面向状态的互联网应用心跳流检测算法

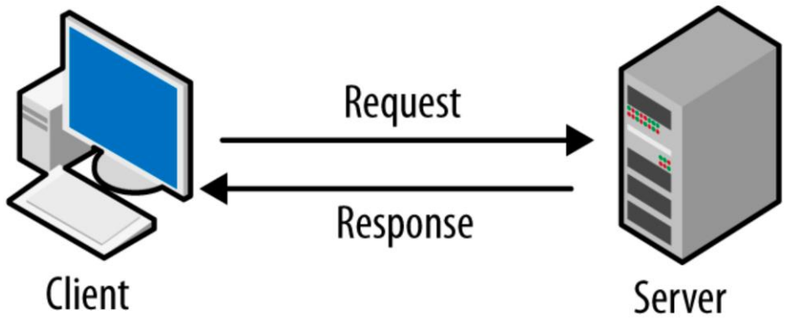
徐亮，郭文仙，丁伟

东南大学网络空间安全学院 江苏 南京 211189

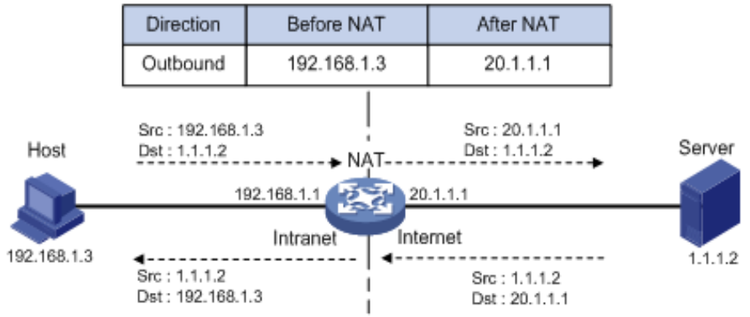
目录

01 背景介绍

心跳机制和心跳流



Client/Server模式



NAT地址转换基本过程

- **心跳机制**：应用程序基于保活需求而设计的机制
- **心跳报文**：应用程序周期性地发送特定数据包
- **心跳流**：应用基于心跳机制产生的报文序列

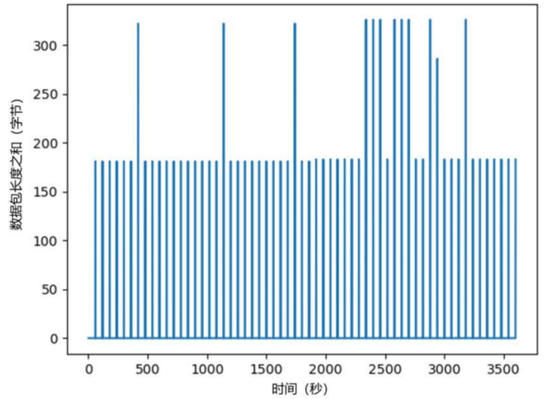
