面向智能手机的隐私保护系统

参赛队员：闵大为 任勇勋 张梦月　指导老师：付安民

南京理工大学

目录

背景

在人们都越来越依赖于手机的强大功能时，有没有人想过，它实际并不安全。**它就像是一个时时贴近我们的小偷，悄悄的将我们的个人隐私、手机通讯录、身份信息向外送走。**

——《央视曝光安卓手机应用偷偷上传用户隐私》

一些解决方案

- **方案一：安装时注意权限**

  不足：一次性判断，如果用户不小心安装了，则无法提供保护；强迫安装

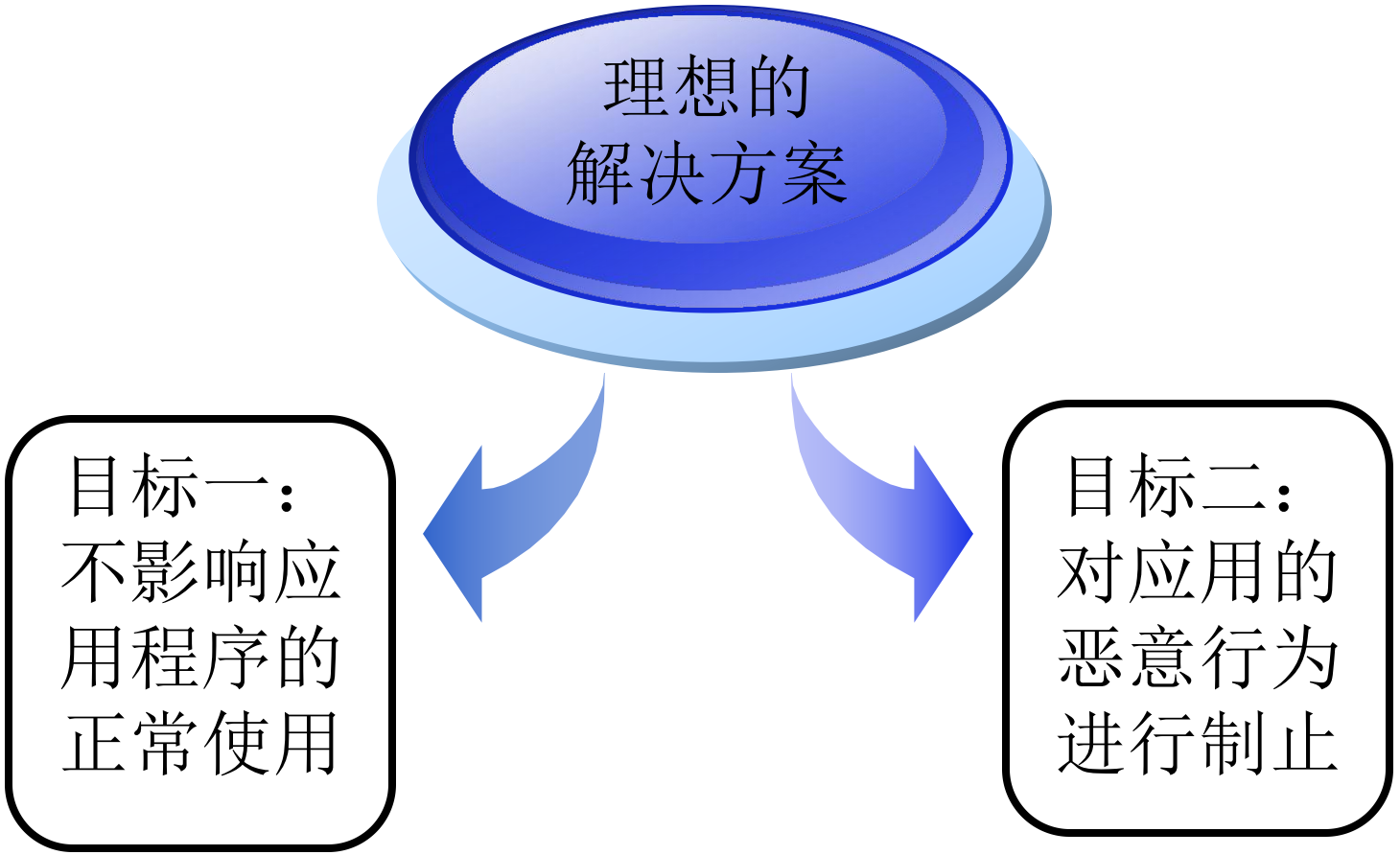
- **方案二：加密信息**

  不足：使用不便，如要保护通讯录则需要加密通讯录，要保护短信又必须加密短信，使用加密信息时需要用相关软件进行解密

- **方案三：禁止应用上网**

  不足：不让应用访问网络，应用无法实现自己的功能，例如即时通，浏览器等。所以这种方法不是很合理。

需求



我们的解决方案：

将蠕虫检测的思想

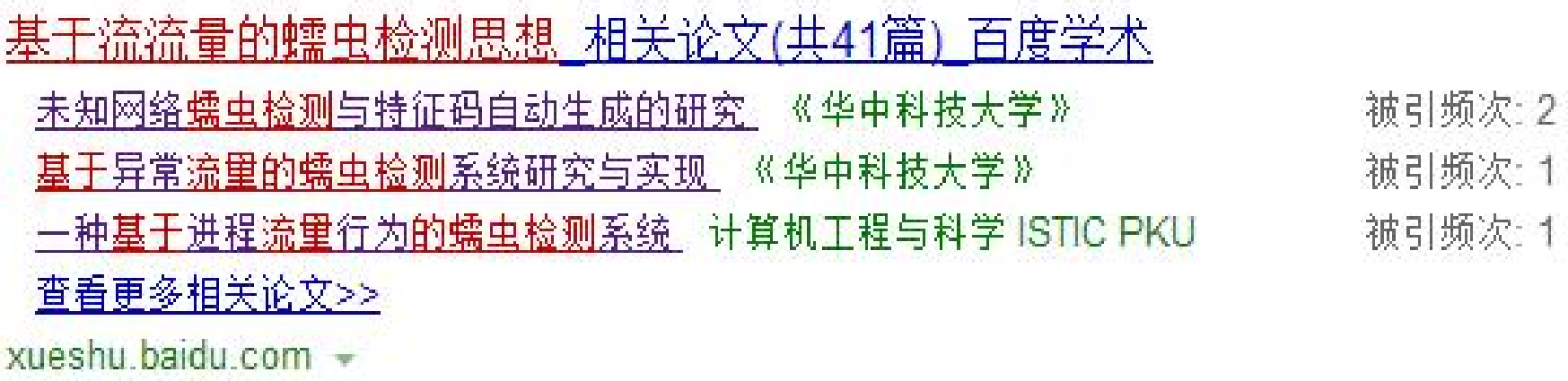
移植到

手机隐私保护上

为什么可以将蠕虫检测的思想应用到手机隐私保护上？

蠕虫检测

- **基本原理：**感染蠕虫后，网络流量会出现异常变化，以此进行检测和防护。
- **成熟性：**在这个方面已经有大量的研究，该检测思想已有一定的积累，比较稳定可靠。



手机上隐私泄露的特点

- 上传数据**一般加密**，且加密方式多种多样。

  （说明基于文本内容的行为分析不太可行）

- 上传隐私数据时，**会提交大量数据**。而大部分的时间内，手机都是从服务器端取数据，如获取网页文字，图片等。不会连续地提交大量数据。

  （说明隐私泄露过程中上传数据量会激增）

相似性

- 蠕虫检测中：由于感染蠕虫，网络流量会激增
- 手机隐私保护中：由于要上传隐私数据，相对于平常，提交的数据会激增

基于流量变化的检测模型

我们给出了手机端隐私泄露行为的定义：

在一定时间内，手机端向某个远程服务器提交的数据包总数超过正常值（阈值）时，则认定该手机中有应用在上传用户隐私数据。

模型一 算法

算法：

for each Δt

1 capture and analyze the data

2（1） calculate $\sum_{t=T}^{T+\Delta t} \mathrm{SIP}_{i,t} \quad (i = 1,2 \cdots n)$

