Mini-Projeto 2: Medição de tráfego Internet

Os operadores de redes de computadores coletam tráfego diretamente dos seus routers com objetivo de analisar os padrões de comunicação dos serviços que correm na infraestrutura, por exemplo, para perceber se os recursos estão a ser usados de forma eficiente ou para gerir a largura de banda de forma a evitar congestão ou para detetar ataques à infraestrutura.
No MP2 os alunos vão fazer o papel de um operador de rede, através da análise de um trace de tráfego real capturado num router do backbone da Internet. Um trace é um ficheiro que contém os dados de uma captura de tráfego (tal como uma captura Wireshark ou tcpdump). O objetivo do trabalho MP2 é estudar as propriedades do tráfego capturado para perceber os seus padrões, de forma a extrair conclusões acerca do comportamento da rede.

Descrição

Cada grupo vai ter um trace diferente para analisar. Cada trace contém um minuto de tráfego coletado num router na cidade de Chicago (US). O tráfego foi anonimizado para proteger a privacidade dos utilizadores. Podem descarregar o trace do vosso grupo aqui:

Shared Folder Link
Password: RC-LEI_2021-2022.

O vosso trace tem o nome **MP2_GrupoXX.csv.tar.gz**, onde **XX** é o número do vosso grupo. Os registos usam o formato CSV, como neste exemplo que mostra as cinco primeiras linhas de um trace:

"No.","Time","Source","Destination","SourcePort","Destination Port","Protocol","Length","Flags"
"1","0.000000","32.238.69.98","63.8.52.229","61885","80","TCP","50","0x010"
"2","0.000013","199.45.252.79","86.62.91.76","","","NTP","400",""
"3","0.000018","150.227.89.173","211.23.106.124","","","UDP","62",""
"4","0.000022","83.87.9.183","63.8.48.222","61213","1935","TCP","1458","0x018"

A primeira linha mostra o cabeçalho do trace, identificando cada uma das colunas. Cada uma das linhas seguintes representa um pacote capturado. A primeira coluna representa o número de captura (o número “1” é o primeiro pacote capturado, o número “2” o segundo, etc.). A segunda mostra o momento (relativo) em que o pacote foi capturado (por exemplo, o segundo pacote foi capturado t=0.000013 segundos depois do primeiro). Nas duas colunas seguintes surge a informação do endereços IPs emissor e destinatário, respetivamente. Neste caso aparecem apenas endereços IPv4, mas o trace pode incluir endereços IPv6, que têm um formato diferente. De seguida, aparece informação da camada de transporte: o porto emissor e o porto destino, respetivamente (por exemplo, o destino do primeiro pacote é o porto 80, o que significa que foi enviado para um servidor Web). Na 7a coluna surge informação sobre o protocolo da última camada Internet (isto é, se diz “TCP” é um segmento TCP, encapsulado num datagrama IP, encapsulado, provavelmente, numa trama Ethernet). Segue-se o tamanho do pacote, em bytes, e a última coluna representa os 9 bits das flags TCP. Por exemplo, o primeiro pacote é um ACK (010000), e o pacote número 4 é um PUSH/ACK (011000), de acordo com as flags TCP.

Por meio de uma análise do trace, os alunos devem responder às perguntas seguintes:

Q1. [1 valor]: Qual o número de pacotes que possuem emissor e recetor IPv4?

Q2. [1 valor]: Qual o número de pacotes que possuem emissor e recetor IPv6?

Q3. [2 valores]: Quantos hosts IPv4 receberam pacotes? Por outras palavras, quantos destinatários com endereços IPv4 únicos é que apareceram no trace?

Q4. [3 valores]: Quantos portos TCP origem únicos apareceram no trace?

Q5. [4 valores]: Qual o tamanho médio, máximo e mínimo dos pacotes do trace?

Q6. [4 valores]: Assumindo que o envio de um RST representa uma tentativa de comunicação falhada, indique quantas tentativas comunicações TCP falharam neste trace.