心跳流检测研究现状

聚类思想

- He等提出了一种周期性序列检测技术
- 易军凯等提出基于数据流分簇处理的方法

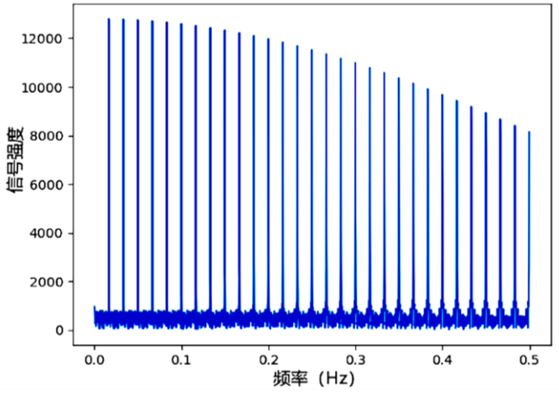
信号处理方式

- 傅里叶变换算法：通过频域的特性来进行心跳流的检测



心跳数据流时域信号序列

傅里叶变换

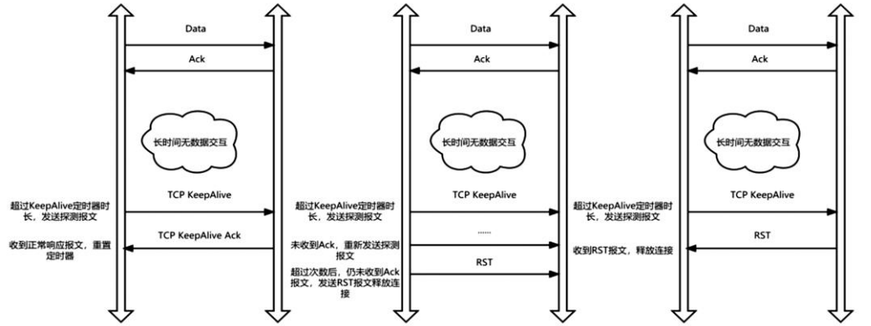


傅里叶变换后得到的频域信号序列

心跳机制的设计

基于公共协议的心跳机制

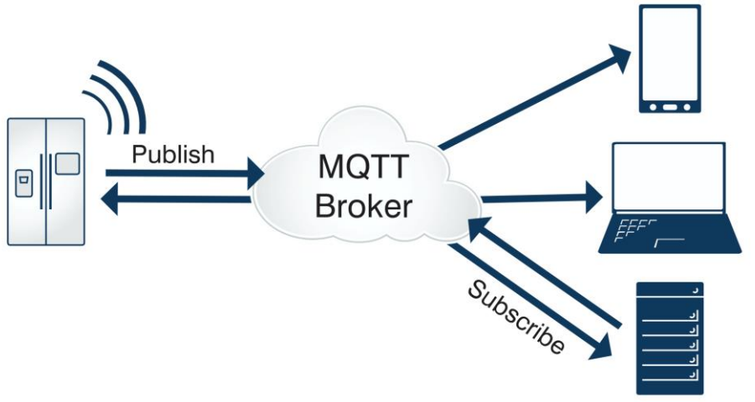
- TCP KeepAlive 心跳机制
- MQTT KeepAlive 心跳机制



