# Classificação multinível de tráfego de rede com base em amostragem

João Cunha, Ricardo Silva, João M. C. Silva e Solange Rito Lima
Departamento de Informática, Universidade do Minho, Braga, Portugal
Email: pg28504@alunos.uminho.pt, pg28501@alunos.uminho.pt, joaomarco@di.uminho.pt,
solange@di.uminho.pt

***Resumo*—A classificação e a caracterização do tráfego de rede são tarefas essenciais para o correto planeamento e gestão das actuais 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.**

**Neste contexto, este artigo tem como principal objetivo a exploração e comparação da acurácia de diferentes estratégias de classificação de tráfego de rede quando conjugadas com diferentes técnicas de amostragem. Recorrendo à ferramenta TIE - Traffic Identification Engine [1], que disponibiliza estratégias de classificação de tráfego operando em distintos níveis da pilha protocolar, e a uma framework de amostragem que implementa técnicas clássicas e emergentes de amostragem de tráfego [2], é analisado o impacto da estratégia de classificação e amostragem na correta identificação dos fluxos de rede. A base dos testes realizados recorre a tráfego de rede real coletado no Instituto Nacional de Estatística, posteriormente submetido a diferentes estratégias de classificação e amostragem. Desta forma, pretende-se contribuir com diretivas para a obtenção de uma classificação realista do tráfego de rede com o menor overhead em termos de tráfego coletado e analisado.**

## I. Introdução

A relevância e os desafios da classificação de tráfego tem crescido na última década na comunidade das redes e telecomunicações, tornado-se numa tarefa crítica para apoio a aspetos que vão desde o planeamento de redes, à auditoria de contratos de serviços e à deteção de ataques (DoS Attacks). No entanto, aspetos como o aumento da diversidade e volume de aplicações que circulam na rede, o encapsulamento de aplicações sobre distintos protocolos aplicacionais, a utilização de protocolos de segurança (IPSec, TTLS, HTTPS, etc.) na cifra dos dados, etc. colocam desafios acrescidos a uma correta classificação do tráfego. Ter a noção de que tipo de aplicações são mais utilizadas, quem as utiliza e com que fins, são razões que levam a que hoje em dia haja um maior controlo sobre o tráfego na Internet e a classificação desse mesmo tráfego de uma forma prática e ágil é necessária para obter esse controlo.

Para fazer face a estes problemas, foi criada uma ferramenta de software aberta à comunidade que permite fazer classificação de tráfego, denominada Traffic Identification Engine (TIE) [1] [3]. O TIE foi desenvolvido por uma equipa de investigadores da Universidade de Nápoles [4], que se basearam em observações dos problemas anteriormente referidos para fornecer à comunidade uma plataforma de fácil desenvolvimento e integração de técnicas de classificação de tráfego. Esta ferramenta é flexível e facilmente conciliável com outras aplicações, logo extensível a outros ramos como, por exemplo, a amostragem de tráfego. A amostragem surge, neste contexto, como a estratégia utilizada para tornar tratável o volume de dados a coletar, analisar e armazenar, já que apenas um subconjunto de tráfego da rede será usado na classificação. No entanto, idealmente, a análise desse subconjunto deverá ser capaz de permitir inferir com realismo o que se passa na rede em termos de fluxos de tráfego.

Assim, o principal objetivo deste artigo é estudar o impacto de diferentes técnicas de classificação na correta identificação dos fluxos de tráfego quando apenas parte do tráfego da rede é utilizado na análise. O subconjunto de tráfego analisado é obtido usando técnicas clássicas e recentes de amostragem de tráfego, tais como técnicas sistemáticas, aleatórias e adaptativas. A prova de conceito recorre ao auxílio da ferramenta TIE, de uma framework de amostragem desenvolvida na Universidade do Minho [2] que implementa as técnicas de amostragem referidas, e a tráfego real obtido na rede do Instituto Nacional de Estatística.

Este artigo encontra-se organizado da seguinte forma: na secção II é abordado o trabalho relacionado com o tema em estudo, nomeadamente em classificação e amostragem; na Secção III serão apresentados alguns conceitos fundamentais sobre as ferramentas e técnicas de amostragem utilizadas; na Secção IV é apresentada a metodologia de testes e algumas ferramentas desenvolvidas no seu âmbito por forma a agilizar a interação com a ferramenta TIE e auxiliar na interpretação dos resultados obtidos; na Secção V são discutidos os resultados obtidos; por fim, na Secção VI, são apresentadas as considerações finais e possíveis ideias para trabalho futuro.

