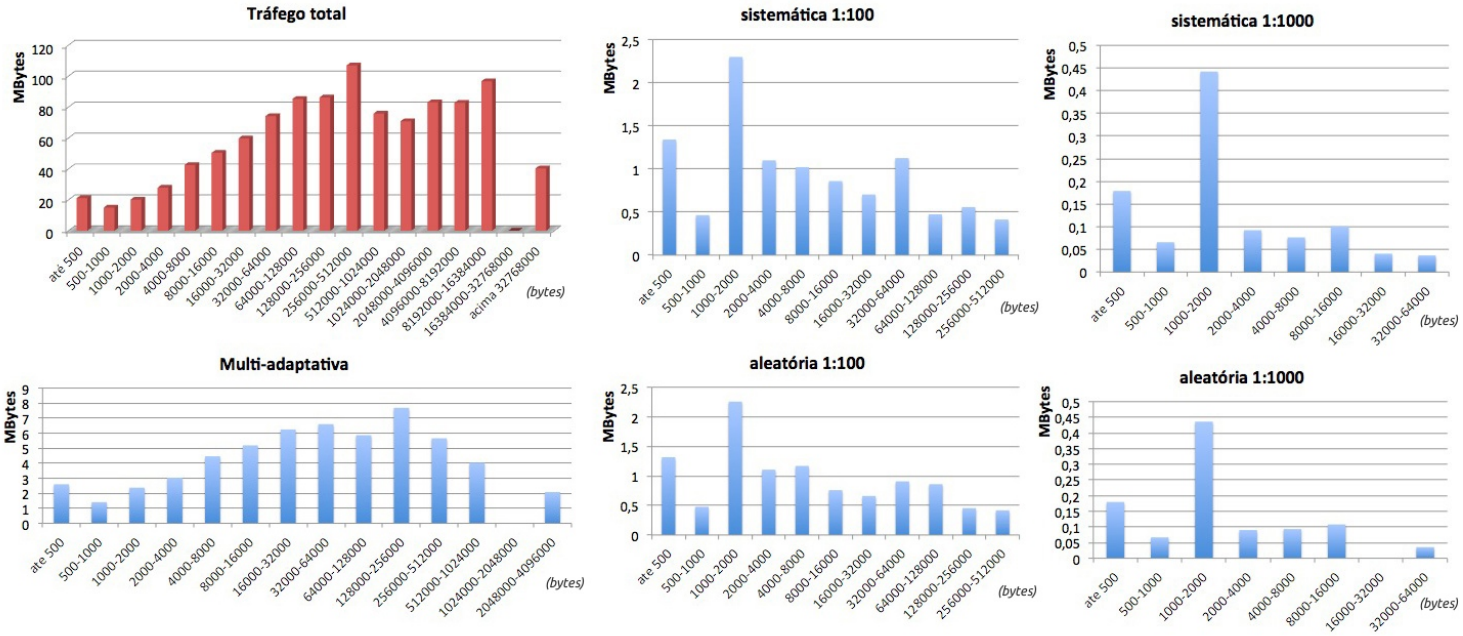

Figura 4.8: *Comparação da distribuição de fluxos após amostragem - Sigcomm*

Para se verificar se esta semelhança está ou não relacionada com o volume de dados resultantes, que no caso da técnica Multi-adaptativa é significativamente maior que o volume obtido das técnicas Sistemática e Aleatória, efectuou-se nova medição para cada uma destas técnicas, mas com a taxa de amostragem a 1:18, ou seja,

de forma que o volume de pacotes resultante seja igual a aproximadamente 5% do tráfego total original, como acontece para o caso da técnica Multi-adaptativa.

O resultado, como ilustrado nas Figuras 4.9 e 4.10, indica-nos que mesmo com um nível de redução de dados equivalenteque à técnica Multi-adaptativa, em termos estatísticos, esta última assemelha-se mais ao comportamento do tráfego original.

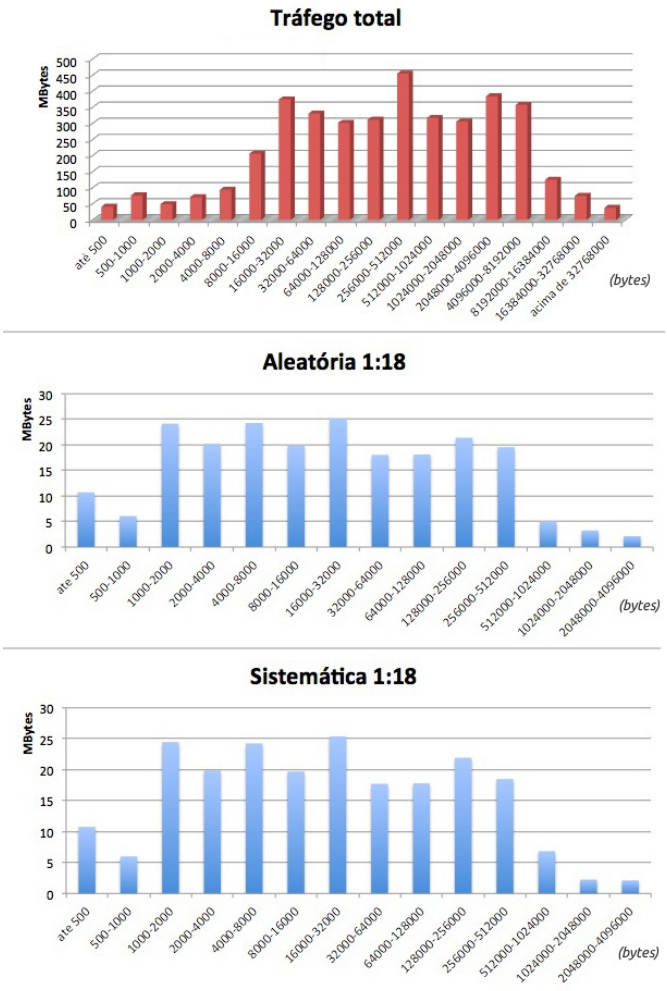

Figura 4.9: *Comparação da distribuição de fluxos com taxa de amostragem 1:18 - OC48*

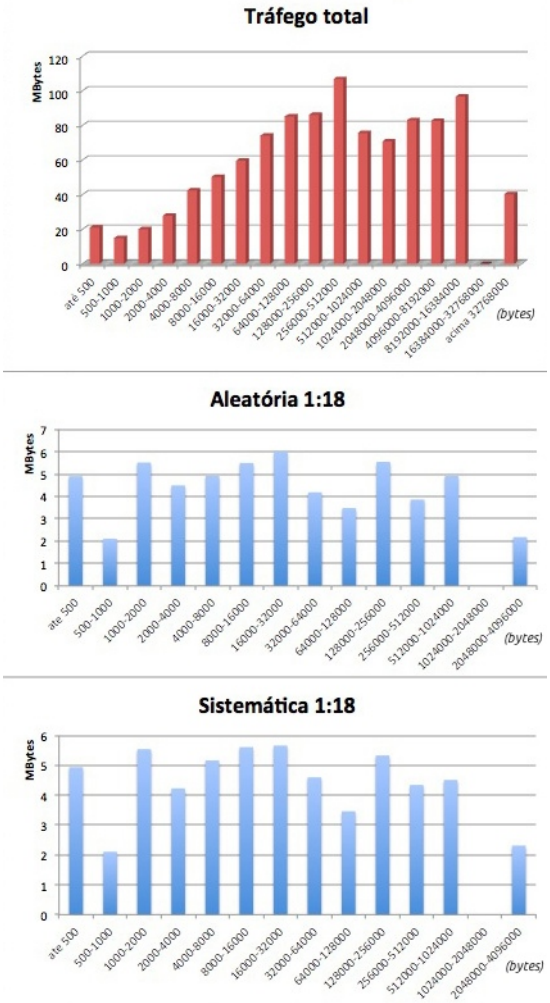

Figura 4.10: *Comparação da distribuição de fluxos com taxa de amostragem 1:18 - Sigcomm*

Uma razão que poderá influenciar este resultado pode advir do facto de que através da técnica Multi-adaptativa as amostras podem conter vários pacotes consecutivos ao contrário da Sistemática e Aleatória (que pode ter no máximo 2 pacotes consecutivos, inda que com baixa probabilidade). Isto poderá ser verificado analisando o nível de acerto da técnica Sistemática, configurada para captar 5 pacotes a cada 100 (5:100), e efectuando a mesma comparação com a técnica Multi-Adaptativa.

4.2.2 Proporções de protocolos classificados

Protocolos de transporte

De seguida é efectuada a contabilização dos fluxos pelo protocolo de transporte que os originou TCP, UDP e Outros. Do conjunto denominado 'Outros', fazem parte todos os restantes protocolos identificados nos fluxos (ICMP, IGMP, SCTP, etc.). O tráfego de cada protocolo será apresentado através das suas proporções de número de pacotes, *bytes* e fluxos, relativos ao conjunto total de dados obtido. Cada conjunto concerne os dados obtidos a partir da aplicação de cada técnica de amostragem. As proporções são então comparadas com as resultantes do conjunto referente ao tráfego total.

Através da análise das Tabelas 4.7, 4.8, 4.9 e 4.10, e respectivos gráficos nas figuras 4.11, 4.12, 4.13 e 4.14, pode verificar-se que todas as técnicas alcançam um nível de acurácia alto, ou seja, a diferença entre as proporções de tráfego identificados como TCP, UDP e Outros, é menor que 3%.