TCP心跳机制

应用自定义心跳机制

- 基于TCP或UDP协议自行设计心跳机制

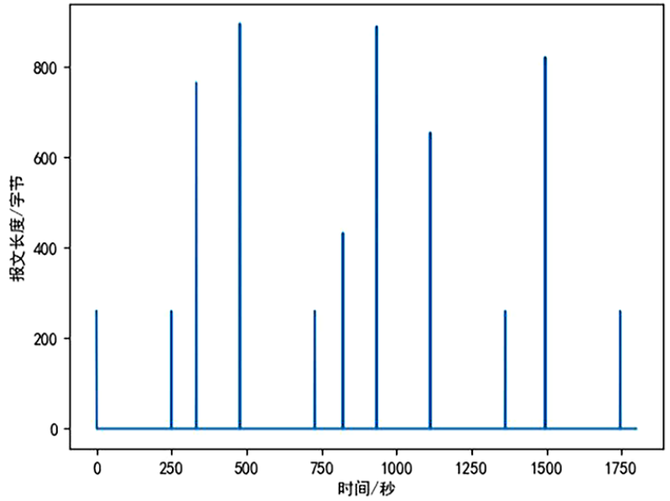


MQTT基本结构图

研究意义

现有方法的局限性

- 容易受到自适应心跳机制或者其他报文的影响



有状态心跳流报文时域图

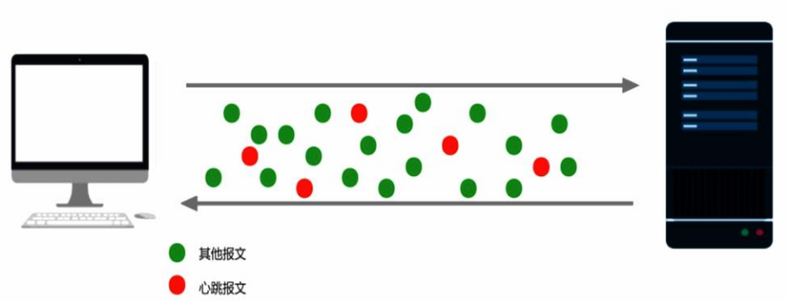
研究意义

- **应用分类**：能够极大地减轻流量分类的工作量和难度
- **网络管理**：有利于管理者掌握和监管辖区内网络情况
- **网络安全**：有助于检测各种恶意攻击行为

02 心跳机制分类

心跳流定义

- **IP流**：在一段时间内有关联关系的报文序列
- **心跳报文**：一种以保活为基础需求，支持即时传输、反向通讯的IP报文
- **心跳流**：包含心跳报文的IP流