Q7. [5 valores]: Imprima o CCDF do tamanho dos pacotes TCP que aparecem no trace. Usa as escalas que pareçam mais apropriadas. Nota: no Anexo 1 é explicado como se cria um gráfico deste tipo.

Entrega

Todas as perguntas devem ser respondidas diretamente no ficheiro **Respostas_GrupoXX.csv** (à exceção da pergunta Q7, ver a seguir), mantendo o formato definido no ficheiro exemplo **resultados_teste.csv**. Para a pergunta Q7 os alunos devem entregar o gráfico no formato PDF: **Q7_GrupoXX.pdf**. Os alunos devem entregar também os programas/scripts que desenvolveram para responder a todas as perguntas. Todos estes ficheiros devem ser incluídos num zip único: **MP2-GrupoXX**.zip.

O trabalho deverá ser entregue até ao **final do dia 19 de Novembro de 2021**.

Referências de apoio ao projeto

Incluímos um trace de teste (**testing_trace.csv.tar.gz**) para poderem testar o vosso código, juntamente com o ficheiro com o output esperado (**resultados_teste.csv**). Como os traces são relativamente grandes aconselhamos os alunos a testarem os seus scripts/programas usando inicialmente apenas um subconjunto do trace.
No zip que contém este documento encontram o programa **TrafficAnalysis.java** que podem usar como base, e que inclui a) uma análise básica do trace de teste (comprimido) e b) o código para criação de um script gnuplot simples. Inclui-se também uma folha de cálculo **CDF.xlsx** com um exemplo de criação de um gráfico CDF, e um **makefile** (que deve ser alterado se necessário para poder compilar os programas dos alunos).
O código java fornecido pode ser compilado e executado através do ficheiro makeFile fornecido, usando para tal a linha de comandos e o comando *make* para compilar e o comando *make run* para executar.
Quem quiser usar o gnuplot para criar o gráfico da pergunta Q.7 tem material de apoio na página oficial [2]. Além disso, há vários tutoriais simples disponíveis online (e.g., [3,4]).

Avaliação do MP2

O MP2 vai ser avaliado pelos docentes da disciplina com base em 3 critérios:
1. Scripts de teste para verificação das respostas, e análise do gráfico CCDF.
2. Análise de código (programas e scripts), incluindo como critérios de avaliação o estilo de programação e o uso apropriado de comentários no código (opcional);
3. Discussão com os alunos (opcional).

Dúvidas e comunicação

Como neste projeto se procura fomentar a autonomia dos alunos, por regra os docentes não responderão a dúvidas na sala de aula. Assim, os alunos devem utilizar o Fórum da disciplina ou o horário de atendimento dos docentes para tentarem esclarecer as dúvidas que vão surgindo.
Nota importante: Não é permitido os alunos partilharem código com soluções, ainda que parciais, a nenhuma parte do MP2, com alunos de outros grupos (nem através do Fórum, nem por qualquer outro meio). Se isso acontecer são anulados os projetos de todos os alunos envolvidos.

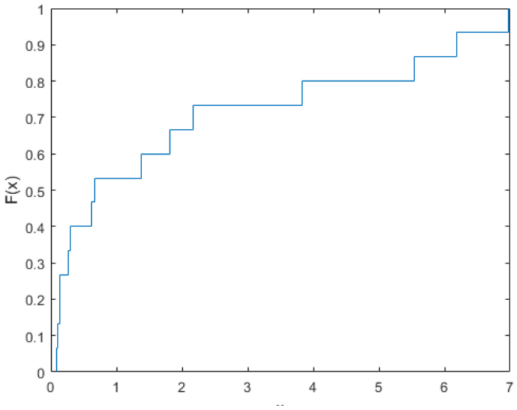
Referências

[1] https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml
[2] http://www.gnuplot.info/
[3] http://physics.ucsc.edu/~medling/programming/gnuplot_tutorial_1/index.html
[4] http://lowrank.net/gnuplot/intro/basic-e.html