## II. Trabalho Relacionado

O encapsulamento de aplicações em protocolos conotados com serviços diferentes, o aumento da utilização de protocolos de segurança, o aumento de aplicações que alocam dinamicamente portas de comunicação, tornam as técnicas de classificação imprecisas quando são baseadas unicamente em protocolos de transporte e nas portas de origem/destino. Por esse motivo, existem vários métodos alternativos para a classificação de tráfego, nomeadamente: (i) análise do *payload*, e.g., por *pattern matching* [5] e análise numérica [6], que são técnicas bastante efetivas na taxa de acerto, mas pesadas e não aconselháveis para classificar tráfego encriptado; (ii) análise

do comportamento dos sistemas ou terminais (*host-behavior*) [6], em que a classificação é feita através da análise do comportamento dos sistemas terminais em vez dos fluxos de tráfego, evitando o processamento do *payload*; e (iii) análise de fluxos (*flow-behavior*), em que se assume que cada aplicação tem as suas próprias propriedades estatísticas.

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 diferentes técnicas de classificação de tráfego. Porém, devido ao crescimento do volume de tráfego é cada vez mais difícil fazer uma análise ao tráfego total, 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. De facto, a utilidade da amostragem de tráfego tem sido explorada em diferentes áreas das redes de telecomunicações, nomeadamente: segurança de redes - deteção de anomalias e intrusões, botnet e identificação de DDoS [7]; determinação da conformidade dos SLA's e controlo de QoS - para estimação dos parâmetros, tais como atraso de pacotes, perda e jitter [8]; engenharia de tráfego - auxiliar a classificação e caracterização de tráfego [9].

No contexto deste trabalho, diferentes técnicas de amostragem vão ser conjugadas com distintas estratégias de classificação no sentido de avaliar o seu desempenho conjugado na correta deteção e classificação de fluxos. Desta forma, pretende-se também inferir de que forma a análise do tráfego de rede pode ser efectuada de forma mais eficiente.

## III. Classificação e amostragem de tráfego

### A. Ferramenta TIE

O TIE é uma ferramenta escrita em linguagem C e pensada para ambientes Unix. Pode ser executado em modo offline ou modo online. O modo offline consiste na leitura de fluxos de tráfego previamente coletados, enquanto que o modo online lê em tempo real os fluxos de tráfego que são capturados da rede à qual o dispositivo está ligado. A aplicação consegue correr dinamicamente vários plugins, que representam técnicas de classificação de tráfego implementadas em software (pode correr um ou mais plugins ao mesmo tempo). Na figura 1 está representada a arquitetura da ferramenta [3].

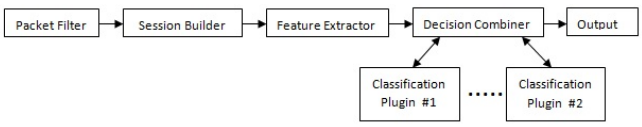



Figura 1. Componentes do TIE

O *Packet Filter* captura tráfego em tempo real ou lê *traces* que foram capturados previamente. Após o input na aplicação, os pacotes são agregados em sessões separadas do tipo flow, biflow, etc. pelo *Session Builder*. 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 output final num formato em que é possível verificar a classificação do tráfego de forma estatística, que permite a comparação de múltiplas abordagens.