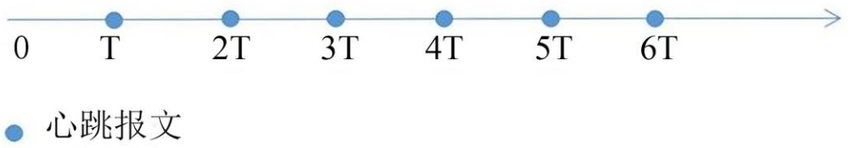


交互型应用流量一般组成图

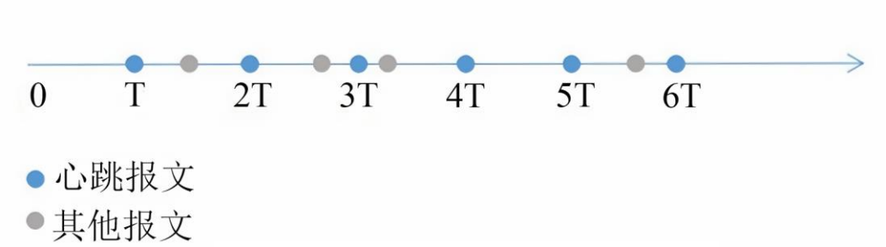
心跳机制分类

无状态

- 单纯无状态心跳机制
- 非单纯无状态心跳机制



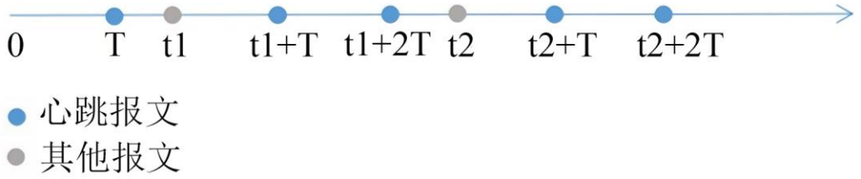
单纯无状态心跳行为图



非单纯无状态心跳行为图

有状态

- 不完全有状态心跳机制
- 完全有状态心跳机制



不完全有状态状态心跳行为图

03 心跳流检测算法

心跳周期的计算

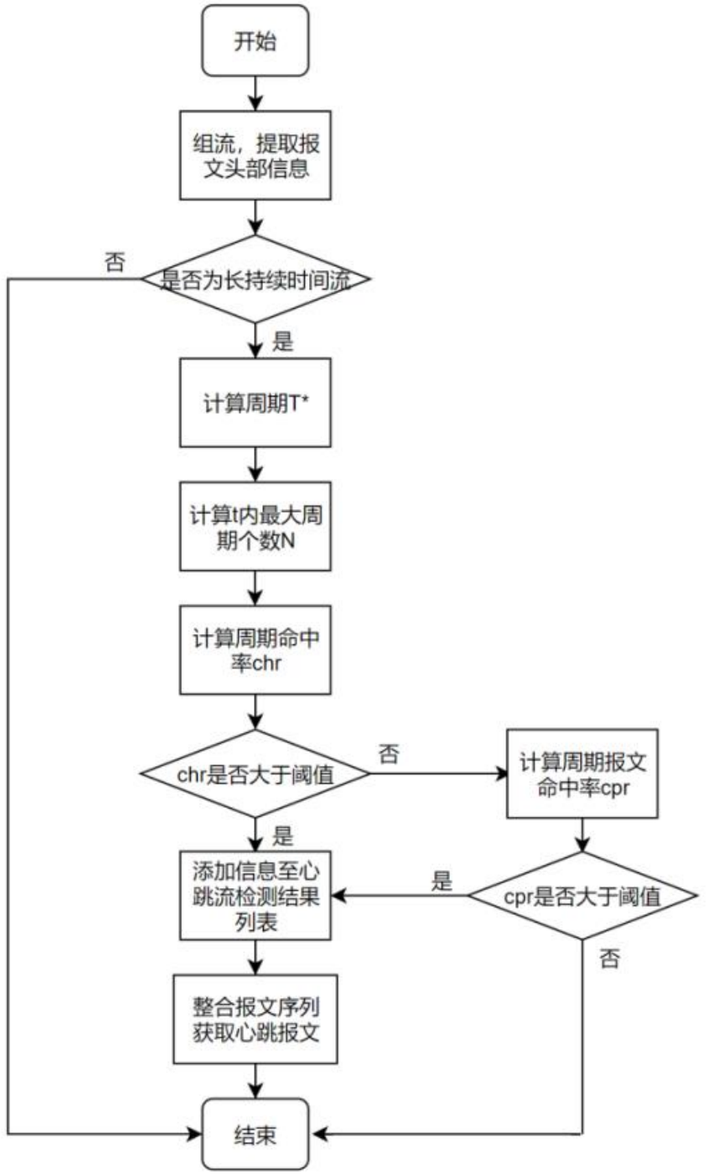
最大报文时间间隔

$$T^* = \max\{\Delta t \mid \Delta t \times c\}$$

Δt 表示同一条流中，相邻2个报文的时间间隔；
c 表示该时间间隔出现的次数

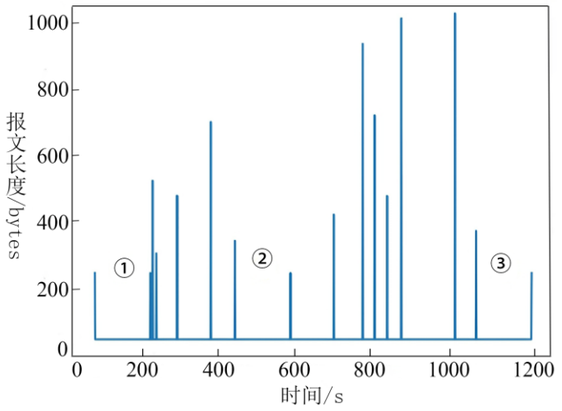
统计算法

- 最大周期个数 $N = [\frac{t}{T^*}]$
- 周期命中率 $C_{hr} = \frac{n_1}{N}$
- 周期报文命中率 $C_{pr} = \frac{n_2}{N}$

