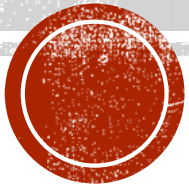
网络流量分类进展



前一段工作总结

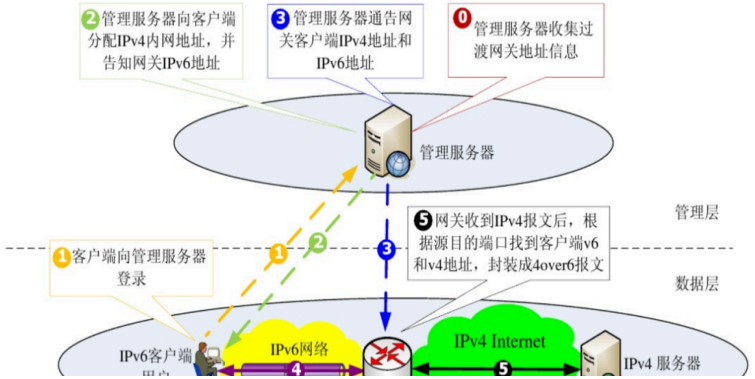
- 开发pcap抓包程序
- 中国高校计算机大赛——网络技术挑战赛
- 与公司合作抓取了半个小时的马上六数据
- Python解析pcap基本信息，确保其性能
- IP地址匿名化处理，cryptopANT-1.2.0

基于机器学习的流量数据分析

清华大学 计算机科学与技术系
队员：郭诚 李映辉 指导老师：徐明伟

一、 问题背景

IPv6 是下一代互联网协议，它的设计用以解决 IPv4 中的许多问题（最重要的一个就是解决了 IP 地址不够分配的问题）为目标。但是，在互联网大规模支持 IPv6 之前，IPv4 到 IPv6 如何平滑过渡是需要人们仔细考虑的一个问题。例如，由于互联网服务提供商的现有网络设备大部分都只支持 IPv4，我们有必要设计一种机制允许 IPv6 用户访问历史遗留的 IPv4 服务，而 IPv4 over IPv6 就是这样的一种机制。其大致原理是采用了“IPv4 over IPv6”隧道方案来实现，IPv6 网络中的用户在希望访问 IPv4 服务时，将 IPv4 包封装在 IPv6 包中发送给一台能够直接访问 IPv4 服务的服务器。该服务器将进行实际 IPv4 资源的访问，然后再将得到的 IPv4 响应封装在 IPv6 包中发送回用户。为了能够通过 IPv4 地址将 IPv4 响应返回给相应的用户，IPv4 over IPv6 服务器需要维护一个 IPv4 和 IPv6 地址的映射表并为需要 IPv4 over IPv6 服务的用户分配 IPv4 地址。原理大致如下图所示：



当前工作

- 调研具体的分类方案
- 基于端口号的分类方法
- 基于有效负载的分类方法
- 基于主机行为的分类方法
- 基于机器学习的分类方法

基于端口号的分类方法

- 传统且实现简单
- 熟悉端口号&动态端口&保留端口号
- 将熟知的端口号进行映射来识别不同的应用类型
- 非熟悉端口号&动态分配端口（P2P）

基于有效负载的分类方法

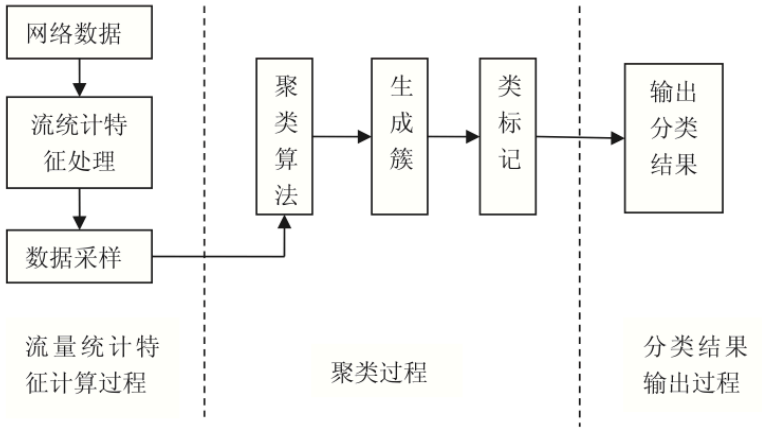
- 避免对端口号的过分依赖
- 分析包的有效负载是否包含已知应用的特殊签名
- 只能识别已知的非加密流量
- 对未知流量无能为力
- 计算代价太大

基于主机行为的分类方法

- 分析主机在传输层的行为模式来进行流量分类
- 无需解读负载
- 无需知道端口号
- 在路由器上采集NetFlow信息
- 不适合我们的项目（数据经过NAT处理）

基于机器学习的分类方法

- 无监督机器学习（聚类）
  - K-Means(划分)
  - DBSCAN（密度）
  - AutoClass（概率）
  - 性能：K-Means和DBSCAN更快
  - 准确率：AutoClass算法的准确率最高
- 有监督机器学习（分类）
  - Label的获取
  - 训练->(验证)->测试分类
- 半监督机器学习（√）
  - 已标记样本和未标记样本



总结

分类方法	基于端口号的流量分类方法	基于有效负载的流量分类方法	基于主机行为的流量分类方法	基于机器学习的流量分类方法
准确性	低	极高	较高	较高
使用场景	较简单的网络流量结构，使用IANA中注册端口号的传统网络应用。	数据包内容为明文的流量分类与识别	大流量的骨干网络中（汇聚的行为主机越多，流量识别准确率越高）	传统及新兴的网络应用，其他几种方法无法应对的场景。
优点	技术简单、易操作、计算开销较低，分类速度快。	分类准确率很高，并且可以对P2P等新型网络应用进行准确的分析。	能够对加密数据流量进行分析，开销较低、容易实施。	准确率较高、可扩展性强，能够对加密数据流量进行分析。
缺点	若网络应用没有使用IANA中注册的端口号，或者使用动态端口，则无法使用该方法。	计算开销太大，涉及到隐私侵犯问题，并且无法解决私有协议及加密流量的分析问题。	无法识别一些特定应用的子类型，当传输层数据被加密时无法使用。	耗费资源较多，需要事先了解数据样本集，应用于在线环境具有一定困难。
单独使用时对高速网络流量分类的适用性	不适合单独使用	适用范围有限	比较适用，但需其他方法配合使用	不太实用，处于实验阶段
单独使用时是否适用于加密流量分类	否	否	是	是
单独使用时是否适用于精细化流量分类	否	是	是	是
单独使用时是否适用于动态变化协议的流量分类	否	否	处于试验阶段	处于试验阶段

整体方案

- 流量数据获取（√）
- 解析获取信息（√）
- IP地址匿名化（√）
- 调研不同分类方法（正在进行）
- 包特征提取
- 基于端口号进行流量分类
- 基于半监督机器学习流量分类
- 总结实验结果到论文中

THANKS