### B. Técnicas de amostragem

A amostragem de tráfego consiste na coleta de subconjuntos do tráfego de rede, obtidos aplicando ao tráfego total técnicas de amostragem. As técnicas de amostragem avaliadas neste trabalho correspondem às principais técnicas atualmente utilizadas em ferramentas de medição de rede, i.e., *systematic count-based*, *systematic time-based* e *random count-based* [10]. Além disso, duas técnicas adaptativas são avaliadas, i.e., *predição linear adaptativa* [11] e *amostragem multi-adaptativa*[12]. Essas técnicas são a seguir sumariamente descritas: 1) Systematic count-based (SystC): a seleção de pacotes é efetuada através de uma função deterministica e invariável baseada na posição de pacotes, usando contadores [10]. Como exemplificado na Figura 2 (a), cada quinto pacote é selecionado e capturado pelo processo de amostragem. 2) Systematic time-based (SystT): o processo de selecção de pacotes segue uma função determinista baseado no tempo de chegada ao ponto de medição [10]. Nesta técnica, o tamanho da amostra e o tempo entre amostras são definidas no início e permanecem inalteradas ao longo do processo de amostragem, como apresentado na Figura 2 (b). Como ilustrado, todos os pacotes que chegam ao ponto de medição ao longo de um período de 100 ms são selecionados para a amostra, considerando que todos os pacotes seguintes ao longo de 200ms são ignorados para fins de medição. 3) Random count-based (RandC): seleciona os pontos iniciais dos intervalos de amostragem de acordo com um processo aleatório. Como apresentado na Figura 2 (c), na aproximação aleatória n-out-of-N , n pacotes são selecionados aleatoriamente da população que consiste em N pacotes [10].

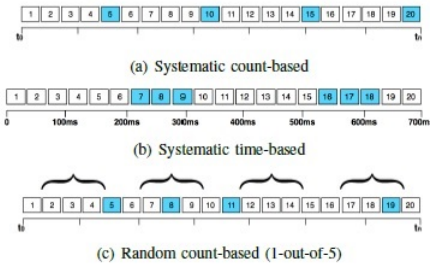



Figura 2. Exemplos de técnicas de amostragem

4) Adaptive linear prediction (LP): esta técnica baseada no tempo, usa predição linear para identificar alterações na actividade da rede, alterando a frequência de amostragem de acordo com o observado (i.e. reduzindo o intervalo entre amostras quando se observa mais tráfego que o previsto, para melhor deteção do novo padrão de tráfego) [11].

5) Multiadaptive sampling (MuST): esta técnica baseada no tempo usa mecanismos similares à anterior para identificar o nível de atividade da rede, no entanto, quer o intervalo entre

amostras quer o tamanho de cada amostra são parâmetros ajustáveis para melhorar o desempenho da amostragem [12].

Neste trabalho, as amostras obtidas usando as diversas técnicas descritas são processados pela ferramenta TIE para posterior classificação do tráfego. Comparando os resultados com os obtidos usando o tráfego global é possível avaliar *overhead* e acurácia, e selecionar qual a melhor técnica de amostragem para inferir corretamente sobre a classificação do tráfego total. A secção seguinte aborda em mais detalhe a metodologia seguida para a concretização deste objetivo.

## IV. Metodologia de Testes

A Figura 3 ilustra os aspectos considerados nos testes realizados, tendo como base *traces* de tráfego real capturados na rede do Instituto Nacional de Estatística (INE).

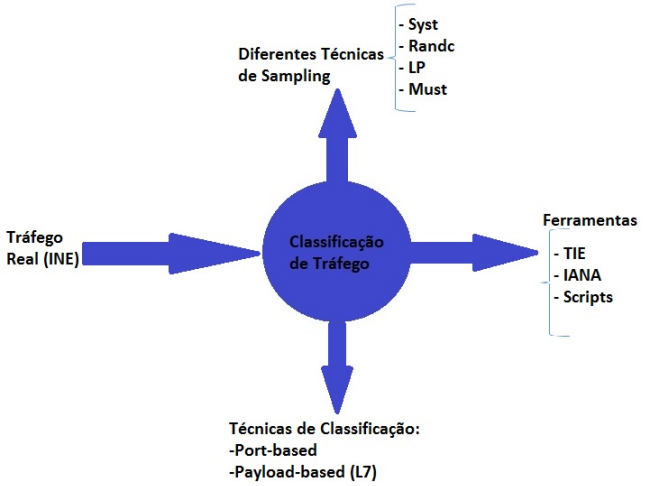



Figura 3. Arquitetura de testes

### A. Classificação e amostragem

Na classificação de tráfego foi utilizada a ferramenta TIE e os seus plugins que implementam diferentes técnicas de classificação, nomeadamente port-based e *payload*-based, permitindo também a sua conjugação. Posteriormente, foi implementada uma ferramenta para interpretação dos resultados do output do TIE e, por consequência da análise do volume de tráfego não classificado, foi adicionada à ferramenta de interpretação dados fornecidos por parte da entidade reguladora da atribuição dos números de portas e endereços IP à escala mundial IANA - Internet Assigned Numbers Authority[1].