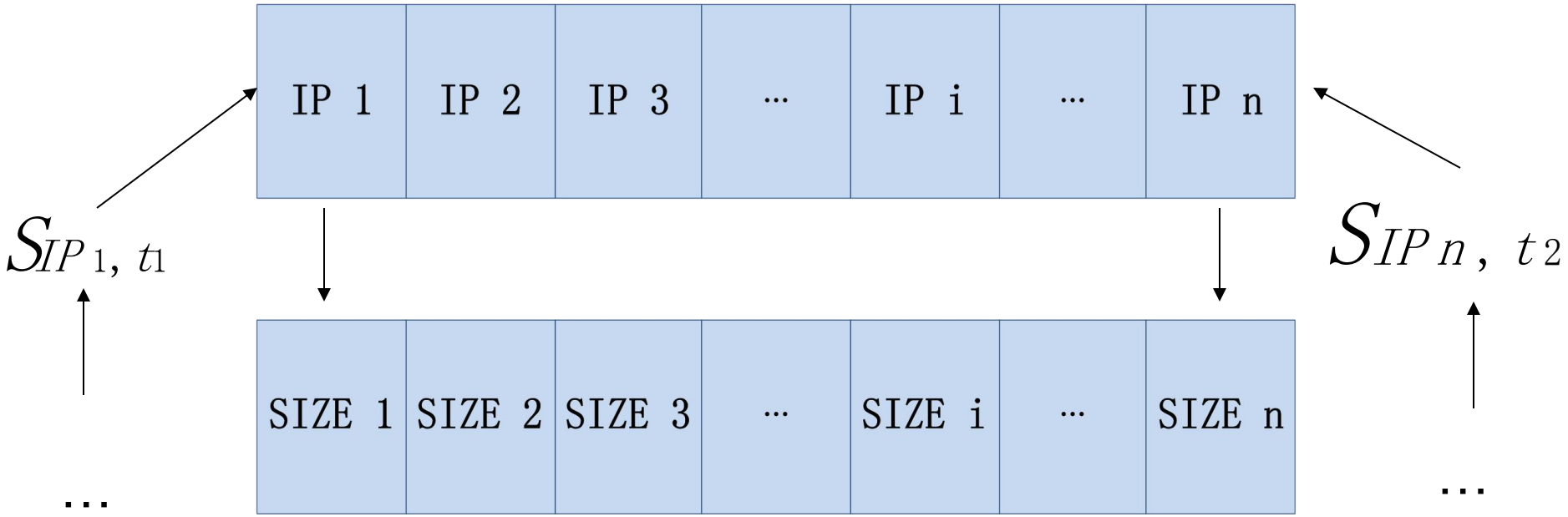
$(2)\, if\ \exists j(1 \le \mathrm{j} \le \mathrm{n}), \sum_{t=T}^{T+\Delta t} \mathrm{SIP}_{j,t} \ge T$

then take action

next Δt

IP地址和对应数据包大小变量储存格式

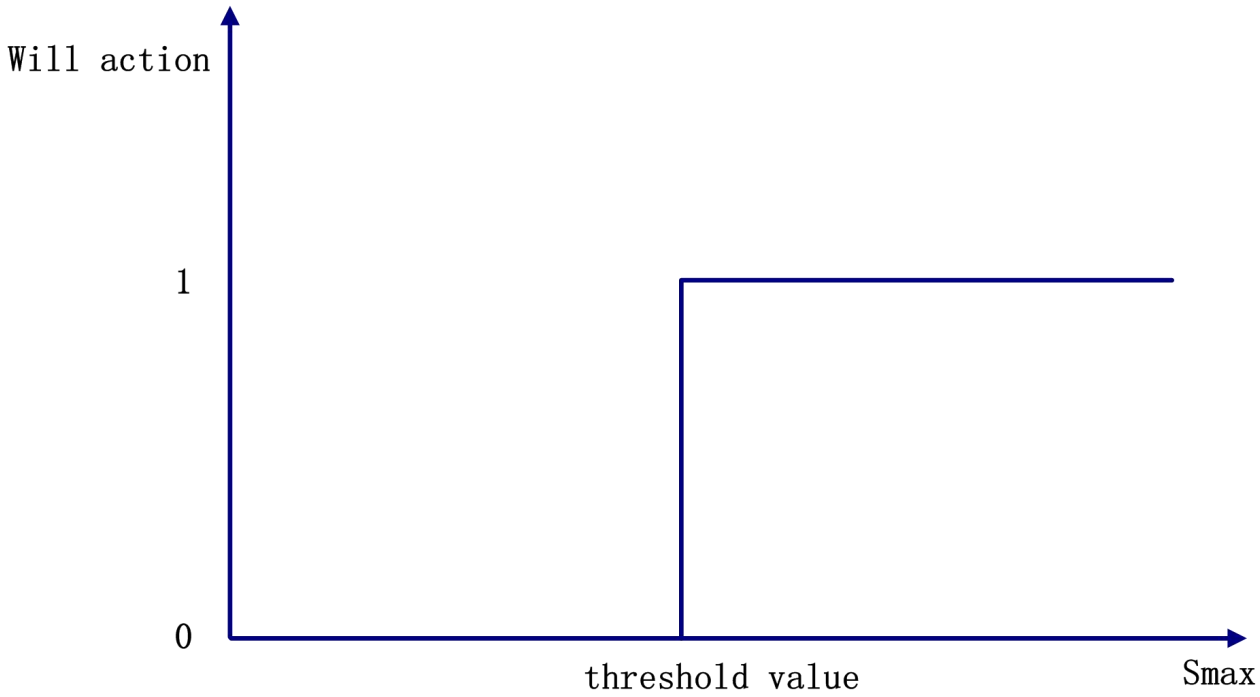
$$\sum_{t=T}^{T+\Delta t} S_{IP_i,\,t} \quad (i = 1,2\cdots n)$$



