MAC6958
Tópicos Avançados em Ciência de Dados para Redes de Computadores

- **Prof. Daniel Macêdo Batista**
  Prof. Roberto Hirata
- **DCC-IME-USP**

  - 5/3/2020

Classificação de tráfego

- **Processo preferencialmente automatizado que categoriza tráfego de redes de acordo com vários parâmetros dentro de um número limitado de classes de tráfego**
- **Útil para aplicar marcação de tráfego que influencia no roteamento, na alocação de recursos e na QoS (Qualidade de Serviço)**
  - Importante porque ninguém vai confiar no que o usuário diz sobre o tráfego dele
  - Pode ser o passo antes do tráfego chegar em um escalonador na rede (prioridades), no controle de admissão (aceita ou não aceita) ou no alocador de recursos (preparação da rede)
- **Útil para detectar anomalias**
  - Falhas de equipamentos/software
  - Ataques

Classificação de tráfego

- **No geral 3 classes:**
  - Tráfego sensível: aplicações que precisam de alta vazão, baixo atraso e/ou baixo jitter (multimídia principalmente mas no futuro: redes veiculares e telemedicina)
  - Tráfego de melhor esforço: aplicações que só precisam que os dados cheguem do outro lado mas sem muitos requisitos (E-mail, HTTP se não for vídeo, etc...). Geralmente o tráfego aqui sofre traffic shapping (Leaky Bucket por exemplo. No linux, aplicativo cbq)
  - Tráfego indesejado: precisa ser bloqueado (spam, worms, botnets, DoS, etc…)
    - Análise posterior pode ser realizada por exemplo se houve um ataque (análise forense)

Classificação de tráfego

- **Arquitetura Internet e informações úteis das camadas presentes nos cabeçalhos**
  - Enlace (802.3, 802.11): Endereço MAC de origem, endereço MAC de destino, tamanho do quadro
  - Rede (IPv4, IPv6): Endereço IP de origem, endereço IP de destino, TTL, DSCP (Differentiated Services Code Point), ECN (Explicit Congestion Notification)
  - Transporte (TCP, UDP): Porta fonte, porta destino, flags (ACK, RST, SYN, FIN)

- Classificação simples baseada nesses dados que costumam representar um fluxo
- Mas na prática não é garantido (portas diferentes, DSCP ignorado, NAT utilizado, etc...)

Classificação de tráfego

- **Arquitetura Internet e informações úteis do payload**
  - Identificação da aplicação mais garantida (HTTP, FTP, HTTPs, SMTP, DNS)
  - Classificação mais complexa porque diferente dos cabeçalhos, pode ter tamanho indefinido (muitos dados)
  - Para ter o dado completo pode ser necessário unir vários fragmentos
  - Privacidade
  - Não funciona muito bem se usar criptografia (pode desconfiar sobre o protocolo mas não é garantido)
  - Até aqui, possível utilizar para uma análise “inline” e em tempo real

Classificação de tráfego

- **Arquitetura Internet e informações úteis do payload**
  - Ir para um nível mais alto agindo como o usuário (se o usuário consegue ler, um mecanismo consegue ler)
  - Redes sociais – os provedores fornecem APIs para acessar os dados
  - Classificação mais complexa porque tem que lidar com PLN (muitos dados)
  - Não funciona muito bem se usuário restringir acesso a sua conta
  - Não funciona muito bem se usar API gratuita (GNIP → $$)

Classificação de tráfego

- **Arquitetura Internet**
  - Difícil quando a classificação é para tráfego malicioso principalmente se o ataque é composto por vários fluxos ou ainda se é um ataque 'zero-day'
  - Classificação mais complexa porque tem que lidar com falsos positivos/negativos

Classificação de tráfego

- **Cenário ideal para classificar**
  - Ter acesso ao pacote `raw' (.pcap)
  - Criar modelos que analisem quais propriedades (features) implicam em um dado tipo de tráfego
  - Treinamento constante do modelo

  - Difícil por conta de privacidade, quantidade grande de features e definição de quando retreinar
  - O ineditismo do tráfego pode definir a complexidade da classificação

Classificação de tráfego

- **Em resumo:**
  - Informações dos cabeçalhos: rápido, não depende de muitos recursos para ser realizado, mas não confiável
  - DPI (Deep Packet Inspection): lento, depende de muitos recursos para ser realizado, mas afetam privacidade e não funciona se há criptografia
  - Classificação estatística: olha atributos do fluxo como tamanho dos pacotes e tempo entre chegadas. É onde se costuma aplicar técnicas de aprendizado de máquina. **Difícil é descobrir as features de interesse e ter bons datasets**

Referências

- T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," in IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56-76, Fourth Quarter 2008. doi: 10.1109/SURV.2008.080406
- Raouf Boutaba, Mohammad A. Salahuddin, Noura Limam, Sara Ayoubi, Nashid Shahriar, Felipe Estrada-Solano, Oscar M. Caicedo, "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities". Journal of Internet Services and Applications, 2018, 9:16. doi: 10.1186/s13174-018-0087-2
- CAIDA, "Internet Traffic Classification", http://www.caida.org/research/traffic-analysis/classification-overview/
- Wikipedia, "Traffic classification", https://en.wikipedia.org/wiki/Traffic_classification