Volume em pacotes

Tabela 4.7: *Percentagem de volume (pacotes) - OC48*

Protocolo	Tráfego Total	Multi Adaptativa	Sistemática 100	Aleatória 1-100	Sistemática 1000	Aleatória 1-1000
TCP	90,99	90,83	90,96	91,12	91,20	91,49
UDP	7,68	7,77	7,70	7,52	7,45	7,45
Outros	1,33	1,41	1,34	1,36	1,36	1,06

Tabela 4.8: *Percentagem de volume (pacotes) - Sigcomm*

Protocolo	Tráfego Total	Multi Adaptativa	Sistemática 100	Aleatória 1-100	Sistemática 1000	Aleatória 1-1000
TCP	66,90	69,79	67,03	67,39	67,94	66,56
UDP	32,33	29,50	32,21	31,98	31,27	32,26
Outros	0,77	0,71	0,76	0,63	0,79	1,18

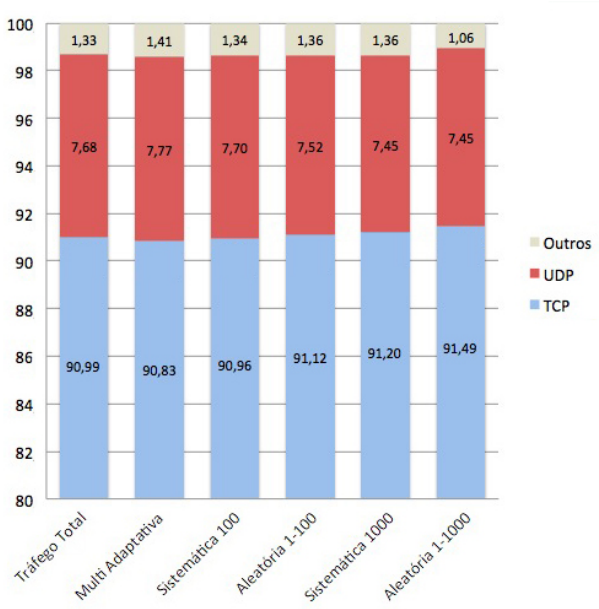

Figura 4.11: *Percentagem de volume (pacotes) - OC48 (escala de 80% a 100% para melhor visualização)*

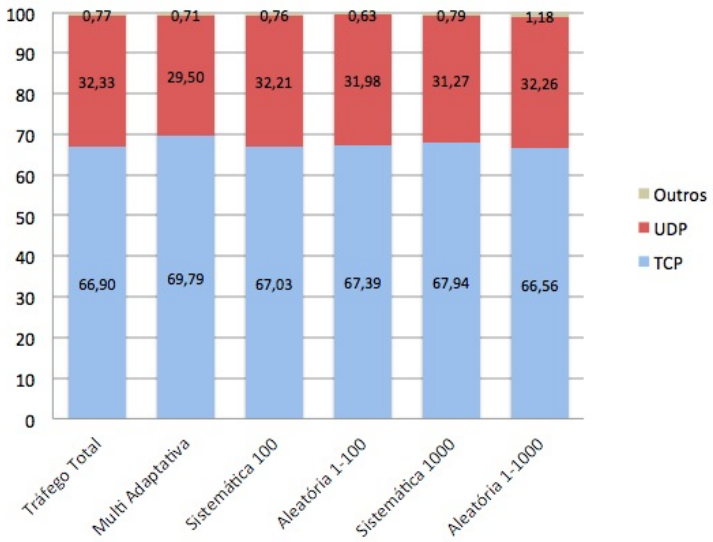

Figura 4.12: *Percentagem de volume (pacotes) - Sigcomm*

Volume em *bytes*

Tabela 4.9: *Percentagem de volume (bytes) - OC48*

Protocolo	Tráfego Total	Multi Adaptativa	Sistemática 100	Aleatória 1-100	Sistemática 1000	Aleatória 1-1000
TCP	96,48	95,95	96,46	96,48	96,79	96,7
UDP	2,77	3,04	2,81	2,79	2,48	2,77
Outros	0,75	1,01	0,73	0,73	0,73	0,53

Tabela 4.10: *Percentagem de volume (bytes) - Sigcomm*

Protocolo	Tráfego Total	Multi Adaptativa	Sistemática 100	Aleatória 1-100	Sistemática 1000	Aleatória 1-1000
TCP	80,29	82,39	80,40	80,31	80,30	80,76
UDP	19,41	17,34	19,32	19,45	19,43	18,83
Outros	0,31	0,27	0,28	0,25	0,27	0,41

Comparando as técnicas Sistemática e Aleatória, não existem diferenças significativas entre as proporções, quer para as técnicas de amostragem com taxa de amostragem 1:100 quer para as técnicas com frequência de amostragem 10 vezes mais baixa (1:1000).

Comparando a técnica de amostragem Multi-adaptativa com as restantes técnicas, tanto em termos de quantidade de pacotes como volume em *bytes*, o erro ainda que baixo, é ligeiramente maior que o das restantes técnicas, mesmo partindo de um volume de dados maior em termos de amostras colectadas.

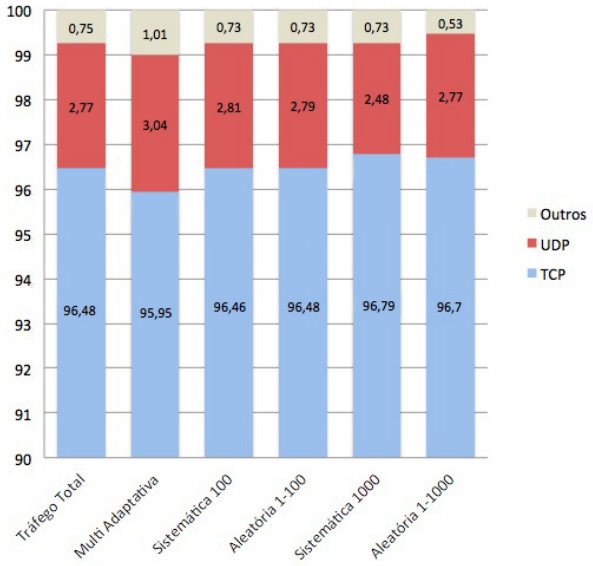

Figura 4.13: *Percentagem de volume (bytes) - OC48 (escala de 90% a 100% para melhor visualização)*

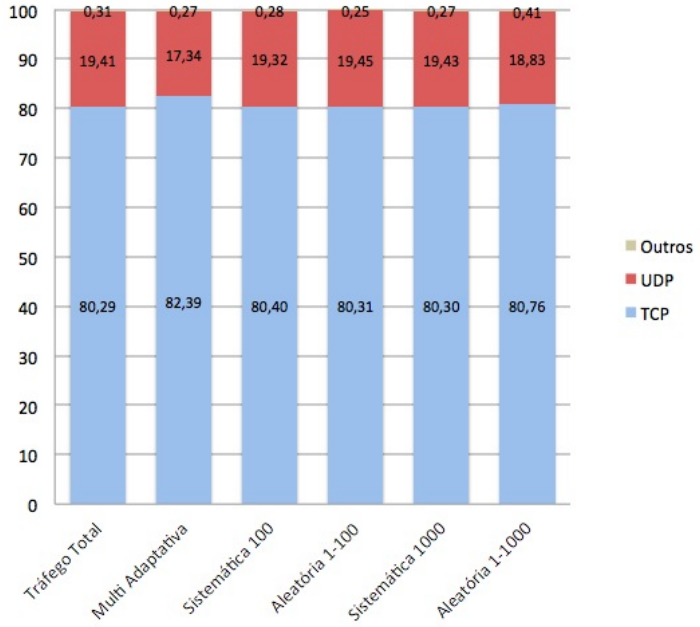

Figura 4.14: *Percentagem de volume (bytes) - Sigcomm*

Volume em fluxos

Verifica-se através das Tabelas 4.11 e 4.12 e dos gráficos das Figuras 4.15 e 4.16 que, após aplicação de amostragem, o número de fluxos correspondente ao protocolo TCP tende a aumentar relativamente ao tráfego total e consequentemente o número de fluxos UDP e Outros a diminuir. Trata-se, portanto, de um erro de sobrestimação da quantidade de fluxos TCP.

Tabela 4.11: *Percentagem de volume (fluxos) - OC48*

Protocolo	Tráfego Total	Multi Adaptativa	Sistemática 100	Aleatória 1-100	Sistemática 1000	Aleatória 1-1000
TCP	76,77	89,58	93,39	93,62	93,51	93,56
UDP	20,61	9,13	5,75	5,50	5,76	5,80
Outros	2,62	1,29	0,86	0,87	0,72	0,64

Tabela 4.12: *Percentagem de volume (fluxos) - Sigcomm*

Protocolo	Tráfego Total	Multi Adaptativa	Sistemática 100	Aleatória 1-100	Sistemática 1000	Aleatória 1-1000
TCP	34,77	56,11	73,87	74,93	75,29	72,65
UDP	63,56	42,41	25,26	24,22	23,90	26,18
Outros	1,67	1,48	0,87	0,85	0,81	1,16

Para este caso, comparando a técnica Multi-adaptativa às restantes, apesar de também se verificar um aumento na percentagem de fluxos TCP e diminuição nos fluxos UDP e Outros, em relação ao tráfego original, o erro da acurácia é significativamente menor, de 14% a 17% aproximadamente.

Relativamente às técnicas Sistemática e Aleatória, as proporções diferem muito pouco, mesmo comparando as técnicas de frequencias 1:100 com as de 1:1000.

