COE241 - Estatística e Modelos Probabilísticos
Segundo Semestre de 2015 - Professora: Rosa Maria Meri Leão

Projeto do Curso

1 Objetivo

O objetivo deste trabalho é analisar dois logs de pacotes coletados na Internet usando o comando tcpdump. O tcpdump é uma ferramenta utilizada para monitorar os pacotes trafegados numa rede de computadores. O primeiro log contém pacotes que foram gerados pelo protocolo TCP e o segundo log contém pacotes gerados pelo protocolo UDP. O principal objetivo da análise dos logs é estudar o tempo de chegada dos pacotes e o tamanho dos pacotes.

2 Formato do Log

Cada linha do log de ações possui dados de um pacote TCP ou UDP. O formato do log está descrito abaixo.

1. **timestamp do tempo de chegada do pacote**

   Para o primeiro pacote, este campo é o timestamp do pacote. Para os pacotes seguintes é o offset da parte inteira do primeiro pacote. Por exempo se os timestamps dos três primeiros pacotes são **187.2, 188.9, 191.3**, no log teremos **187.2, 1.9 (188.9-187), 4.3 (191.3-187)**.

2. **IP renumerado do host de origem**

3. **IP renumerado do host de destino**

4. **porta do host de origem**

5. **porta do host de destino**

6. **tamanho da parte de dados do pacote** O log dos pacotes UDP não possui este último campo.

3 Análises a serem realizadas

3.1 Histograma

Você deve calcular o histograma para as seguintes variáveis aleatórias:

(1) timestamp do tempo de chegada do pacote TCP,

(2) timestamp do tempo de chegada do pacote UDP,

(3) tamanho do pacote TCP.

Lembre-se que na hora de calcular o histograma, você deve escolher o tamanho do *bin* adequado. Descreva no seu relatório como você calculou o tamanho do *bin* do seu histograma.

3.2 Função Distribuição Empírica

Você deve calcular a função distribuição empírica para as seguintes variáveis aleatórias:

(1) timestamp do tempo de chegada do pacote TCP,

(2) timestamp do tempo de chegada do pacote UDP,

(3) tamanho do pacote TCP.

Faça o gráfico da distribuição complementar empírica (x x $1 - F_X(x)$) com o eixo das ordenadas em escala log e observe o comportamento da cauda da distribuição. Comente sobre o resultado no seu relatório.

3.3 Média, Variância e BoxPlot

Você deve calcular a média, variância e BoxPlot para as seguintes variáveis aleatórias:

(1) timestamp do tempo de chegada do pacote TCP,

(2) timestamp do tempo de chegada do pacote UDP,

(3) tamanho do pacote TCP.

Descreva o que você pode observar através das medidas obtidas.

3.4 Parametrizando distribuições

Neste item o objetivo é você parametrizar um conjunto de distribuições da literatura usando os dados do log. Após a parametrização você irá verificar se alguma das variáveis aleatórias do log pode ser representada por uma distribuição da literatura. Utilize o método dos momentos ou o método da máxima verossimilhança para parametrizar as seguintes distribuições: exponencial, gaussiana, lognormal, weibull.

As variáveis aleatórias que você deve considerar são: **(1) timestamp do tempo de chegada do pacote TCP**, **(2) timestamp do tempo de chegada do pacote UDP**, **(3) tamanho do pacote TCP**. Ou seja, você deve obter para cada dessas variáveis o valor dos parâmetros das distribuições citadas acima.

Após a obtenção dos valores dos parâmetros, você deve fazer um gráfico para cada uma das variáveis aleatórias com a função distribuição empírica (obtida na seção 3.2) e as quatro distribuições que você parametrizou. Observando o gráfico você deve identificar se existe ou não uma distribuição da literatura que poderia ser usada para representar a variável aleatória. Note que esta é só uma comparação visual. Para ter certeza de poder usar é necessário formular um teste de hipótese.

3.5 Gráfico QQplot ou ProbabilityPlot

Os gráficos QQplot ou ProbabilityPlot servem para comparar a distribuição de duas variáveis aleatórias. Você deve traçar os gráficos para os seguintes casos:

1. **timestamp do tempo de chegada do pacote do pacote TCP** x **timestamp do tempo de chegada do pacote do pacote UDP** O objetivo é observar se o tempo de chegada dos pacotes gerados pelos dois protocolos possui semelhança com relação a distribuição de probabilidade.
2. **timestamp do tempo de chegada do pacote do pacote TCP** x cada uma das distribuições parametrizadas (exponencial, gaussiana, lognormal, weibull).
3. **timestamp do tempo de chegada do pacote do pacote UDP** x cada uma das distribuições parametrizadas (exponencial, gaussiana, lognormal, weibull).

4. **(3) Tamanho do pacote TCP** x cada uma das distribuições parametrizadas (exponencial, gaussiana, lognormal, weibull).

   Neste item você pode observar qual a melhor distribuição para cada uma das variáveis aleatórias.

4 Relatório

Você deve fazer um relatório contendo todos os resultados que você obteve e explicando como você os obteve. É importante comentar cada um dos resultados e explicar como o resultado que você obteve poderá influenciar no planejamento e desempenho da rede observada. A avaliação do projeto será feita com base na qualidade do relatório.