模型一 阈值

$$if\ \exists j\left(1 \le \mathrm{j} \le \mathrm{n}\right), \sum_{\mathrm{t=T}}^{\mathrm{T+\Delta t}} \mathrm{SIP}_{j,t} \ge T$$

$$then\ \ take\,action$$



模型一 新的问题

有些应用在使用的过程中会上传一定量的数据包，例如微信。

但是用户希望信任该程序，也就是说如果是微信在上传大量数据，我们不采取行动。

如何考虑用户信任的情况？

难点分析

数据包总和变量是属于哪个应用的？

要信任某个程序，我们必须知道超过阈值的数据包总和变量是属于哪个应用的，只有这样才可以选择是否忽略。但是**抓到的数据包并不提供与应用对应的信息**。

IP 1	IP 2	IP 3	…	IP i	…	IP n

SIZE 1	SIZE 2	SIZE 3	…	SIZE i	…	SIZE n

我们的解决方案：

将机器学习的思想

应用到

SIZE 变量和应用的对应上

机器学习的思想

- 机器学习的一种定义：

  学习是系统所作的适应性变化，使系统在下一次完成同样或类似的任务时更为有效。

- 特点：1. 每次行动都会学习

  2. 下次更为有效

启发：学习的过程

- 借鉴机器学习的两个特点：

  1. 每次行动都会学习

  2. 下次更为有效

- 我们给出了基于学习的解决方案，该方案不是直接给出变量SIZE i对应的应用，而是**不断学习，下一次给出更好的答案**。

模型二 问题分析

问题转换：由于SIZE与IP是一一对应的， SIZE 和应用的对应问题就等价于IP与应用的对应问题。

关系分析：

（1）一个应用可以使用多个IP地址：

$$G(App)=\{Ip_1,Ip_2,\cdots,Ip_n\}$$