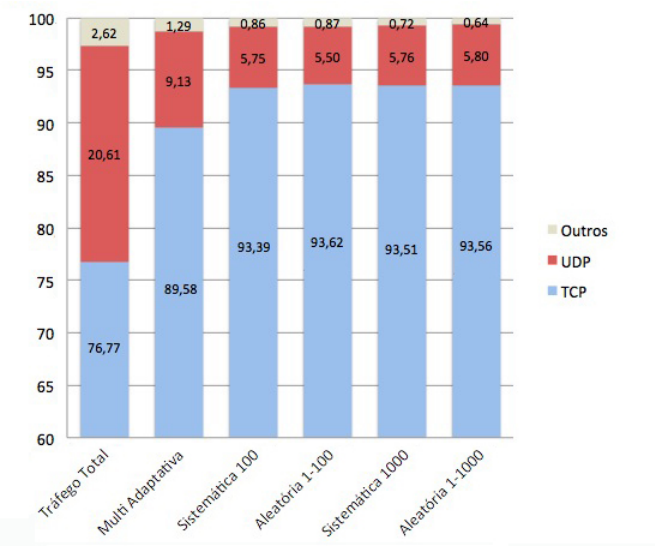

Figura 4.15: *Percentagem de volume (fluxos) - OC48 (escala de 60% a 100% para melhor visualização)*

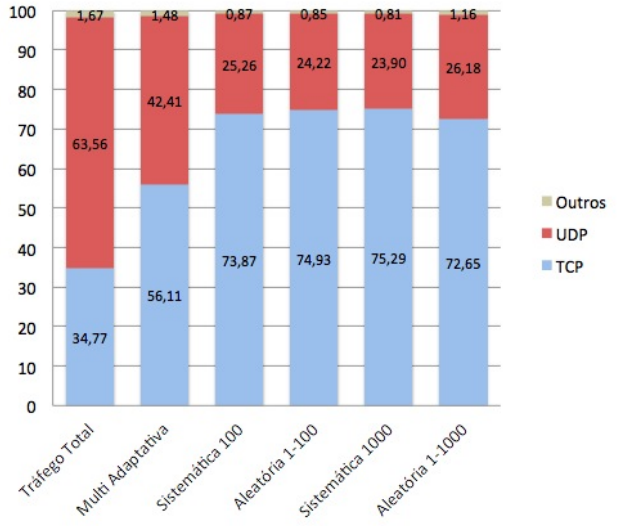

Figura 4.16: *Percentagem de volume (fluxos) - Sigcomm*

Apesar do comportamento verificado anteriormente em relação às proporções de pacotes e *bytes*, nos caso das proporções de fluxos, verifica-se que o níveis de acurácia baixam. A acurácia relativamente às técnicas Multi-adaptativa e Sistemática/Aleatória invertem-se, ou seja, para os casos anteriores relativamente às proporções de número de pacotes e *bytes*, as técnicas Sistemática e Aleatória apresentam uma acurácia ligeiramente maior que a Multi-adaptativa. Já neste caso as proporções de fluxos obtidas segundo estas técnicas, possuem um erro maior que o obtido pela Multi-adaptativa, que por sua vez também possui um erro significativo em relação ao tráfego total.

Efectuando a comparação entre as técnicas Sistemática e Aleatória, pode verificar-se que as proporções não se alteram consoante a alteração da frequência de amostragem. Ou seja, técnicas de amostragem com frequências de amostragem 1:100 e 1:1000 apresentam resultados muito próximos em termos das proporções identificadas como pertencendo aos conjunto TCP, UDP e Outros. Utilizando taxas de amostragem maiores, os resultados não diferem signficativamente, em relação aos obtidos pelas técnicas Sistemática e Aleatória (verificação efectuada utilizando a taxa de amostragem 1:18, que resulta na selecção de aproximadamente 5% dos pacotes, como no caso da técnica Multi-adaptativa).

Protocolos aplicacionais

De seguida são contabilizados os fluxos da mesma forma que anteriormente mas desta vez, dispondo os resultados segundo o protocolo aplicacional que gerou cada fluxo. A classificação tem por base apenas a informação relativa ao valor das portas

dos cabeçalhos TCP ou UDP. Em comparação com a análise anterior, esta abordagem resulta num aumento da granularidade através da qual são classificados os fluxos.

Devido à grande variedade de protocolos aplicacionais identificados e à pouca relevância em termos de volume de tráfego de grande parte destes, os protocolos ilustrados nos gráficos e utilizados para comparação são apenas aqueles cuja soma total do volume de dados (*bytes*) representa pelo menos 80% do tráfego total. O restante será incluído num grupo designado 'Outros'.

Neste caso apenas são apresentados os resultados em termos do volume em *bytes*. Relativamente ao número de fluxos e pacotes o comportamento observado é análogo ao observado em TCP, UDP e Outros (secção 4.2.2), ou seja, há uma aproximação elevada em termos das proporções de pacotes, e baixa em termos das proporções de fluxos. Os resultados são apresentados nas Tabelas 4.13 e 4.14, correspondentes aos gráficos das Figuras 4.17 e 4.18.

Analisando os gráficos, verifica-se que mesmo efectuando a classificação com maior granularidade relativamente à classificação em protocolos de transporte, obtém-se uma acurácia considerável em termos das proporções das quantidades do tráfego comparativamente com o tráfego original.

Tabela 4.13: *Percentagem de volume (bytes) - OC48*

Protocolo	Tráfego Total	Multi-adaptativa	Sistemática 1:100	Aleatória 1:100	Sistemática 1:1000	Aleatória 1:1000
http	60,68	60,78	60,49	60,15	62,17	60,32
p2p	12,82	12,65	12,85	13,19	12,34	13,03
ftp	1,35	1,40	1,31	1,35	1,53	1,58
smtp	1,97	1,86	1,94	2,11	2,32	2,42
valisys	0,89	0,86	0,92	0,89	0,80	0,88
games	0,60	0,61	0,61	0,61	0,49	0,52
https	0,84	0,90	0,81	0,92	0,55	0,78
ms-streaming	1,04	1,00	1,04	1,07	0,92	0,97
Outros	19,80	19,94	20,03	19,71	18,87	19,49

Tabela 4.14: *Percentagem de volume (bytes) - Sigcomm*

Protocolo	Tráfego Total	Multi-adaptativa	Sistemática 1:100	Aleatória 1:100	Sistemática 1:1000	Aleatória 1:1000
http	44,85	46,38	45,08	44,29	45,11	44,70
https	12,40	11,67	12,43	12,61	11,47	11,96
imap	12,66	14,50	12,45	12,75	12,04	13,54
rtmp	1,39	0,62	1,20	1,38	0,97	1,52
ipsec	8,28	7,33	8,22	8,51	8,52	7,00
mdns	1,76	1,54	1,77	1,68	1,61	1,94
ssh	4,10	4,69	4,10	4,33	5,29	3,69
netbios	6,76	6,29	6,76	6,65	6,45	7,10
Outros	7,80	6,98	7,97	7,80	8,54	8,56

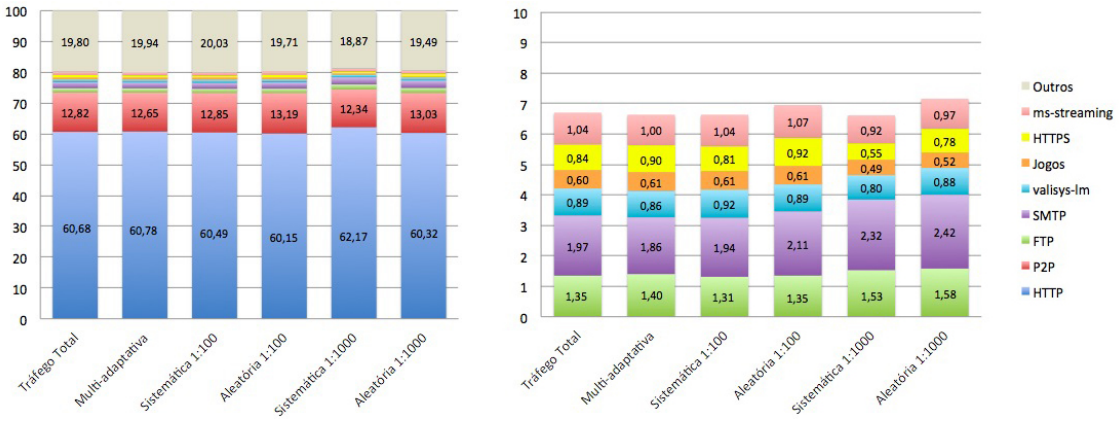

Figura 4.17: *Percentagem de volume (bytes) - OC48 (à direita os fluxos menos significativos, com uma escala diferente para melhor visualização)*

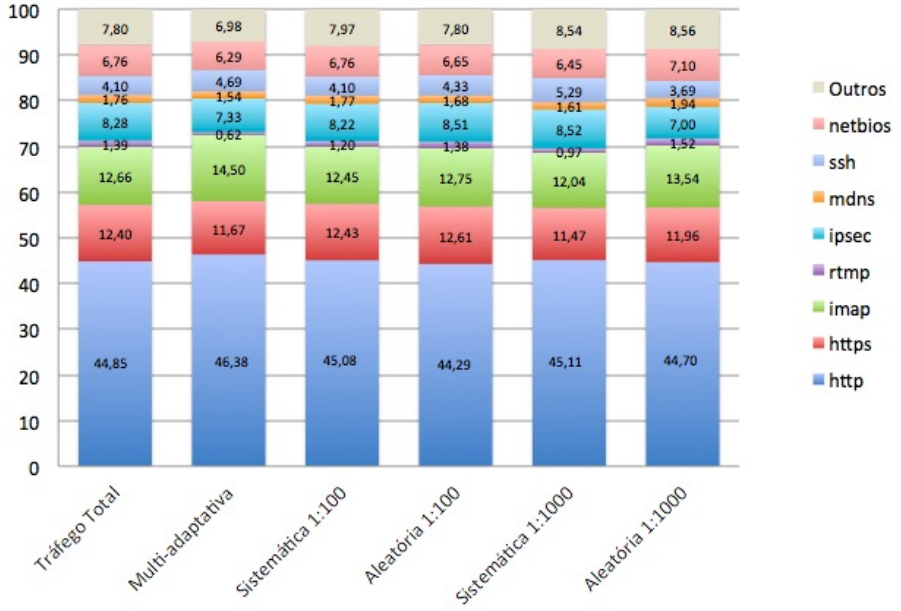

Figura 4.18: *Percentagem de volume (bytes) - Sigcomm*

Em ambos os *traces* de tráfego analisados, a técnica Sistemática 1:100 foi a que, no geral, obteve resultados com maior acurácia, após calculada a soma dos erros relativos da proporção de cada protocolo resultante, face ao tráfego total original (Tabelas 4.15 e 4.16). Comparando as técnicas Sistemática e Aleatória de menor frequência de amostragem (1:1000), a técnica Aleatória é a que apresenta maior acurácia nas proporções para ambos os *traces* de tráfego.