Face à necessidade de comparar o impacto da amostragem na classificação do tráfego total, inicialmente, o *trace* de tráfego do INE é injetado numa framework de amostragem cujo output são os vários *traces* com o tráfego amostrado através de cada uma das técnica de amostragem mencionadas. Neste ponto, foi desenvolvida uma script na linguagem de programação C para agilizar a interação entre os *traces* amostrados e o TIE para a classificação de tráfego, obtendo-se os resultados gerados pelo TIE para todos os *traces* de tráfego amostrado que serão, por fim, interpretados com o auxílio da ferramenta de interpretação de resultados desenvolvida. Esta sequência de passos é ilustrada na Figura 4.

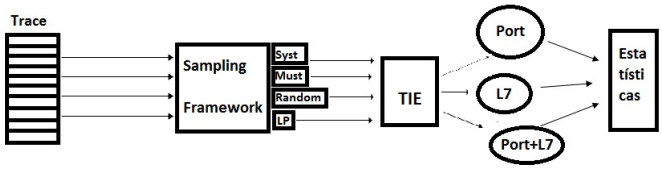



Figura 4. Esquema dos testes realizados

### B. Interpretação de resultados

A ferramenta de interpretação de resultados foi desenvolvida na linguagem JAVA e tem como objetivo apresentar os resultados da classificação do tráfego numa forma mais intuitiva. Numa primeira instância, a ferramenta lê o output gerado pelo TIE e combina esse output com a lista de aplicações reconhecida pelo TIE. Dado que a taxa de tráfego desconhecido se mantinha elevada, foi criada uma segunda versão da ferramenta que, em vez de consultar a lista de aplicações do TIE, consulta uma lista fornecida pelo IANA permitindo uma classificação mais atualizada em termos de aplicações. Ambas as versões da ferramenta geram o output em forma de gráficos.

## V. Discussão dos Resultados

Os testes realizados, numa primeira fase, focam a classificação multinível utilizando tráfego total e, posteriormente, o impacto das diferentes técnicas de amostragem. Este último aspecto é analisado sobre diversas perspectivas: (i) a capacidade de identificação de fluxos existentes e Heavy Hitters (i.e., os fluxos mais significativos em termos de volume de dados); (ii) a acurácia na identificação dos protocolos da camada de transporte e aplicação. São ainda apresentados resultados das técnicas de amostragem com diferentes frequências de amostragem, para avaliar o compromisso entre *overhead* e acurácia na classificação. Os parâmetros estatísticos comparativos utilizados incluem: número de fluxos processados e identificados, taxa de identificação de aplicações, grau de confiança da classificação do tráfego, erro médio absoluto (MAE) e erro quadrático (MSE).

### A. Análise de tráfego Total

Nos vários resultados apresentados a seguir foi utilizado o *trace* do INE correspondente ao tráfego total[2]. Na Figura 5 apresentam-se os dados referentes à classificação do tráfego pelo plugin *port* e na Figura 6 os dados referentes à classificação pelo plugin *L7*. As diferenças da classificação feita pelos plugins *port* e *L7* em termos globais e de grau de confiança não são muito significativas, notando-se maiores diferenças em termos das aplicações identificadas de baixo volume (representadas na figura por Others). Relativamente

[1]IANA - http://www.iana.org/

[2]Este *trace* corresponde a coletas de 20 min em diferentes periodos de atividade da rede. No global, o *trace* possui 252.087 fluxos individuais unidireccionais em cerca de 3 milhões de pacotes.

