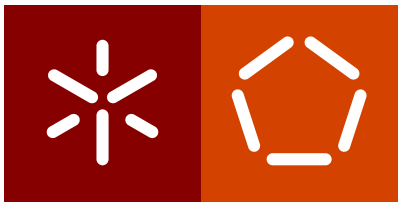
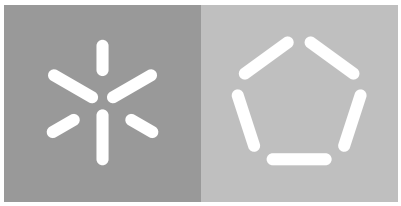


Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Marcelo Silva Cunha

Classificação e Monitorização Escalável de Serviços de Vídeo

Novembro de 2016



Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Marcelo Silva Cunha

Classificação e Monitorização Escalável de Serviços de Vídeo

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob orientação de
Solange Rito Lima
João Marco Silva

Novembro de 2016

AGRADECIMENTOS

O que torna a espécie humana tão única e, de certo modo, tão fascinante é a sua inigualável capacidade de raciocinar e de se organizar em comunidade, que nos possibilita a realização de feitos, por vezes, inimagináveis e a contínua conquista e inovação. Apesar de cada um de nós ser diferente à sua maneira, todos temos algo em comum, que é o facto de necessitarmos uns dos outros para atingirmos os nossos objetivos e a felicidade plena.

Tal como num sistema de redes existem interligações que permitem a transmissão eficaz de informação e o alcançar de um objetivo, também eu careci do apoio de outras pessoas para alcançar o meu sucesso, não perdendo o foco.

Deste modo, começo por agradecer à Professora Solange Rito Lima (orientadora) e ao Professor João Marco Silva (coorientador), pela permanente disponibilidade em partilhar dicas e ajudas, que foram essenciais para a realização deste trabalho com sucesso. Agradeço, também, o conhecimento e as vivências que partilharam comigo e fizeram de mim um indivíduo mais rico e conhecedor.

De seguida, agradeço, indubitavelmente, aos meus Pais pelo amor, carinho e apoio incondicional indispensáveis neste período tão difícil e determinante da minha vida. Eles foram um pilar essencial, fazendo-me ver sempre o caminho certo, apesar das ”pedras”que nele surgiram e que me tentavam derrubar.

Também, um agradecimento especial à Rita Conde pela sua constante disponibilidade em ajudar na realização deste trabalho, mesmo percebendo pouco do seu contexto, e pelo inesgotável apoio e motivação que me deu nos momentos mais difíceis que tive de enfrentar, que com as suas palavras mais carinhosas me fez nunca perder a rota do caminho da minha felicidade.

Sem esquecer, os Serviços de Comunicação da Universidade do Minho, em especial o Engenheiro Amândio pela sua disponibilidade e tempo dispensado no processo da coleta do tráfego.

Por fim, mas não menos importante, agradeço à minha restante família, amigos e colegas, que com pequenos gestos, me ajudaram a encarar o curso com mais tranquilidade e leveza, tendo levado ao meu sucesso académico.

ABSTRACT

Given the growth of video traffic volume in the Internet, the effective monitoring of video services is presented as a major challenge for the management of current and next-generation networks, where multiple services, protocols and access technologies coexist and compete for resources.

The efficient monitoring of network services involves, not only accurate measurement of parameters of interest, but also the lowest possible impact on the normal network operation. In this way, traffic sampling techniques aim to obtain information about all traffic, only considering a subset of packets traversing the network, presents itself as a scalable solution to face the challenges posed by monitoring video services. In this context, this master work has as main objective to analyze the performance of sampling-based network monitoring in the correct classification and characterization of video services.

RESUMO

Face ao crescimento do volume de tráfego de vídeo na *Internet*, a monitorização eficiente de serviços de vídeo apresenta-se como um desafio de grande importância para a gestão das redes atuais e de próxima geração, onde múltiplos serviços, protocolos e tecnologias de acesso coexistem e competem por recursos.

A monitorização eficiente de serviços envolvem a medição precisa de parâmetros de interesse e o menor impacto possível na operação normal da rede. Neste sentido as técnicas de amostragem de tráfego procuram obter informações sobre todo o tráfego considerando apenas um subconjunto dos pacotes em trânsito na rede, apresentando-se como uma solução escalável para os desafios impostos pela monitorização de serviços de vídeo. Neste contexto, o presente trabalho de mestrado tem como principal objetivo analisar o desempenho da monitorização baseada em amostragem de tráfego na correta classificação e caracterização de serviços de vídeo.

CONTEÚDO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÓNIMOS

MEI Mestrado em Engenharia Informática
UM Universidade do Minho
SCOM Serviços de Comunicação da Universidade do Minho
P2P *Peer-to-Peer*
QoE *Quality of Experience*
QoS *Quality of Service*
ISP *Internet Service Provider*
IP *Internet Protocol*
NAT *Network Address Translation*
SLA *Service Level Agreement*
TV Televisão
HTTP *Hypertext Transfer Protocol*
HTTPS *Hypertext Transfer Protocol Secure*
UDP *User Datagram Protocol*
TCP *Transport Control Protocol*
TLS *Transport Layer Security*
SSL *Secure Sockets Layer*
RTMP *Real Time Messaging Protocol*
TSTAT *TCP STatistic and Analysis Tool*
TIE *Traffic Identification Engine*
SystC *Systematic Count-based*
SystT *Systematic Time-based*
RandC *Random Count-based*
LP *Adaptive Liner Prediction*
MuST *Multiadaptive Sampling*
HH *Heavy-Hitters*
DPI *Deep Packet Inspection*
RTP *Real-time Transport Protocol*
RTCP *Real-time Transport Control Protocol*
SIP *Session Initiation Protocol*
DASH *Dynamic Adaptive Streaming over HTTP*
CDN *Content Distribution Networks*
OSI *Open Systems Interconnection*

NTSC *National Television System(s) Committee*
PAL *Phase Alternating Line*
IPTV *Internet Protocol Television*
CAIDA *Center for Applied Internet Data Analysis*
RDP *Remote Desktop Protocol*
VoD *Video on Demand*
C2S *Client to Server*
S2C *Server to Client*
DNS *Domain Name System*

1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

De acordo com os últimos resultados publicados pela *Cisco Visual Networking Index* [3], nos últimos 5 anos o tráfego de *Internet* global quintuplicou e avalia-se que nos próximos 5 anos deverá ainda aumentar 3 vezes mais. Com isso, estima-se que em 2016 ultrapassará 1 *zettabyte* ($1x10^{21}$ Bytes) de dados a circular na rede de *Internet* em todo o mundo. Para se atingir e aumentar estes números elevados, muito têm contribuído o tráfego proveniente de tecnologias móveis (como telemóveis, *smartphones*, *tablets*, computadores portáteis, etc.), tecnologias baseadas em *P2P* (*Peer-to-Peer*) ou tecnologias de VoD (*Video on Demand*), que tiveram nos últimos anos uma evolução e aumento de popularidade muito elevado, levando a que o aumento do tráfego proveniente de serviços de vídeo se tornasse muito difícil de monitorizar. "Serão necessários 5 milhões de anos para que uma pessoa possa visualizar toda a quantidade de vídeo que irá atravessar as redes IP (*Internet Protocol*) por mês, em 2019"[3], logo resume a importância que os serviços de vídeo obtiveram na sociedade e no mundo em geral, prevendo-se que em 2019 o tráfego de vídeo será entre 80-90% do tráfego global.

Video Streaming e os seus serviços de distribuição têm requisitos específicos no que toca às características da rede, especificamente em termos da latência, *jitter*, bem como o perfil de banda larga disponível para fluxos de dados individuais. Estes tipos de requisitos precisam de novos protocolos de transporte de dados orientados para o transporte de fluxos de dados de grandes dimensões. Estes tipos de protocolos estão em constante desenvolvimento e dependem dos conteúdos distribuídos, da capacidade de memória existente nos dispositivos, da localização dos dados mais próximos da ligação e do utilizador final, e da capacidade de transporte em redes de longa distância. Isto também coloca certas responsabilidades nos operadores de rede, pois agora não têm apenas que lidar com o rápido aumento do volume de tráfego nas suas redes, como também com novos requisitos de qualidade de serviço (QoS - *Quality of Service*). Os operadores precisam de mecanismos eficientes capazes de monitorizar fluxos de tráfego através das suas redes, permitindo controlar ligações individuais, diversificação de parâmetros de qualidade associados a fluxos de dados distin-

tos (por exemplo em vídeo, primeiramente latência e *jitter*), e o ajuste dinâmico de recursos para permitir uma melhor qualidade de experiência ao utilizador final (QoE - *Quality of Experience*). Portanto, o desafio advém do *overhead* existente nos métodos de monitorização, devido a não estarem adaptados para lidar com o súbito aumento da quantidade de fluxos de tráfego individuais. Daí que os métodos baseados em amostragem de tráfego, que criam estatisticamente uma imagem real da rede e do tráfego que a atravessa, são cada vez mais de uma importância crítica. Por isso, a adoção de mecanismos de monitorização de tráfego, conjugados com mecanismos de amostragem de tráfego estabelecem um meio de resolução do problema da monitorização de grandes volumes de dados.

1.2 Motivação e objetivos

A classificação e a caracterização do tráfego de rede são tarefas essenciais para o correto planeamento e gestão das atuais redes de comunicações. No entanto, face ao elevado volume de tráfego envolvido, essas tarefas podem beneficiar largamente do recurso a tráfego amostrado, desde que este permita obter uma visão da rede realista através de pequenas porções de tráfego.

A evolução tecnológica ao nível das redes IP levou a que a quantidade de dados dos serviços multimédia tenha aumentado de forma abrupta nos últimos tempos (por exemplo, o *streaming* de vídeo). Sendo esses dados, geralmente, pesados em termos de quantidade e de processamento, é primordial que se apliquem métodos e procedimentos de monitorização do tráfego proveniente de serviços de vídeo de forma a reduzir o *overhead* e o peso computacional na coleta e tratamento do tráfego sem comprometer a acurácia na estimação dos parâmetros de interesse. Embora existam alguns estudos que comparam o impacto da utilização de técnicas de amostragem em contextos diversos, a classificação e caracterização de serviços de vídeo com base em amostragem é um tema pouco focado, principalmente considerando a utilização de um leque variado de técnicas de amostragem clássicas e recentes na comparação.

Portanto, o principal objetivo na realização deste trabalho de mestrado consiste em analisar estratégias eficientes de monitorização de serviços de vídeo, através de técnicas de amostragem de tráfego, sendo para isso necessário ter em conta os seguintes aspetos:

- classificar os diferentes tipos de serviços de vídeo e identificar o conjunto de parâmetros de interesse para a sua monitorização;
- estudar as diferentes abordagens de amostragem de tráfego, suas características e implicações na monitorização de serviços;

- efetuar um estudo comparativo da aplicação das principais técnicas de amostragem de tráfego na monitorização de serviços de vídeo, considerando a relação entre a redução do uso de recursos e a acurácia na estimação dos parâmetros de interesse;
- propor eventuais métodos de ajuste de erros de estimação causados pela amostragem de tráfego.

Para se atingir os objetivos definidos anteriormente, devem ser utilizadas coletas de tráfego provenientes de ambientes ou cenários de redes reais.

1.3 Estrutura da Dissertação

Em adição a este capítulo introdutório, esta dissertação contém mais quatro capítulos essenciais para reflexão sobre o trabalho desenvolvido, face aos objetivos propostos anteriormente.

No Capítulo 2 é apresentado o estado da arte, isto é, são abordados os principais desenvolvimentos nas áreas presentes no domínio desta dissertação, expondo as principais ideias e trabalhos já realizados na área em estudo.

No Capítulo 3 são apresentados os diferentes componentes definidos com vista à implementação da metodologia de testes utilizada nesta dissertação. São também discutidas as diferentes fases de processamento das ferramentas utilizadas e as suas respetivas técnicas implementadas.

No Capítulo 4 são apresentados os resultados obtidos após a aplicação da metodologia de testes utilizada. Inicialmente, são apresentados resultados globais da classificação do tráfego de rede da UM (Universidade do Minho), e são ainda caracterizados os fluxos de vídeo encontrados nesse tráfego. Seguidamente, são apresentados e discutidos os resultados referentes à aplicação de diferentes técnicas de amostragem de tráfego.

Finalmente, no Capítulo 5 são apresentadas as conclusões deste trabalho de mestrado com uma avaliação e reflexão de acordo com os resultados obtidos e os objetivos inicialmente determinados. São também abordadas possíveis vertentes para trabalho futuro associado ao tema desenvolvido.

2

ESTADO DA ARTE

Neste capítulo serão apresentados alguns dos principais desenvolvimentos e estudos relacionados com o tema abordado nesta dissertação. Na pesquisa realizada foram levantados quatro temas principais capazes de contextualizar o leitor com o trabalho desenvolvido. Inicialmente, são apresentados os temas da classificação e amostragem de tráfego, bem como os seus diversos métodos que permitem realizar uma classificação e amostragem de tráfego diferenciada e diversificada. O estado da arte referente ao tema dos serviços de vídeo disponíveis atualmente é bastante variado, por isso, este tema é dividido em duas secções, em que a primeira aborda os desenvolvimentos nas tecnologias de serviços de *streaming* de vídeo através da *Internet*, e a segunda complementa o estudo dos serviços de vídeo, apresentando outras vertentes além da enunciada anteriormente. Numa última abordagem, são tratados diversos parâmetros de interesse que permitem estimar a QoS ou QoE de um determinado serviço de vídeo.

2.1 CLASSIFICAÇÃO DE TRÁFEGO

A área da medição e monitorização de tráfego de rede tem tido grandes avanços nos últimos anos, quer seja em desenvolvimento de novos métodos de monitorização ou em melhoramentos dos métodos mais tradicionais de forma a adaptá-los às novas necessidades dos serviços que geram tráfego de rede.

É neste ponto que surge a classificação de tráfego, uma importante tarefa associada à monitorização que permite identificar e diferenciar o tráfego de rede de acordo com vários atributos dos pacotes de comunicação (por exemplo, tipo de protocolo da camada de transporte, número de portas, endereços de IP, etc.). Contudo, devido ao desenvolvimento tecnológico das redes e dos protocolos de comunicação, a vertente da classificação de tráfego enfrenta atualmente vários desafios que tendem a dificultar uma correta e astuta classificação do tráfego. Alguns desses desafios são: o encapsulamento de aplicações em protocolos conotados com serviços diferentes; o aumento do número de aplicações que alocam dinamicamente portas de comunicação (é muito comum várias aplicações utilizarem a mesma porta de comunicação da camada de transporte); a utilização de protocolos de

segurança na cifragem dos dados, dificultando a interpretação dos mesmos. Todos os desafios enunciados são provocados acima de tudo pela cada vez maior diversidade e volume de dados das aplicações e serviços que fazem uso das atuais redes de comunicação.

Neste sentido, existem vários métodos de classificação de tráfego que são enumerados e brevemente explicados de seguida:

- **Classificação ao nível das portas de comunicação (*port-based*):** método de classificação mais habitualmente aplicado e popular. Consiste na análise da informação acerca das portas de origem/destino da camada de transporte. Com base nesse aspeto, este método faz a associação das portas de comunicação com determinadas aplicações. Isto considerando que cada aplicação opera com um protocolo específico, ao qual se encontra associado um número de porta. Atualmente, é a técnica de classificação mais rápida e que menos recursos consome, porém muito imprecisa devido aos problemas acima mencionados.

- **Classificação ao nível do *payload* (*payload-based*):** é um método mais avançado que o anterior, e baseia-se na análise do *payload* dos pacotes utilizando técnicas de DPI (*Deep Packet Inspection*), como *pattern matching* [4] e análise numérica [5]. De facto, é um método bastante eficiente na taxa de acerto, mas bastante pesado em termos computacionais e não aconselhável para tráfego encriptado, pois além de ser difícil de processar, poderá também implicar em problemas legais de privacidade de dados [6].

- **Classificação ao nível dos sistemas terminais (*host-behavior*):** os métodos de classificação de tráfego anteriores utilizam heurísticas assentes nos atributos dos pacotes de rede, já este método utiliza uma abordagem distinta, baseada no comportamento dos sistemas recetores dos fluxos de tráfego, ou seja, dos sistemas terminais [2]. Por exemplo, para determinar se um determinado sistema é um prestador de serviços, faz-se uma análise ao nível da observação da quantidade de fluxos que o sistema possuí. Assim, este género de classificação requer bastante tempo de processamento, por isso não é o método de classificação ideal para cenários em tempo real.

- **Classificação ao nível de fluxos (*flow-behavior*):** este método utiliza duas áreas distintas da computação, a área das Redes de Computadores e a área da Inteligência Artificial. Para caracterizar uma aplicação, assume-se que cada aplicação tem as suas próprias propriedades estatísticas, daí que se pode realizar a classificação analisando apenas características dos fluxos de tráfego. Contudo, a desvantagem da utilização deste método é a necessidade de computação específica, o que o torna pouco acessível [2].

Num ponto de vista comparativo, o método de classificação através das portas de comunicação permite uma classificação simples e rápida, utilizando poucos recursos computacionais, porém para aplicações que tenham a capacidade de alocar dinamicamente as portas de comunicação, tem uma taxa de fiabilidade muito baixa, colocando este método num patamar de fiabilidade abaixo dos restantes métodos classificativos. O método baseado em DPI apresenta problemas ao nível da classificação de tráfego cifrado, pois dificilmente terá acesso a chaves e assinaturas do processo de cifragem dos dados, para que dessa forma, não tenha problemas ao nível da privacidade dos dados. Para contornar esse problema existem os métodos de análise dos sistemas terminais e de análise de fluxos, contudo não são adequados para uma classificação em tempo real, devido aos seus elevados custos computacionais.

Para garantir a qualidade e o bom funcionamento dos serviços de rede, é necessário que as ferramentas de análise e classificação de tráfego implementem, de forma escalável e fiável, as técnicas de classificação de tráfego. Assim, a seleção do método de classificação a utilizar deverá ter em conta o cenário de rede em que este será aplicado.

Porém, devido ao crescimento do volume de tráfego é cada vez mais difícil fazer uma análise ao tráfego como um todo, por isso, a adoção de mecanismos de classificação de tráfego conjugados com mecanismos de amostragem de tráfego constituem um cenário cada vez mais útil e necessário.

2.2 Técnicas de amostragem de tráfego

Através da implementação da amostragem de tráfego, 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 [7].

Essa informação é proveniente de um processo de captura anteriormente realizada ou feita em tempo real, que pode ser analisada e armazenada em diferentes granularidades: ao nível do pacote e ao nível do fluxo. Uma monitorização ao nível do pacote permite obter uma análise com grande detalhe dos eventos que ocorreram na rede. No caso de uma monitorização ao nível do fluxo, o nível de detalhe da rede não é tão abrangente como ao nível do pacote, porém os dados são dispostos num formato mais leve e flexível que se adapta a cenários de redes com alto débito [8].

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

2.2.1 *Amostragem Sistemática*

A técnica de amostragem sistemática envolve a seleção individual de pacotes segundo uma função determinística. Essa seleção de pacotes pode ser *count-based* (seleção feita consoante uma função determinística baseada na posição dos pacotes), *time-based* (seleção feita consoante uma função determinística baseada no tempo de chegada dos pacotes ao ponto de medição), ou *event-based* (seleção feita aquando da ocorrência de determinados eventos nos conteúdos que atravessam a rede). A Figura 2.1 a) e a Figura 2.1 b) representam, respetivamente, a amostragem sistemática *count-based* [9] e a amostragem sistemática *time-based* [9]. No primeiro caso, cada quinto pacote é selecionado e capturado pelo processo de amostragem. No segundo caso, todos os pacotes que chegam ao ponto de medição ao longo de um período de 100 milissegundos são selecionados para uma amostra, considerando que todos os pacotes de entrada ao longo de 200 milissegundos são ignorados para fins de medição.

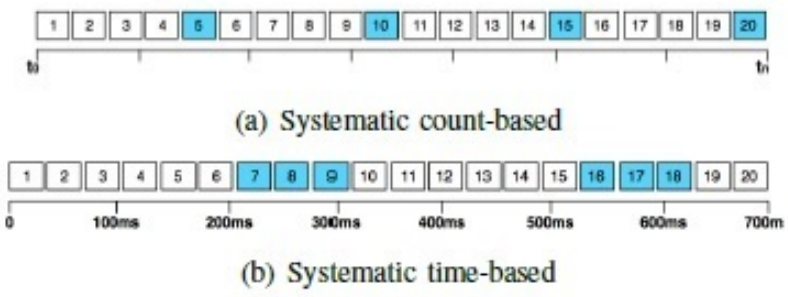


Figura 2.1.: Seleção de pacotes pelas técnicas *count-based* e *time-based* [1]

As técnicas sistemáticas acabam por ser bastante simples em termos de implementação, mas requerem um esforço computacional bastante grande ao nível da sua execução [10], especialmente quando é necessária a utilização de uma abordagem baseada em eventos específicos do tráfego de rede na qual os pacotes necessitam de ser processados para identificar o evento definido, que pode ser simplesmente um campo do cabeçalho desses pacotes.

2.2.2 *Amostragem Aleatória*

A amostragem aleatória tem uma abordagem bastante diferenciada em comparação com a amostragem sistemática, derivado de que a amostragem aleatória tem como objetivo reduzir o risco de amostras tendenciosas. Esta técnica implementa uma função aleatória a um determinado conjunto de pacotes, conseguindo assim tomar a decisão referente à seleção do pacote.

Dentro da amostragem aleatória existe diferentes técnicas implementadas, desde técnicas que usam probabilidades predefinidas [9], a técnicas que combinam a aleatoriadade com

a amostragem sistemática. Um desses casos é a técnica aleatória *n-Out-of-N*, que combina o intervalo fixo entre amostras da técnica de amostragem sistemática, com a amostragem aleatória ao selecionar n pacotes para amostra, dentro de cada intervalo N. Na Figura 2.2 pode-se visualizar como se processa a seleção de pacotes dessa técnica, neste caso está a ser selecionado aleatoriamente um pacote a cada 5 pacotes.

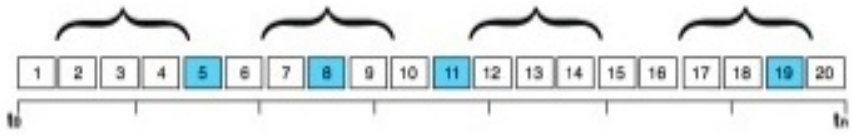


Figura 2.2.: Seleção de pacotes pela técnica *n-Out-of-N* [1]

Apesar de tudo, estas técnicas têm as suas limitações devido à aleatoriedade, por exemplo, não é possível garantir que o mesmo pacote seja capturado em distintas operações de monitorização. Por isso, é difícil garantir uma monitorização da rede fiável ao nível dos vários parâmetros de *QoS* como o *jitter*, *delay* ou perda de pacotes. Simplesmente, não existe garantia de que os mesmos pacotes serão capturados em dois pontos da rede distintos, isto porque, ambos têm uma forma aleatória de seleção de pacotes. Assim, este tipo de técnica é mais desejável para ambientes simples de monitorização.

2.2.3 *Amostragem Adaptativa*

A amostragem adaptativa foi desenvolvida com o objetivo de eliminar algumas limitações das técnicas apresentadas anteriormente, porém tirando proveito de alguns pontos dessas técnicas, as técnicas adaptativas permitem a observação otimizada de um parâmetro específico do tráfego de rede.

As técnicas de medição adaptativas geralmente são desenvolvidas para a estimação de um parâmetro específico, como o *delay*, *jitter* [10] ou a perda de pacotes [11]. Assim sendo, as atuais técnicas de amostragem adaptativa têm apresentado grandes evoluções na capacidade de representar o comportamento estatístico do tráfego de rede.