4.2.3 Análise do tempo entre chegada de pacotes

De seguida, serão analisados os fluxos quanto à média do tempo entre chegada de pacotes (*mean inter arrival times - mean* IAT). Esta é uma métrica frequentemente utilizada para caracterização de fluxos de tráfego. Está relacionada com a duração

Tabela 4.15: *Erros relativos por protocolo - OC48*

Téc. de amostragem	HTTP	P2P	FTP	SMTP	valisys-lm	Jogos	HTTPS	ms-streaming	Erro relativo (Soma)
Multi-adaptativa	0,002	0,013	0,037	0,057	0,036	0,017	0,078	0,046	0,286
Sistemática 1:100	0,003	0,003	0,029	0,018	0,033	0,025	0,035	0,009	0,155
Aleatória 1:100	0,009	0,029	0,001	0,068	0,008	0,023	0,098	0,027	0,261
Sistemática 1:1000	0,025	0,037	0,133	0,173	0,100	0,176	0,345	0,122	1,111
Aleatória 1:1000	0,006	0,016	0,171	0,228	0,013	0,135	0,072	0,069	0,709

Tabela 4.16: *Erros relativos por protocolo - Sigcomm*

Téc. de amostragem	HTTP	HTTPS	IMAP	RTMP	Ipsec	MDNS	SSH	netbios	Erro relativo (Soma)
Multi-adaptativa	0,034	0,059	0,145	0,551	0,114	0,125	0,143	0,070	1,241
Sistemática 1:100	0,005	0,002	0,016	0,133	0,007	0,005	0,000	0,001	0,170
Aleatória 1:100	0,012	0,017	0,007	0,005	0,028	0,046	0,056	0,016	0,188
Sistemática 1:1000	0,006	0,075	0,049	0,300	0,029	0,087	0,289	0,046	0,881
Aleatória 1:1000	0,003	0,035	0,069	0,092	0,155	0,100	0,100	0,050	0,604

temporal destes e é útil para a identificação de fluxos de tráfego segundo as aplicações que os originaram.

Para isso são estudados fluxos específicos, identificados como originados por determinados protocolos/aplicações. O objectivo é estudar o impacto do processo de amostragem sobre esta métrica, segundo os diferentes protocolos/aplicações. A *mean* IAT será então analisada tendo em conta a acurácia, ou seja, o erro em comparação com os mesmos fluxos do conjunto total de tráfego (sem amostragem).

Os fluxos que serão analisados, são descritos nas Tabelas 4.17 e 4.18 por ordem de volume de cada um dos *traces*. Cada linha representa os dados referentes apenas a um fluxo individual.

Tabela 4.17: *Volumes em pacotes e bytes de fluxos individuais - OC48*

	Tráfego Total		Multi Adaptativa		Random 1-100		Sistemática 100		Random 1-1000		Sistemática 1000	
Protocolo	pacotes	bytes	pacotes	bytes	pacotes	bytes	pacotes	bytes	pacotes	bytes	pacotes	bytes
http	25222	37841757	1375	2064760	244	366976	276	415104	34	51136	27	40608
ftp	13096	18939046	807	1170138	129	188082	132	190766	17	24494	9	13122
valisys	10778	13369134	556	700674	79	99826	111	139426	17	21599	10	12898
p2p	4802	7202235	250	375508	44	66176	44	66176	5	7520	3	4512
smtp	5574	5772018	176	179621	57	56238	47	49989	8	5758	6	6480
https	2027	3030920	226	334008	21	31584	20	28649	3	4512	2	3008
ms-streaming	2042	2355449	95	100730	16	18889	19	25165	2	1973	-	-
games	3120	2974400	179	167370	33	30654	38	35454	4	3588	3	2130

Tabela 4.18: *Volumes em pacotes e bytes de fluxos individuais - Sigcomm*

	Tráfego Total		Multi Adaptativa		Random 1-100		Sistemática 100		Random 1-1000		Sistemática 1000	
Protocolo	pacotes	bytes	pacotes	bytes	pacotes	bytes	pacotes	bytes	pacotes	bytes	pacotes	bytes
https	25927	40223088	1334	2082353	264	411290	267	413137	26	41126	24	36691
ipsec	16703	15450608	833	865376	168	157712	148	139728	16	12832	17	13664
http	8099	13535760	344	576416	91	152516	83	138612	10	16760	11	18436
imap	6575	10399236	396	620390	64	101032	63	101568	8	12290	5	8206
ssh	7540	8864376	390	481740	83	90592	79	93588	10	13864	11	14092
rtmp	4520	7500581	147	242773	43	69596	38	63688	3	5028	2	3352
netbios	12139	3083522	556	141224	111	28194	114	28956	12	3048	15	3810
mdns	1019	796147	39	26932	13	9309	5	3054	-	-	1	774

Segundo os dados apresentados nas Tabelas 4.17 e 4.18, há um fluxo em cada um dos conjuntos de dados que não foi identificado no processo de amostragem: no *trace* OC48, o fluxo mais relevante do protocolo *ms-streaming* não foi identificado pela técnica Sistemática 1:1000. Enquanto no *trace Sigcomm* o fluxo mais relevante do protocolo MDNS não é identificado apenas pela técnica 1:1000. A probabilidade de isto acontecer é significativa principalmente nos fluxos menores, uma vez que nas técnicas de frequência de amostragem 1:1000, a redução dos dados é de cerca de 99,9% do tráfego total original.

As Tabelas 4.19 e 4.20 mostram a percentagem de dados que resulta individualmente em cada fluxo, relativamente ao tráfego total, após redução resultante da aplicação de amostragem.

Tabela 4.19: *Percentagem de dados resultantes por fluxo - OC48*

	Tráfego Total		Multi Adaptativa		Random 1-100		Sistemática 100		Random 1-1000		Sistemática 1000	
Protocolo	pacotes	bytes	pacotes (%)	bytes (%)	pacotes (%)	bytes (%)	pacotes (%)	bytes (%)	pacotes (%)	bytes (%)	pacotes (%)	bytes (%)
http	25222	37841757	5,45	5,46	0,97	0,97	1,09	1,10	0,13	0,14	0,11	0,11
ftp	13096	18939046	6,16	6,18	0,99	0,99	1,01	1,01	0,13	0,13	0,07	0,07
valisys	10778	13369134	5,16	5,24	0,73	0,75	1,03	1,04	0,16	0,16	0,09	0,10
p2p	4802	7202235	5,21	5,21	0,92	0,92	0,92	0,92	0,10	0,10	0,06	0,06
smtp	5574	5772018	3,16	3,11	1,02	0,97	0,84	0,87	0,14	0,10	0,11	0,11
https	2027	3030920	11,15	11,02	1,04	1,04	0,99	0,95	0,15	0,15	0,10	0,10
ms-streaming	2042	2355449	4,65	4,28	0,78	0,80	0,93	1,07	0,10	0,08	-	-
games	3120	2974400	5,74	5,63	1,06	1,03	1,22	1,19	-	-	0,10	0,07
	Média		5,83	5,77	0,94	0,93	1,00	1,02	0,13	0,12	0,09	0,09

Tabela 4.20: *Percentagem de dados resultantes por fluxo - Sigcomm*

	Tráfego Total		Multi Adaptativa		Random 1-100		Sistemática 100		Random 1-1000		Sistemática 1000	
Protocolo	pacotes	bytes	pacotes (%)	bytes (%)	pacotes (%)	bytes (%)	pacotes (%)	bytes (%)	pacotes (%)	bytes (%)	pacotes (%)	bytes (%)
https	25927	40223088	5,15	5,18	1,02	1,02	1,03	1,03	0,10	0,10	0,09	0,09
ipsec	16703	15450608	4,99	5,60	1,01	1,02	0,89	0,90	0,10	0,08	0,10	0,09
http	8099	13535760	4,25	4,26	1,12	1,13	1,02	1,02	0,12	0,12	0,14	0,14
imap	6575	10399236	6,02	5,97	0,97	0,97	0,96	0,98	0,12	0,12	0,08	0,08
ssh	7540	8864376	5,17	5,43	1,10	1,02	1,05	1,06	0,13	0,16	0,15	0,16
rtmp	4520	7500581	3,25	3,24	0,95	0,93	0,84	0,85	0,07	0,07	0,04	0,04
netbios	12139	3083522	4,58	4,58	0,91	0,91	0,94	0,94	0,10	0,10	0,12	0,12
mdns	1019	796147	3,83	3,38	1,28	1,17	0,49	0,38	-	-	0,10	0,10
	Média		4,65	4,70	1,05	1,02	0,90	0,89	0,11	0,11	0,10	0,10

As percentagens de redução resultantes da aplicação das técnicas de amostragem Sistemática e Aleatória, são mais estáveis dada a sua fixa taxa de amostragem. Os

valores são sempre muito próximos de 1% para o caso da frequência de amostragem 1:100, e 0.1% para o caso 1:1000.

Já no caso da técnica Multi-adaptativa há maior variação dada a natureza dinâmica da taxa de amostragem, que depende do comportamento do tráfego. Ainda assim a redução de dados anda por volta dos 95%, como se pode verificar nas Tabelas 4.19 e 4.20. Estes valores e sua variação são importantes, uma vez que, para estimar os valores da *mean* IAT, são utilizadas a duração total dos fluxos e a percentagem de dados obtidos após amostragem.

Assim, a duração destes fluxos, ou seja, o intervalo de tempo em segundos entre o primeiro e último pacote de cada fluxo, é apresentada nas Tabelas 4.21 e 4.22. Cada uma das colunas referentes a cada técnica de amostragem, contém também o erro relativo à diferença entre a duração obtida por cada técnica e a duração dos mesmos fluxos no tráfego total. Os valores de erro médio são ilustrados no gráfico da Figura 4.19.