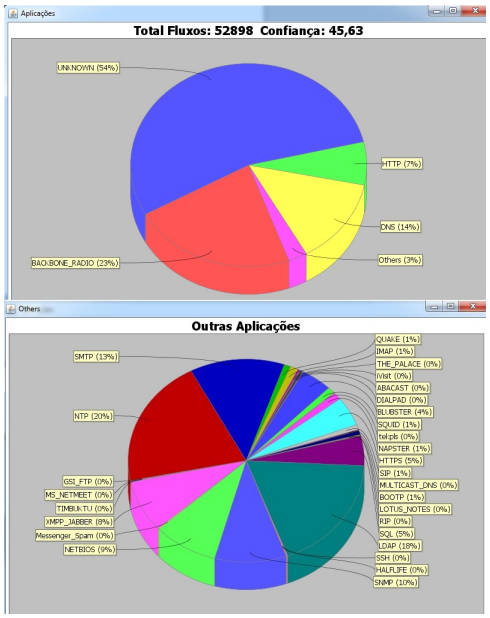



Figura 5. Classificação do tráfego *port-based*

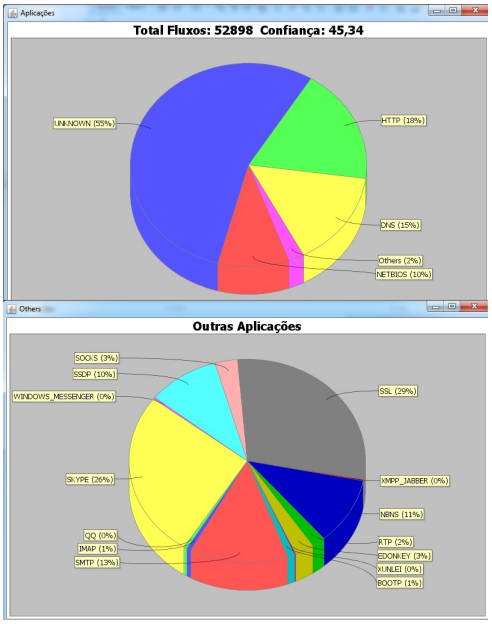



Figura 6. Classificação do tráfego *payload-based*

à aplicação conjunta dos dois plugins, a taxa de tráfego desconhecido diminuiu ligeiramente, salientando-se que as classificações *port+L7* e *L7+port* apresentam diferenças (na documentação da ferramenta TIE[1], encontra-se indicado que a ordem pela qual os plugins estão ativos tem impacto na classificação final).

*1) Análise com IANA:* Como referido anteriormente, no sentido de aumentar o volume de aplicações identificadas por porta, a base de dados de aplicações fornecida pelo TIE foi complementada com a lista de aplicações fornecida pela IANA. Pela análise da Figura 7, verifica-se que na classificação

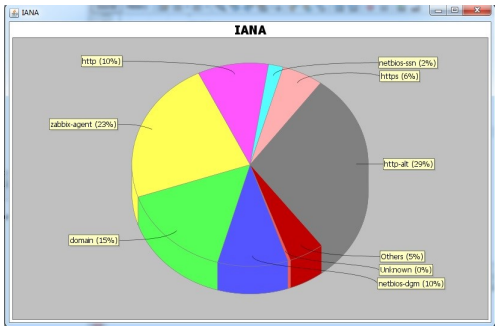



Figura 7. Classificação do tráfego com informação da IANA

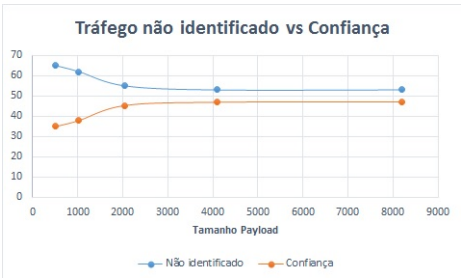



Figura 8. Impacto do tamanho do *payload*

por portas da IANA se obtém 0% de tráfego não identificado, sendo reconhecidas um muito maior número de aplicações.

*2) Impacto do tamanho do payload na classificação:* Um dos parâmetros exigidos na execução do plugin L7 é o tamanho (em bytes) do *payload* que irá ser analisado para a classificação de tráfego, sendo o tamanho padrão 2048 bytes. No sentido de avaliar o impacto desse valor no volume de tráfego identificado, realizaram-se experiências com 5 tamanhos alternativos de *payload*, mais concretamente 512, 1024, 2048, 4096, 8192 bytes. A Figura 8 compara o tamanho do *payload* com a percentagem de aplicações identificadas no tráfego e o grau de confiança obtido. Como ilustrado, o aumento do tamanho do *payload* analisado reduz o número de fluxos de tráfego com aplicações desconhecidas e aumenta o grau de confiança dos resultados obtidos. Estas variáveis tendem a estabilizar após um *payload* de cerca de 3000 bytes, em que o aumento de dados analisados se torna infrutífero.