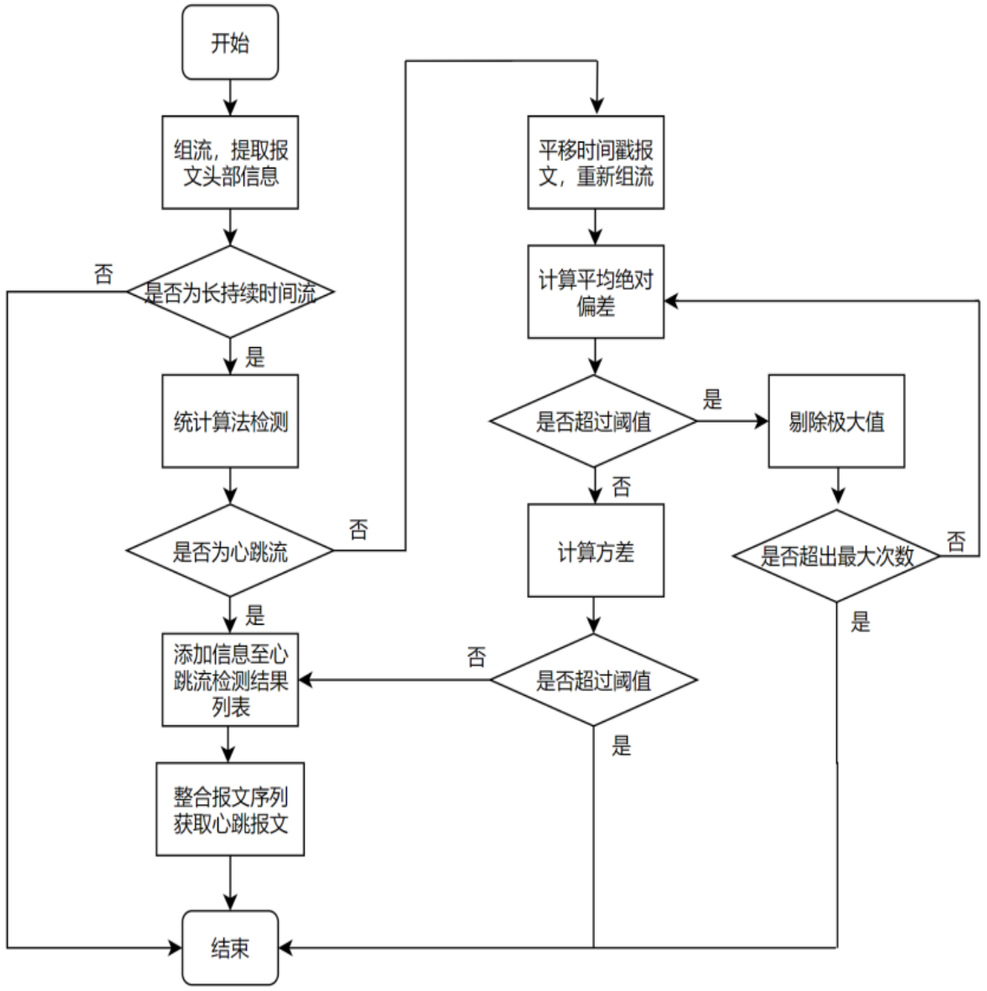


偏离指标算法

时序平移方法



有状态心跳报文时域图



计算公式

- 方差

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

- 平均绝对偏差

$$MAD=\frac{\sum_{i=1}^{n}|x_i-\mu|}{n}$$

04 实验结果与分析

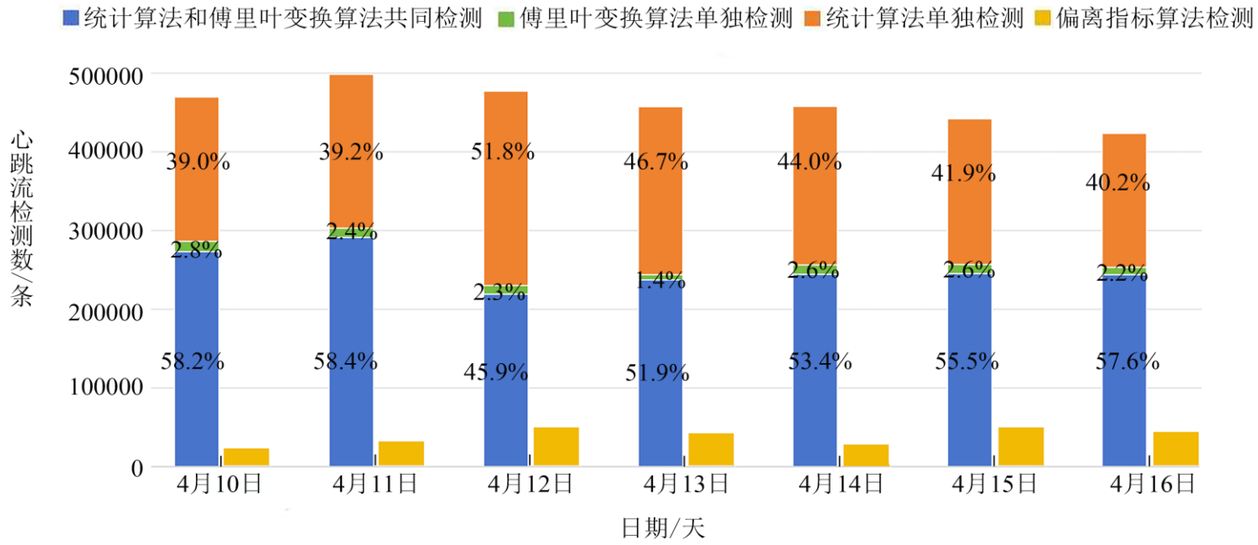
实验环境和参数设置

- **实验平台**：CERNET南京主节点网络边界的长流检测与分类平台
- **数据源**：某211高校校园网7d控制流
- **周期参数**：30 min

心跳流检测实验参数

参数	参数含义	参数取值
t /min	流持续时间	25 ~ 30
T/s	心跳间隔	5 ~ 600
c	心跳报文数	≥ 3
L /B	心跳报文长度	< 500
C_{hr}	周期命中率	0.5
C_{pr}	周期报文命中率	0.5
MAD	平均绝对偏差	10
S^2	方差	10

结果分析



2023年4月10日至16日统计算法和傅里叶变换算法检测心跳流数

	4月10日	4月11日	4月12日	4月13日	4月14日	4月15日	4月16日
两种算法共同检测数	273117	291257	219195	237326	244450	245250	243880
统计算法单独检测数	183306	195396	247022	213502	201616	185092	170306
傅里叶变换算法单独检测数	13251	11795	10972	6564	11816	11637	9376
总数	469674	498448	477189	457392	457882	441979	423562

2023年4月10日至16日统计算法和傅里叶变换算法检测心跳流数与总数比值

	4月10日	4月11日	4月12日	4月13日	4月14日	4月15日	4月16日