As propostas mais comuns baseiam-se em dois tipos, na lógica *Fuzzy* e em Predição Linear. As técnicas de amostragem baseadas na lógica *Fuzzy* [12], ajustam automaticamente a taxa de amostragem, consoante dados de situações semelhantes passadas, determinando a ação mais adequada para uma determinada condição de tráfego [13]. Como é percetível, este tipo de abordagem é bastante dispendioso computacionalmente ao nível da capacidade de memória, pois é necessário manter constantemente uma base de dados histórica para se calcular o ajuste da taxa de amostragem.

As técnicas baseadas em Predição Linear [14] [15] tentam prever o comportamento da rede ou dos padrões de tráfego com base num parâmetro observado em amostras recentes. Dessa forma, se uma previsão se confirma, a taxa de amostragem pode ser reduzida, enquanto que previsões incorretas indicam uma mudança no comportamento da rede e requerem um aumento na taxa de amostragem para determinar o novo padrão [13].

Estas técnicas de amostragem adaptativa são muito úteis e conseguem de forma segura e fiável representar o comportamento total do tráfego de rede. Porém, estas técnicas trazem problemas ao nível do *overhead* do volume de dados envolvido nos processos de medição.

2.2.4 *Amostragem Multi-Adaptativa*

A técnica de amostragem Multi-Adaptativa [16] é uma técnica adaptativa, em que o termo 'multi' está relacionado com a forma de adaptação da técnica. Utiliza predição linear para reduzir o *overhead* associado aos processos de monitorização mantendo a acurácia na estimação do comportamento da rede. Esta redução é obtida considerando o tamanho da amostra como um segundo fator adaptativo, além do intervalo entre amostras. Para isso, durante o processo de seleção é considerado o nível de atividade na rede, ou seja, caso o nível de atividade na rede aumente, a frequência de amostragem tende a aumentar também, mas para evitar uma sobrecarga no momento da medição, reduz-se o tamanho das amostras a coletar. Por outro lado, se o nível de atividade na rede diminuir a frequência de amostragem também diminui, mas aumenta o tamanho das amostras.

Com esta abordagem, é possível obter uma diminuição significativa no *overhead* associado ao volume de dados presentes no processo de monitorização. Mesmo com essa redução a técnica é capaz de manter a acurácia das estimativas [1].

2.3 *streaming* de vídeo

Em [17] é apresentado um estudo sobre os principais desenvolvimentos e protocolos desenvolvidos na área do *streaming* de vídeo nas duas últimas décadas. Existem três etapas fundamentais na história do *streaming* de vídeo, sendo: *Streaming* de vídeo pelo modelo Cliente-Servidor, *Streaming* de vídeo por *P2P* e *Streaming* de vídeo por *HTTP*.

2.3.1 *Streaming de vídeo pelo modelo Cliente-Servidor*

Nos primeiros anos de desenvolvimento de tecnologias de *streaming*, foram dados os primeiros passos na construção de protocolos de comunicação orientados, exclusivamente, para o transporte de grandes quantidades de dados. Daí a necessidade de serem desenvolvidas soluções orientadas exclusivamente para *streaming*, em que sejam considerados e

tratados de forma diferenciada alguns parâmetros importantes nos serviços de vídeo, como o *delay* ou *jitter*. Por conseguinte, foram desenvolvidos protocolos como o RTP (*Real-time Transport Protocol*) [18], especificamente para entrega de dados em tempo real, tais como o áudio e vídeo interativos. Tendo por base uma ideologia assente essencialmente no modelo cliente-servidor, inicialmente este protocolo era incorporado nas aplicações responsáveis por receber e reproduzir *streams* de vídeo enviados pelos servidores de *streaming* através da *Internet*.

O protocolo RTP usa o protocolo UDP (User Datagram Protocol) ao nível da camada de transporte e implementa dois tipos de canais no momento da comunicação, um canal para o envio dos dados, e um outro para o envio exclusivamente de mensagens de controlo do protocolo RTCP (Real-time Transport Control Protocol) [18], que são úteis para monitorizar estatísticas da rede para controlo da *QoS* ou para se conseguir sincronização quando se transporta múltiplas *streams*.

Neste ponto, as tecnologias de comunicação por *streaming* em modelos cliente-servidor obtiveram grandes avanços, por isso, o aparecimento de novas soluções foi uma consequência natural, como o desenvolvimento do protocolo SIP (*Session Initiation Protocol*) [19]. Este protocolo de sinalização foi concebido com o intuito de auxiliar as aplicações a lidar com sessões de múltiplos participantes, pois permite criar, modificar e terminar sessões com um ou mais participantes. Essas sessões incluem chamadas de telefone através da *Internet*, distribuição de dados multimédia e conferências multimédia.

2.3.2 *Streaming de vídeo por P2P*

No paradigma apresentado anteriormente, o consumo de recursos (por exemplo, largura de banda) do cliente e servidor são sempre dissociados, contudo num modelo *P2P streaming* cada nó é ao mesmo tempo um consumidor e fornecedor de serviços. A filosofia *P2P* tira vantagem do facto de cada *host* ou nó final participar num grupo *multicast* [17] para, dessa forma, contribuírem com as suas capacidades de *upload* para o envio dados que os outros nós ainda não tenham recebido. Assim, cada nó não está apenas a fazer *download* de *streams* de video, está também a fornecer a outros nós, que estejam a visualizar o mesmo conteúdo, estes *streams*. Esta filosofia não requer suporte das infraestruturas de rede, por isso, tem baixos custos de utilização e implementação.

Esta abordagem de *P2P* é ainda hoje muito usada por aplicações famosas de distribuição de todo o tipo de conteúdo, como o BitTorrent, porém uma nova geração de protocolos de *P2P video streaming* emergiu, como *CoolStreaming* [20]. Neste protocolo, os nós trocam, periodicamente, informações com os seus vizinhos sobre que segmentos de vídeo cada um tem nos seus *buffers*, e se algum faltar terá de ser explicitamente pedido e transferido de algum dos vizinhos que o tenha. Este tipo de abordagem é muito mais orientada para

as necessidades do mundo real, porém quando existe elevada diferença temporal entre o tempo de pedido de segmentos e a chegada desses segmentos ao nó final, pode originar casos de *delay* extremo.

2.3.3 *Streaming de vídeo por HTTP*

Apesar das soluções *P2P* terem provado ser bastante eficientes na transferência de dados de vídeo, para utilizadores regulares não são muito convenientes. Devido à necessidade de satisfazer os processos de *P2P* descritos anteriormente, os utilizadores terão que instalar aplicações externas e manter portas *TCP* (*Transport Control Protocol*) ou *UDP* abertas em *firewalls* e outros sistemas de proteção. Daí a necessidade de passar para o desenvolvimento de outros paradigmas de *streaming*.

A visualização de conteúdos em tempo real na *Internet* tornou-se uma prática corrente nos dias atuais com o desenvolvimento de *Web browsers* cada vez mais evoluídos. *Dynamic Adaptive Streaming over HTTP* (DASH) [21] e as *Content Distribution Networks* (CDN) tiveram grande impacto, para que o *HTTP streaming* se tornasse num dos modelos de visualização de vídeo mais utilizados nos dias de hoje.

O DASH utiliza *streaming* adaptativo, isto é, a qualidade do vídeo transmitido adapta-se, automaticamente, à capacidade da ligação de rede. As CDNs disponibilizam múltiplos servidores em várias localizações geográficas, distribuídas em diversos ISPs (*Internet Service Providers*), permitindo assim que os utilizadores usem *streams* de videos provenientes de servidores perto da sua localização. Portanto, o *HTTP video streaming* pode ser considerado como o *download* progressivo de segmentos de vídeo dos servidores através do protocolo *HTTP* (*Hypertext Transfer Protocol*), por isso, as aplicações que suportem *HTTP* podem procurar posições arbitrárias no fluxo do vídeo através da realização de solicitações de determinados intervalos de *bytes* ao servidor.

Com o aparecimento de serviços como *Netflix* e *YouTube*, que são cada vez mais populares e abrangem milhões de utilizadores todos os dias, fica demonstrado que este género de abordagens foram extensivamente adotadas pela indústria.

2.4 OUTROS SERVIÇOS DE VÍDEO

As aplicações que utilizam *video streaming* via HTTP serão os principais alvos de estudo neste trabalho, porém não se deve deixar de notar que não são as únicas responsáveis pelo tráfego de vídeo presente nas redes atuais. Daí que neste ponto serão introduzidos outros serviços de vídeo que utilizam, maioritariamente, redes IP para suportar a comunicação.

2.4.1 *Vídeo Conferência*

É um serviço que permite o contacto visual e sonoro entre indivíduos que se encontram em pontos ou locais distintos. Pode ser uma comunicação ponto a ponto ou multiponto em que abrange vários tipos de dados como texto, áudio e vídeo, num ambiente de comunicação em tempo real.

As necessidades de comunicação tornaram-se, nos dias de hoje, uma parte fundamental no dia a dia de qualquer pessoa ou organização. Por causa disso, as tecnologias de vídeo conferência tiveram um grande aumento de popularidade, o que levou a que as redes IP, que suportam o transporte entre vários pontos deste tipo de comunicações, viessem a passar por alguns problemas, tais como [22]:

- Endereçamento e conectividade: a espontânea escassez de endereços de rede provocou desafios ao nível do endereçamento e conectividade, levando a uma utilização generalizada de endereços privados e de NAT (*Network Address Translation*);
- Condições heterogéneas da rede: as condições de rede, atualmente, diferem de utilizador para utilizador. Devido à diferenciação de SLAs (*Service Level Agreements*) que os ISPs fazem de cliente para cliente e de contrato para contrato, as condições de banda larga ou de latência de comunicação, normalmente, são heterogéneas entre os utilizadores da comunicação por vídeo conferência. Por isso, é um grande desafio criar condições para reunir vários tipos de utilizadores dentro da mesma vídeo conferência;
- Requisitos de tempo real: como referido anteriormente, a comunicação por vídeo conferência é, maioritariamente, feita em tempo real, por isso, necessita de um rigoroso controlo do tempo de latência da comunicação. Em casos de períodos de latência elevados, a experiência do utilizador numa chamada por vídeo conferência pode se tornar impraticável. Este requisito de latência pode ser aplicado tanto à componente de voz/áudio como de vídeo.

O valor da comunicação e das formas de comunicação é incalculável e, hoje em dia, é tão natural utilizar-se esta tecnologia tanto a nível pessoal como empresarial que os serviços que disponibilizam estes tipos de tecnologias são inúmeros e variados. Esses serviços podem ser pagos ou gratuitos, consoante o género de utilização que o utilizador tenha com esses serviços.

No mesmo estudo [22], são investigados vários tipos de serviços que disponibilizam vídeo conferência. Mas a principal conclusão é que nenhum desses serviços consegue ter o maior proveito das redes dos utilizadores. Os serviços baseados em P2P são aqueles que conseguem obter uma melhor relação entre QoS e QoE do utilizador, por exemplo quando um determinado ponto não tem uma taxa de transferência que consiga suportar a comunicação entre todos os outros pontos finais, ele pode solicitar a outros pontos

próximos, sejam ou não pertencentes à mesma sessão de vídeo conferência, que auxiliem na função de entrega dos dados. Contudo, os problemas anteriormente enunciados não são totalmente resolvidos com esta abordagem tecnológica, porém os serviços P2P (por exemplo, *Skype*) estão a ser os de utilização mais comum nos dias de hoje e conseguem obter novos aperfeiçoamentos e estudos constantes.

2.4.2 *Câmara IP (Vídeo vigilância)*

Os sistemas de vídeo vigilância são sistemas muito importantes e eficazes na administração de segurança. Combinando as várias câmaras de segurança com tecnologias de vídeo em rede, estes sistemas podem comprimir e enviar imagens em tempo real, permitindo assim que remotamente se visualize as áreas de atividades protegidas que estão a ser alvo de vigilância. Assim, as câmaras IP tornaram-se num dispositivo chave nos sistemas de vídeo vigilância.

Existem dois tipos de câmaras IP, as centralizadas e as descentralizadas. Um sistema de vigilância de câmaras IP centralizado é uma rede de várias câmaras ligadas. Contudo, essas câmaras são apenas os olhos do sistema de segurança, porque existe um servidor de gravação, ao qual as câmaras estão ligadas por rede, que tem a responsabilidade de receber e guardar as imagens que as câmaras fornecem. Este tipo de sistemas asseguram que caso alguma câmara seja alvo de algum tipo de vandalismo, os dados por ela gravados estejam seguros e previamente guardados. Um sistema de vigilância de câmaras IP descentralizado é um sistema completamente auto suficiente. Não necessita de qualquer servidor ou outro tipo de sistema computacional auxiliar, devido a ser capaz de fazer a gravação em tempo real no próprio sistema da câmara IP, armazenando os dados localmente. Este tipo de sistemas asseguram que existem muito poucas hipóteses de se perder todos os dados da vigilância de uma vez só.

Uma arquitetura base para este tipo de sistemas é apresentada em [23], onde é utilizada uma câmara de filmar comum para a captura de imagens em tempo real. O sinal de vídeo capturado é transmitido pelos sistemas NTSC (*National Television System Committee*) ou PAL (*Phase Alternating Line*) através de serviços de *Internet*, *unicast* ou *multicast*. Ao nível da camada de transporte, os protocolos utilizados podem ser TCP ou UDP, dependendo do tipo de protocolo de nível superior utilizado, tais como HTTP ou RTP/RTCP. Assim, permite-se ao utilizador que visualize o vídeo capturado em *web browsers* ou através de programas de visualização de media.

2.4.3 *IPTV (Internet Protocol Television)*

O IPTV fornece serviços de televisão digital, através do protocolo de comunicação IP, para utilizadores pessoais ou comerciais.

O funcionamento do IPTV é bastante simples, em vez da receção dos programas de TV (Televisão) dos sinais analógicos/digitais tradicionais que utilizam antenas de satélite para receber os respetivos sinais, a receção é feita através de *streaming* (carregamento e visualização do vídeo em simultâneo) através da conexão com a *Internet*. Permite a entrega de áudio e vídeo com alta qualidade e depende de uma conexão de rede de 4 Mbps, pelo menos.

Apesar de utilizar funcionalidades *streaming*, o IPTV não pode ser considerado um serviço de *streaming*, pois ao contrário dos tipos de *streaming* apresentados anteriormente, em IPTV existe garantia da qualidade no momento da entrega e visualização dos dados (garantia de QoE do utilizador).

O modo de visualização dos conteúdos pode ser muito variado, dependendo do dispositivo final que irá permitir a visualização. Um computador ou um dispositivo móvel pode ser um desses dispositivos, porém é possível visualizar estes conteúdos numa TV convencional, mas para isso é necessário *hardware* adicional capaz de interpretar o sinal recebido através da rede e convertendo-o num formato compatível com a TV. Este tipo de *hardware* é denominado de *set-top box* ou *powerbox*.

Em [24], são abordados alguns dos principais serviços que o IPTV fornece, tais como a difusão do sinal em alta qualidade das televisões comerciais, *video on demand*, serviços *triple play* (pacotes onde os ISPs fornecem vídeo, voz e dados em simultâneo), Voz sobre IP (VoIP), acesso *Web/email*, etc.

Dos vários serviços apresentados, existem três que se destacam pela sua popularidade e elevada taxa de utilização, sendo eles:

- Video on Demand (VoD): é um serviço que permite visualizar áudio ou vídeo que seja concretamente selecionado pelo utilizador. Com isto, o utilizador pode visualizar filmes ou programas de TV que estejam disponíveis a uma determinada taxa monetária;

- *Time-shifted* IPTV: este serviço permite ao utilizador ver programas que já tenham sido difundidos anteriormente. Assim, não é estritamente necessário que o programa que inicia em determinado momento tenha de ser visto nesse exato momento, podendo ser visualizado posteriormente numa altura que seja mais conveniente para o utilizador.

- *Simulcast* IPTV: permite fazer envio dos mesmos sinais de vídeo e áudio através da *internet* para vários tipos de serviços ao mesmo tempo. Assim, o utilizador pode usufruir de vários serviços simultaneamente utilizando o mesmo dispositivo. Por

exemplo, pode visualizar um determinado evento num canal televisivo e ouvir o som de um canal de rádio ao mesmo tempo, tudo num mesmo computador ou *set-top box*.

Com todas estas aplicações, o IPTV tornou-se numa grande e emergente solução da *web* com potencial para aumentar de forma exponencial o acesso à *Internet* e, consequentemente, a quantidade de novo tráfego que os ISP's terão de controlar.

2.5 Parâmetros de serviços de vídeo

Por definição, o tráfego de rede é caracterizado por um determinado número de fluxos associados pela transferência de pacotes entre dois pontos finais, tipicamente separados geograficamente [25]. Contudo, existem vários parâmetros de interesse que podem estar relacionados com uma determinada transferência de ficheiros, por exemplo, não só a taxa de transferência é o único parâmetro de interesse numa sessão de transferência de ficheiros, também é importante perceber que existem determinados parâmetros que podem provocar alterações na entrega dos dados. *Delay*, *jitter*, perda de pacotes, tamanho da sessão, são alguns dos vários parâmetros que possibilitam analisar a qualidade de uma determinada sessão de transferência de ficheiros. De seguida serão apresentadas as principais métricas de tráfego de rede e o seu relacionamento com os serviços de vídeo.

2.5.1 *Taxa de transferência*

Este parâmetro descreve a taxa efetiva de dados que são transferidos na rede entre dois sistemas terminais [26]. A taxa de transferência pode também ser observada através do tipo de serviço (por exemplo, serviços de vídeo), através dos *streams* de dados (por exemplo, *download* de vídeo), ou até de uma sequência especifica do vídeo, assumindo-se que os dados não estão encriptados e é possível serem examinados os detalhes dos mesmos.

2.5.2 *Ocupação da ligação*

Este parâmetro descreve a ocupação de uma determinada ligação (*Link Ocupation*) de transferência de dados em determinada altura do tempo [25], isto é, indica a percentagem de toda a capacidade de ligação que está a ser ocupada, em determinado momento, por determinado tipo de serviço (por exemplo, serviços de vídeo), ou por um fluxo de dados (por exemplo, download de vídeo). Tipicamente, os serviços de vídeo têm limites para este parâmetro, sendo que a partir de determinada percentagem de utilização da ligação a *QoE* fica comprometida.

2.5.3 *Atraso*

Este parâmetro expressa o atraso observado na transmissão de pacotes entre o ponto inicial e o ponto de destino [27]. Isto é, indica a diferença de tempo entre o momento em que o primeiro *bit* do pacote sai do ponto inicial e o momento em que o último *bit* do mesmo pacote chega ao ponto de destino (*delay*). Efetivamente, este parâmetro descreve a latência de uma ligação de dados até que a troca de pacotes, entre a fonte e o destino para um determinado tipo de serviço, esteja concluída.

Por vezes, o atraso não é observado pelo utilizador final, em casos de sessões de VoD onde o conteúdo é enviado através do protocolo UDP e é feito um pré-armazenamento dos dados, isto para casos em que a ligação se mantenha estável durante toda a sessão. Todavia, é possível que se perceba um pequeno *delay* entre o momento em que o conteúdo é selecionado e o momento em que começa a visualização.

Porém, em casos de vídeo conferências o atraso é um parâmetro crítico, já que se houver um atraso excessivo são introduzidas quebras na conversação. Isto implica que os participantes tenham que esperar um certo tempo para que os pacotes de dados gerados pelos outros participantes na conversação cheguem até si.

2.5.4 *Variação do atraso*

A variação do atraso (*jitter*) descreve a variação do atraso observado para um determinado *stream* de dados ou entre dois pacotes consecutivos [28]. Isto é frequentemente causado pela variação da carga na ligação de dados ou pelo recalculo de novas rotas. O *jitter* pode ser causado também pela topologia da rede, ou seja, a rede pode permitir que existam múltiplos caminhos entre a fonte e o destino, cada um com diferentes taxas de *delay*. Portanto, uma solução desse género pode-se tornar indesejável para sistemas de distribuição de conteúdos de vídeo, pois tendencialmente haverá uma maior variação dos *delays* da ligação entre o servidor e o cliente.

2.5.5 *Perda de pacotes*

A perda de pacotes é traduzida pela ocorrência de um ou mais pacotes de dados, que navegam através de uma rede de comunicações não atingirem o seu destino [29]. Existem várias razões para a ocorrência deste tipo de problema, entre as quais a congestão da ligação de rede. Isto significa que quando a ligação de dados tem uma elevada congestão (ou que as condições físicas da ligação sejam desfavoráveis), pode ocorrer perda de pacotes levando a que a qualidade de serviço seja severamente comprometida.

Este parâmetro afeta maioritariamente serviços com base em transferências através do protocolo UDP. No caso dos serviços de vídeo, este problema é sentido quando existe uma perda de qualidade na visualização da imagem ou até quando o não é possível visualizar-se o vídeo, devido a uma perda total de pacotes. Isto provoca sérios problemas na QoE no momento de utilização do serviço de vídeo.

2.5.6 *Tamanho da sessão*

Este parâmetro descreve a duração total de uma sessão de vídeo, expresso em unidades de tempo [25]. O tamanho da sessão de vídeo é variável e não existem dados referente à média de duração de uma sessão de vídeo. Contudo, em prática, é possível observar que o tamanho da sessão afeta vários aspetos dos serviços de vídeo, tais como a perda de pacotes (quanto maior a sessão, maior a probabilidade de haver perda de pacotes) e *delay* (quanto maior a sessão, maior a probabilidade de haver congestionamento da rede).

2.5.7 *Relação entre parâmetros e serviços*

Qualquer um dos parâmetros abordados anteriormente tem um grau de importância diferente para cada um dos vários serviços de vídeo. Na Tabela 2.1 estão associados a três tipos de serviço de vídeo (vídeo conferência, VoD e *video streaming*), os parâmetros de interesse segundo uma ordem de importância. Essa ordem é de cima para baixo, correspondente ao mais importante para o menos importante [30].

No caso da vídeo conferência, sendo este um serviço bidirecional e com grande impacto em tempo real o *delay* e *jitter* tomam um importante papel em relação à perda de pacotes, pois numa vídeo conferência, para os utilizadores é preferível perder alguma qualidade de imagem, do que estar a ter quebras e falhas durante a ligação.

Numa experiência de não tempo real, como VoD, a importância do *delay* ou *jitter* deixa de ser tão relevante, devido ao facto de numa sessão de VoD, os utilizadores esperam que os conteúdos pretendidos cheguem com o máximo de qualidade possível no menor tempo possível. Portanto, a taxa de transferência, a perda de pacotes ou o tamanho da sessão têm maior relevo para este tipo de serviço.

De volta a serviços em tempo real, desta vez existe uma maior relevância para a perda de pacotes, devido aos utilizadores tipicamente esperarem *streams* de vídeo de boa qualidade e uma sincronização perfeita entre o som e a imagem. Por isso, a perda de pacotes ser mais importante que o *delay* do serviço, pois o utilizador prefere esperar que o vídeo demore algum tempo a ser carregado e visualizar esse vídeo de forma corrente, do que ter de esperar constantemente que o vídeo seja carregado para os *buffers* da sua aplicação.

Tabela 2.1.: Serviços de vídeo e parâmetros de tráfego

Serviços de Vídeo	Parâmetros de Tráfego
Vídeo conferência	Delay Jitter Perda de pacotes Taxa de transferência Ocupação da ligação Tamanho da sessão
VoD	Taxa de Transferência Perda de pacotes Tamanho da sessão Ocupação da ligação Delay Jitter
Video streaming	Jitter Perda de pacotes Delay Taxa de transferência Ocupação da ligação Tamanho da sessão

2.6 SUMÁRIO

Neste capítulo foram apresentados vários temas que são relevantes para a concretização deste trabalho de mestrado. Com a informação apresentada pretende-se facilitar a compreensão da solução proposta e apresentada nos capítulos seguintes. Dois dos temas abordados foram a classificação e amostragem do tráfego de rede, bem com os diferentes métodos e técnicas de classificação e amostragem.