### *B. Análise de tráfego Amostrado*

A amostragem de tráfego tem como principal objetivo reduzir o impacto da monitorização no funcionamento da rede, mantendo a acurácia na estimação de parâmetros referentes ao seu comportamento estatístico. Nas secções seguintes são apresentados os resultados alcançados através da comparação dessas técnicas no contexto da classificação, nomeadamente na análise aos fluxos identificados e *Heavy Hitters*, na identificação de protocolos de transporte e aplicações. Numa primeira fase é utilizada a técnica port-based e, posteriormente, é analisado o impacto da alteração da técnica de classificação, considerando-se classificação port-based e *payload*-based.

*1) Identificação de fluxos existentes e Heavy Hitters:* Previsivelmente, a frequência de amostragem é diretamente proporcional à acurácia das estimativas, dado que o aumento do tráfego coletado resulta num maior conjunto de dados para análise estatística. Enquanto isto pode ser verdade na análise de uma técnica de amostragem, quando se altera o método de amostragem, o compromisso entre o aumento do tráfego amostrado e a acurácia das estimativas pode ter diferenças significativas. Esta secção foca a análise efectuada com a técnica SystC com várias frequências de amostragem e, posteriormente, compara as diferentes técnicas entre si. São usados dois parâmetros comparativos, nomeadamente: (i) a quantidade de fluxos identificados; e (ii) a percentagem de fluxos *heavy hitters* (HH) identificados, em que a noção de *heavy hitter* se refere aos fluxos mais significativos em termos de volume de dados (neste artigo considerados como os fluxos maiores que representam 20% do volume de tráfego).

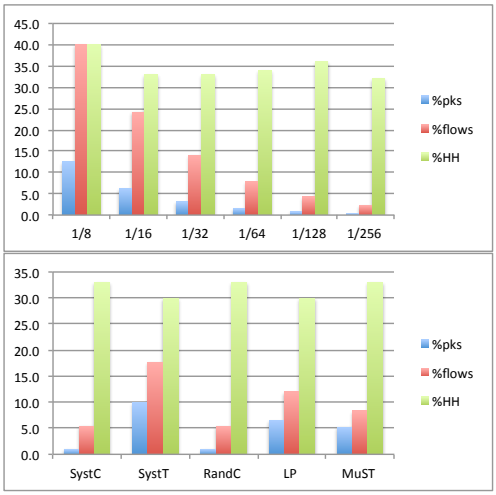
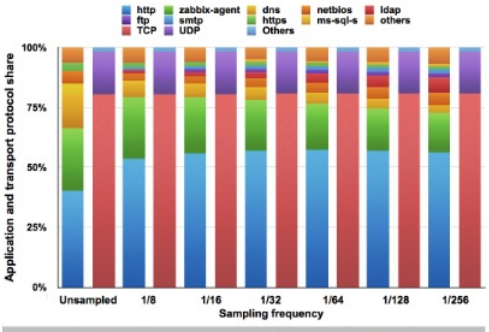



Figura 9. Identificação de fluxos - análise comparativa

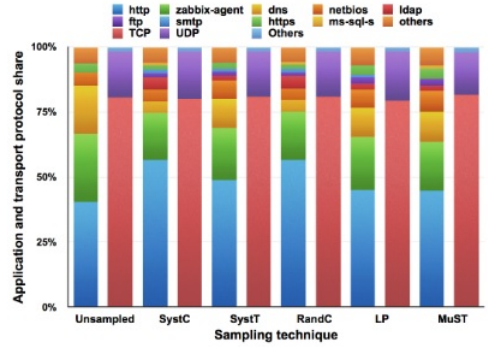
Analisando a Figura 9 verifica-se que com a técnica de amostragem SystC diminuindo o número de pacotes coletados, diminui também o número de fluxos identificados. Como exemplo, a frequência SystC 1/8 representa que é selecionado 1 pacote em cada 8. Numa análise mais pormenorizada do parâmetro dos fluxos *heavy hitters* conclui-se que a redução da frequência de amostragem não implica uma diferença significativa na identificação destes fluxos.