Tabela 4.21: *Percentagem de erro de duração por fluxo - OC48*

Protocolo	Tráfego Total	Multi Adaptativa	Erro (%)	Aleatória 1:100	Erro (%)	Sistemática 1:100	Erro (%)	Aleatória 1:1000	Erro (%)	Sistemática 1:1000	Erro (%)
http	119,59	113,40	5,18	116,40	2,67	115,53	3,40	114,88	3,94	94,18	21,25
ftp	119,58	116,19	2,84	117,11	2,07	119,37	0,18	103,02	13,85	80,70	32,52
valisys	119,67	116,15	2,94	113,31	5,32	119,28	0,33	105,92	11,49	108,77	9,11
p2p	110,51	106,98	3,20	107,39	2,83	109,21	1,18	87,60	20,73	65,21	41,00
smtp	31,53	30,00	4,86	25,74	18,36	25,06	20,52	15,85	49,74	21,89	30,57
https	12,66	7,03	44,45	3,58	71,68	7,38	41,69	1,37	89,18	1,09	91,36
ms-streaming	119,61	111,22	7,02	101,64	15,03	99,98	16,41	71,94	39,85	-	-
games	119,67	116,22	2,88	107,55	10,13	111,45	6,87	76,48	36,09	14,68	87,73
Erro médio (%)			9,17		16,01		11,32		33,11		44,79

Tabela 4.22: *Percentagem de erro de duração por fluxo - Sigcomm*

Protocolo	Tráfego Total	Multi Adaptativa	Erro (%)	Aleatória 1:100	Erro (%)	Sistemática 1:100	Erro (%)	Aleatória 1:1000	Erro (%)	Sistemática 1:1000	Erro (%)
https	5393,06	4684,61	13,14	2571,97	52,31	5373,83	0,36	1914,30	64,50	1727,72	67,96
ipsec	4906,32	4424,69	9,82	4356,67	11,20	4311,01	12,13	3654,31	25,52	4203,96	14,32
http	93,90	74,82	20,32	77,75	17,20	78,48	16,42	66,26	29,44	72,44	22,85
imap	2687,43	2681,10	0,24	2677,21	0,38	2682,96	0,17	2339,98	12,93	2357,62	12,27
ssh	4636,91	4205,63	9,30	2858,79	38,35	4578,05	1,27	505,30	89,10	377,66	91,86
rtmp	135,37	123,88	8,49	131,20	3,08	130,26	3,77	46,97	65,30	46,30	65,79
netbios	16125,23	16114,27	0,07	15830,92	1,83	15006,47	6,94	15611,06	3,19	13646,74	15,37
mdns	22021,61	21603,48	1,90	14932,64	32,19	9893,89	55,07	-	-	0,00	100,00
Erro médio (%)			7,91		19,57		12,02		41,43		48,80

Observando as Tabelas 4.21 e 4.22, verifica-se geralmente que para todas a técnicas de amostragem, os registos de fluxos com menor duração apresentam um erro significativamente maior comparativamente aos restantes fluxos, o que indica que para o cálculo de estimativas, quanto menor for a duração do fluxo, menor é a acurácia que se obtém, provavelmente devido à menor probablidade de ser amostrado.

Apesar das diferenças, em termos de redução de dados processados, impostas por cada técnica de amostragem, analisando os valores médios dos erros percentuais

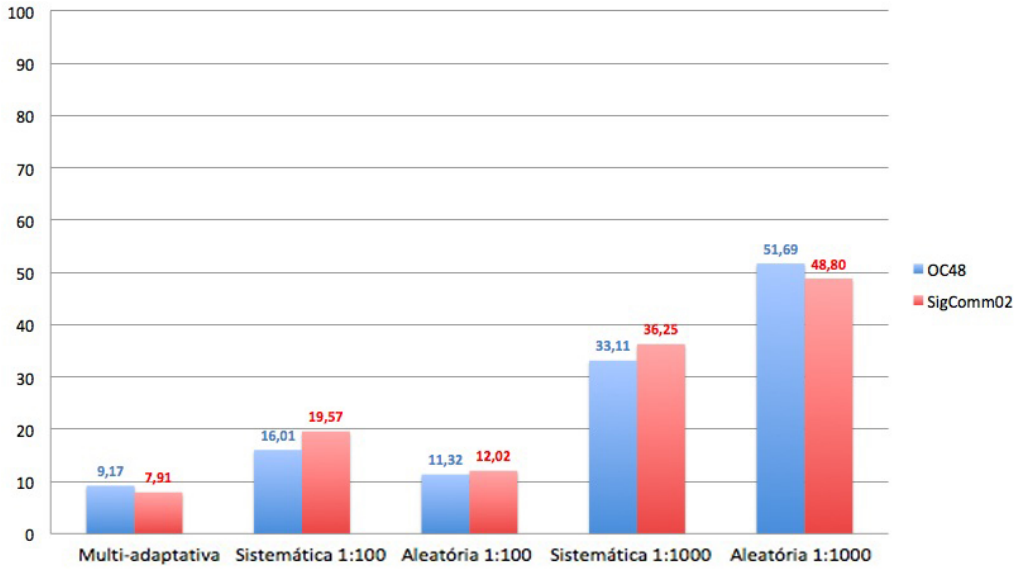

Figura 4.19: *Percentagens de erro de duração (média) por técnica de amostragem*

na Figura 4.19, não se verifica uma diferença significativa entre os resultados das técnicas Multi-Adaptativa e Sistemática/Aleatória com frequência de amostragem 1:100. Inclusivé vários fluxos relativos a estas técnicas (Sistemática e Aleatória), apresentam erros de duração muito baixos (próximos de 0%), como se verifica nas Tabelas 4.21 e 4.22.

A percentagem média dos erros obtidos em cada fluxo, segundo cada técnica de amostragem é representada nos gráficos das seguintes Figuras 4.20 e 4.21.

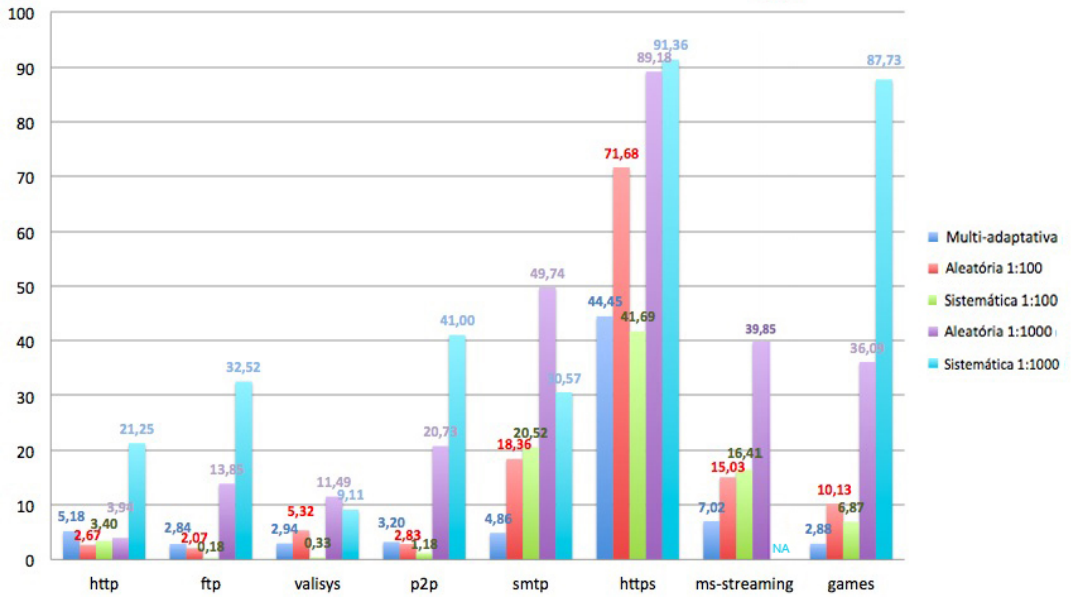

Figura 4.20: *Percentagem de erro de duração por fluxos individuais - OC48*

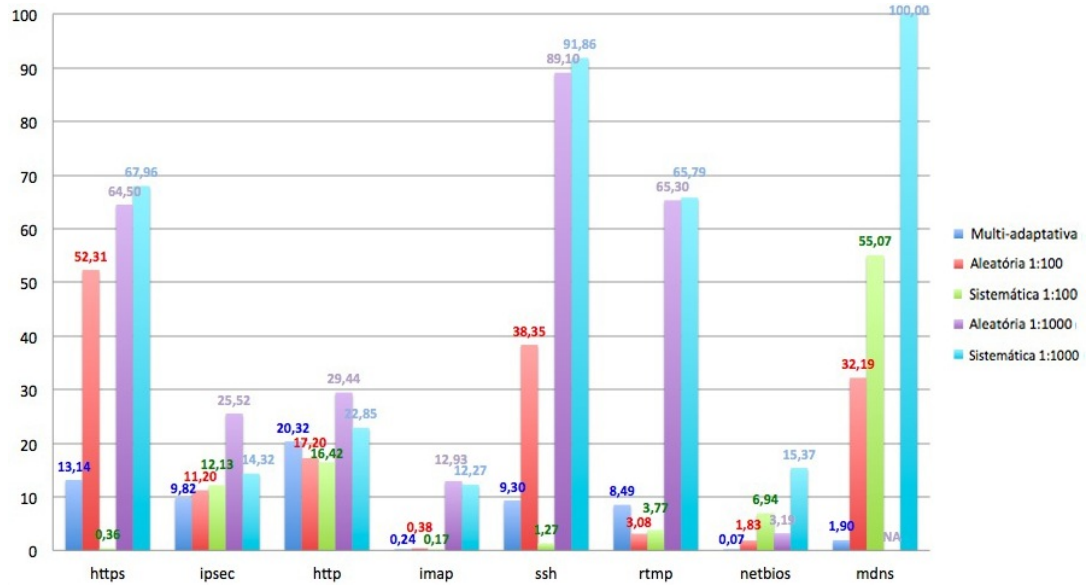

Figura 4.21: *Percentagem de erro de duração por fluxos individuais - Sigcomm*

Como expectável, o maior erro acontece para as técnicas com taxa de amostragem menor (1:1000). O erro de 100% que se pode verificar para o protocolo MDNS, significa que deste fluxo apenas 1 pacote foi seleccionado, sendo portanto impossível efectuar qualquer estimativa temporal.

Estimativas de *meam* IAT

Para efectuar estimativas da *mean* IAT é efectuada a divisão da duração do fluxos pelo número de pacotes menos 1 (nºde intervalos), como é ilustrado no exemplo da Figura 4.22.