De seguida, é abordado o tema do *Streaming* de vídeo onde é demonstrado um estudo sobre diferentes abordagens de *streaming*. Contudo, nem todos os serviços de vídeo que utilizam *internet* são baseados em *streaming*. Por isso, é importante apresentar algumas considerações sobre esses mesmos serviços.

Por fim, foram apresentados diferentes parâmetros de interesse de serviços de vídeo e a relação entre eles.

3

COLETA E PROCESSAMENTO DOS DADOS

Neste capítulo serão documentados os passos para a construção do sistema de testes que será a base de produção de dados, que após analisados, servirão de ponte para se atingir os objetivos definidos neste trabalho de mestrado.

Um dos principais pontos de arranque deste trabalho passa pela identificação de um meio ou método de classificação de tráfego que consiga fazer uma diferenciação clara entre o tráfego de vídeo e o restante tráfego de rede. Um cenário idealista seria realizar o estudo sobre uma rede que contivesse exclusivamente tráfego de vídeo a circular nos seus terminais, contudo seria um estudo completamente desfasado da realidade, porque redes com essa característica são incomuns no mundo real.

Sendo o passo da classificação de tráfego fundamental, não deixa de ser menos importante, identificar uma forma de auxiliar a monitorização das redes que suportam serviços deste tipo. Para isso, será abordado o tema da amostragem de tráfego, através de uma ferramenta identificada para o efeito.

Com esses dois conceitos, a classificação e amostragem de tráfego, a serem aplicados a tráfego de rede real, poderão ser obtidas conclusões que permitem atingir os objetivos propostos para este trabalho.

3.1 FERRAMENTAS DE CLASSIFICAÇÃO DE TRÁFEGO DE VÍDEO

Nesta secção serão apresentadas as ferramentas de classificação de tráfego utilizadas no âmbito deste trabalho.

3.1.1 *TSTAT (TCP STatistic and Analysis Tool)*

Um dos passos fundamentais na realização deste trabalho passa por encontrar uma forma de identificar o tráfego de vídeo. Para isso, recorreu-se à ferramenta TSTAT [31] que foi construída com o objetivo de caracterizar o comportamento de conexões TCP/IP. Porém, foi sendo evoluída e atualizada para uma ferramenta capaz de realizar uma monitorização

e análise de qualquer tipo de tráfego proveniente de qualquer tipo de rede IP. O TSTAT é capaz de interpretar *traces* de tráfego capturado em tempo real, utilizando o *hardware* da máquina em que a ferramenta se encontra a correr, ou pode analisar *traces* de tráfego capturado previamente, que estejam em determinados formatos de ficheiros. No âmbito deste trabalho, foi feita uma análise de *traces* capturados previamente que utilizam o formato suportado pelo *libpcap* (i.e., *.pcap*).

O resultado do processamento pelo TSTAT resulta num conjunto de ficheiros de texto *txt* (*logs*). Cada um desses ficheiros contém estatísticas importantes que caracterizam os fluxos de rede IP. Cada linha presente no ficheiro representa um fluxo e cada coluna representa uma variável que desse fluxo. Em [32] é possível encontrar uma lista dos vários *logs* retornados e quais as variáveis associadas.

Os ficheiros estão divididos por categorias, tais como *chat*, tcp, udp, vídeo, etc. Contudo, no âmbito deste trabalho, o ficheiro principal analisado é o ficheiro da categoria de vídeo, que indica quais os fluxos de vídeo foram identificados como tal.

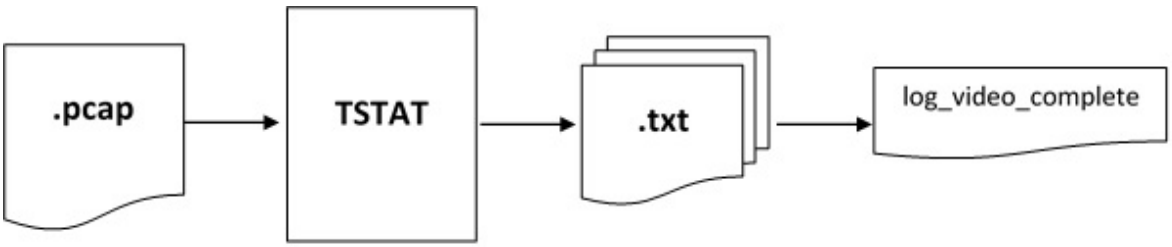


Figura 3.1.: Diagrama de processamento de fluxos pelo TSTAT

Como representado na Figura 3.1, o TSTAT começa por receber como *input* um ficheiro *.pcap* que contém os pacotes do tráfego de rede que foram capturados. Após esse passo, o TSTAT divide os pacotes por fluxos de tráfego e identifica esses fluxos dividindo-os em diferentes categorias (diferentes *logs*). Como referido anteriormente, para este trabalho é importante analisar os dados presentes no ficheiro *log_video_complete* que indica cada fluxo de vídeo que o TSTAT conseguiu detetar.

Em [33] é explicado como esse processo de deteção ou classificação é feito. Portanto, é dito que são registadas todas as conexões TCP, que são detetados como vídeo. Assim, na situação de estar perante conexões de vídeo RTMP (*Real Time Messaging Protocol*) ou conexões TLS (*Transport Layer Security*) associadas ao *Youtube*, a classificação é feita diretamente sobre o *handshake* dessas conexões. Porém, no caso mais comum de conexões via HTTP, o TSTAT tem duas abordagens distintas:

1. Correspondência do URL do pedido HTTP – O TSTAT classifica conexões de serviços mais conhecidos através da correspondência do URL do endereço desse serviço;

2. Operação do TSTAT de classificação de *Streaming* – este processo consiste em dois tipos de classificação diferentes:
    - pelo valor do *Content-Type* presente no cabeçalho do protocolo HTTP;
    - pela identificação da assinatura no *payload* dos dados que permite identificar o servidor dos dados de vídeo.

Portanto, com esta ferramenta consegue-se fazer a diferenciação entre o tráfego de vídeo e o restante tráfego de rede e realizar a sua classificação consoante diferentes parâmetros. Todos os parâmetros presentes no ficheiro de *log* referente a vídeo podem ser consultados em [32].

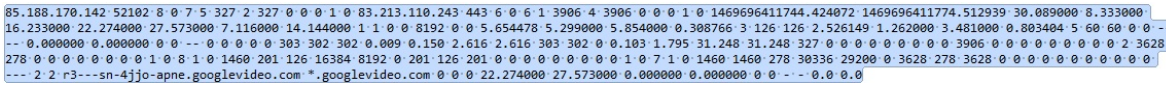


Figura 3.2.: Exemplo de um fluxo processado pelo TSTAT

Cada fluxo de vídeo identificado pelo TSTAT (ver exemplo na Figura 3.2) é caracterizado através de um total de 183 parâmetros, no entanto, apenas alguns destes são relevantes para a geração de resultados no âmbito deste trabalho. Os valores numéricos apresentados dentro de parêntesis que seguem o nome do parâmetro correspondem à entrada (coluna) da tabela dos *logs* do TSTAT [32]. Os parâmetros selecionados são:

- Número de pacotes processados: quantidade total de pacotes processados pelo TSTAT;
- Número de fluxos identificados: quantidade de fluxos total processados pelo TSTAT;
- Número de pacotes [C2S[1]](3): número de pacotes enviados pelo cliente para o servidor;
- Número de pacotes [S2C[2]](17): número de pacotes enviados pelo servidor para o cliente;
- *Server Name* (169): identifica o nome do servidor indicado pelo cliente durante a extensão das mensagens de *Hello* [34], no caso de fluxos TLS.

No caso dos dois primeiros parâmetros, o TSTAT fornece essa informação na *shell* de execução do comando, após a execução do mesmo.

Os dois parâmetros que indicam o número de pacotes, permitem fazer uma análise consoante a direção da comunicação entre o servidor e o cliente. Contudo, em determinados

1 *Client to Server*
2 *Server to Client*

pontos deste trabalho, a quantidade de pacotes será analisada como um fluxo bidirecional, ou seja, a soma entre estes dois parâmetros.

O parâmetro *server name* é processado no momento de *handshake* inicial do estabelecimento da comunicação entre o cliente e o servidor de vídeo. No estudo realizado em [2] é explicado o processo de extração da assinatura do servidor e é nesse momento que o TSTAT retira o valor do parâmetro *server name*, que permitirá identificar os prestadores de serviço, no âmbito deste trabalho. A Figura 3.3 apresenta as diferentes fases do *handshake* pelo protocolo TLS/SSL.

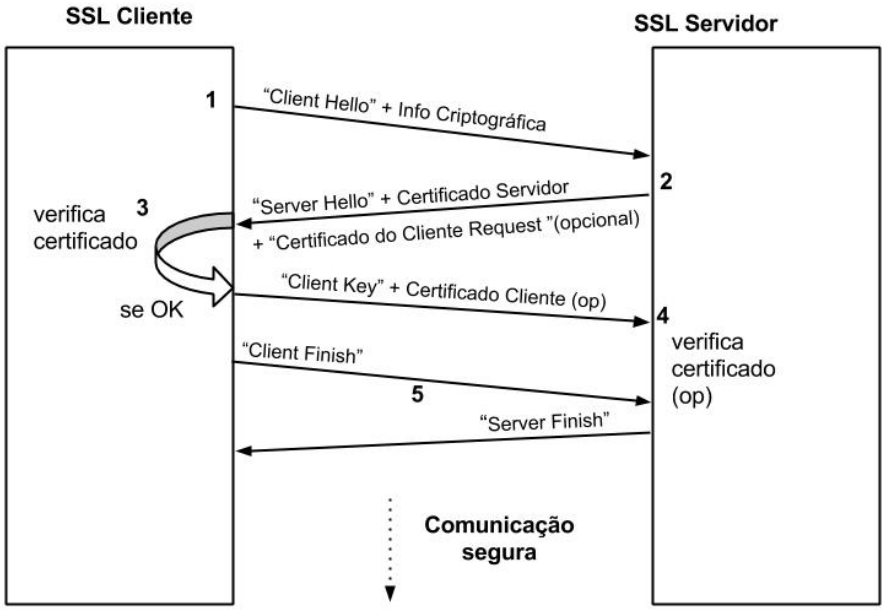


Figura 3.3.: Fases do estabelecimento de uma comunicação TLS/SSL [2]

Devido à cifragem dos dados requerida pelo protocolo SSL (*Secure Sockets Layer*), o processo de *handshake* contém várias fases de comunicação, em que o servidor e o cliente trocam mensagens com informações criptográficas sobre certificados de cliente e servidor. Numa das extensões dos certificados trocados está presente o parâmetro *server name*, mais concretamente na mensagem *client hello*, fase 1 da Figura 3.3. Nesta fase o cliente solicita o estabelecimento de uma comunicação segura enviando uma mensagem *client hello* conjun-

tamente com a informação criptográfica. Assim, neste ponto o TSTAT consegue retirar o valor do parâmetro *server name* nas extensões da mensagem.

3.1.2 *TIE (Traffic Identification Engine)*

O TIE [35] é um projeto de *software* aberto desenvolvido por uma equipa de investigadores da Universidade de Nápoles com o intuito de realizar classificação de tráfego de rede. Esta ferramenta fornece uma plataforma de fácil desenvolvimento e integração de técnicas de classificação de tráfego. Na Secção 2.1 foram apresentadas algumas dessas técnicas de classificação, implementadas e disponibilizadas pelo TIE, nomeadamente as técnicas de classificação ao nível das portas de comunicação e ao nível da análise do *payload* (DPI).

Esta ferramenta consegue correr dinamicamente vários *plugins*, que consistem na implementação das várias técnicas de classificação de tráfego (pode correr um ou mais *plugins*, simultaneamente). Permite também dois modos de execução distintos, o modo *online* e o modo *offline*. O modo *online* lê em tempo real os fluxos de tráfego que passam pela *interface* de rede do dispositivo, enquanto que o modo *offline* consiste na leitura e processamento de *traces* de tráfego de rede previamente coletados.

No artigo de apresentação da ferramenta [35], é apresentada a arquitetura da própria, que está ilustrada na Figura 3.4.

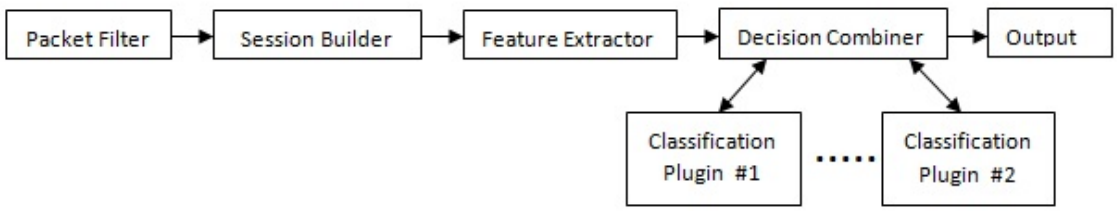


Figura 3.4.: Arquitetura da ferramenta TIE

Cada um dos componentes da Figura 3.4 tem as suas próprias funções. O *Packet Filter* tem a função de aplicar os modos de execução já referenciados, ou seja, faz a captura de tráfego em tempo real ou lê *traces* que foram capturados previamente. Após este *input* na ferramenta, os pacotes são agregados em sessões separadas do tipo *flow*, *biflow*, etc. Pelo *Session Builder*, consoante o *5-tuple* de cada fluxo (endereços IP de origem e destino, portas de origem e destino e tipo de protocolo em uso). O *Feature Extractor* realiza a extração dos campos dos pacotes que são necessários para a classificação. O tipo de classificação é feita pelo *Decision Combiner* que gere a execução dos diferentes *plugins* ativos. O *Output* gera o

resultado final num formato em que é possível verificar a classificação do tráfego de forma estatística, que permite a comparação em diferentes abordagens.

O *output* final é composto por vários parâmetros informativos dos fluxos identificados. Assim, cada fluxo é caracterizado por um total de 15 parâmetros, que são apresentados na Tabela 3.1.

Tabela 3.1.: TIE - Parâmetros resultantes do processamento

Parâmetro	Descrição
id	Identificador do fluxo
5-tuple	Endereços IP fonte/destino, Portas fonte/destino e protocolo
timestart	Momento de início da sessão
timeend	Momento de fim da sessão
pkt-up	Quantidade de pacotes enviados
pkt-dw	Quantidade de pacotes recebidos
bytes-up	Quantidade de dados enviados (em Bytes)
bytes-dw	Quantidade de dados recebidos (em Bytes)
app_id	ID da aplicação identificada
app_subid	SubId da aplicação identificada
confidence	Nível de confiança do processo de classificação

O TIE realiza uma classificação dos fluxos consoante o *5-tuple* dos fluxos, agrupando os fluxos que tenham essa característica idêntica. Depois, consoante os métodos de classificação de tráfego ativos, consegue realizar a classificação pelas heurísticas desses métodos e apresentar os valores dos parâmetros representados na Tabela 3.1. Dos parâmetros apresentados, os mais utilizados para análise neste trabalho são os ID's da aplicação identificada. Será maioritariamente sobre eles que as análises feitas neste trabalho se centrarão. Com esses parâmetros estima-se que se consiga realizar uma classificação de tráfego que possibilite fazer a diferenciação entre o tráfego de vídeo e o restante tráfego.

A Figura 3.5 apresenta um fluxograma que representa o modo de funcionamento do processo do TIE neste trabalho. Assim, a ferramenta começa por receber como *input* um *trace* de tráfego previamente coletado no formato *.pcap*, realizando o processamento e classificação desse tráfego conforme os *plugins* de classificação ativos. Os *plugins* de classificação abordados neste trabalho serão o *port-based* e *L7*, classificação ao nível das portas de comunicação e ao nível da análise do *payload* (DPI), respetivamente. Assim, serão feitas três tipos de classificação de tráfego distintas em que cada uma terá os seus próprios resultados presentes num ficheiro de texto *class.txt*. De reparar que o terceiro tipo de classificação é a

junção dos dois métodos de classificação de tráfego abordados, que o TIE permite realizar simultaneamente.

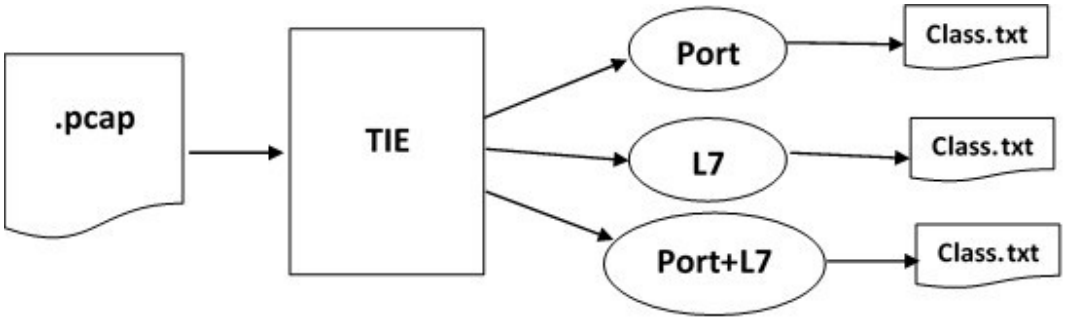


Figura 3.5.: Diagrama de processamento dos fluxos pelo TIE

3.2 FERRAMENTAS DE AMOSTRAGEM DE TRÁFEGO

Na Secção 2.2 desta dissertação foi introduzido o tema da amostragem de tráfego, apresentando-se as principais técnicas de amostragem de tráfego de rede existentes.

3.2.1 *Framework de amostragem de tráfego*

Consoante os objetivos propostos para este trabalho, a amostragem de tráfego e aplicação das técnicas de amostragem a tráfego de rede real são o principal foco de estudo neste trabalho, portanto foi necessário encontrar uma forma simples e prática de aplicar essas técnicas de amostragem ao tráfego de rede coletado. Para isso, foi selecionada uma *framework* de amostragem de tráfego de rede apresentada em [36].

O funcionamento desta *framework* é bastante simples, com uma *interface* bastante intuitiva (ver Figura 3.6), em que o utilizador seleciona a técnica de amostragem que pretende utilizar e aplica ao ficheiro alvo essa mesma técnica de amostragem. De seguida, será solicitada a frequência de amostragem que a técnica de amostragem aplicará aos pacotes de tráfego presentes no ficheiro selecionado.

O desenvolvimento desta *framework* está assente numa arquitetura de várias camadas, em que cada camada inferior fornece serviços à camada superior, escondendo os detalhes de funcionamento desses serviços. De forma a ser uma *framework* multiplataforma, foi implementada usando *libpcap* que fornece os métodos necessários de captura ou leitura dos pacotes de rede a processar [36].

Com esta *framework*, consegue-se atingir um dos principais objetivos do trabalho, que é aplicar várias técnicas de amostragem de tráfego a cenários de tráfego real e perceber o seu impacto em várias métricas de interesse. Com isso, é possível fazer a avaliação de

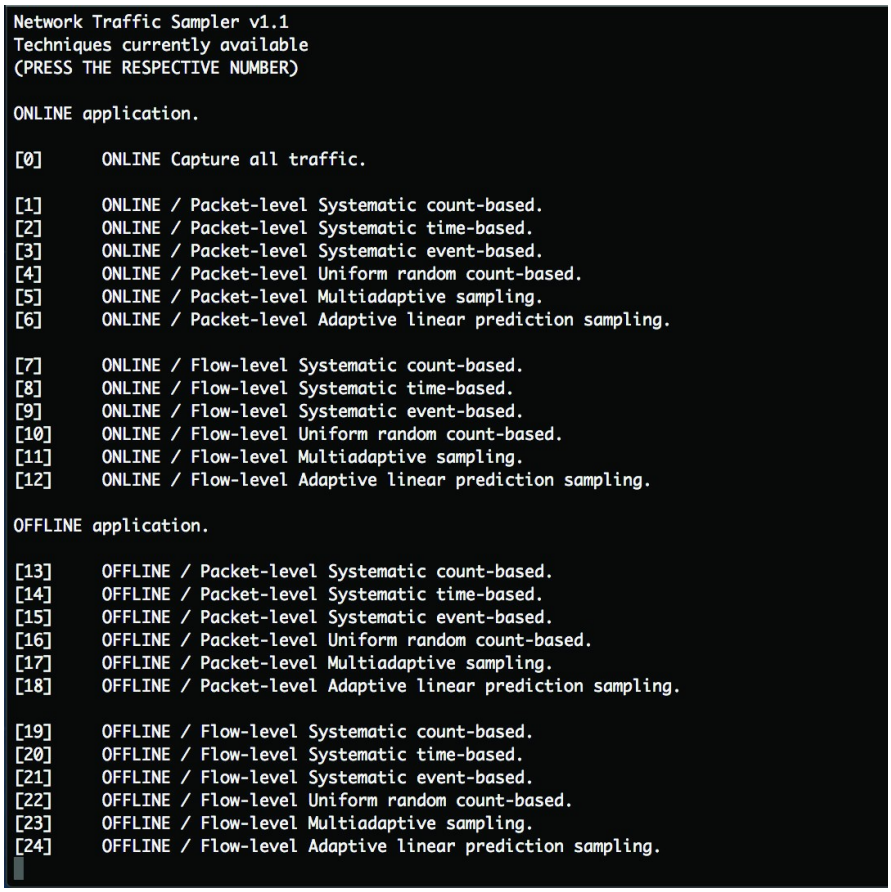


Figura 3.6.: *Interface* com as técnicas de amostragem disponíveis

diferentes técnicas de amostragem e comparar todos os resultados com o tráfego total, de forma a perceber quais as técnicas que são mais aconselháveis para amostragem de tráfego de vídeo.

3.2.2 *Técnicas de amostragem abordadas*

Neste ponto, é importante ressaltar que a utilização de diferentes técnicas e parâmetros de frequência levam a resultados distintos em diferentes atividades de monitorização. Identificar técnicas que melhor se aplicam ao contexto da classificação e caracterização de tráfego de vídeo é parte fundamental deste trabalho. Assim, é necessário criar um ambiente que permita comparar o desempenho de várias técnicas e suas frequências de amostragem. Tendo em conta que as diferentes formas de amostragem, podem correr a diferentes frequências de amostragem, tomou-se como base as técnicas e frequências avaliadas em [37].

Assim, serão abordadas três técnicas de amostragem mais populares presentes em vários estudos, tais como: *Systematic Count-based*, *Uniform Random Count-based* e *Systematic Time-based*. Em adição, serão analisadas duas técnicas mais complexas que as anteriores, conhecidas como técnicas adaptativas: *Adaptive Linear Prediction* e *Multiadaptive sampling*.

A apreciação dos resultados das técnicas de amostragem estará divida consoante o tipo de técnica. Inicialmente serão analisadas as técnicas *Systematic Count-based* no sentido de identificar o impacto que diferentes frequências de amostragem terão na estimação de parâmetros; posteriormente será aplicado o mesmo estudo do primeiro caso, porém desta vez a técnicas *Systematic Time-based*; por fim, será feita uma comparação global das diferentes técnicas de amostragem.

A forma de funcionamento destas técnicas foi apresentada na Secção 2.2, porém, de seguida, será efetuada uma breve explicação do funcionamento das diferentes técnicas consoante as frequências selecionadas.

- ***Systematic count-based* (*SystC*):** a seleção de pacotes é feita através de uma função determinística e invariável baseada na posição dos pacotes, usando contadores. Utilizando a frequência 1 em cada 100 pacotes (*SystC* 1/100), assumindo x como a posição do último pacote selecionado, o próximo pacote a ser selecionado ocorrerá na posição x+99. Esta configuração de frequência é a que será utilizada para comparação com outra técnicas de amostragem, como sugerido em [38]. As restantes frequências (ver Tabela 3.2) permitem uma comparação do impacto da frequência na acurácia da estimação.