Relativamente às diferentes técnicas de amostragem, denota-se que um maior número de pacotes implica um maior número de fluxos identificados. No entanto, as técnicas *count-based* são mais eficientes, pois para o mesmo número de pacotes amostrados a percentagem dos fluxos identificados é significativamente superior. Isto acontece devido às políticas de seleção de pacotes distintos em uso, ou seja, o processo de seleção de pacotes baseado nas técnicas *count-based* aumenta a probabilidade de captura de fluxos distintos, ao contrário das técnicas *time-based* em que os pacotes são selecionados sequencialmente.




| | 1/8 | 1/16 | 1/32 | 1/64 | 1/28 | 1/56 |
|---|---|---|---|---|---|---|
| MAE | 0.0099 | 0.0464 | 0.1738 | 0.1337 | 0.1232 | 0.3045 |
| MSE | 0.0017 | 0.0026 | 0.0383 | 0.0262 | 0.0189 | 0.1383 |

Figura 10. Análise na camada de transporte e aplicação - SystC




| | SystC | SystT | RandC | LP | MuST |
|---|---|---|---|---|---|
| MAE | 0.2464 | 0.0999 | 0.0462 | 0.7521 | 0.7634 |
| MSE | 0.0817 | 0.0112 | 0.0330 | 0.7992 | 0.7671 |

Figura 11. Análise na camada de transporte e aplicação - Comparação entre técnicas

*2) Identificação de protocolos de transporte e aplicação:* Considerando uma análise ao nível da camada de transporte, a redução do número de pacotes amostrados promovido pela diminuição da taxa de frequência de amostragem da técnica SystC, não afeta a acurácia da classificação de tráfego, como apresentado na Figura 10 e confirmado pelos baixos erro médio absoluto (MAE) e erro quadrático (MSE) em todas as frequências. Porém, considerando a camada de aplicação, a distribuição das aplicações é bastante afetada, resultando numa elevada quantidade de tráfego *HTTP* identificado.

Embora a alteração da técnica de amostragem não tenha impacto na acurácia na classificação da camada de transporte (ver Figura 11), em termos da camada aplicacional isso já não se verifica. Como apresentado também na Figura 11, as técnicas baseadas em tempo conseguem ter um desempenho mais realista na distribuição da taxa de utilização da camada aplicacional, em que a técnica *MuST* obtém os melhores resultados. Globalmente, os resultados evidenciam que pequenas frações do tráfego de rede permitem obter uma visão bastante útil e realista dos fluxos de rede na pilha protocolar.

*3) Impacto da técnica de classificação:* De seguida, encontram-se os resultados da classificação do tráfego com as técnicas de amostragem acima referidas e as diferentes

| Técnicas de Amostragem | Numero de Fluxos | Unknown (%) | Grau de Confiança (%) |
|---|---|---|---|
| Systc1-8 | 26382 | 84 | 15,64 |
| Systc1-16 | 16767 | 87 | 13,21 |
| Systc1-32 | 10207 | 88 | 11,66 |
| Systc1-64 | 5871 | 89 | 11,12 |
| Systc1-128 | 3252 | 89 | 10,58 |
| Systc1-256 | 1768 | 89 | 11,09 |
| Systc1-512 | 960 | 88 | 12,29 |
| Systc1-1000 | 487 | 91 | 9,45 |

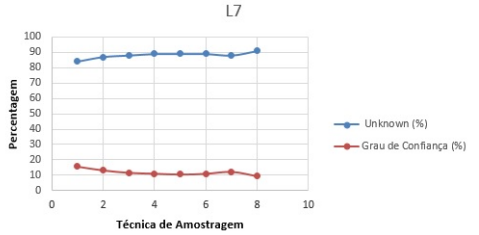



Figura 12. Amostragem sistemática com classificação *payload-based*

| Técnicas de Amostragem | Numero de Fluxos | Unknown (%) | Grau de Confiança (%) |
|---|---|---|---|
| Systc1-100 | 4032 | 89 | 10,99 |
| Systt | 9784 | 68 | 31,67 |
| Randc | 3975 | 89 | 11,37 |
| LP | 6918 | 67 | 32,58 |
| MUST | 4088 | 65 | 34,89 |

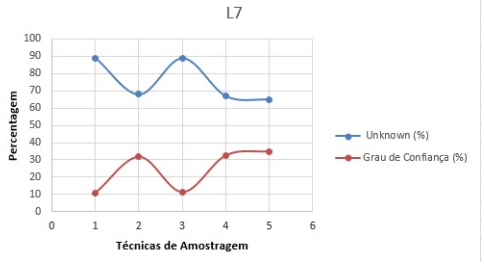



Figura 13. Técnicas de amostragem com classificação *payload-based*

estratégias de classificação port-based e L7. A técnica SystC é aplicada para as frequências *Systc1-8, Systc1-16, Systc1-32, Systc1-64, Systc1-128, Systc1-256, Systc1-512 e Systc1-1000.* num dos *traces* mais significativos do INE.