两种算法共同检测数占比	0.582	0.584	0.459	0.519	0.534	0.555	0.576
统计算法单独检测数占比	0.390	0.392	0.518	0.467	0.440	0.419	0.402
傅里叶变换算法单独检测数占比	0.028	0.024	0.023	0.014	0.026	0.026	0.022

案例分析

4月11日合并后心跳流数TOP10的应用信息

端口	协议	$T_{心跳周期}$/s	L/bytes	关联域名	所属公司/应用	数量
443	TCP	45,60,120	339,396,337,392	*.wns.windows.com	微软	44 859
21111	TCP	120,60	93,105,80	*.gnway.cc	金万维	18933
3478	UDP	10	108	stun.*.bilibili.com	哔哩哔哩	16 117
6003	TCP	45	93,215,216	*.push.126.net	网易	10846
11114	TCP	120,60,30	93,105,104,80	*.wps.cn	金山软件	9 820
8829	UDP	28	-	*.bcsp2p.baidu.com	百度云	8139
80	TCP	15	80	*.pan.baidu.com	百度网盘	5 768
9993	UDP	5,10	165,182	*.zerotier.com	ZeroTier	5 560
80	TCP	45	-	*.janguoyun.com	坚果云	2667
3000	UDP	60	104,52	*.oray.net	向日葵	2 564

05 结论与展望

结论

- 算法能有效应对其他报文干扰
- 能够实时性检测
- 可以获得纯粹的心跳流报文
- 实验结果证实算法有用性，为应用分类提供支持

展望

- 进一步探索改进有状态心跳流的检测方法
- 更深入研究基于心跳流的应用分类工作

谢谢大家

联系信息：

徐亮（研究生）：lxu@njnet.edu.cn

丁伟教授：wding@njnet.edu.cn