  Será de esperar que, devido a esta técnica ter um custo de esforço computacional diretamente proporcional à frequência de amostragem utilizada, os resultados obtidos dependam também da frequência de amostragem.

- ***Systematic time-based* (*SystT*):** neste caso a seleção de pacotes segue uma função determinística baseada no tempo. Os parâmetros, que são inicialmente recebidos por esta técnica, não têm qualquer alteração durante todo o processo de amostragem, tal como na técnica *SystC*.

  Como a técnica de amostragem *SystC*, esta técnica é analisada em duas vertentes diferentes, comparando a técnica consoante diferentes frequências, de forma a perceber o impacto que a frequência terá na estimação, e com outras técnicas de amostragem (ver Tabela 3.2). Nessa última vertente os parâmetros definidos para essa comparação são 100 ms para o tamanho da amostra e 1000 ms para o intervalo de tempo entre amostras (*SystT* 100/1000), como sugerido em [37]. O seu funcionamento é baseado num determinado ponto de medição, em que todos os pacotes durante 100 ms são selecionados para a amostra e os pacotes seguintes durante 900 ms são descartados.

- ***Random count-based* (*RandC*):** esta técnica funciona à base de uma função aleatória, que seleciona um determinado número de pacotes dentro de um conjunto dos mesmos. Esta técnica será comparada com as restantes técnicas e os parâmetros que recebe são 1 pacote dentro de um conjunto de 100 pacotes (*RandC* 1/100), como proposto em [38].

- ***Adaptive linear prediction* (*LP*):** baseada na técnica *SystT*, esta técnica usa funções de predição linear para prever alterações na atividade da rede conseguindo ajustar a frequência de amostragem, enquanto o tamanho da amostra se mantém inalterado.

  Neste trabalho, esta técnica será alvo de comparação com outras técnicas de amostragem para se tentar perceber se ajustar a amostragem com o nível de atividade da rede (através do débito instantâneo), terá impacto significativo na estimação de parâmetros de interesse. As frequências de amostragem utilizadas foram 100 ms para o tamanho fixo da amostra e 200 ms para o intervalo de tempo inicial entre as amostras (*LP* 100/200) [13].

- ***Multiadaptive* (*MuST*):** tendo em conta o processo explicado na técnica anterior, a técnica *MuST* acrescenta a variação do tamanho da amostra, ou seja, numa fase de pico de tráfego esta técnica aumenta a frequência de amostragem, mas diminuí o tamanho da amostra. Em períodos de menor atividade da rede, é diminuída a frequência, mas aumentado o tamanho da amostra. Tudo isto, para impedir que em momentos de ação crítica não exista sobrecarga computacional, que possa levar a perdas de informação ou que o processo de monitorização não interfira com o funcionamento normal da rede. Para a comparação entre técnicas, o valor dos parâmetros usado foi de 200 ms para o tamanho da amostra inicial e 500 ms para o intervalo de tempo inicial entre amostras (*MuST* 200/500), como sugerido em [37].

A Tabela 3.2 apresenta um resumo das técnicas estudadas neste trabalho e suas respetivas frequências de amostragem.

De forma a aplicar e a relacionar os conceitos descritos anteriormente, foi utilizado no âmbito deste trabalho um processo de testes que permite aplicar diferentes técnicas de amostragem a cenários de tráfego real e assim permitir, posteriormente, comparar resultados entre o tráfego amostrado e o tráfego não amostrado. Esse processo está representado na Figura 3.7. Inicialmente, é processado um ficheiro *.pcap*, pela *framework* de amostragem de tráfego anteriormente referenciada na Secção 3.2.1, que irá aplicar as diferentes técnicas e, por conseguinte, resultará em vários ficheiros *.pcap* referentes ao tráfego amostrado consoante as técnicas e suas frequências de amostragem. Esses ficheiros serão, posteriormente, aplicados às ferramentas de classificação para efeito de comparação dos seus desempenhos quanto à identificação de fluxos de vídeo, como será explicado no capítulo referente à metodologia de testes.

Tabela 3.2.: Técnicas de amostragem avaliadas e suas frequências

Nome	Tipo de técnica	Frequência de amostragem
***SystC* 1/100**	*Systematic count-based*	1 a cada 100 pacotes
***SystC* 1/8**	*Systematic count-based*	1 a cada 8 pacotes
***SystC* 1/16**	*Systematic count-based*	1 a cada 16 pacotes
***SystC* 1/32**	*Systematic count-based*	1 a cada 32 pacotes
***SystC* 1/64**	*Systematic count-based*	1 a cada 64 pacotes
***SystC* 1/128**	*Systematic count-based*	1 a cada 128 pacotes
***SystC* 1/256**	*Systematic count-based*	1 a cada 256 pacotes
***SystT* 100/1000**	*Systematic time-based*	100ms a cada 1000ms
***SystT* 100/500**	*Systematic time-based*	100ms a cada 500ms
***SystT* 200/500**	*Systematic time-based*	200ms a cada 500ms
***SystT* 200/1000**	*Systematic time-based*	200ms a cada 1000ms
***SystT* 500/1500**	*Systematic time-based*	500ms a cada 1500ms
***SystT* 500/2500**	*Systematic time-based*	500ms a cada 2500ms
***SystT* 500/3500**	*Systematic time-based*	500ms a cada 3500ms
***RandC* 1/100**	*Uniform random count-based*	1 entre cada 100 pacotes
***LP* 100/200**	*Adaptive linear prediction*	Tamanho fixo da amostra: 100ms Intervalo inicial entre amostras: 200ms
***MuST* 200/500**	*Multiadaptive*	Tamanho inicial da amostra: 200ms Intervalo inicial entre amostras: 500ms

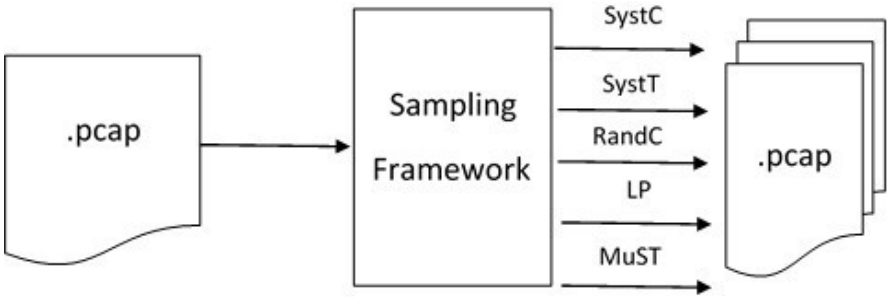


Figura 3.7.: Fluxograma do processo de amostragem de tráfego de rede

3.3 COLETA DE FLUXOS DE TRÁFEGO

De forma a testar as ferramentas de classificação de tráfego, anteriormente referenciada, inicialmente foram realizadas pequenas recolhas de tráfego que contivessem concretamente fluxos de tráfego de vídeo. Contudo, para se atingir os objetivos desta dissertação eram necessários *traces* mais abrangentes da realidade atual das redes de comunicações, algo que com uma coleta simples de tráfego não se consegue satisfazer.

Após uma breve pesquisa por *traces*, foram encontrados alguns *datasets* disponibilizados no *website* da CAIDA (*Center for Applied Internet Data Analysis*) [39], que são uma coleção de dados provenientes de diversos pontos geograficamente e topologicamente diferentes. Estes dados são coletados em diferentes períodos de tempo, sendo a coleta disponibilizada referente a um mês de coleta de 4 em 4 meses. Durante esse mês os dados coletados referem-se a 1 minuto de coleta de 5 em 5 minutos. Contudo, devido ao processo de remoção do *payload* dos pacotes tornou impossível a utilização desses *traces* no âmbito deste trabalho. Assim, foi decidido solicitar aos SCOM (Serviços de Comunicação da Universidade do Minho) uma recolha de tráfego mais refinada, do tráfego de rede de todo o campus da UM.

A recolha dos fluxos, efetuada pelos SCOM, foi realizada entre os dias 28 e 29 de julho de 2016, em diferentes horas do dia consoante o cenário de pico de utilização da rede nesses dias. Estes períodos ocorreram durante as 10, 11 e 15 horas desses dias. Devido à elevada quantidade de tráfego gerado durante esses períodos de pico, seria difícil e moroso processar traces com tão elevada quantidade de dados. Assim, a coleta é feita em espaços curtos de tempo, ou seja, durante a hora de pico são coletados 30 segundos de tráfego a cada 5 minutos. A Figura 3.8 demonstra os períodos de coleta de tráfego em cada um dos dias. Assinalado a verde são os períodos em que a coleta é feita, e no restante tempo até ao próximo ponto de coleta, não foram coletados quaisquer dados. A vermelho encontra-se os 30 segundos imediatamente a seguir ao período de coleta, que são o início do período de tráfego não coletado.

Da recolha de tráfego foram gerados vários ficheiros no formato *pcap*. Cada um dos ficheiros corresponde ao tráfego capturado em cada um dos períodos de coleta de tráfego. Devido à elevada quantidade de tráfego e tamanho dos ficheiros capturados, todo o processamento dos dados, no ambiente de testes, será feito ficheiro a ficheiro. Na Tabela 3.3 são apresentados os números referentes aos ficheiros recolhidos, sendo esses números a quantidade de ficheiros coletados e a soma do tamanho desses ficheiros, em cada hora. Assim, verifica-se que foram gerados um total de 72 ficheiros com 100,2 *GBytes*, resultantes da coleta de tráfego realizada.

A Figura 3.9 ajuda a perceber o estado de utilização da rede no momento em que houve a captura dos dados e a interpretar mais facilmente os dados da Tabela 3.3. No geral, verifica-

Figura 3.8.: Ilustração dos períodos de coleta de tráfego

Tabela 3.3.: Dados recolhidos por dia e horas

Dia	Horário (em horas)	Quantidade de ficheiros	Tamanho (Gbytes)
28 Jul	10	12	16,0
	11	12	17,1
	15	12	16,4
29 Jul	10	12	17,1
	11	12	17,8
	15	12	15,8
Total		72	100,2

se que o período das 11 horas obtém uma maior quantidade de dados coletados, em ambos os dias.

Contudo, é de notar que estes dados foram coletados fora do período letivo de atividades da UM, o que pode ter impacto nos resultados obtidos neste estudo, nomeadamente no volume de tráfego de vídeo identificado. Assim, posteriormente, tomou-se a decisão de se realizar uma nova coleta de dados, numa época em que o período letivo de atividades da UM decorra com normalidade. Assim, foi realizada uma coleta adicional de tráfego de rede no dia 10 de outubro de 2016, com exatamente as mesmas características da coleta anteriormente realizada.

Na Tabela 3.4 pode-se verificar os números referentes aos dados desta nova coleta, sendo que a quantidade de ficheiros recolhidos se manteve inalterada nos 12 ficheiros por cada

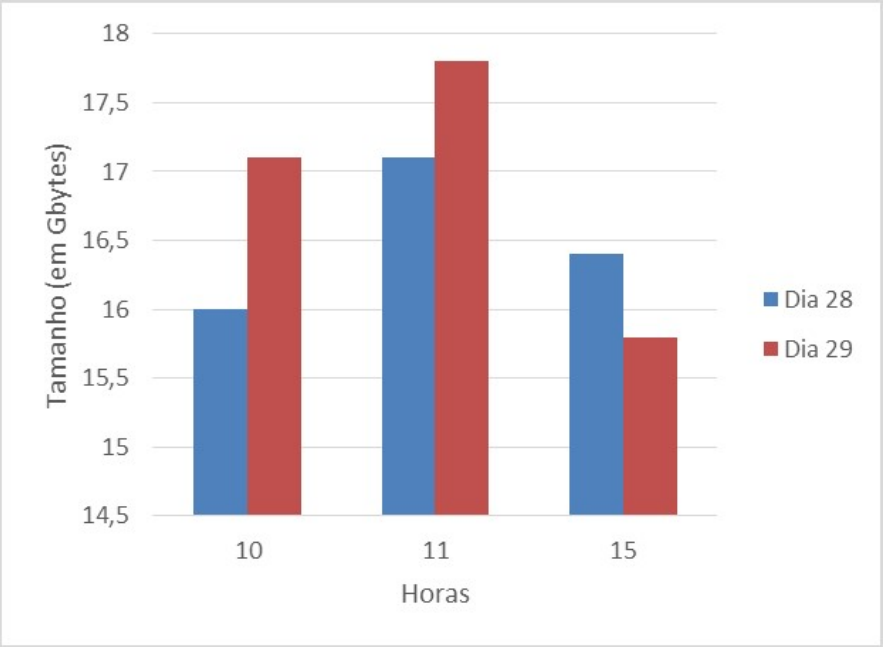


Figura 3.9.: Gráfico sobre a quantidade de dados por dia e hora

hora (provenientes da coleta de tráfego de 5 em 5 minutos), dando um total de 36 ficheiros de tráfego recolhidos. Por conseguinte, esta nova coleta originou 48,56 *GBytes* de dados.

Tabela 3.4.: Dados adicionais recolhidos por dia e horas

Dia	Horário (em horas)	Quantidade de ficheiros	Tamanho (Gbytes)
10 Out	10	12	17,35
	11	12	16,74
	15	12	14,47
Total		36	48,56

A Figura 3.10 permite fazer uma comparação mais intuitiva das várias recolhas. Como se pode verificar, não existe uma grande diferença entre as várias coletas realizadas. Contudo, a coleta de tráfego adicional obtém uma maior quantidade de tráfego no período das 10 horas, comparativamente com as outras coletas. Nas restantes horas teve resultados ligeiramente inferiores.

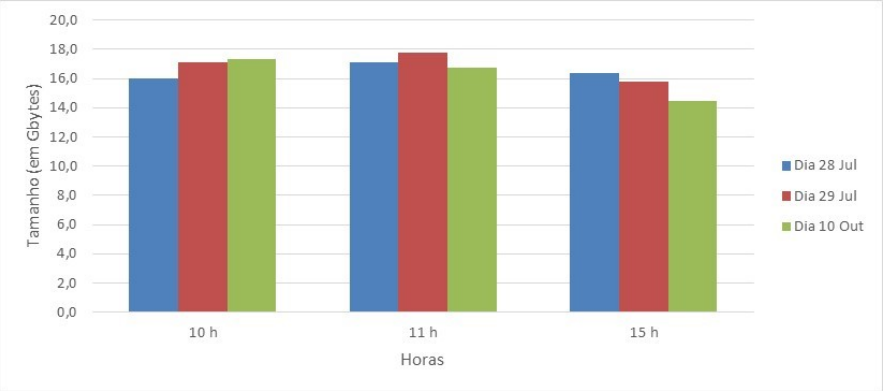


Figura 3.10.: Gráfico sobre a quantidade de dados por dia e hora - coleta adicional

3.4 METODOLOGIA DE TESTES E PARÂMETROS COMPARATIVOS

No contexto da monitorização escalável de serviços de vídeo e de forma a atingir os objetivos propostos para este trabalho de mestrado, a metodologia de testes consiste em recolher várias amostras de tráfego de rede, previamente coletado, recorrendo a várias técnicas de amostragem. Através de uma *framework* que aplica essas técnicas de amostragem, consegue-se obter o tráfego amostrado que será alvo de métodos de classificação, para assim ser possível avaliar, em que medida se pode extrapolar e generalizar a monitorização da rede no suporte a serviços de vídeo. Devido a isso, foi necessário produzir um ambiente de testes capaz de suportar os objetivos traçados.

Para isso, a metodologia foi agilizada com o desenvolvimento de vários *scripts* que permitem auxiliar a interpretação dos resultados obtidos. A sua utilização, bem como a disponibilização de todas as ferramentas e *scripts* utilizadas na metodologia de testes deste trabalho, encontram-se no Anexo A.

Durante este capítulo têm vindo a ser apresentados vários componentes que irão fazer parte do ambiente de testes deste trabalho. Neste ponto, irá ser explicado como se interligam todos estes componentes e como é feita a interpretação dos seus *outputs* de forma a obter-se os resultados finais que serão apresentados e analisados posteriormente nesta dissertação.

3.4.1 *Processamento com TSTAT*

Tendo em conta que se está a lidar com diferentes propósitos de processamento de dados e que são sequencialmente dependentes entre si, o ambiente de testes terá que ser dividido em várias fases. Portanto, na Figura 3.11 está representado um fluxograma composto por todos os componentes já referenciados e a respetiva interpretação final dos resultados.

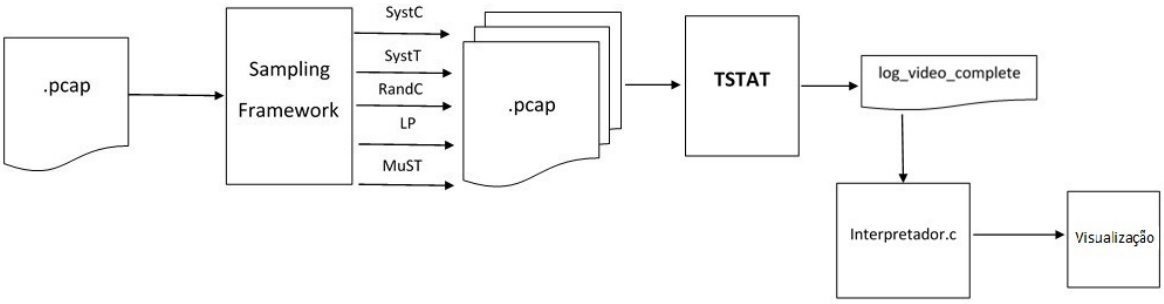


Figura 3.11.: Fluxograma do ambiente de testes pelo TSTAT

Inicialmente, é feito um processamento de amostragem do tráfego, como já explicado na Secção 3.2, em que através de uma ferramenta de aplicação de técnicas de amostragem a tráfego de rede, se obtém parcelas mais pequenas desse tráfego que serão alvo de análise e interpretação neste trabalho. Neste ponto é necessário salientar que esta fase da metodologia de testes não foi necessária para o processamento das estatísticas referentes ao tráfego total, sendo o processo iniciado na fase seguinte.

Numa segunda fase, os dados são processados ao nível da classificação de tráfego de vídeo, de forma a diferenciar o tráfego proveniente de serviços de vídeo do restante tráfego de rede. Para isso, foi utilizada uma ferramenta capaz de fazer essa diferenciação, neste caso o TSTAT, contudo todos os pormenores e detalhes desta fase de processamento são explicados na Secção 3.1.1.

Por último, na terceira fase de processamento dos dados, desenvolveu-se um programa utilizando a linguagem de programação C para auxiliar na interpretação e produção dos resultados. Esses resultados estão presentes no ficheiro *log_video_complete* que é dado como *output* do TSTAT. Este programa é capaz de, através do ficheiro de *output* do TSTAT, identificar a quantidade de fluxos processados, quantidade de pacotes de cada fluxo consoante a direção de comunicação, cálculo de fluxos *Heavy Hitters* (HH), etc.

De forma a tornar a interpretação e comparação dos resultados mais intuitiva, foi utilizado um programa de visualização estatística capaz de produzir gráficos estatísticos consoante os resultados interpretados pelo programa de interpretação de dados.

Parâmetros comparativos

Para se proceder à comparação das várias técnicas de amostragem e perceber qual o impacto das mesmas no tráfego de rede, foram considerados diferentes parâmetros de fluxos (parâmetros resultantes do processamento pelo TSTAT), nomeadamente: a quantidade de fluxos de vídeo identificada; a quantidade de pacotes presente nos fluxos de vídeo; a percentagem de fluxos de vídeo no tráfego total; a percentagem de pacotes de vídeo nas amostras; a percentagem de fluxos de vídeo nas amostras; a percentagem de fluxos *heavy-hitters*,

onde a noção de *heavy-hitter* se refere a 20% dos fluxos de vídeo identificados como mais significativos, em termos de da quantidade de pacotes de cada fluxo [37].

A identificação de fluxos de vídeo é a principal componente a ser analisada neste trabalho, contudo analisar as técnicas de amostragem, apenas pela quantidade de fluxos identificados não permite ter uma ideia real do tráfego de vídeo, porque os fluxos em si podem ser diferentes. Ou seja, basta um pacote de tráfego pertencente a um determinado fluxo de vídeo não ser selecionado para amostra, considera-se a classificação desse fluxo diferente da classificação feita no tráfego total. Por isso, é interessante estimar a similaridade entre os fluxos identificados no tráfego amostrado e os fluxos identificados no tráfego total, para perceber a capacidade do tráfego amostrado obter os mesmos dados do tráfego total, em termos de fluxos de tráfego.

Com a classificação de fluxos feita pelo TSTAT e utilizando o parâmetro *server name*, será estimada a capacidade de fornecer dados suficientes para a identificação dos servidores de vídeo dos fluxos identificados.

Na comparação de técnicas de amostragem, os parâmetros comparativos apresentados estarão sempre presentes nos resultados, podendo na presença de resultados pouco expectáveis se recorrer a outros parâmetros de análise.

3.4.2 *Processamento com TIE*

O ambiente de testes do processamento do tráfego pelo TIE é idêntico ao ambiente apresentado para o caso do TSTAT. Inicialmente, na fase do processo de amostragem, o processo é exatamente idêntico ao apresentado anteriormente. Posteriormente, os *traces* de tráfego amostrados são processados pelo TIE, num processo também já explicado anteriormente na Secção 3.1.2. A Figura 3.12 representa o fluxograma do ambiente de testes, com os processos referenciados interligados entre si.

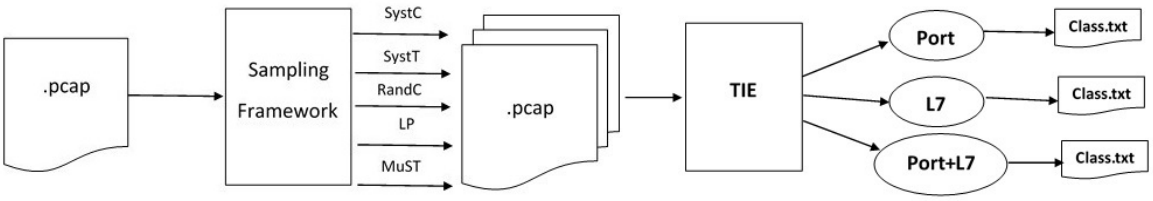


Figura 3.12.: Fluxograma do ambiente de testes pelo TIE

No fim deste processamento, é necessário analisar o *output* gerado. Para isso, foi desenvolvido um programa na linguagem de programação *Java*, capaz de interpretar e apresentar os resultados presentes no *output* do TIE. Esse programa, ao processar o ficheiro, retira para cada fluxo o valor do parâmetro *app_id*, para combinar com a lista de aplicações reconhecida

e fornecida pelo TIE. Com isso, o programa gera gráficos estatísticos com as percentagens de aplicações identificadas.

3.5 SUMÁRIO

Neste capítulo encontram-se documentados os passos e as decisões que foram tomadas para a realização com sucesso dos objetivos definidos para este projeto de mestrado. Em primeiro lugar houve a necessidade de identificar uma metodologia de classificação de tráfego de vídeo, ou seja, foi realizada uma pesquisa e análise de ferramentas de classificação de tráfego que se mostrassem potencialmente capazes de fazer a diferenciação de tráfego de vídeo com o restante tráfego.

Um segundo ponto passa pela explicação de um outro processo importante, a amostragem de tráfego. Foram abordadas questões como a ferramenta utilizada nesse processo, as técnicas de amostragem abordadas e a metodologia utilizada para a conclusão deste processo.

Identificadas as ferramentas, foi necessário proceder à coleta de tráfego de rede significativo. Assim, foram documentados os passos e pormenores de relevância sobre este processo.

Por fim, foi feita a junção de todos os componentes explicados, de forma a se obter uma metodologia de testes geral, que servirá de base para os resultados apresentados posteriormente nesta dissertação, consoante determinados parâmetros comparativos explicados nesta secção.