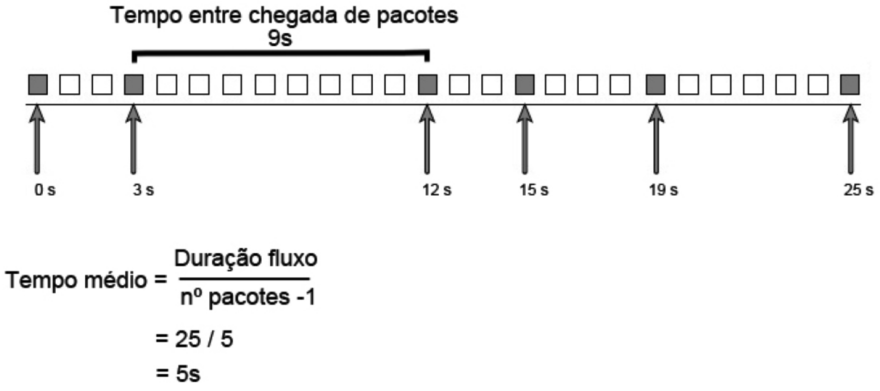

Figura 4.22: *Exemplo e fórmula de cálculo do tempo médio entre chegadas de pacotes - mean IAT*

Esta métrica é bastante influenciada durante a aplicação dos mecanismos de amostragem. Assim para minimizar esse efeito, da mesma forma que no trabalho

apresentado em [32], a *mean* IAT é dimensionada de acordo com a frequência de amostragem utilizada.

Assim, os valores da *mean* IAT obtidos, são multiplicados pela percentagem de redução de pacotes obtida através da amostragem. Por exemplo, para o *trace* OC48, o fluxo HTTP, obtido da técnica de amostragem Sistemática 1:100 tem a duração de 115,53 segundos (Tabela 4.21) referente à captura de 276 pacotes (Tabela 4.17). Efectuando o cálculo da mesma forma como na Figura 4.22, obtém-se uma *mean* IAT igual a 0.42s. Considerando o valor da percentagem de pacotes resultante igual a 1.09% segundo a Tabela 4.19, multiplicando o valor de *mean* IAT a este valor, obtém-se o valor estimado de 0.00458 (0.42*0.0109=0.00458), sendo o valor referente ao tráfego total igual a 0.0047.

As Tabelas 4.23 e 4.24 apresentam os valores obtidos através do calculo da estimativa da *mean* IAT, calculado da forma como o exemplo anterior. O erro percentual da estimativa do valor *mean* IAT em comparação com o valor *mean* IAT do tráfego total, é apresentado para cada fluxo, nas Tabelas 4.25 e 4.26.

Tabela 4.23: *Estimativas dos valores da mean IAT - OC48*

Protocolo	Tráfego Total	Multi-adaptativa	Sistemática 1:100	Aleatória 1:100	Sistemática 1:1000	Aleatória 1:1000
http	0,0047	0,0045	0,0046	0,0046	0,0039	0,0047
ftp	0,0091	0,0089	0,0092	0,0090	0,0069	0,0084
valisys	0,0111	0,0108	0,0112	0,0106	0,0112	0,0104
p2p	0,0230	0,0224	0,0233	0,0229	0,0204	0,0228
smtp	0,0057	0,0054	0,0046	0,0047	0,0047	0,0032
https	0,0062	0,0035	0,0038	0,0019	0,0011	0,0010
ms-streaming	0,0586	0,0550	0,0517	0,0531	-	0,0705
games	0,0384	0,0377	0,0367	0,0355	0,0071	0,0327

Tabela 4.24: *Estimativas dos valores da mean IAT - Sigcomm*

Protocolo	Tráfego Total	Multi-adaptativa	Sistemática 1:100	Aleatória 1:100	Sistemática 1:1000	Aleatória 1:1000
https	0,2080	0,1808	0,2080	0,0996	0,0695	0,0768
ipsec	0,2938	0,2652	0,2599	0,2624	0,2674	0,2334
http	0,0116	0,0093	0,0098	0,0097	0,0098	0,0091
imap	0,4088	0,4088	0,4146	0,4136	0,4482	0,4067
ssh	0,6151	0,5592	0,6150	0,3838	0,0551	0,0745
rtmp	0,0300	0,0276	0,0296	0,0297	0,0205	0,0156
netbios	1,3285	1,3299	1,2472	1,3160	1,2045	1,4029
mdns	21,6322	21,7586	12,1368	15,8754	0,0000	-

Tabela 4.25: *Erro de estimativas (percentagem) dos valores da mean IAT - OC48*

Protocolo	Multi-adaptativa	Sistemática 1:100	Aleatória 1:100	Sistemática 1:1000	Aleatória 1:1000
http	5,1180	3,0565	2,2804	18,2272	1,0352
ftp	2,7233	0,5756	1,3127	24,0858	8,4765
valisys	2,7699	0,5708	4,1082	0,9834	5,9621
p2p	2,8296	1,0958	0,5898	11,5130	0,9351
smtp	4,3332	18,8073	16,9172	16,6921	42,5706
https	44,2341	38,6554	70,2821	82,7243	83,7733
ms-streaming	6,0729	11,8132	9,4051	-	20,2388
games	1,6651	4,3800	7,3525	81,6063	14,8144
Erro médio	8,7183	9,8693	14,0310	33,6903	22,2257

Tabela 4.26: *Erro de estimativas (percentagem) dos valores da mean IAT - Sigcomm*

Protocolo	Multi-adaptativa	Sistemática 1:100	Aleatória 1:100	Sistemática 1:1000	Aleatória 1:1000
https	13,0748	0,0140	52,1304	66,5726	63,0861
ipsec	9,7134	11,5411	10,6764	8,9654	20,5575
http	20,0951	15,4117	16,2883	15,1467	21,6076
imap	0,0019	1,4285	1,1858	9,6431	0,5050
ssh	9,0800	0,0171	37,6035	91,0421	87,8934
rtmp	7,8784	1,1905	0,7908	31,6039	47,9595
netbios	0,1039	6,1221	0,9408	9,3327	5,6038
mdns	0,5841	43,8950	26,6123	100,0000	-
Erro médio	7,5664	9,9525	18,2785	41,5383	35,3161

As técnicas Sistemática e Aleatória com frequência de amostragem de 1:1000 apresentam na maioria dos protocolos classificados um erro significativo, cuja média do total dos protocolos é de cerca de 33.6% e 22.2% respectivamente, no caso do *trace* OC48, e 41.5% e 35.3% no caso do *trace Sigcomm*. Isto apesar de em casos pontuais, como HTTP e P2P em OC48, e IMAP em *Sigcomm*, a estimativa estar bastante próxima do tráfego original.

Assim, dada a média de erros obtido por estas técnicas com menor frequência de amostragem, pode considerar-se que estas não são as mais adequadas para estimação desta métrica, devido ao facto, mencionado anteriormente, de que a grande maioria dos fluxos classificados são obtidos de quantidades de dados muito reduzidos, face ao tráfego original (99.9% de redução). Daí as elevadas percentagens de erro observadas nas Tabelas anteriores (4.23, 4.24, 4.25 e 4.26).

Relativamente às mesmas técnicas mas com frequência de amostragem mais alta (1:100), comparando o erro médio obtido, verifica-se que a técnica Sistemática 1:100 obtém resultados com maior acurácia em relação à Aleatória 1:100, o que era de certa forma esperado, face às diferenças do erro de duração verificadas anteriormente nas Tabelas 4.21 e 4.22.

A diferença dos resultados apresentados pela técnica Multi-adaptativa e Sistemática 1:100 é ligeira, com uma pequena vantagem para a primeira. Pode concluir-se então que, fundamentalmente, estes erros estão relacionados mais directamente com o erro de estimação de duração dos fluxos, ilustrado anteriormente no gráfico da Figura 4.19.

4.3 Resumo

Neste capítulo foram expostos e discutidos os resultados obtidos através dos critérios descritos na secção 3.6. Os resultados mostram que o desempenho das técnicas de amostragem, comparadas entre si, podem produzir maior ou menor acurácia de estimativas do tráfego real, dependendo das características de tráfego em análise. No próximo capítulo serão apresentadas as conclusões finais e perspectivas de trabalho futuro.

Capítulo 5

Conclusões

Neste trabalho foi apresentado um estudo sobre o impacto dos processos de medição de tráfego baseados em amostragem, nas características estatísticas extraídas dos conjuntos de tráfego em análise, mais especificamente, tráfego organizado sob a forma de registos de fluxos.

Este estudo é motivado principalmente pela necessidade crescente que as ferramentas que suportam actividades relacionadas com a classificação e caracterização de tráfego têm em manter o seu funcionamento escalável, com os cada vez maiores volumes de dados que circulam nas redes de alto débito.

Desta forma, pretende-se apresentar dados que permitam efectuar análises, e obter conclusões, acerca dos níveis de acurácia que são possíveis de obter das estimativas efectuadas após aplicação de diferentes técnicas de amostragem. As características em estudo, são extraídas dos conjuntos de tráfego sob forma de fluxos e são comummente úteis no âmbito da classificação e caracterização de tráfego.

Os *traces* de tráfego utilizados para a validação do trabalho contêm a informação organizada ao nível dos pacotes IP, e são resultantes de capturas reais de tráfego em diferentes cenários, com *links* de características distintas. Estes *traces* são de acesso aberto ao público, o que pode ser útil a trabalhos futuros na área, ou extensões relacionadas com este trabalho, cujos resultados possam ser comparados aos aqui obtidos.

As técnicas de amostragem estudadas, são definidas em [36] e são comummente utilizadas em ferramentas bastante populares como o *NetFlow* desenvolvido pela *Cisco Systems*. As técnicas são designadas Sistemática e Aleatória, e são aplicadas segundo diferentes frequências de amostragem comummente utilizadas.

Outra técnica de amostragem utilizada - Multi-adaptativa - é uma técnica adaptativa baseada em predição linear, desenvolvida recentemente com uma alteração

inovadora relativamente às técnicas baseadas nos mesmos princípios. A alteração consiste em dotar o seu carácter adaptativo segundo dois parâmetros - tamanho da amostra e intervalo entre amostras. Isto permite à técnica funcionar com níveis de *overhead* mais baixos sem perder acurácia na obtenção de estimativas [8].