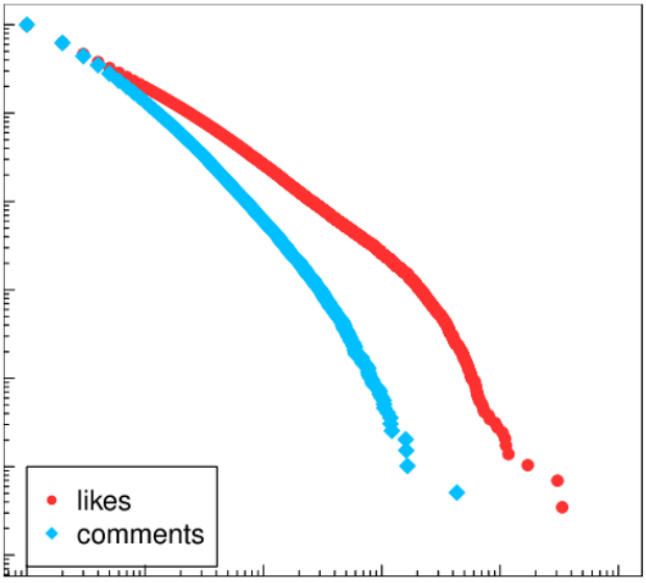
Anexo I: Sumarizando um conjunto de dados

Os investigadores normalmente tentam sumarizar largas quantidades de dados usando funções distribuição. Como exemplo, imaginem que têm uma lista de páginas Web com vários tamanhos (em número de bytes). A função distribuição cumulativa (Cumulative Distribution Function, CDF) dos tamanhos das páginas conteria, no eixo das ordenadas (yy) a "fração de páginas Web que têm tamanho menor ou igual a x bytes" e um eixo das ordenadas (xx) com o número de bytes. O gráfico começaria em y=0 (nenhuma página Web tem 0 bytes ou menos) e chegaria a y=1 quando o x atingir o valor da página maior.

Vejam a Figura 1 como exemplo, e imaginem que representa o CDF do tamanho das páginas Web de um dado conjunto de páginas Web. O eixo dos xx representa, assim, o tamanho de uma página Web, em kB. Neste CDF é possível observar que as páginas mais pequenas têm cerca de 100 bytes, ou menos, e que a mediana (quando y = 0.5) é de cerca de 0.75 kB. As páginas maiores (menos de 10% do total) têm 7 kB.



Por vezes também é importante fazer a pergunta oposta, e aí pode imprimir-se o CCDF (Complementary Cumulative Distribution Function, CCDF), que, dando o mesmo exemplo, é "a fração de páginas Web maior do que x bytes". O gráfico desta vez começaria em y=1, já que todas as páginas têm mais de 0 bytes, e iria decrescer gradualmente até chegar a y=0, quando o valor de x está no tamanho da página maior. Os eixos podem ser representados em escalas lineares ou logarítmicas, para se poder ver em algum detalhe determinadas regiões da curva. Vejam a Figura 2, que apresenta um CCDF do número de likes e comentários de um conjunto empírico de dados da rede social Facebook, usando escalas logarítmicas.



Neste mini-projeto, os alunos vão desenvolver e imprimir CCDFs, usando escalas lineares ou logarítmicas (a escolha é dos alunos, que devem decidir qual a melhor forma de apresentar dado resultado).
Para os gráficos podem usar qualquer ferramenta à vossa escolha (e.g., gnuplot, Matlab, Excel, R). Para a geração de distribuições muito software tem já toolboxes e há também bibliotecas específicas para esse propósito em várias linguagens de programação. Também não é complicado fazer um pequeno programa para gerar as distribuições que vão necessitar de criar neste MP. Nós fornecemos o código para gerar um script gnuplot, que podem usar como base.