4

ANÁLISE DE RESULTADOS

Neste capítulo serão apresentados os resultados relativamente à comparação entre as diferentes técnicas de amostragem e o seu impacto na classificação de tráfego de vídeo, utilizando a metodologia de testes e todos os seus componentes apresentados no capítulo anterior. Essa metodologia será utilizada com o intuito de, com os parâmetros comparativos definidos, realizar uma comparação entre o tráfego obtido com as técnicas de amostragem e o tráfego total, para assim diferenciar a acurácia das técnicas na estimação dos parâmetros de interesse.

Inicialmente, serão apresentados resultados gerais de todo o tráfego capturado não amostrado, consoante a ferramenta de classificação de tráfego utilizada. De seguida, serão apresentados os resultados referentes ao tráfego amostrado, consoante as diferentes perspetivas de análise baseadas nas diferentes técnicas de amostragem.

Contudo, tendo em conta que o principal objetivo deste trabalho é realizar um estudo referente a tráfego de vídeo, os resultados serão ainda assentes na análise de parâmetros de interesse desse tipo de tráfego.

4.1 ANÁLISE DO TRÁFEGO TOTAL E MÉTODOS DE CLASSIFICAÇÃO DE TRÁFEGO

4.1.1 *Análise do tráfego total com TSTAT*

Para ser possível realizar o estudo comparativo entre o tráfego amostrado e o tráfego não amostrado, é necessário ter em consideração os valores referentes ao tráfego total. Por isso, neste ponto inicial são apresentadas algumas estatísticas referentes ao tráfego total que servirão como base de comparação para a análise das técnicas de amostragem abordadas.

Nesta fase, será efetuada uma análise mais representativa do estado da rede e do tráfego de vídeo nos períodos de coleta de tráfego apresentados na Secção 3.3. Por intermédio da heurística de classificação de tráfego do TSTAT, em termos gerais, a proporção de tráfego de vídeo estimada, presente no tráfego coletado, é de 2%. Este valor reduzido pode ser justificável, pelo facto da coleta de tráfego ter sido realizada num ambiente académico como a UM e num período fora do pico de atividades curriculares. A veracidade desta

afirmação foi investigada recorrendo a *traces* adicionais coletados em períodos de aulas, consoante discutido na Secção 4.1.3.

Uma outra razão, tem a ver com o relacionamento de dois fatores: a metodologia de classificação de tráfego de vídeo implementada pelo TSTAT (ver Secção 3.1.1); e o tempo da coleta de tráfego. Ou seja, a metodologia de classificação de tráfego realizada pelo TSTAT assenta, essencialmente, sobre o momento de *handshake* inicial das ligações dos fluxos de vídeo. Portanto, os pacotes trocados entre o cliente e o servidor nesse momento são fundamentais para a ferramenta conseguir realizar a classificação de tráfego de vídeo. Assim, tendo em conta que o período da coleta de tráfego realizada é de 30 segundos, estima-se que apenas as sessões de tráfego de vídeo iniciadas dentro desse período são classificadas e caracterizadas pelo TSTAT como fluxos de tráfego de vídeo. Note-se, no entanto, que períodos de coleta da ordem de grandeza utilizada são usuais em cenários de monitorização de redes operacionais, como é o caso da CAIDA [39].

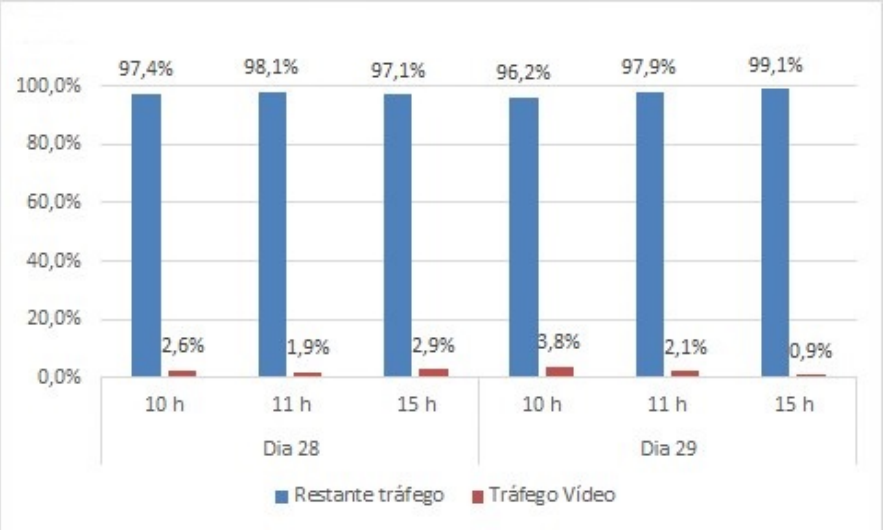


Figura 4.1.: Quantidade de tráfego de vídeo por dia e hora

Atendendo a que a coleta de tráfego foi realizada em períodos distintos, a Figura 4.1 ilustra através de um gráfico de barras a percentagem de tráfego de vídeo identificado em cada um dos dias e horários da coleta realizada. Analisando a figura, conclui-se, na generalidade, aquilo que já se referenciou anteriormente, a pouca quantidade de tráfego de vídeo identificada. De uma forma geral, os valores percentuais de tráfego de vídeo identificado estão entre 1% e 3,8% e, consequentemente, o restante tráfego entre 96% e 99%. De referir que estes dados são baseados na quantidade de pacotes de rede coletada.

Fluxos de vídeo identificados e Heavy Hitters

Um fluxo de dados é uma agregação compacta de dados, que consiste numa coleção incremental de estatísticas sobre um determinado grupo de pacotes de rede, que têm determinadas características comuns [40]. Essas estatísticas podem ser relacionadas com valores pertencentes aos cabeçalhos dos pacotes, tais como os endereços de IP de fonte e destino [41], ou com distribuições temporais, por exemplo todos os pacotes que atravessem um determinado ponto de medição a determinado instante de tempo. Assim, os fluxos identificados por ferramentas de classificação permitem perceber, com base num determinado conjunto de pacotes, determinadas características ou acontecimentos que se passam numa determinada rede de comunicação.

Sendo o principal foco no tráfego de vídeo, na Tabela 4.1 são apresentados valores estatísticas retirados da classificação feita pelo TSTAT. De todo o tráfego coletado, foram identificados 4989 fluxos de tráfego de vídeo distintos que contêm um total de 2863092 pacotes de tráfego.

Dia	Hora	Total Pacotes	Numero Fluxos Video	Pacotes video	Média pacotes/fluxo
28	10	19717144	983	521598	531
	11	21631888	723	404916	560
	15	17403832	1064	511266	481
29	10	20875760	1006	786006	781
	11	21780840	831	459260	553
	15	19463896	382	180046	471
Total		120873360	4989	2863092	

Tabela 4.1.: Quantidade de pacotes e fluxos de vídeo

A Figura 4.2 demonstra a relação entre a quantidade de fluxos de vídeo identificados e a média de pacotes por fluxo de vídeo. Ajuda também a visualizar os valores apresentados na Tabela 4.1 e as suas variações ao longo do tempo.

Através destes resultados observa-se que um maior número de pacotes de vídeo não implica que haja um maior número de fluxos. Por exemplo, no período das 15 horas do dia 28, é o período com menor quantidade de tráfego coletado, em termos da quantidade de pacotes, porém é o período onde existe a maior quantidade de fluxos de tráfego de vídeo identificados, cerca de 1064 fluxos. Em termos da quantidade de pacotes de vídeo não é também o que tem a maior quantidade. De destacar que o período das 15 horas do dia 29 é o que contém uma menor quantidade de fluxos de vídeo identificados, com 382 fluxos.

Com a Figura 4.2 verifica-se a média entre o número de pacotes de vídeo e o número fluxos de vídeo identificados. De salientar que o período das 10 horas do dia 29 tem a maior média de cerca de 781 pacotes por fluxo. De referir também, no período das 15 horas do dia 29 a média de pacotes por fluxo ultrapassa a quantidade total de fluxos de vídeo identificados.

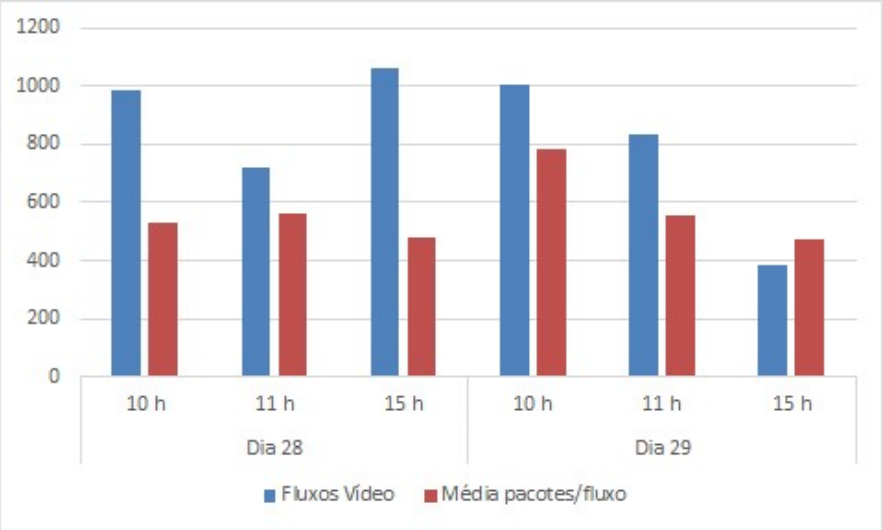


Figura 4.2.: Relação entre a quantidade de fluxos de vídeo e a média de pacotes por fluxo de vídeo

O facto de não haver uma relação direta de proporcionalidade entre o número de pacotes e o número de fluxos é justificado, pelo facto de, um fluxo de dados não ter uma quantidade de pacotes fixa, ou seja, é igualmente possível um fluxo ter apenas um pacote, como é possível ser composto por todos os pacotes presentes no tráfego analisado. Por isso, a estimação dos fluxos HH ser um importante método comparativo, porque permite perceber a relevância que um pequeno conjunto dos fluxos mais significativos tem na totalidade do tráfego, consoante a sua quantidade de pacotes. A noção de HH está explicada na Secção 3.4.1.

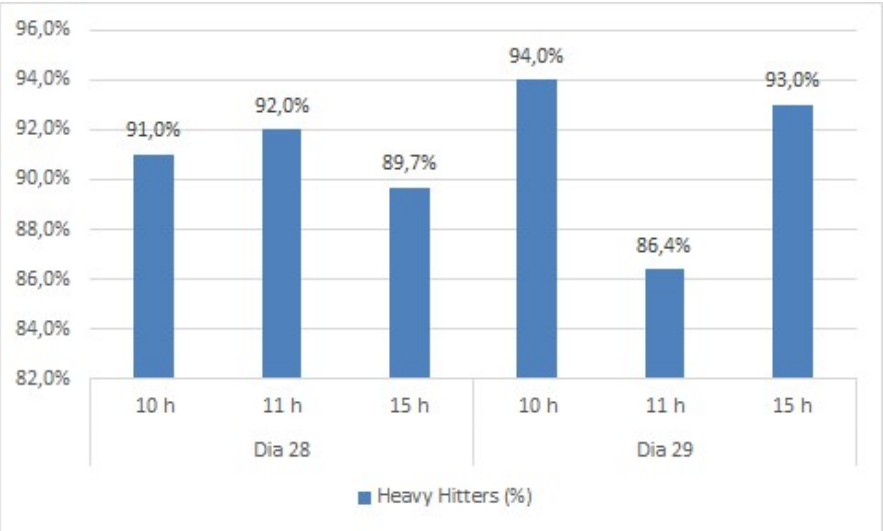


Figura 4.3.: Heavy-Hitters por período de coleta

A Figura 4.3 demonstra a percentagem de HH identificada por cada um dos períodos da coleta de tráfego. O período de coleta de tráfego das 10 horas do dia 29 é o que consegue obter uma maior percentagem de fluxos HH, cerca de 94%, que equivale ao período em que existe um maior volume de tráfego de vídeo, como referenciado anteriormente. Este valor significa que com apenas 20% dos fluxos de vídeo identificados consegue-se obter o equivalente a 94% do tráfego de vídeo nesse período. No período das 11 horas do dia 29 é visível o pico mais baixo em termos da percentagem de HH (86,4%), isto significa que, neste período, existe uma distribuição mais equitativa dos pacotes do tráfego de vídeo pelos vários fluxos, comparativamente com os restantes períodos de coleta. De destacar também, o valor do período das 15 horas do dia 29. Visto que este é o período com menor quantidade de tráfego e de fluxos de vídeo identificados, tem o segundo maior valor de fluxos HH (93%). Em média, entre todos os períodos de coleta de tráfego, existe 91% de fluxos HH.

Análise por sentido dos fluxos de vídeo

Como referido anteriormente, a noção de fluxo de tráfego diz que um fluxo é composto por um conjunto de pacotes de tráfego. A forma como esses pacotes são agrupados em fluxos pode provocar ligeiras alterações na classificação, porque cada ferramenta usa as suas próprias metodologias de classificação de tráfego. Por exemplo, uma forma de diferenciar essas metodologias tem a ver com a forma como é tratada a direcionalidade da comunicação, isto é, os fluxos de tráfego podem ser bidirecionais ou unidirecionais. No primeiro caso os pacotes são agrupados como se a comunicação fosse de apenas um só sentido, sem haver diferenciação do sentido de comunicação; no caso dos fluxos unidirecionais a diferenciação do sentido da comunicação (S2C ou C2S) é fundamental.

Posto isto, o TSTAT trata e classifica os fluxos como bidirecionais, porém dentro dos parâmetros que se obtém do *output*, existem alguns que fazem a diferenciação da direção de comunicação, tal como a quantidade de pacotes, que é dada consoante a direção C2S e S2C. Por isso, na Figura 4.4 são apresentadas estatísticas referentes à quantidade de pacotes e HH, consoante a direção S2C ou C2S.

Uma característica dos serviços de vídeo é que, tipicamente, a maior parte das suas implementações assentam sobre modelos cliente-servidor. Por essa razão, o tráfego, em quantidade de pacotes, com origem nos clientes será menor que o tráfego no sentido com origem no servidor. Isso é consequência do cliente realizar pedidos de dados ao servidor, que por sua vez, envia os dados pretendidos pelo cliente que são, naturalmente, em muito maior quantidade. É o que se verifica na análise à Figura 4.4, em cada um dos períodos de coleta de tráfego, a quantidade de pacotes no sentido S2C é sempre superior a 70% do total de pacotes de dados.

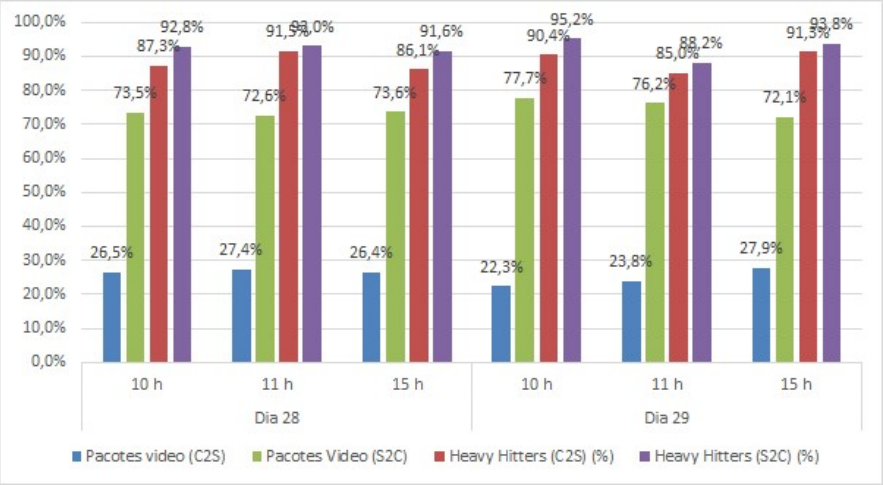


Figura 4.4.: Estatísticas segundo a direção de comunicação

Contudo, em termos de HH evidencia-se que a diferença entre os dois sentidos é mínima, isto porque os HH não são uma estatística diretamente relacionada com a característica de tráfego explicada anteriormente.

Identificação de prestador de serviço

Neste ponto será feita uma análise do tráfego classificado como vídeo por prestador de serviço. A identificação do prestador de serviço é efetuada com base no parâmetro *server name*, já explicado na Secção 3.1.1. O objetivo de agregar os fluxos pelos diferentes prestadores de serviços, permite que seja efetuada uma análise comparativa entre eles. Tendo em conta que o tráfego de vídeo é o principal objeto de estudo deste trabalho, a análise será referente a prestadores de serviço desse tipo de tráfego.

A Figura 4.5 permite verificar quais os servidores de vídeo mais utilizados, pelos utilizadores da UM, no momento da coleta de tráfego. Cerca de 55% do tráfego identificado pertence a servidores da *Googlevideo*, mais de metade de todo o tráfego identificado. Em 2º lugar encontram-se os servidores do *Youtube* com 26% do tráfego identificado, se considerarmos que estes dois servidores pertencem ao mesmo prestador de serviço, a *Google*, ficariamos com um tráfego agregado de 81% da totalidade do tráfego identificado. Contudo, o TSTAT não conseguiu identificar os servidores de alguns dos fluxos, significando cerca de 16% do tráfego. Por último, o servidor identificado como *Gvt1.com* corresponde ao restante tráfego, cerca de 3%.

Todas as análises feitas nesta Secção 4.1.1, referem-se a estatísticas sobre todo o tráfego coletado. Estas estatísticas servirão agora de base para os processos comparativos, entre as várias técnicas de amostragem.

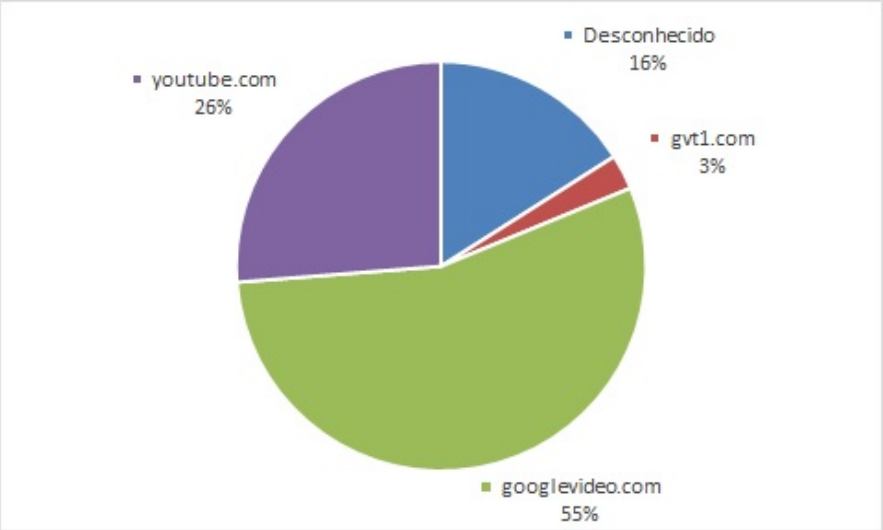


Figura 4.5.: Servidores de tráfego de vídeo identificados

4.1.2 *Análise do tráfego total com TIE*

Anteriormente, na Secção 3.1.2, apresentou-se a ferramenta TIE, que realiza uma classificação aplicando diferentes métodos de classificação de tráfego.

Os métodos de classificação de tráfego que serão abordados neste ponto, são o método de classificação ao nível das portas de comunicação utilizadas pelo tráfego (*port-based*) e o métodos de análise de *payload*, DPI (*payload-based*).

Estes métodos de classificação permitem ao TIE inferir determinados parâmetros que caracterizam os fluxos de rede presentes no tráfego analisado. Dentro desses parâmetros (ver Tabela 3.1), existe o parâmetro que permite identificar, a nível aplicacional, o tipo de tráfego desse fluxo.

De forma a interpretar mais intuitivamente os resultados, foi desenvolvido um programa que interpreta esses resultados e os demonstra graficamente, como as figuras apresentadas a seguir. Essas figuras ilustram os dados do parâmetro *App_id* para cada método de classificação.

A Figura 4.6 apresenta as percentagens das principais aplicações identificadas com a técnica *port-based*. Existe uma grande percentagem de tráfego em que a ferramenta não conseguiu identificar qualquer aplicação, cerca de 58%, contudo, os fluxos baseados no protocolo HTTP têm uma grande margem de aplicações identificadas, com o HTTP com 9% e o HTTPS (*Hypertext Transfer Protocol Secure*) com 24%. Por fim, 8% tráfego é proveniente do serviço DNS (*Domain Name System*) e 2% de tráfego pertence a uma quantidade de aplicações com fluxos de menor taxa.

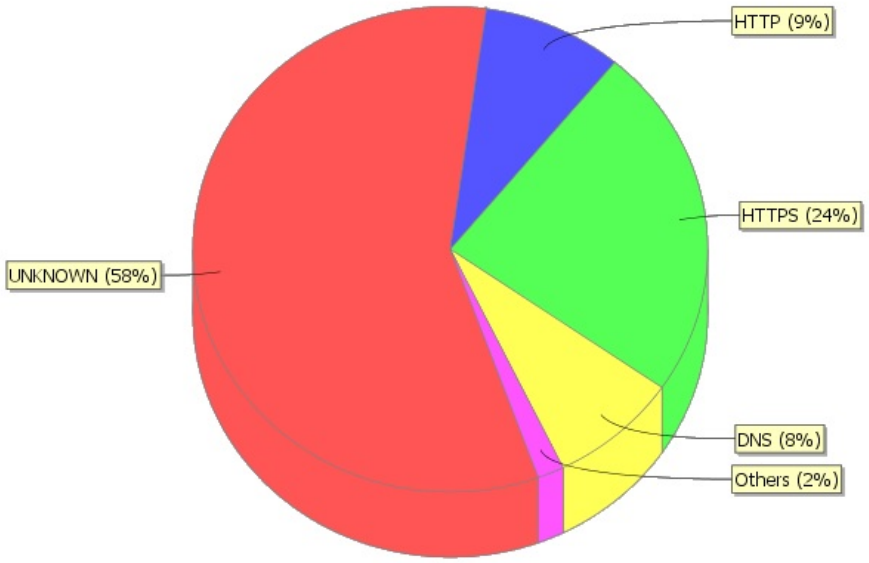


Figura 4.6.: Classificação do tráfego por *port-based*

A Figura 4.7 apresenta as percentagens das principais aplicações identificadas com a técnica *payload-based*. Neste caso, a percentagem de tráfego desconhecido é ainda maior, cerca de 71%. Com 11% do tráfego identificado como SSL, 10% como DNS e 6% como HTTP, estas aplicações são as que têm uma maior taxa de tráfego identificado. Novamente, com 2% estão identificadas outras aplicações que têm uma taxa mais residual.

Neste método de classificação, que tem por base a análise dos de dados dos fluxos, é obtida uma maior taxa de tráfego desconhecido. Isto acaba por ser uma consequência natural, devido a haver cada vez mais tráfego encriptado ao nível do *payload*, o que torna os dados mais difíceis de analisar para este tipo de métodos de classificação. Por exemplo, de notar que não houve qualquer percentagem de tráfego HTTPS identificado, a razão prende-se por esse protocolo utilizar encriptação nos seus dados. Contudo, sendo essa encriptação realizada pelo protocolo SSL/TLS, existe uma parte do tráfego que é identificado como tal.

A Figura 4.8 apresenta as percentagens das principais aplicações identificadas com as técnicas *port-based* e *payload-based*. Com a combinação destas técnicas obtém-se uma menor taxa de tráfego desconhecido, 48%. Com isso, nota-se o aparecimento novamente do protocolo HTTPS, com os mesmos 24% de tráfego identificado, anteriormente vistos no método *port-based*. Os protocolos aplicacionais HTTP e DNS aparecem nesta classificação com a mesma percentagem, 11% para cada um. O protocolo SSL tem uma taxa de 3%. Por fim, a taxa de outras aplicações tem um aumento para 3%, comparativamente com a análise dos métodos anteriores.