As características estatísticas extraídas dos conjuntos de tráfego, são relativas quer a fluxos de tráfego agregado, quer a fluxos individuais. Para isso, foi desenvolvida uma ferramenta que permite seleccionar e submeter conjuntos de tráfego, da qual resultam os subconjuntos de dados já organizados sob forma de fluxos, consoante a técnica de amostragem escolhida e parâmetros definidos.

Na análise dos resultados, são apresentadas as características estatísticas referentes ao tráfego total, dispondo os dados de forma a possibilitar efectuar a comparação entre estes e os obtidos por cada uma das técnicas de amostragem. O impacto é então analisado e é comparado o desempenho das técnicas de amostragem aplicadas ao tráfego total original quanto às características em análise.

As técnicas são analisadas comparativamente quanto à representatividade que cada uma obtém do tráfego total original, em termos de tamanho dos fluxos, proporções de volumes por protocolo de transporte e protocolo aplicacional, e também relativamente à conservação de propriedades relacionadas com a temporalidade dos fluxos como a *mean* IAT. Nesta última foram analisados os fluxos mais significativos de cada aplicação classificada.

As comparações efectuadas levam em conta o nível de redução dos dados processados durante o mecanismo de amostragem.

5.1 Síntese de resultados

Analisando a informação obtida, para a caracterização do tráfego baseada na informação extraída dos fluxos, quanto maior a quantidade de informação que cada fluxo conservar em si, como número de pacotes, duração, etc., mais acurácia terão as estimativas a efectuar.

É possível observar que dos conjuntos de dados obtidos após aplicação das técnicas de amostragem Sistemática e Aleatória, resulta em termos percentuais, maior quantidade de fluxos quando em comparação com a técnica Multi-adaptativa. No entanto, grande parte destes fluxos são bastante afectados pelo processo de amostragem, ou seja, grande quantidade de fluxos possui muito poucos dados, como 1 pacote ou pouco mais, o que dificulta a estimação de características estatísticas em grande parte dos fluxos resultantes.

Comparando as técnicas, pode verificar-se que as técnicas Sistemática e Aleatória são significativamente mais susceptíveis a este tipo de perda de informação que a técnica Multi-adaptativa.

Assim, com base nesta informação específica, pode concluir-se que para fins de classificação protocolar, principalmente baseada na informação de fluxos individuais, a técnica Multi-adaptativa prevê-se mais eficaz. Isto é reforçado também pelo facto de que a distribuição dos fluxos por volume de dados, quando em comparação com a distribuição do tráfego total apresenta maior similaridade, mesmo quando comparado com as técnicas Sistemática e Aleatória com alteração das frequências de amostragem para obtenção de quantidade de dados equivalente à obtido pela técnica Multi-adaptativa.

Já em termos das estimativas das proporções de tráfego, quer a nível de protocolo de transporte quer protocolo aplicional, quando se pretende obter informação acerca dos volumes de tráfego face ao tráfego total, todas as técnicas apresentam níveis de acurácia considerados elevados.

Em especial, as técnicas Sistemática e Aleatória, mesmo com frequências de amostragem muito baixas (1:1000) proporcionando uma redução do volume de dados de cerca de 99.9%, têm um desempenho assinalável.

A técnica Multi-adaptativa, surpreendentemente, foi a técnica cujas estimativas mais se afastam do tráfego total, ainda que ligeiramente. Esta informação permite concluir que para este tipo de estimativas, os níveis de redução de informação não têm um efeito significativo na acurácia.

Relativamente às estimativas relacionadas com a duração dos fluxos, mais precisamente os erros de duração, o senso comum diz-nos que a degradação das características possíveis de extrair é maior quanto maior for a redução de dados após amostragem.

Comparando os resultados das técnicas de amostragem, através da análise aos fluxos mais significativos identificados como pertencendo a cada aplicação/protocolo, a técnica Multi-adaptativa, sendo aquela que se baseia em maior quantidade de informação extraída do tráfego total, possui um erro de duração menor, à volta de 9%. No entanto se comparando com o erro obtido através da técnica sistemática 1:100, que possui uma redução de dados aproximadamente 5 vezes maior que a técnica Multi-adaptativa, não existe uma diferença significativa, o erro é cerca de 12%.

Adicionalmente, a técnica Multi-adaptativa, apesar de ser a que, em termos das estimativas relacionadas com a duração dos fluxos, apresenta resultados mais preci-

sos, tem contra si o facto de dificultar o cálculo das estimativas *mean* IAT devido ao seu comportamento dinâmico, sendo difícil prever com exactidão a percentagem de redução de dados, uma vez que esta está dependente das variações de comportamento do tráfego. Já o mesmo não acontece com as técnicas de amostragem que possuem frequências de amostragem fixa, em que nos casos estudados, a variação da percentagem de redução de dados é quase nula em torno dos valores de 99%, para as técnicas com frequência 1:100, e 99.9% para as técnicas com frequências 1:1000.

Outro facto que se observa quanto ao comportamento das técnicas de amostragem relativamente às técnicas Sistemática e Aleatória, é que, apesar da utilização da última ser mais recomendada para amostragem de pacotes [5] e por permitir evitar medições tendenciosas relacionadas com possíveis semelhanças entre o padrão de amostragem e o padrão da característica de interesse observada [37], praticamente não são detectadas diferenças assinaláveis entre os resultados apresentados pelas duas, excluindo a análise temporal, em que os erros de duração são significativamente menores na técnica Sistemática com frequência de amostragem 1:100 (16% e 11% para o *trace* OC48, e 19% e 12% para o *trace Sigcomm*), e maiores para o caso da frequência de amostragem 1:1000 (33% e 44% para o *trace* OC48, e 41% e 48% para o *trace Sigcomm*), para ambos os *traces* analisados (Tabelas 4.21 e 4.22).

Relativamente aos resultados obtidos através da técnica Multi-Adaptativa em comparação com os resultados obtidos através da aplicação das restantes técnicas, verifica-se que, apesar da aplicação da técnica Multi-adaptativa gerar conjuntos de dados partindo da amostragem de maior quantidade de pacotes comparativamente com as restantes técnicas, esta não demostra uma melhoria significativa na acurácia das estimativas analisadas, quando em comparação com as técnicas Sistemática ou Aleatória de frequência de amostragem maior (1:100).

Relativamente a estas técnicas mas com frequências de amostragem muito baixa (1:1000), apesar de algumas estimativas apresentarem níveis de degradação elevados causada pela grande redução dos dados após amostragem, estas também se mostram úteis, mas para domínios de aplicação mais específicos, como mostram os resultados referentes à representatividade da classificação em termos das proporções dos volumes de tráfego.

5.2 Trabalho futuro

Este trabalho teve como fim contribuir para o estudo do impacto das técnicas de amostragem para fins de classificação e caracterização do tráfego.

Perspectivas de trabalho futuro visam possibilitar expandir este estudo através

de análises a quantidades mais vastas e diversificadas de tráfego, representativas de variados cenários deste, tais como, tráfego VoIP, Multimédia, Web e aplicações P2P, e também através do estudo de maior quantidade de métricas e critérios de comparação a utilizar para avaliação de técnicas de amostragem.

Este tipo de estudo será importante tendo em vista o desenvolvimento de uma ferramenta que permita obter dados concretos e fiáveis acerca da acurácia de determinada técnica de amostragem de tráfego, visando a classificação e caracterização deste através de métodos eficazes. Isto consoante o ajuste de vários parâmetros como, características de *links* de transmissão, diferentes categorias de tráfego, e especificidades das próprias técnicas de amostragem.

A abordagem do estudo relacionado com a temporalidade dos fluxos, presente na secção 4.2.4, é recente, mas o nível de acurácia de alguns resultados obtidos, suscitam um estudo mais específico sobre a obtenção de estimativas de diferentes métricas, considerando o nível de redução de dados que se obtém após aplicação das técnicas de amostragem.

Referências Bibliográficas

[1] Traffic Classification cisco wan and application optimization solution guide. http://www.cisco.com/en/US/docs/nsite/enterprise/wan/wan_optimization/chap05.html. Accessed: 2013-01-20. 2, 11

[2] David Moore, Ken Keys, Ryan Koga, Edouard Lagache, and K. C. Claffy. The coralreef software suite as a tool for system and network administrators. In *Proceedings of the 15th USENIX conference on System administration*, LISA '01, pages 133–144, Berkeley, CA, USA, 2001. USENIX Association. 2, 12

[3] Manuel Crotti, Maurizio Dusi, Francesco Gringoli, and Luca Salgarelli. Traffic classification through simple statistical fingerprinting. *SIGCOMM Comput. Commun. Rev.*, 37(1):5–16, January 2007. 2, 12, 14

[4] S.K. Thompson and G.A.F. Seber. *Adaptive sampling*. Wiley Series in Probability and Statistics. Wiley, 1996. 3

[5] Netflow services solutions guide. http://www.cisco.com/en/US/docs/ios/solutions_docs/netflow/nfwhite.html. Accessed: 2013-01-01. 3, 4, 58

[6] Edwin A. Hernandez, Matthew C. Chidester, and Alan D. George. Adaptive sampling for network management. *J. Netw. Syst. Manage.*, 9(4):409–434, December 2001. 3, 17, 23, 24

[7] Hongbo Wang, Yu Lin, Yuehui Jin, and Shiduan Cheng. Easily-implemented adaptive packet sampling for high speed networks flow measurement. In *Proceedings of the 6th international conference on Computational Science - Volume Part IV*, ICCS'06, pages 128–135, Berlin, Heidelberg, 2006. Springer-Verlag. 3

[8] João Marco C. Silva and Solange Rito Lima. Multiadaptive sampling for lightweight network measurements. In *Computer Communications and Networks (ICCCN), 2012 21st International Conference on*, pages 1 –7, 30 2012-aug. 2 2012. 3, 5, 17, 22, 23, 24, 56

[9] Paul Barford and David Plonka. Characteristics of network traffic flow anomalies. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, IMW '01, pages 69–73, New York, NY, USA, 2001. ACM. 4

[10] B. Claise. Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. RFC 5101 (Proposed Standard), January 2008. 4