Os resultados obtidos com a classificação port-based de *payload*-based foram bastante similares. Pela análise da Figura 12 concluí-se que qualquer uma das técnicas de classificação de tráfego tem uma elevada taxa de tráfego desconhecido, o que já acontecia considerando o tráfego total. Ao considerar um estudo mais minucioso dos gráficos e das curvas de tráfego desconhecido e grau de confiança observa-se que estas duas métricas estatísticas são inversamente proporcionais.

Quando comparando as várias técnicas de amostragem (ver Figura 13), mais uma vez, é notório que as curvas dos parâmetros estatísticos são inversamente proporcionais. Numa análise mais pormenorizada denota-se que a técnica de amostragem com melhores resultados é a amostragem multi-adaptativa (*MUST*) pois obtém sempre taxas de tráfego desconhecido inferiores e graus de confiança maiores que as outras técnicas. Também se salienta que a classificação de tráfego *port+L7* obtém os melhores resultados, independentemente da técnica de amostragem utilizada. A combinação da técnica de amostragem (*MUST*) com a técnica de classificação de tráfego *port+L7* tem os melhores resultados, obtendo um grau de aplicações identificadas superior a 50%.

## VI. Conclusões

Neste artigo foi apresentada uma análise comparativa do impacto da utilização de diferentes técnicas de classificação e amostragem na classificação de tráfego de rede. Nesse contexto, avaliou-se a aplicabilidade real das atuais e emergentes técnicas de amostragem para a identificação e classificação de fluxos, tendo em conta que este é um importante passo para tornar essas tarefas mais tratáveis evitando estimativas menos realistas da utilização da rede. Neste aspeto, as técnicas *count-based* revelaram ser mais eficientes uma vez que permitem identificar mais fluxos, com a mesma quantidade de tráfego amostrado. Adicionalmente, é notório que uma maior frequência de amostragem não conduz necessariamente a uma maior acurácia nas estimativas e, dependendo dos objetivos, pode ser vantajoso o recurso a uma amostragem de baixa frequência ou uma técnica de amostragem específica quando se pretende melhorar o compromisso entre o excesso de dados a processar e armazenar, e a acurácia dos resultados quando se estuda o comportamento dos fluxos de rede.

*AGRADECIMENTOS* - Este trabalho é suportado pela FCT - Fundação para a Ciência e Tecnologia no âmbito do projeto UID/CEC/00319/2013.

## Referências

[1] W. de Donato, A. Pescapé, and A. Dainotti, "Traffic identification engine: an open platform for traffic classification," *Network, IEEE*, vol. 28, no. 2, pp. 56–64, 2014.

[2] 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.

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

[4] A. Dainotti, W. De Donato, A. Pescapé, A. Botta, G. Aceto, and G. Ventre, "Tie - traffic identification engine," http://tie.comics.unina.it/doku.php?id=topmenu:home.

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

[6] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: multilevel traffic classification in the dark," in *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 4. ACM, 2005, pp. 229–240.

[7] G. Androulidakis, V. Chatzigiannakis, and S. Papavassiliou, "Network anomaly detection and classification via opportunistic sampling," *Network, IEEE*, vol. 23, no. 1, pp. 6 –12, 2009.

[8] Y. Gu, L. Breslau, N. Duffield, and S. Sen, "On passive one-way loss measurements using sampled flow statistics," in *INFOCOM 2009, IEEE*, april 2009, pp. 2946 –2950.

[9] D. Tammaro, S. Valenti, D. Rossi, and A. Pescapé, "Exploiting packet-sampling measurements for traffic characterization and classification," *International Journal of Network Management*, pp. 451–476, 2012.

[10] T. Zseby, M. Molina, and N. Duffield, "Sampling and Filtering Techniques for IP Packet Selection," RFC 5475, Internet Engineering Task Force, Mar. 2009. [Online]. Available: http://datatracker.ietf.org/doc/rfc5475/

[11] 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.

[12] 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.