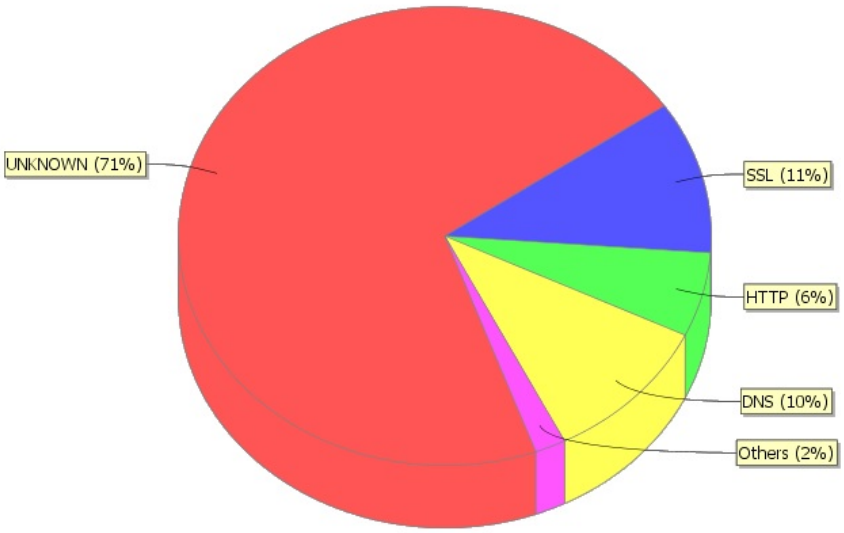


Figura 4.7.: Classificação do tráfego por *payload-based*

Como se pode verificar, consoante cada tipo de classificação utilizada, variam os diferentes valores de tráfego aplicacional identificado. Juntando a isto, o facto de não haver uma diferenciação exata do tráfego de vídeo, os resultados e análises realizadas ao nível da utilização de técnicas de amostragem, ou seja, a análise abordada nesta dissertação a partir da Secção 4.2, serão realizadas recorrendo à ferramenta TSTAT.

Como referido anteriormente, o TSTAT regista todas as conexões TCP, que são detetadas como conexões de vídeo. Assim, tendo em conta que existe uma proporção bastante considerável de tráfego HTTP, consequentemente TCP, a limitação da ferramenta, em classificar tráfego de vídeo que não seja TCP, não interfere gravemente com este estudo, devido a que a proporção desse tráfego é muito pequena.

4.1.3 *Análise do tráfego de coleta adicional*

Como referido anteriormente, foi realizada uma coleta adicional de tráfego de rede nos SCOM. Pelos resultados da primeira coleta, verificou-se que a quantidade de tráfego de vídeo identificada era relativamente baixa, apenas 2%. Uma das razões que se poderia apontar para esse facto, tem a ver com a altura em que a coleta de tráfego foi realizada.

A primeira coleta foi realizada em dois dias, dias 28 e 29 de julho de 2016, por isso, devido a ser uma época de baixa atividade na UM, poderia ser uma razão que contribuísse para a baixa quantidade de tráfego de vídeo detetada. Assim, decidiu-se proceder a uma nova coleta de tráfego para se investigar se os resultados da primeira coleta estariam desfasados

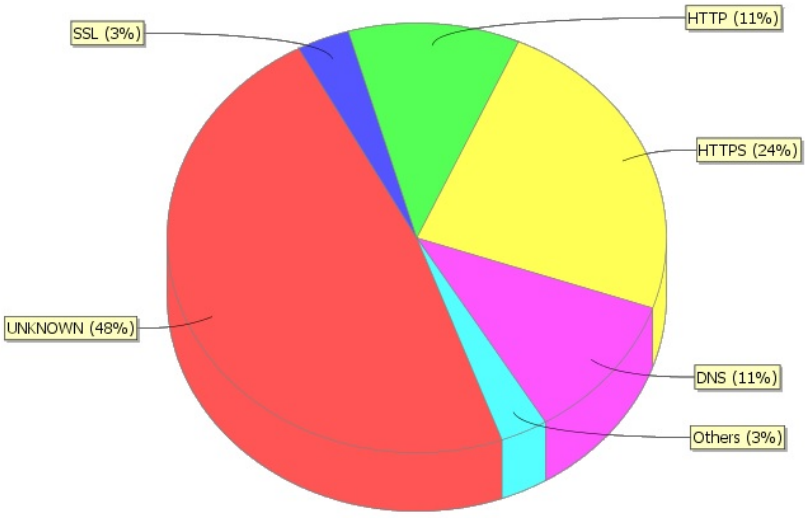


Figura 4.8.: Classificação do tráfego por *port-based* e *payload-based*

da realidade. Daí, essa coleta adicional foi realizada no dia 10 de outubro de 2016 nas mesmas condições da primeira coleta.

Relembrando os resultados sobre o volume de dados coletados, a nova coleta de tráfego teve, de uma forma geral, resultados idênticos aos obtidos na coleta anteriormente realizada. No caso da percentagem de tráfego de vídeo presente na nova coleta, verifica-se uma diminuição comparativamente com a coleta anterior, muito perto dos 0% de tráfego de vídeo identificado, enquanto que nas primeiras coletas o tráfego de vídeo rondava os 2%.

De referir que estes resultados são referentes a tráfego proveniente da rede da UM, onde existe um ambiente puramente académico. Por isso, existem interesses dos utilizadores ligeiramente diferentes, comparativamente com uma rede mais genérica. Adicionalmente, por causa da metodologia de classificação utilizada pelo TSTAT, apenas se consideram fluxos de vídeo que tenham início durante os 30 segundos em que a coleta de tráfego é realizada.

Assim, com estes resultados obtidos, conclui-se que a primeira coleta de tráfego obtém resultados representativos, em relação ao tráfego de rede da UM. Portanto, será conveniente utilizar a primeira coleta nos próximos estudos realizados, devido a obter-se uma maior quantidade de tráfego de vídeo nessa coleta e, com isso, resultados mais interessantes no âmbito deste trabalho.

4.2 ANÁLISE DA TÉCNICA *systematic count-based*

Nesta secção será abordada a técnica de amostragem de tráfego *SystC*, de forma a avaliar o seu impacto na classificação de tráfego de rede proveniente de serviços de vídeo. Com a coleta de tráfego realizada na UM, o processo de amostragem (a primeira fase da metodologia de testes apresentada na Secção 3.4) para esta técnica é realizado a diferentes frequências, para, dessa forma, se comparar o impacto das frequências de amostragem na classificação de tráfego de vídeo.

O método de amostragem desta técnica é baseado na posição dos pacotes, que são selecionados consoante os parâmetros do tamanho da amostra e do intervalo entre amostras indicados no início do processo de amostragem. Na Tabela 3.2 foram apresentadas as frequências de amostragem que serão avaliadas para esta técnica. Assim, após serem aplicadas as frequências de amostragem referidas, foram determinados os resultados que são apresentados na Tabela 4.2, em termos do tamanho dos dados amostrados (em *GBytes*), do número total de fluxos identificados, do número total de pacotes presentes nos fluxos e do número médio de pacotes por cada fluxo.

Tabela 4.2.: Estatísticas gerais - *SystC*

Técnica e frequência	Tamanho (*GBytes*)	Total Fluxos	Total pacotes	Pacotes/fluxo
SystC 1/8	13,15	289769	15109172	52
SystC 1/16	6,67	161899	7669304	47
SystC 1/32	3,32	89306	3813289	43
SystC 1/64	1,64	49502	1885699	38
SystC 1/128	0,82	28562	948174	33
SystC 1/256	0,41	16844	474069	28

Intuitivamente, a frequência de amostragem é diretamente proporcional à quantidade de dados, de fluxos ou de pacotes. Isto é, quanto maior for a frequência utilizada, maiores serão os parâmetros enunciados. Essencialmente, a Tabela 4.2 confirma essa relação para todos os casos analisados. A frequência *1/8* é a maior de todas as analisadas, pegando 1 pacote a cada 8, com um total de dados de mais de 13 GBytes, com 289769 fluxos de dados distintos identificados, um total de pacotes 15109172 e em média cerca de 52 pacotes por fluxo. Nas restantes frequências nota-se que todos os parâmetros vão decrescendo, a cada diminuição da frequência. Em alguns casos, para cerca de metade do valor anterior.

Essa última característica é mais percetível na Figura 4.9, onde estão representadas as percentagens de tráfego amostrado para as diferentes frequências, sendo essas percentagens baseadas nos dados do tráfego total. Os parâmetros que representam essa percentagem

referem-se ao volume de dados do tráfego e à quantidade de pacotes selecionados para amostra.

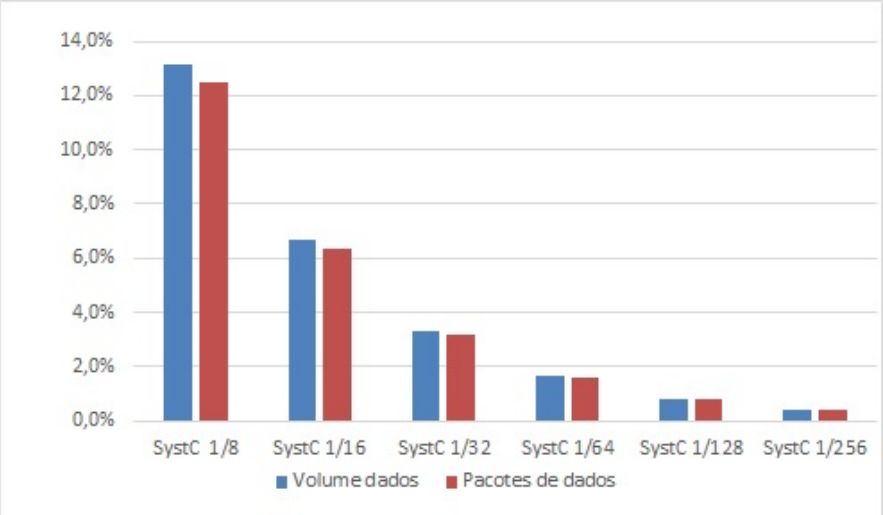


Figura 4.9.: Percentagem do tráfego amostrado no tráfego total

Genericamente, verifica-se que na frequência de amostragem 1/8 obtem a maior percentagem, isto é, a frequência de amostragem 1/8 consegue obter cerca de 13% do tráfego total coletado. Por conseguinte, os valores vão decrescendo gradualmente sempre que se diminui a frequência de amostragem.

4.2.1 *Fluxos de vídeo identificados*

Após apresentadas e interpretadas as estatísticas referentes à amostragem do tráfego, deve-se apresentar as estatísticas referentes à identificação do tráfego de vídeo no tráfego amostrado. O principal objetivo consiste em perceber o impacto que a técnica de amostragem *SystC* tem na identificação de tráfego de vídeo. Assim, será feita uma análise sobre esse impacto.

Por conseguinte, na fase de obtenção dos resultados verificou-se um facto diferente em relação a esta técnica de amostragem, como se pode verificar na Tabela 4.3. Considerando o valor *n/N* presente na figura, relacionado com cada frequência da técnica *SystC* e com cada um dos períodos da coleta de tráfego, n é o número de fluxos de vídeo identificados e N a quantidade de pacotes presentes nos n fluxos.

Como se pode verificar, em praticamente todos os períodos de coleta e todas as frequências de amostragem, não são identificados quaisquer fluxos de vídeo e, consequentemente, pacotes de tráfego de vídeo. Excecionalmente, no período do dia 28 das 11 horas, para a frequência 1/8, foi identificado 1 fluxo de vídeo com 8 pacotes.

	Dia 28 - 10 horas	Dia 28 - 11 horas	Dia 28 - 15 horas	Dia 29 - 10 horas	Dia 29 - 11 horas	Dia 29 - 15 horas
SystC 1/8	0/0	1/8	0/0	0/0	0/0	0/0
SystC 1/16	0/0	0/0	0/0	0/0	0/0	0/0
SystC 1/32	0/0	0/0	0/0	0/0	0/0	0/0
SystC 1/64	0/0	0/0	0/0	0/0	0/0	0/0
SystC 1/128	0/0	0/0	0/0	0/0	0/0	0/0
SystC 1/256	0/0	0/0	0/0	0/0	0/0	0/0

Tabela 4.3.: Quantidade de fluxos e pacotes de vídeo identificados por frequência de amostragem e período de coleta - *SystC*

De forma a tentar perceber a razão da não identificação de tráfego de vídeo, utilizando esta técnica de amostragem, foi necessário isolar e estudar o único fluxo que foi identificado. Para isso, foi utilizada uma aplicação de análise de tráfego de rede, o *Wireshark* [42]. Dessa forma, foi possível isolar o fluxo de vídeo em questão, com ajuda dos parâmetros identificados pelo TSTAT (tais como: ip origem/destino ou portas origem/destino).

A Figura 4.10 confirma os 8 pacotes que pertencem a esse fluxo, mostrando a visão na aplicação de análise escolhida.

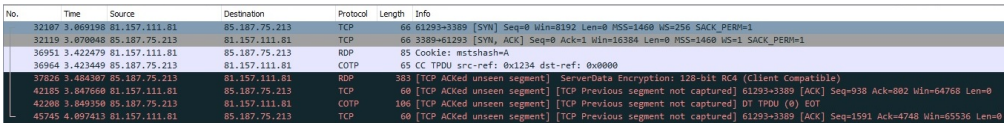


Figura 4.10.: Pacotes presentes no fluxo de vídeo identificado na técnica *SystC*

Através dos dados ilustrados na Figura 4.10 verifica-se que este fluxo advém da utilização da aplicação de controlo remoto *Windows Remote Desktop*. Isto conclui-se devido à presença do protocolo RDP (*Remote Desktop Protocol*) [43] no fluxo. Este protocolo foi propositadamente produzido pela *Microsoft* para satisfazer serviços de controlo remoto dos seus sistemas operativos. Este tipo de serviços faz o transporte de dados multimédia entre dois pontos, por isso faz sentido que o TSTAT classifique este fluxo como vídeo.

Para proceder a essa classificação, o TSTAT faz uso dos pacotes de *handshake* TCP presentes no fluxo, os pacotes *[SYN]* e *[SYN, ACK]*. Estes dois pacotes em conjunto com o pacote RDP, permitem ao TSTAT identificar este fluxo como vídeo.

Chegou-se a essa conclusão devido a alguns testes realizados, que consistiram em verificar se o TSTAT conseguiria identificar o fluxo se algum dos 3 pacotes em questão fosse retirado da análise. Para isso, foram produzidos *traces* em que os pacotes em questão não estavam presentes. Deste modo, verificou-se que o fluxo era identificado, apenas quando os 3 pacotes estavam simultaneamente presentes no *trace*. Por conseguinte, tendo em conta o processo de classificação de tráfego de vídeo utilizado pelo TSTAT (ver Secção 3.1.1), estima-se que a presença de pacotes responsáveis pelo estabelecimento da ligação inicial, sejam fundamentais para a ferramenta conseguir identificar os fluxos como tráfego de vídeo.

Por isso, considerando as frequências utilizadas e a percentagem de tráfego de vídeo presente no tráfego global, cerca de 2%, estima-se que a probabilidade de ser selecionado um pacote de vídeo é reduzida. Portanto, considerando-se que são necessários vários pacotes para se proceder à classificação de fluxos de vídeo, a probabilidade de um fluxo ser identificado como vídeo é ainda menor. Dessa forma, são justificáveis os resultados da classificação do tráfego amostrado nas frequências de amostragem avaliadas.

Este tipo de resultados não são expectáveis de acontecer em técnicas que façam uma seleção de pacotes sequencial (por exemplo, técnicas *Time-based*), porque dessa forma aumenta a probabilidade de se selecionar pacotes que sejam dependentes entre si para o processo de classificação de tráfego, porque, geralmente, o espaçamento entre esses pacotes é bastante curto em termos temporais e posicionais.

Similarmente, deve-se ponderar a utilização de técnicas de amostragem baseadas em eventos específicos, que selecionam pacotes aquando da ocorrência de determinados eventos nos conteúdos que atravessam a rede. Assim, seria possível selecionar os pacotes mais importantes para a classificação de tráfego de vídeo do TSTAT. Por outro lado, implicaria a necessidade de processar todos os pacotes em busca do evento em questão, o que terá impacto direto no custo computacional, principalmente em redes de alto débito, como a rede em que foram feitas as capturas para análise.

4.3 Análise da técnica *systematic time-based*

Nesta secção será abordada a técnica de amostragem de tráfego *Systematic Time-based*, de forma a avaliar o seu impacto na classificação de tráfego de rede proveniente de serviços de vídeo. Mais uma vez, serão comparadas diferentes frequências de amostragem e, consoante determinados parâmetros, avaliado o desempenho desta técnica quando é executada nessas frequências. A definição e seleção das frequências que serão alvo deste estudo, encontram-se na Tabela 3.2.

Por conseguinte, a estratégia de seleção de pacotes desta técnica, baseia-se no tempo em que os pacotes chegam ao ponto de amostragem, ou seja, após receber como parâmetros o tamanho da amostra e o intervalo entre amostras, numa certa unidade de tempo, esta técnica faz a coleta de tráfego no tempo especificado como tamanho da amostra e excluí os pacotes sequentes, durante o intervalo de tempo seguinte que é indicado como segundo parâmetro. Mais detalhes sobre esta técnica já foram apresentados na Secção 3.2.

Sendo assim, foram aplicadas todas essas frequências ao tráfego coletado nos SCOM que serve como base de estudo neste trabalho. Na Tabela 4.4 estão apresentadas as estatísticas gerais sobre a quantidade de dados (em *GBytes*), o número total de fluxos identificados, o número total de pacotes presentes nesses fluxos e a quantidade média de pacotes por fluxo, para cada frequência.

Tabela 4.4.: Estatísticas gerais - *SystT*

Técnica e frequência	Tamanho (*GBytes*)	Total Fluxos	Total pacotes	Pacotes/fluxo
***SystT* 100/500**	20,81	371723	23922536	64
***SystT* 100/1000**	10,21	193618	11773204	61
***SystT* 200/500**	41,95	675718	48228817	71
***SystT* 200/1000**	20,77	352495	23893716	68
***SystT* 500/1500**	34,46	536536	39604351	74
***SystT* 500/2500**	20,5	341928	23516439	69
***SystT* 500/3500**	14,16	240198	16273960	68

Analisando a Tabela 4.4, a técnica *SystT 200/500* é a que contém uma maior quantidade de dados, cerca de 42 *GBytes* de dados. A ordem de forma decrescente das frequências referidas, em termos do tamanho total dos dados, é a seguinte: *SystT 200/500; SystT 500/1500; SystT 100/500; SystT 200/1000; SystT 500/2500; SystT 500/3500; SystT 100/1000*.

Considerando os fluxos identificados, obtemos exatamente a mesma ordem entre as frequências, anteriormente enunciada para o caso do tamanho dos dados. Ao contrário das técnicas *Count-based*, neste caso a seleção de pacotes segue, até um determinado limite, uma ordem sequencial. Assim, é possível aumentar a probabilidade de selecionar os pacotes do *handshake* TCP necessários para a identificação de fluxos de vídeo.

Contudo, se olharmos para a média de pacotes por fluxo, denota-se que não existe a mesma característica. Por exemplo, a frequência que detém uma maior quantidade de pacotes por fluxo é a *SystT 500/1500*, com cerca de 74 pacotes por fluxo.

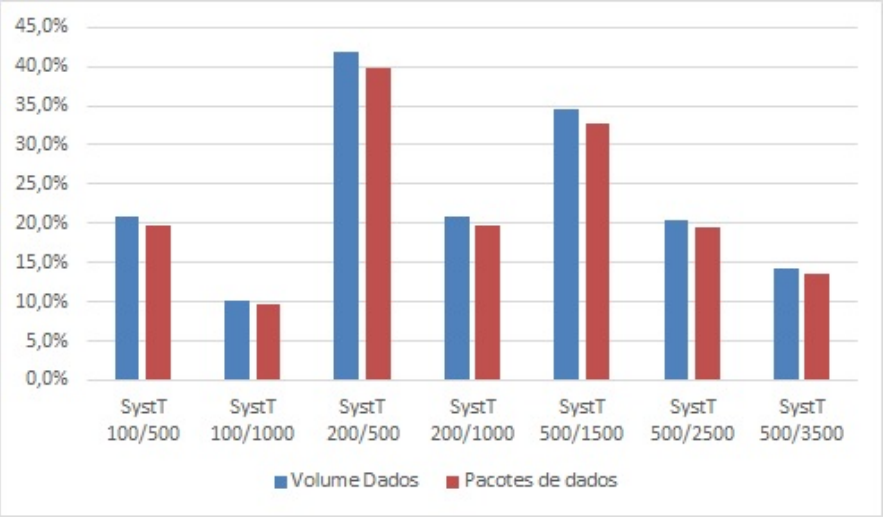


Figura 4.11.: Percentagem de tráfego amostrado - *SystT*

A Figura 4.11 auxilia a interpretação dos dados da Tabela 4.4. A frequência *SystT 200/500* consegue obter mais de 40% dos dados de tráfego total e obtém os mesmos 40%, se analisarmos em termos da quantidade de pacotes de dados. Como anteriormente referido, é expectável que com o aumento da frequência de amostragem haja um aumento da quantidade de dados amostrados. No entanto, essa característica não é tão linear para esta técnica como na técnica *SystC*, porque o estado de utilização da rede não é linear e isso influencia o volume de dados amostrados por períodos de tempo.

4.3.1 *Fluxos de vídeo identificados e Heavy Hitters*

Após uma introdução sobre a técnica de amostragem *SystT* e ilustração dos seus efeitos sobre o tráfego capturado, deve-se fazer uma análise ao tráfego de vídeo e às implicações que esta técnica de amostragem provoca sobre este tipo de tráfego.

Tendo em conta que o processo de coleta do tráfego foi realizado em diferentes períodos, em dois dias distintos, os resultados apresentados são baseados em todos esses momentos, contudo para não se verificar uma análise demasiado extensa, os resultados de todos os períodos são agregados num só gráfico estatístico. Como verificado na Secção 4.1.1, a quantidade de tráfego e de fluxos identificados é diferenciada para todos os períodos da coleta. No entanto, não se deve deixar de reparar que os parâmetros comparativos utilizados para análise são baseados nos valores identificados na estatística do tráfego total.

A Figura 4.12 ilustra os resultados estatísticos consoante 3 parâmetros comparativos distintos, sendo eles: a percentagem de pacotes de vídeo selecionados para amostra; a percentagem de fluxos de vídeo identificados no tráfego amostrado; e a percentagem dos fluxos *heavy-hitters*.

Analisando a Figura 4.12 verifica-se que existe uma tendência para que quanto maior a frequência de amostragem, em quantidade de dados amostrados, maior a quantidade de fluxos e de pacotes serão identificados como vídeo. Por exemplo, a frequência 200/500 apresenta os maiores resultados em termos de percentagem de pacotes e de fluxos de vídeo identificados, 13,1% e 28,4% respetivamente. Como visto anteriormente, tem cerca de 42 *GBytes* de dados amostrados, sendo a frequência de amostragem com maior quantidade de dados. Próximo destes valores encontra-se a frequência 500/1500, com 8% de pacotes e 25,7% de fluxos de vídeo identificados. Confirma-se também a análise feita anteriormente, em como esta frequência é a segunda com maior quantidade de dados, por isso é também a segunda com maior quantidade de pacotes e fluxos de vídeo identificados.