[11] Rui Castro, Mark Coates, Gang Liang, Robert Nowak, and Bin Yu. Network tomography: recent developments. *Statistical Science*, 19:499–517, 2004. 10

[12] Nick Duffield. Sampling for passive internet measurement: A review. *Statistical Science*, 19:472–498, 2004. 11

[13] Les Cottrell. Passive vs. active monitoring. *http://www.slac.stanford.edu/comp/net/wan-mon/passive-vs-active.html*, 2001. 11

[14] Internet traffic classification. http://www.caida.org/research/traffic-analysis/classification-overview/. Accessed: 2013-02-02. 11

[15] Internet assigned numbers authority. http://www.iana.org/. Accessed: 2013-02-12. 12, 29

[16] Andrew W. Moore and Konstantina Papagiannaki. Toward the accurate identification of network applications. In *In PAM*, pages 41–54, 2005. 12

[17] Alessandro Finamore, Marco Mellia, Michela Meo, Maurizio M. Munafo, P. D. Torino, and Dario Rossi. Experiences of internet traffic monitoring with tstat. *IEEE Network*, pages 8–14, 2011. 12

[18] Martin Roesch. Snort - lightweight intrusion detection for networks. In *Proceedings of the 13th USENIX conference on System administration*, LISA '99, pages 229–238, Berkeley, CA, USA, 1999. USENIX Association. 12, 13

[19] Alok Madhukar and Carey Williamson. A longitudinal study of p2p traffic classification. In *Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation*, MASCOTS '06, pages 179–188, Washington, DC, USA, 2006. IEEE Computer Society. 12

[20] Sebastian Zander, Thuy Nguyen, and Grenville Armitage. Automated traffic classification and application identification using machine learning. In *Proceedings of the The IEEE Conference on Local Computer Networks 30th Anniversary*, LCN '05, pages 250–257, Washington, DC, USA, 2005. IEEE Computer Society. 12

[21] L7 filter - application layer packet classifier for linux. http://l7-filter.sourceforge.net/. Accessed: 2012-11-02. 13, 18

[22] Wireshark. http://www.wireshark.org/. Accessed: 2012-11-13. 13

[23] S. Cherry. The voip backlash. *IEEE Spectr.*, 42(10):61–63, October 2005. 13

[24] Niccolò Cascarano, Luigi Ciminiera, and Fulvio Risso. Improving cost and accuracy of dpi traffic classifiers. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 641–646, New York, NY, USA, 2010. ACM. 13

[25] Sailesh Kumar, Sarang Dharmapurikar, Fang Yu, Patrick Crowley, and Jonathan Turner. Algorithms to accelerate multiple regular expressions matching for deep packet inspection. *SIGCOMM Comput. Commun. Rev.*, 36(4):339–350, August 2006. 13

[26] Neelam Goyal, Justin Ormont, Randy Smith, Karthikeyan Sankaralingam, and Cristian Estan. Signature Matching in Network Processing Using SIMD/GPU Architectures. Technical Report TR1628, Department of Computer Sciences, The University of Wisconsin-Madison, Madison, WI, 2008. 13

[27] Jun-Sang Park, Sung-Ho Yoon, and Myung-Sup Kim. Software architecture for a lightweight payload signature-based traffic classification system. In *Proceedings of the Third international conference on Traffic monitoring and analysis*, TMA'11, pages 136–149, Berlin, Heidelberg, 2011. Springer-Verlag. 13

[28] Tingwen Liu, Yong Sun, and Li Guo. Fast and memory-efficient traffic classification with deep packet inspection in cmp architecture. In *Proceedings of the 2010 IEEE Fifth International Conference on Networking, Architecture, and Storage*, NAS '10, pages 208–217, Washington, DC, USA, 2010. IEEE Computer Society. 13

[29] Thomas Karagiannis, Andre Broido, Michalis Faloutsos, and Kc claffy. Transport layer identification of p2p traffic. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, IMC '04, pages 121–134, New York, NY, USA, 2004. ACM. 13

[30] Thomas Karagiannis, Konstantina Papagiannaki, and Michalis Faloutsos. Blinc: multilevel traffic classification in the dark. In *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '05, pages 229–240, New York, NY, USA, 2005. ACM. 13, 27

[31] T. T.T. Nguyen and G. Armitage. A survey of techniques for internet traffic classification using machine learning. *Commun. Surveys Tuts.*, 10(4):56–76, October 2008. 14

[32] P. M. Santiago del Rio, D. Corral, and Jose Luis Garcia-Dorado andJavier Aracil. On the impact of packet sampling on skype traffic classification. *High Performance Computing and Networking Universidad Autonoma de Madrid, Spain*, 2012. 14, 18, 52

[33] Yu Wang and Shun-Zheng Yu. Supervised learning real-time traffic classifiers. *JNW*, 4(7):622–629, 2009. 14

[34] J. Erman, A. Mahanti, and M. Arlitt. Qrp05-4: Internet traffic identification using machine learning. In *Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE*, pages 1–6, 2006. 14

[35] Andrew W. Moore and Denis Zuev. Internet traffic classification using bayesian analysis techniques. *SIGMETRICS Perform. Eval. Rev.*, 33(1):50–60, June 2005. 15

[36] T. Zseby, Fraunhofer Fokus, M. Molina, N. Duffield, S. Niccolini, and F. Raspall. Sampling and filtering techniques for ip packet selection", rfc 5475, 2009. 15, 16, 55

[37] A. Dogman, R. Saatchi, and S. Al-Khayatt. An adaptive statistical sampling technique for computer network traffic. In *Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th International Symposium on*, pages 479–483, 2010. 17, 23, 58

[38] R. Serral-Gracia, A. Cabellos-Aparicio, and J. Domingo-Pascual. Packet loss estimation using distributed adaptive sampling. In *Network Operations and Management Symposium Workshops, 2008. NOMS Workshops 2008. IEEE*, pages 124–131, 2008. 17

[39] Qingshan Jiang, R. Srinivasan, and D. Slonowsky. Measurement based traffic prediction using fuzzy logic. In *Electrical and Computer Engineering, 2002. IEEE CCECE 2002. Canadian Conference on*, volume 2, pages 834–840 vol.2, 2002. 17

[40] Yiyi Lu and Chen He. Resource allocation using adaptive linear prediction in wdm/tdm {EPONs}. *{AEU} - International Journal of Electronics and Communications*, 64(2):173 – 176, 2010. 17

[41] Yongtao Wei, Jinkuan Wang, and Cuirong Wang. A traffic prediction based bandwidth management algorithm of a future internet architecture. In *Proceedings of the 2010 Third International Conference on Intelligent Networks and Intelligent Systems*, ICINIS '10, pages 560–563, Washington, DC, USA, 2010. IEEE Computer Society. 17

[42] Joao Marco C. Silva and Solange Rito Lima. Optimizing network measurements through self-adaptive sampling. In *Proceedings of the 2012 IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems*, HPCC '12, pages 794–801, Washington, DC, USA, 2012. IEEE Computer Society. 17, 18

[43] Davide Tammaro, Silvio Valenti, Dario Rossi, and Antonio Pescapé. Exploiting packet-sampling measurements for traffic characterization and classification. *Int. J. Netw. Manag.*, 22(6):451–476, November 2012. 18, 27

[44] Valentin Carela-Espanol, Pere Barlet-Ros, Albert Cabellos-Aparicio, and Josep Sole-Pareta. Analysis of the impact of sampling on netflow traffic classification. *Comput. Netw.*, 55(5):1083–1099, April 2011. 18, 27

[45] P. M. Santiago del Rio, J. Ramos, Jose Luis Garcia-Dorado, Javier Aracil, Antonio Cuadra SÃąnchez, and Maria del Mar Cutanda-Rodriguez. On the processing time for detection of skype traffic. In *IWCMC'11*, pages 1784–1788, 2011. 18

[46] Nick Duffield, Carsten Lund, and Mikkel Thorup. Learn more, sample less: Control of volume and variance in network measurement. *IEEE TRANSACTIONS IN INFORMATION THEORY*, 51:1756–1775. 25

[47] Nick Duffield, Carsten Lund, and Mikkel Thorup. Flow sampling under hard resource constraints. *SIGMETRICS Perform. Eval. Rev.*, 32(1):85–96, June 2004. 25

[48] R. Jurga and M. Hulboj. Packet sampling for network monitoring. Technical report, 2007. 25

[49] Netflow performance analysis. http://www.cisco.com/en/US/technologies/tk543/tk812/technologies_white_paper0900aecd802a0eb9.html. Accessed: 2012-12-11. 26

[50] Anukool Lakhina, Mark Crovella, and Christiphe Diot. Characterization of network-wide anomalies in traffic flows. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, IMC '04, pages 201–206, New York, NY, USA, 2004. ACM. 27

[51] DongJin Lee, Brian Carpenter, and Nevil Brownlee. Media streaming observations: Trends in udp to tcp ratio. 2011. 28

[52] Analyzing udp usage in internet traffic. http://www.caida.org/research/traffic-analysis/tcpudpratio/. Accessed: 2013-04-03. 28

[53] Speedguide.net :: Broadband tweaks, tools and info. http://www.speedguide.net/. Accessed: 2013-03-16. 29

[54] Aaron Schulman, Dave Levin, and Neil Spring. CRAWDAD trace umd/sigcomm2008/pcap/ethernet (v. 2009-03-02). Downloaded from http://crawdad.cs.dartmouth.edu/umd/sigcomm2008/pcap/Ethernet, March 2009. 29

[55] Trace statistics for caida passive oc48 and oc192 traces. Downloaded from http://www.caida.org/data/passive/trace_stats/. 29

[56] Larry Peterson. Inter-as traffic patterns and their implications. In *in Proc. IEEE GLOBECOM*, 1999. 37

[57] Anja Feldmann, Albert Greenberg, Carsten Lund, Nick Reingold, Jennifer Rexford, and Fred True. Deriving traffic demands for operational ip networks: methodology and experience. *IEEE/ACM Trans. Netw.*, 9(3):265–280, June 2001. 37

[58] Mark E. Crovella and Azer Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Trans. Netw.*, 5(6):835–846, December 1997. 38