De destacar que, na análise feita ao tráfego total, verificou-se em certos pontos que uma maior quantidade de pacotes não significa uma maior quantidade de fluxos de vídeo identificados. Essa característica é, também, verificada nesta técnica de amostragem, mais concretamente entre as frequências 500/2500 e 500/3500. Obviamente, a frequência 500/2500 é

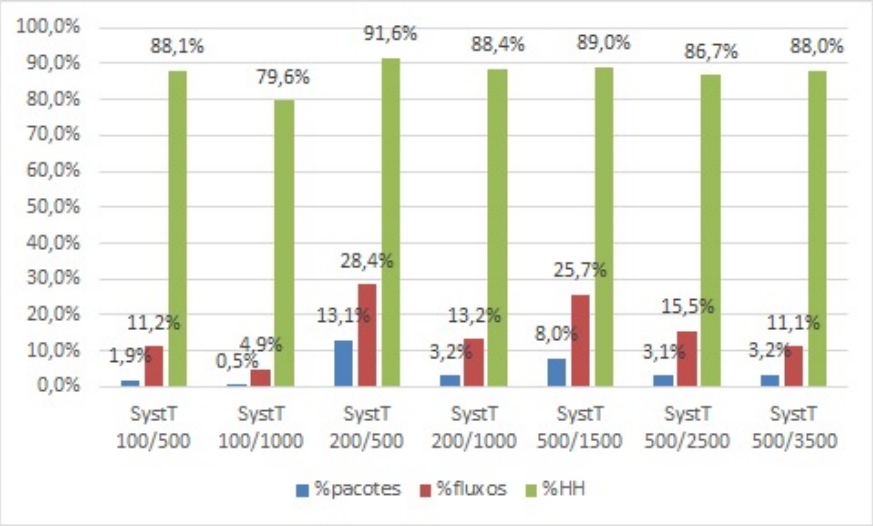


Figura 4.12.: Identificação de tráfego de vídeo - *SystT*

maior que a frequência 500/3500, por isso, na Figura 4.11 confirma-se que o primeiro caso tem maior quantidade de dados que o segundo. Ainda assim, isso não quer dizer que serão identificados mais dados como vídeo, como demonstra a Figura 4.12, em que no primeiro caso temos 3,1% de pacotes de vídeo identificados e 3,2% de pacotes de vídeo no segundo caso. Porém, em termos de fluxos de vídeo são identificados 15,5% e 11,1% no primeiro e segundo casos, respetivamente. Assim, identifica-se a mesma característica já identificada e explicada na Secção 4.1.1.

Em termos da percentagem de HH confirma-se a mesma tendência, embora as diferenças entre as frequências sejam mínimas. Excetuando, a frequência 100/1000 que com 79,6% de HH obtém uma diferença de cerca de 8% para as restantes frequências. Comparativamente com o tráfego total, que teve em média 91% de fluxos HH, apesar das diferenças serem mínimas, a frequência 200/500 foi a que mais se aproximou desse valor.

Fluxos de vídeo distintos

Como abordado anteriormente, devido a serem aplicadas técnicas de amostragem ao nível da seleção de pacotes de tráfego, é evidente que haverão fluxos de tráfego distintos, entre os fluxos identificados no tráfego amostrado e os fluxos identificados no tráfego total.

Neste ponto, serão abordadas as estatísticas referentes a essa diferenciação entre fluxos, contudo os fluxos estudados neste caso referem-se aos fluxos de vídeo, sendo os seus valores apresentados na Tabela 4.5.

Os números na tabela são calculados com base nos fluxos de vídeo identificados para cada uma das diferentes frequências de amostragem. A técnica que conseguiu obter um maior número de fluxos iguais foi a técnica *SystT 500/1500* com 49 fluxos iguais e 1274 flu-

Tabela 4.5.: Fluxos de vídeo distintos - *SystT*

Técnica e frequência	Fluxos distintos	Fluxos Iguais
SystT 100/500	554	4
SystT 100/1000	241	1
SystT 200/500	1372	45
SystT 200/1000	646	15
SystT 500/1500	1234	49
SystT 500/2500	733	40
SystT 500/3500	536	20

xos distintos. Isto significa que, 49 dos fluxos de vídeo identificados no tráfego amostrado pela técnica *SystT 500/1500* têm exatamente as mesmas características (por exemplo, quantidade de pacotes do fluxo) que as que foram identificadas na classificação do tráfego total. Em segundo lugar, ficou a frequência 200/500 com 1372 fluxos diferentes e 45 fluxos iguais. De notar que, a frequência 500/1500, que tem uma menor quantidade de dados amostrados que a frequência 200/500, com menos fluxos de vídeo identificados, conseguiu obter um maior número de fluxos iguais, comparativamente com o tráfego total. Assim, analisando os valores da tabela, é possível verificar um padrão referente ao tamanho da amostra coletada, verificando-se que quanto maior o tamanho da amostra, maior é a taxa de fluxos iguais identificados. Tendo em conta que se está a quantificar diretamente a informação recolhida, é natural que quanto mais informação houver, mais próximo se fica dos valores totais.

4.3.2 *Análise por sentido dos fluxos de vídeo*

Quando se está a avaliar tráfego de determinadas características, é natural que alguns resultados sejam, em condições normais, expectáveis. Quando é abordado tráfego de vídeo é expectável que, na análise de fluxos unidirecionais haja um maior volume de dados no sentido da receção do cliente. Isto acontece devido a estes serviços serem assentes, maioritariamente, em metodologias Cliente-Servidor.

A Figura 4.13 apresenta resultados que demonstram essa conclusão, em termos da quantidade de pacotes. Apesar de se estar a lidar com tráfego amostrado, essa característica mantém-se inalterada. Analisando a figura, conclui-se que para todas as frequências de amostragem a taxa de pacotes de vídeo é muito maior no sentido S2C, do que no sentido contrário. Essa característica verifica-se, também, ao nível dos HH, contudo as diferenças não são tão acentuadas como no primeiro caso.

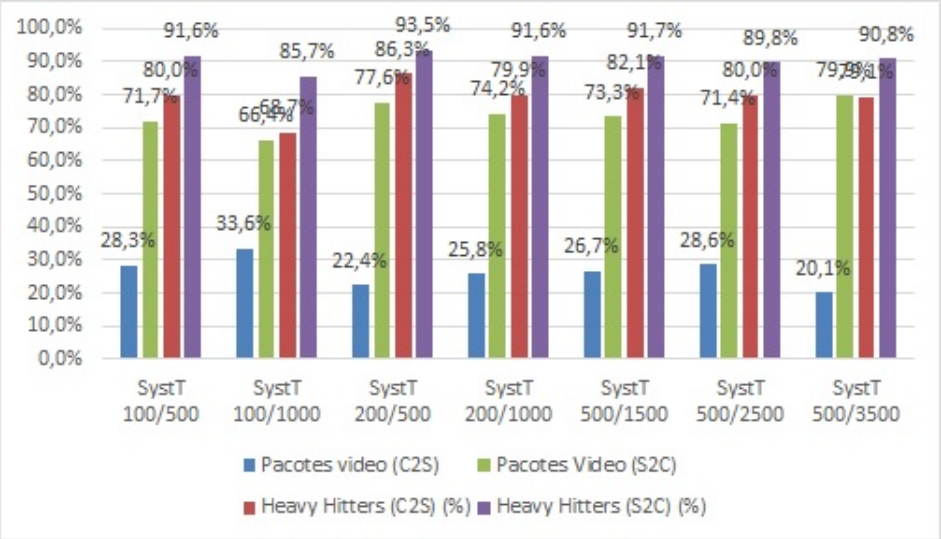


Figura 4.13.: Estatísticas segundo a direção de comunicação - *SystT*

A frequência 500/3500 tem uma maior diferença, em termos da percentagem de pacotes de vídeo, entre os dois sentidos, com 20,1% no sentido C2S e 79,9% no sentido S2C. A frequência 100/1000 obtém valores de 33,6% no sentido C2S e 66,4% no sentido contrário, sendo a frequência com uma menor diferença entre os dois sentidos.

De destacar a frequência 100/1000, que tem uma maior percentagem de pacotes no sentido S2C e obtém a menor taxa de fluxos HH, cerca de 68,7% no sentido C2S e 85,7% no sentido S2C.

4.3.3 *Identificação do serviço*

Devido à classificação do tráfego pelo TSTAT, é possível, através de um parâmetro específico (o parâmetro *server_name*), estimar a quantidade de fluxos de vídeo consoante o prestador de serviço identificado. Para isso, é apresentada a Figura 4.14.

Através da análise da Figura 4.14, entende-se que para todas as frequências, os servidores da *googlevideo.com* são os que têm um maior número de fluxos identificados, sendo mais de 50% do tráfego de vídeo identificado em todas as frequências. Sendo a frequência 100/1000 a que tem uma maior percentagem deste tipo de servidor, 61,8%.

O segundo servidor com maior percentagem de fluxos é o *youtube.com*, com valores entre os 27% e os 33%. E, por último, o servidor *gvt1.com* com valores entre os 2% e os 3%.

Contudo, existe uma parte significativa deste tráfego que é tratado como desconhecido, sendo os valores entre os 6% e os 18%.

Em relação aos fluxos de servidores desconhecidos, deve-se destacar que se verifica um grande aumento desses fluxos, quando são utilizadas frequências de amostragem em que o

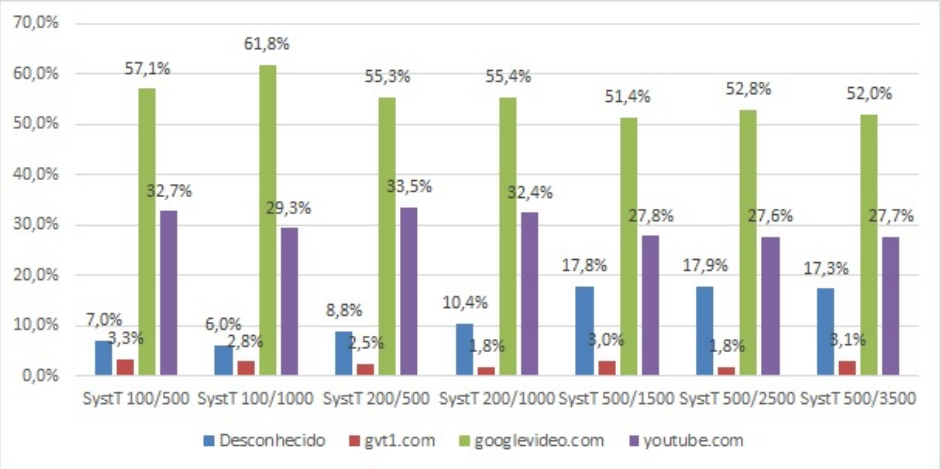


Figura 4.14.: Servidores de tráfego de vídeo identificados - *SystT*

parâmetro tamanho da amostra é mais elevado. De facto, esse género de frequências são as que conseguem obter um resultado mais próximo dos resultados obtidos na classificação do tráfego total. A razão para a ocorrência disso, deve-se ao facto dos pacotes selecionados serem escolhidos com uma maior sequência temporal, logo maior a probabilidade de selecionar pacotes que dependam uns dos outros para realizar uma classificação mais eficaz dos fluxos de vídeo.

4.4 Comparação Entre Diferentes Técnicas de Amostragem

Essencialmente, existem duas formas de amostragem de tráfego de rede distintas, amostragem baseada na posição dos pacotes de tráfego de rede, e amostragem baseada em pontos temporais de seleção de pacotes. Nas secções anteriores foram estudadas essas abordagens, através de várias frequências, e discutidos os resultados do seu impacto na estimação de parâmetros de interesse no tráfego de vídeo.

Nesta secção será efetuado um estudo comparativo entre as diferentes técnicas de amostragem. Cada técnica implementa de forma diferente as abordagens referidas anteriormente. Assim, as técnicas e suas respetivas frequências que serão estudadas nesta secção, são: *SystC 1/100* e *RandC 1/100*, técnicas baseadas na seleção de pacotes segundo a sua posição; *SystT 100/1000*, *LP 100/200* e *MuST 200/500*, técnicas em que a seleção de pacotes é baseada em funções temporais. Na Secção 3.2 encontram-se explicadas as diferentes técnicas de amostragem estudadas, bem como as razões da seleção, para análise, das respetivas frequências de amostragem.

Tomada a decisão em relação aos métodos de amostragem a serem avaliados e suas respetivas frequências, deve-se avaliar os resultados gerais da aplicação dessas técnicas na

metodologia de testes deste trabalho. Na Tabela 4.6 são apresentadas as estatísticas gerais sobre a quantidade de dados (em *GBytes*), o número total de fluxos identificados, o número total de pacotes presentes nesses fluxos e a quantidade média de pacotes por fluxo, para cada técnica.

Tabela 4.6.: Estatísticas gerais - diferentes técnicas

Técnica e frequência	Tamanho (*GBytes*)	Total Fluxos	Total pacotes	Pacotes/fluxo
***SystC* 1/100**	1,05	34551	1208009	35
***SystT* 100/1000**	10,21	193618	11773204	61
***RandC* 1/100**	1,07	35111	1227091	35
***LP* 100/200**	7,6	146615	8759711	60
***MuST* 200/500**	9,59	168308	11790769	70

Avaliando os resultados presentes na Tabela 4.6 conclui-se que, em termos da quantidade de dados amostrados, a técnica *SystT* contém a maior quantidade com mais de 10 *GBytes* de dados amostrados, seguida pelas técnicas adaptativas *MuST* e *LP* com cerca de 9,5 e 7,6 *GBytes*, respetivamente.

A relação entre a média do número total de pacotes e o número total de fluxos identificados indica que a técnica *MuST* tem a maior média com 70 pacotes/fluxo, seguido das técnicas *SystT* e *LP* com 61 e 60 pacotes/fluxo, respetivamente. Estes números demonstram que a técnica *SystT*, apesar de ser a que tem maior quantidade de dados, não tem a melhor relação entre esses dados e o número de fluxos identificados, ao contrário das técnicas adaptativas, que com menos quantidade de dados obtêm uma maior taxa de fluxos identificados. Nas técnicas *RandC* e *SystC* verifica-se que ambas têm a mesma média de pacotes por fluxo, isto é justificado pelo facto de terem a mesma frequência de amostragem. Contudo, os seus números de fluxos identificados são diferentes.

A Figura 4.15 faz uma comparação, visualmente mais intuitiva, do volume de dados de tráfego selecionado para análise. Os resultados estão em forma de percentagem que relaciona o tráfego amostrado com o tráfego total.

4.4.1 *Fluxos de vídeo identificados e Heavy Hitters*

Relativamente aos resultados referentes ao tráfego de vídeo, espera-se que o impacto que a amostragem tem neste tipo de tráfego esteja sempre dependente da quantidade de dados coletados. Apesar de neste trabalho se estar a comparar diretamente diferentes abordagens de amostragem de pacotes de rede, nunca se deve deixar de refletir no facto de quanto mais dados se estiver a analisar, tendencialmente, melhores resultados se irão obter. Posto isto,

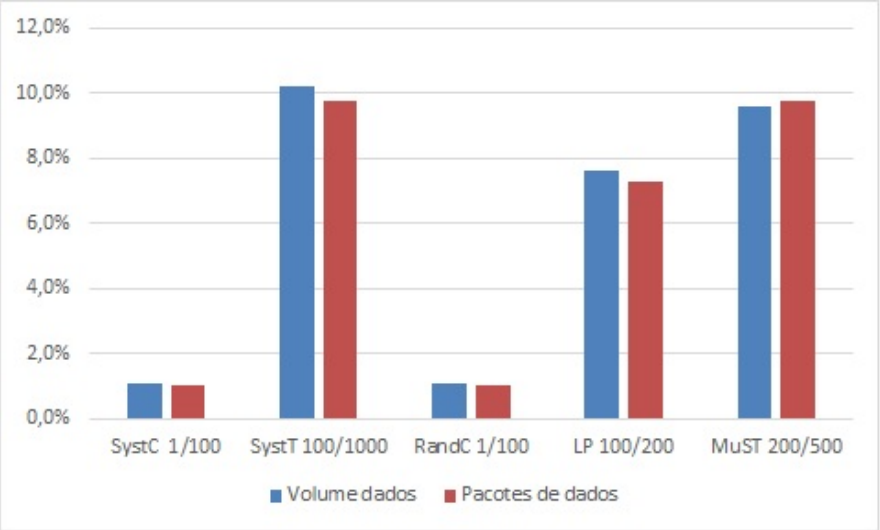


Figura 4.15.: Percentagem de tráfego amostrado - diferentes técnicas

é normal que a amostragem de tráfego tenha resultados mais próximos do tráfego total, quanto maior seja o tráfego amostrado.

Assim, a Figura 4.16 demonstra os resultado do impacto de diferentes técnicas de amostragem na classificação de tráfego de vídeo, segundo três parâmetros comparativos. Esses parâmetros comparativos fazem a relação percentual entre determinadas características do tráfego de vídeo total e o tráfego de vídeo amostrado, tais como: a quantidade de pacotes; a quantidade de fluxos identificados; e a quantidade média de HH.

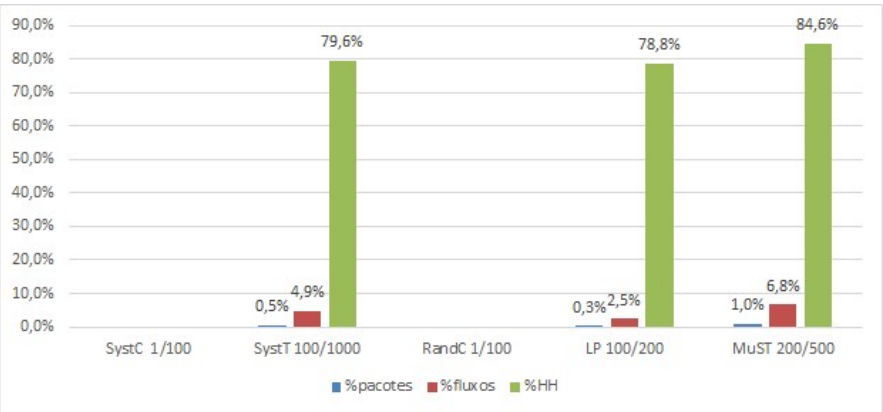


Figura 4.16.: Identificação de tráfego de vídeo - diferentes técnicas

Através da Figura 4.16, confirma-se aquilo que já foi estudado na Secção 4.2, para técnicas que baseiam a seleção de pacotes na sua posição, não se obtém qualquer identificação de fluxos provenientes de tráfego de vídeo.

Contudo, nas restantes técnicas, com 1% de pacotes e 6,8% de fluxos identificados, a técnica *MuST* consegue obter os melhores resultados no que diz respeito à percentagem de pacotes e fluxos. Consegue atingir os 84,6% de fluxos HH, sendo também a técnica com melhor resultado nesse aspeto. Em segundo lugar, encontra-se a técnica *SystT* com 0,5% de pacotes de vídeo do tráfego total, 4,9% de fluxos vídeo identificados e 79,6% de fluxos HH. Por fim, a técnica *LP* que obtém 0,3% de pacotes, 2,5% de fluxos e 78,8% de fluxos HH.

No início desta subsecção referiu-se o facto de que, tendencialmente, a quanto maior a quantidade de dados para análise, mais próximo se estará dos valores do tráfego total. Porém, os resultados enunciados demonstram que o método de amostragem é também importante nessa análise. Anteriormente, verificou-se que a técnica *SystT* era a que, em termos brutos, tinha uma maior quantidade de dados, contudo neste ponto, concluiu-se que essa técnica não é a que atinge um melhor resultado em termos de pacotes e fluxos de vídeo identificados. Inclusivamente, a técnica *LP*, que tem menos 2,6 *GBytes* de dados brutos, consegue obter resultados muito próximos da técnica *SystT*.

Concluindo, a forma como são amostrados os dados é importante para a estimação de parâmetros de tráfego de vídeo, constatando-se que as técnicas adaptativas conseguem obter resultados similares ou superiores a outras técnicas que têm uma maior quantidade de dados. Isso advém do facto de, no caso da técnica *MuST*, adaptar os seus parâmetros consoante o nível de atividade da rede em determinado instante de tempo. Assim, sendo os serviços de vídeo, tendencialmente, geradores de picos de tráfego de rede, os seus pacotes têm uma maior probabilidade de ser selecionados para amostra com técnicas adaptativas do que com as restantes técnicas.

Fluxos de vídeo distintos

Quando se está a analisar tráfego amostrado, é expectável que os fluxos identificados sejam diferentes, em termos dos diversos parâmetros do *output* do TSTAT. Contudo, esta análise permite perceber se mesmo com amostragem, é possível obter a classificação de fluxos de vídeo exatamente igual à classificação dos fluxos de vídeo identificados no tráfego total.

Na Tabela 4.7 verifica-se que algumas técnicas conseguiram obter alguns fluxos iguais.

Tabela 4.7.: Fluxos de vídeo distintos - diferentes técnicas

Técnica e frequência	Fluxos distintos	Fluxos Iguais
SystC 1/100	0	0
SystT 100/1000	241	1
RandC 1/100	0	0
LP 100/200	125	0
MuST 200/500	330	7

A técnica *SystT* conseguiu identificar 1 fluxo, enquanto que a técnica *MuST* conseguiu identificar 7 fluxos iguais.

Esta análise permite demonstrar que mesmo com tráfego amostrado, é possível obter classificação do tráfego de vídeo, exatamente com as mesmas características que os fluxos identificados no tráfego total. Por exemplo, basta no processo de amostragem não selecionar um pacote pertencente a esse fluxo, que no âmbito desta análise esse fluxo será caracterizado como sendo diferente. Por isso, devido à grande probabilidade de algum pacote não ser selecionado, é significativo que se consiga obter fluxos exatamente iguais no tráfego amostrado, em termos da classificação dos mesmos.

4.4.2 *Análise por sentido dos fluxos de vídeo*

No contexto dos serviços de vídeo assentes em modelos Cliente-Servidor, a análise consoante a direcionalidade dos fluxos de tráfego é bastante significativa. Por isso, para verificar se as diferenças entre *upload* e *download* se mantêm quando estamos a lidar com tráfego de rede amostrado, foi projetada a Figura 4.17. Ela apresenta, para as diferentes técnicas de amostragem, a percentagem de pacotes e de fluxos HH consoante os dois sentidos de comunicação.

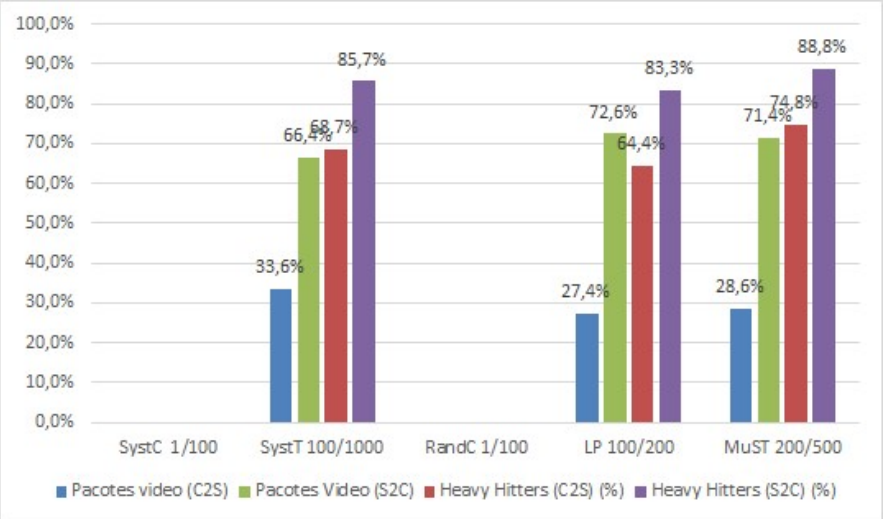


Figura 4.17.: Estatísticas segundo a direção de comunicação - diferentes técnicas

A figura demonstra que a característica acima descrita é verdadeira, independentemente da técnica de amostragem utilizada. A percentagem de pacotes de vídeo no sentido C2S tem um valor mais baixo nas técnicas adaptativas, com 27,4% na técnica *LP* e 28,6% na técnica *MuST*, enquanto que a técnica *SystT* apresenta um valor de 33,6%. No sentido contrário, S2C, os valores são de 72,6% para a técnica *LP*, 71,4% para a técnica *MuST* e 66,4% para

a técnica *SystT*. As técnicas adaptativas conseguem obter uma aproximação mais real dos resultados obtidos no tráfego total. Relativamente aos número obtidos sobre os fluxos HH, a técnica *MuST* é a que consegue obter uma maior percentagem de fluxos HH.

4.4.3 *Identificação do serviço*

Nesta situação é importante estimar os servidores de vídeo presentes no tráfego analisado e perceber se a amostragem implica alterações nesse tipo de classificação. Como referenciado anteriormente, este tipo de análise é possível através de um parâmetro específico proveniente dos resultados do processamento do TSTAT.

Por isso, a Figura 4.18 apresenta os servidores identificados e a sua percentagem no tráfego amostrado em análise.

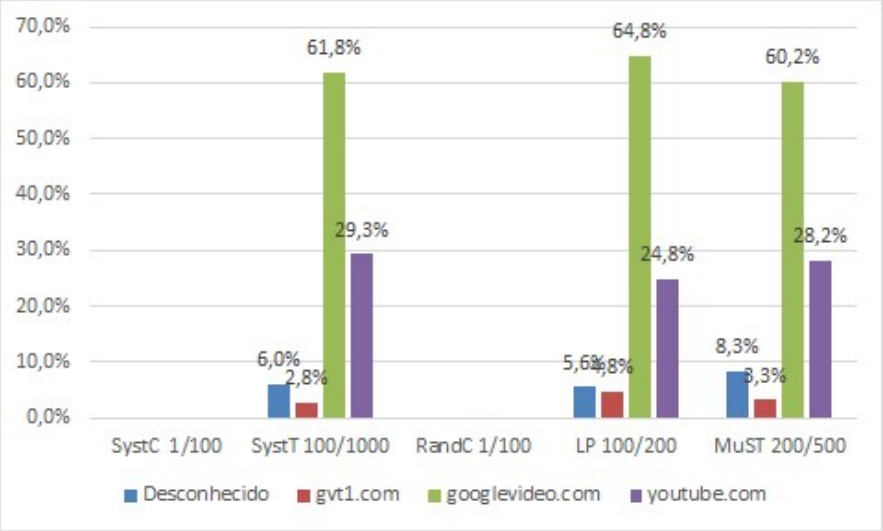


Figura 4.18.: Servidores de tráfego de vídeo identificados - diferentes técnicas

Relembrando da análise efetuada ao tráfego total, o servidor *googlevideo.com* detinha 55% do tráfego de vídeo, em segundo lugar o servidor *youtube.com* equivalia a 26% do tráfego, em 16% do tráfego o servidor era tratado como desconhecido para a ferramenta de classificação e o servidor *gvt1.com* representava os restantes 3%. Estes resultados estão representados na Figura 4.5.

Com esses resultados e os resultados da Figura 4.18, verifica-se que para o servidor *googlevideo.com* a técnica *MuST* é a que mais se aproxima dos resultados totais, com uma representatividade de 60,2% no tráfego amostrado. Seguida pela técnica *SystT*, 61,8%, e por último a técnica *LP* com 64,8%. Assim, verifica-se que para valores mais elevados a amostragem tende a sobrestimar esses valores.

O servidor *youtube.com* tem uma melhor aproximação na técnica *LP* com 24,8%, enquanto que na técnica *MuST* representa 28,2% do tráfego de vídeo e 29,3% na técnica *SystT*.

Para o servidor *gvt1.com* os resultados são bastante aproximados da realidade em qualquer das técnicas, embora a técnica *LP* fique mais afastada do cenário total, com 4,8%, enquanto que as técnicas *SystT* e *MuST* conseguem uma representação de 2,8% e 3,3%, respetivamente.

Quanto ao tráfego desconhecido nota-se uma diferença maior comparando com o cenário do tráfego total. Tendo em conta que a percentagem de tráfego desconhecido, em termos do servidor de vídeo, no tráfego total é de 16%, as técnicas de amostragem obtêm uma representatividade de 8,3%, 6% e 5,6%, respetivamente as técnicas *MuST*, *SystT* e *LP*. Estes valores acabam por estar relacionados com a sobrestimação acima referenciada.

Estes resultados demonstram que não existe um grande impacto das técnicas de amostragem, baseadas em funções temporais, na classificação de tráfego de vídeo consoante os diferentes servidores.

4.5 Síntese dos resultados

Esta secção salienta os principais resultados da análise do tráfego de rede coletado. Inicialmente, analisou-se o tráfego total consoante duas perspetivas de processamento, ou seja, através das duas ferramentas de classificação de tráfego abordadas, o TSTAT e o TIE.

Através do TSTAT verificou-se que durante os dois dias de coleta se obteve 2% de tráfego de vídeo, com picos entre o 1% e os 3,8% nos diferentes períodos da coleta. Essa percentagem de tráfego de vídeo deu origem a um número total de 4989 fluxos de vídeo identificados, verificando-se que não existe uma relação de proporcionalidade direta entre o número de fluxos identificados e a quantidade de pacotes coletados. Em termos de HH, foi identificada uma média de 91% de fluxos HH. Foram identificados 3 prestadores de serviços de vídeo distintos com diferentes representatividades, o *googlevideo.com* com 55%, o *youtube.com* com 26% e *gvt1.com* com 3%, nos restantes 16% do tráfego de vídeo os seus servidores são desconhecidos.

Através do TIE, foram realizadas 3 classificações de tráfego de vídeo distintas em que na sua globalidade se verificou uma percentagem considerável de tráfego HTTP, consequentemente TCP. Assim, tendo em conta esses resultados e a dificuldade em diferenciar o tráfego de vídeo do restante tráfego, foi decidido utilizar apenas a ferramenta TSTAT.

Na análise da coleta de tráfego adicional verificou-se que existiu ainda uma menor taxa de tráfego de vídeo identificado.

Na análise da técnica de amostragem *SystC* verificou-se a incapacidade de classificação de fluxos de vídeo, nas frequências abordadas, identificando apenas 1 fluxo de vídeo com 8 pacotes. Percebeu-se que a heurística de classificação de tráfego de vídeo utilizada pelo

TSTAT necessitava de vários pacotes do momento do estabelecimento da ligação de dados, para conseguir realizar a classificação de tráfego de vídeo. Por conseguinte, a probabilidade desta técnica selecionar esses pacotes simultaneamente é relativamente baixa. Portanto, técnicas que realizem uma seleção sequencial de pacotes ou que selecionem consoante eventos específicos da rede, serão mais indicadas para se proceder à classificação de tráfego de vídeo em tráfego amostrado.

Através da análise à técnica *SystT* verificou-se, ao contrário da técnica *SystC*, a identificação de valores razoáveis de fluxos de tráfego de vídeo. Nesta técnica existe a tendência de quanto mais dados amostrados, mais fluxos de tráfego de vídeo são identificados. Por exemplo a frequência 200/500 tem mais de 40% de todo o tráfego coletado e obtém os melhores resultados em termos da classificação de tráfego de vídeo. De notar que existe a capacidade de identificar integralmente os mesmos fluxos que se identificou no tráfego total. Verificou-se também que nos resultados da identificação do prestador de serviço, as frequências que utilizam um maior valor de tamanho da amostra, conseguem obter um resultado mais aproximado dos resultados do tráfego total.

Na análise comparativa entre diferentes técnicas de amostragem, para técnicas que baseiam a seleção de pacotes segundo a sua posição (*SystC* e *RandC*), verificou-se, novamente, a incapacidade de classificar os fluxos de tráfego de vídeo. Contudo, essas técnicas agregam apenas 1% do tráfego total. As restantes técnicas apresentam valores a rondar os 7,5% e os 10% de tráfego, sendo a técnica *MuST* a que melhores resultados consegue obter em percentagem de pacotes e fluxos de vídeo identificados. Verificou-se uma tendência para se obter melhores resultados, em termos da classificação de fluxos de tráfego de vídeo, com técnicas adaptativas, pois conseguem ter resultados similares ou superiores a outras técnicas que têm uma maior quantidade de dados.

4.6 SUMÁRIO

Neste capítulo foram apresentados os resultados da investigação e metodologia de trabalho exposta no capítulo 3.

Antes de ser abordado o tema da amostragem de tráfego, foram expostos os resultados globais do tráfego coletado, após o seu processamento pelas ferramentas de classificação de tráfego abordadas. Assim, com base nestes resultados, foi possível realizar o estudo comparativo proposto como objetivo neste trabalho.

De seguida, apresentam-se os resultados do tráfego amostrado, consoante as diferentes perspetivas de amostragem abordadas. Inicialmente, os resultados são interpretados relativamente à técnica *SystC* e de seguida à técnica *SystT*. Também, são comparadas diferentes técnicas de amostragem e interpretados os resultados dessa comparação.

Por último são sintetizados os principais resultados e conclusões obtidos no estudo realizado.

5

CONCLUSÕES E TRABALHO FUTURO

Neste capítulo é efetuada uma síntese do trabalho desenvolvido e apresentado, demonstrando a sua relação para com os objetivos propostos para este trabalho de mestrado. São também referidos alguns possíveis tópicos para trabalho futuro.

5.1 RESUMO DO TRABALHO DESENVOLVIDO

O tráfego de rede possui inúmeras características de acordo com os serviços que se pretendam classificar. Desta forma, para se obter uma classificação correta e fiável é necessário o levantamento das particularidades relativas ao tráfego de rede. Assim, tendo em conta que o principal objetivo deste trabalho foi estudar uma forma fiável e escalável de medir o estado da rede no suporte a serviços de vídeo, utilizando amostragem de tráfego, foi necessário fazer um levantamento das características dos fluxos de vídeo.

Relativamente ao método de captura e análise de tráfego de vídeo, utilizou-se medição passiva do tráfego de rede com análise *off-line*, recorrendo, numa instância inicial, a duas ferramentas de classificação de tráfego, TSTAT e TIE.

Neste ponto, deve-se referir que um objetivo interrelacionado deste trabalho passou pelo estudo de diferentes abordagens de amostragem de tráfego, identificando as características e implica-ções de cada uma delas na classificação de serviços de vídeo. Para isso, recorreu-se a uma ferramenta de amostragem de tráfego que aplica, ao tráfego alvo, diferentes técnicas de amostragem do mesmo.

Após a identificação das ferramentas a utilizar para ajudarem a cumprir os objetivos traçados para esta dissertação, chegou o momento de se proceder ao estudo comparativo da aplicação das diferentes técnicas de amostragem de tráfego na monitorização de serviços de vídeo. Porém, antes foi necessário fazer uma análise ao tráfego de uma forma geral, para dessa forma se obter um ponto de partida no estudo comparativo.

Desde logo, verificou-se uma taxa de tráfego de vídeo identificada, relativamente baixa. Por isso, decidiu-se proceder a uma nova coleta de tráfego para se inferir, se haveria implicações do momento da coleta de tráfego na quantidade de tráfego de vídeo identificada.

A grande maioria dos estudos comparativos foram realizados com base na ferramenta TSTAT, isto porque, se veio a verificar que a outra ferramenta de classificação de tráfego não era a mais aconselhável para este estudo, devido à dificuldade de identificação do tráfego de vídeo, na classificação feita pelo TIE.

Por conseguinte, entrou-se na fase de conciliar os resultados globais com os resultados do tráfego amostrado. Foi necessário definir alguns parâmetros comparativos, nomeadamente perceber a quantidade de fluxos de vídeo identificados, fluxos HH, implicação da direcionalidade dos fluxos de vídeo ou identificação do prestador de serviço. Dessa forma, foram abordados três estudos comparativos distintos baseados nas diferentes técnicas de amostragem de tráfego. Primeiramente, abordaram-se as técnicas *SystC* e *SystT* individualmente, comparando as suas frequências de amostragem e o impacto que elas teriam na estimação dos parâmetros definidos. Por fim, comparou-se várias técnicas de amostragem distintas.

Verificou-se que a aplicação de diferentes abordagens de amostragem de tráfego teve, naturalmente, implicações na estimação de características de tráfego de serviços de vídeo. Técnicas *count-based* que tenham um tamanho de amostra relativamente pequeno não são as ideais para se tratar tráfego de vídeo. Porém, a técnica *MuST* apresentou na globalidade os melhores resultados na estimação de tráfego de vídeo. Contudo, prevê-se que uma abordagem *event-based* seja a melhor solução para tratar este tipo de tráfego.

O tema desenvolvido nesta dissertação permitiu obter um estudo comparativo entre diferentes técnicas de amostragem e perceber o seu impacto na estimação de parâmetros de interesse referentes a serviços de vídeo. Os estudos desenvolvidos neste contexto de serviços de vídeo são praticamente inexistentes e, apesar de existirem alguns estudos relativamente à amostragem de tráfego, são também muito poucos aqueles que fazem um estudo comparativo tão abrangente de técnicas de amostragem.

5.2 TRABALHO FUTURO

Relativamente ao trabalho futuro, sugere-se a utilização de formas de classificação de tráfego distintas, como a utilização de ferramentas de classificação de tráfego que sejam capazes de realizar a identificação de parâmetros de vídeo com abordagens complementares às utilizadas pelo TSTAT.

Sugere-se a utilização de *traces* de tráfego de rede provenientes de redes mais representativas, não apenas referentes a um determinado contexto como o da UM. Mas também, que se recorra a uma coleta de tráfego mais extensiva, em termos de períodos de tempo.

Também, se sugere a aplicação deste tipo de estudo para outros tipos de serviços de rede, nomeadamente, serviços de *Cloud* ou de chamadas em tempo real.

Finalmente, sugere-se um estudo mais detalhado sobre a aplicação de técnicas de amostragem *event-based*, de forma a que esses eventos sejam relativos à seleção de pacotes proveniente de tráfego de vídeo.

A

CONFIGURAÇÃO DA METODOLOGIA DE TESTES

No endereço: http://drive.google.com/file/d/0B_ahawKfekTQNXAxY3hCTG92RDQ, estão disponíveis todas as ferramentas e *scripts* de interpretação utilizados na metodologia de testes deste trabalho. Toda a metodologia de testes é composta por 3 ferramentas de classificação ou amostragem de tráfego de rede, que são:

- TSTAT - informações de utilização e ligação para *download* em [31];
- TIE - informações de utilização e ligação para *download* em [44];
- *Sampling Framework* - informações de utilização e ligação para *download* em [36].

Para agilizar o processo de interpretação dos resultados fornecidos pelas ferramentas de classificação, foram criados dois *scripts* que são capazes de realizar esse processo e, assim, serão apresentados de seguida:

- Interpretador de resultados TSTAT - escrito na linguagem de programação C, este *script* lê e processa os ficheiros de resultados originados pelo TSTAT. Demonstra, através da linha de comandos, dados sobre esse ficheiro. Os dados que indica são: quantidade total de fluxos; quantidade total de pacotes; HH; quantidade total de pacotes C2S; HH C2S; quantidade total de pacotes S2C; HH S2C; quantidade de fluxos por servidor. O método de utilização deste *script* realiza-se através da linha de comandos e no momento da sua execução deve ser passado como argumento o ficheiro que será alvo de análise;
- Interpretador de resultados TIE - escrito na linguagem de programação JAVA, permite ser executado em diferentes sistemas operativos. O seu propósito é ler e interpretar os ficheiros de resultados originados pelo TIE. Demonstra informação relativa à classificação por aplicações que o TIE gera, apresentando esses dados num formato gráfico. O ficheiro a processar deve ser colocado na pasta de execução do projeto.

BIBLIOGRAFIA

[1] J. M. C. Silva, P. Carvalho, and S. R. Lima, "Computational weight of network traffic sampling techniques," in *2014 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–6, IEEE, 2014.

[2] D. Oliveira, P. Carvalho, and S. R. Lima, "Towards cloud storage services characterization," in *Computational Science and Engineering (CSE), 2015 IEEE 18th International Conference on*, pp. 129–136, IEEE, 2015.

[3] I. Cisco Systems, "Cisco Visual Networking Index: Forecast and Methodology, 2009-2014," Tech. Rep. 23.09.2010, 2010.

[4] R. Ramaswamy, L. Kencl, and G. Iannaccone, "Approximate fingerprinting to accelerate pattern matching," in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pp. 301–306, ACM, 2006.

[5] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: multilevel traffic classification in the dark," in *ACM SIGCOMM Computer Communication Review*, vol. 35, pp. 229–240, ACM, 2005.

[6] A. Callado, C. Kamienski, G. Szabó, B. P. Gerö, J. Kelner, S. Fernandes, and D. Sadok, "A survey on internet traffic identification," *Communications Surveys & Tutorials, IEEE*, vol. 11, no. 3, pp. 37–52, 2009.

[7] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., sep 2000.

[8] P. Barford and D. Plonka, "Characteristics of network traffic flow anomalies," in *Proceedings of the First ACM SIGCOMM Workshop on Internet Measurement - IMW '01*, (New York, New York, USA), p. 69, ACM Press, nov 2001.

[9] T. Zseby, M. Molina, and N. Duffield, "Sampling and Filtering Techniques for IP Packet Selection." RFC 5475, Mar. 2009.

[10] N. Duffield, "Sampling for passive internet measurement: A review," *Statistical Science*, pp. 472–498, 2004.

[11] 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*, pp. 124–131, IEEE, 2008.

[12] J. Giertl, J. Baca, F. Jakab, and R. Andoga, "Adaptive sampling in measuring traffic parameters in a computer network using a fuzzy regulator and a neural network," *Cybernetics and Systems Analysis*, vol. 44, no. 3, pp. 348–356, 2008.

[13] E. A. Hernandez, M. C. Chidester, and A. D. George, "Adaptive sampling for network management," *Journal of Network and Systems Management*, vol. 9, no. 4, pp. 409–434, 2001.

[14] Y. Lu and C. He, "Resource allocation using adaptive linear prediction in wdm/tdm epons," *AEU-International Journal of Electronics and Communications*, vol. 64, no. 2, pp. 173–176, 2010.

[15] Y. Wei, J. Wang, and C. Wang, "A traffic prediction based bandwidth management algorithm of a future internet architecture," in *Intelligent Networks and Intelligent Systems (ICINIS), 2010 3rd International Conference on*, pp. 560–563, IEEE, 2010.

[16] J. M. C. Silva, P. Carvalho, and S. R. Lima, "A multiadaptive sampling technique for cost-effective network measurements," *Computer Networks*, vol. 57, no. 17, pp. 3357–3369, 2013.

[17] B. Li, Z. Wang, J. Liu, and W. Zhu, "Two decades of internet video streaming: A retrospective view," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 9, no. 1s, p. 33, 2013.

[18] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "Rtp: A transport protocol for real-time applications ietf rfc 3550," *IETF*, 2003.

[19] A. Johnston, S. Donovan, R. Sparks, C. Cunningham, and K. Summers, "Rfc 3665-session initiation protocol (sip) basic call flow examples," *IETF, December*, 2003.

[20] X. Zhang, J. Liu, B. Li, and T.-S. P. Yum, "Coolstreaming/donet: a data-driven overlay network for peer-to-peer live media streaming," in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, vol. 3, pp. 2102–2111, IEEE, 2005.

[21] T. Stockhammer, "Dynamic adaptive streaming over http–: standards and design principles," in *Proceedings of the second annual ACM conference on Multimedia systems*, pp. 133–144, ACM, 2011.

[22] C. Luo, W. Wang, J. Tang, J. Sun, and J. Li, "A multiparty videoconferencing system over an application-level multicast protocol," *IEEE Transactions on Multimedia*, vol. 9, no. 8, pp. 1621–1632, 2007.

[23] G. J. Yang, B. W. Choi, and J. H. Kim, "Implementation of http live streaming for an ip camera using an open source multimedia converter," *International Journal of Software Engineering and Its Applications*, vol. 8, no. 6, pp. 39–50, 2014.

[24] Y. Xiao, X. Du, and J. Zhang, "Internet protocol television (iptv): the killer application for the next-generation internet," in *Institute of Electrical and Electronics Engineers*, 2007.

[25] R. Raghavendra and E. M. Belding, "Characterizing high-bandwidth real-time video traffic in residential broadband networks," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, pp. 597–602, IEEE, 2010.

[26] X. Recommendation, "135-speed of service (delay and throughput), performance values for public data networks when providing packet-switched services," *International Telegraph and Telephone Consultative Committee*, 1993.

[27] G. Almes, S. Kalidindi, and M. Zekauskas, "Rfc 2679: A one-way delay metric for ippm," *Internet Society (Sep. 1999)*, pp. 1–20, 1999.

[28] N. W. Group *et al.*, "Ip packet delay variation metric for ip performance metrics (ippm)," tech. rep., RFC 3393, 2002.

[29] G. Almes, S. Kalidindi, and M. Zekauskas, "Rfc 2680: A one-way packet loss metric for ippm," *Status: Standards Track*, 1999.

[30] K. H. Chan, J. Babiarz, and F. Baker, "Configuration guidelines for diffserv service classes," *RFC 4594*, 2006.

[31] T. N. G. P. di Torino, "Tcp statistic and analysis tool," *Web page: http://tstat.tlc.polito.it/*. Acedido em: outubro 2016.

[32] T.-T. STatistic, "Analysis tool - logs," *Web page: http://tstat.tlc.polito.it/measure.shtml#LOG*. Acedido em: outubro 2016.

[33] M. Mellia, A. Carpani, and R. L. Cigno, "Tstat: Tcp statistic and analysis tool," in *International Workshop on Quality of Service in Multiservice IP Networks*, pp. 145–157, Springer, 2003.

[34] S. Blake-Wilson, M. Nystrom, D. Hopwood, J. Mikkelsen, and T. Wright, "Rfc 3546: Transport layer security (tls) extensions," 2003.

[35] A. Dainotti, W. De Donato, and A. Pescapé, "Tie: A community-oriented traffic classification platform," in *Traffic Monitoring and Analysis*, pp. 64–74, Springer, 2009.

[36] J. M. C. Silva, P. Carvalho, and S. R. Lima, "A modular sampling framework for flexible traffic analysis," in *SoftCOM 2015 - 23rd International Conference on Software, Telecommunications and Computer Networks*, 2015.

[37] J. M. C. d. Silva, "A modular traffic sampling architecture for flexible network measurements," 2016.

[38] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, and J. Solé-Pareta, "Analysis of the impact of sampling on netflow traffic classification," *Computer Networks*, vol. 55, no. 5, pp. 1083–1099, 2011.

[39] U. of California's San Diego Supercomputer Center, "Caida - center for applied internet data analysis," *Web page:* http://www.caida.org/data/overview/. Acedido em: outubro 2016.

[40] M. Grossglauser and J. Rexford, "Passive traffic measurement for ip operations, the internet as a large-scale complex system (chapter)," *Oxford University Press, to appear*, vol. 3, p. 1, 2004.

[41] Z. Qian, H. Li, R. Tang, and K. W. Cheung, "Performance of traffic aggregation versus segregation for optical flow switching networks," in *2011 International Conference on Information Photonics and Optical Communications*, 2011.

[42] G. Combs *et al.*, "Wireshark," *Web page:* http://www.wireshark.org/. Acedido em: outubro 2016.

[43] Microsoft, "Remote desktop protocol," *Web page:* https://msdn.microsoft.com/en-us/library/aa383015(v=vs.85).aspx. Acedido em: outubro 2016.

[44] U. o. N. Computer Science Department, "Traffic identification engine," *Web page:* http://tie.comics.unina.it/. Acedido em: outubro 2016.