（2）一个IP地址只给一个应用使用：

$$F(Ip)=App$$

引入机器学习思想前后对比

- 假设：集合A表示{可疑应用1,可疑应用2...可疑应用n}

  F(Ip) 表示求Ip在集合A中对应的可疑应用

- 引入机器学习思想之前：

  由 Ip ，集合 A，求 F(Ip)　**条件不足，难以解决**。

- 引入机器学习思想之后：

  由 IP ，集合A，以及”以前出现过的对应关系”，求F(Ip)。

  因为“以前出现过的对应关系”这个**条件的内容是在增加**的，所以解决问题的可能性也在增加。

模型二 学习过程

- 在用户使用的过程中，会出现：

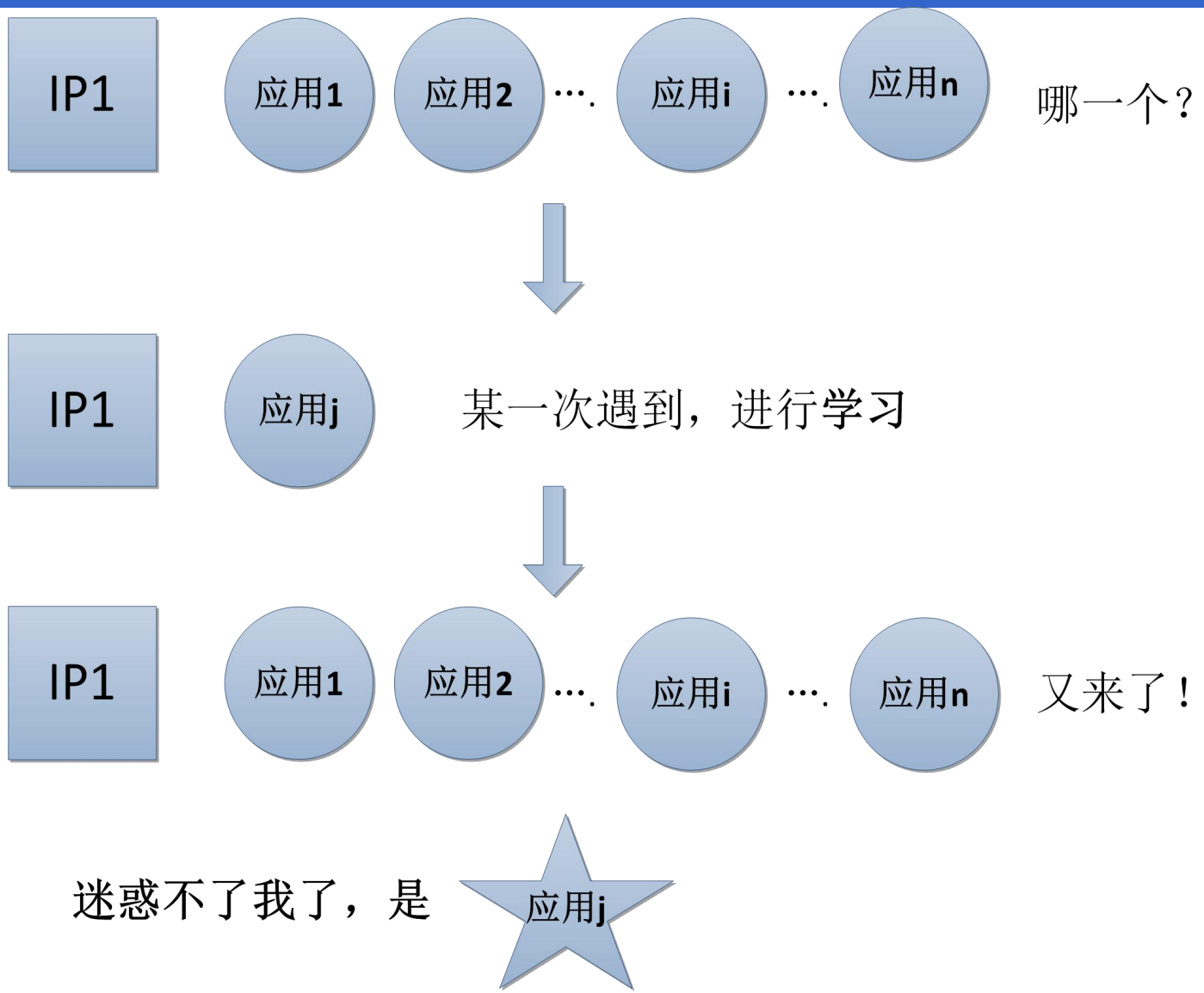
  F(Ip)所在的可疑应用集合A里只有一个应用，

  也就是说出现了**一一对应的情况：**

$$F(Ip) = App$$

- **通过学习，我们可以把这个关系记录下来，**下次再遇到这个Ip的时候就可以排除干扰项，找到目标了。

模型二 学习过程


IP1
应用1
应用2
….
应用i
….
应用n
哪一个？
IP1
应用j
某一次遇到，进行学习
IP1
应用1
应用2
….
应用i
….
应用n
又来了！
迷惑不了我了，是
应用j

模型二 算法

for each Δt

1 capture and analyze the data

2 (1) calculate $\sum_{t=T}^{T+\Delta t} S_{IP_i,t} \quad (i = 1,2 \cdots n)$

(2) *if* $\exists j(1 \le j \le n), \sum_{t=T}^{T+\Delta t} S_{IP_j,t} \ge T\{$

learn to find App

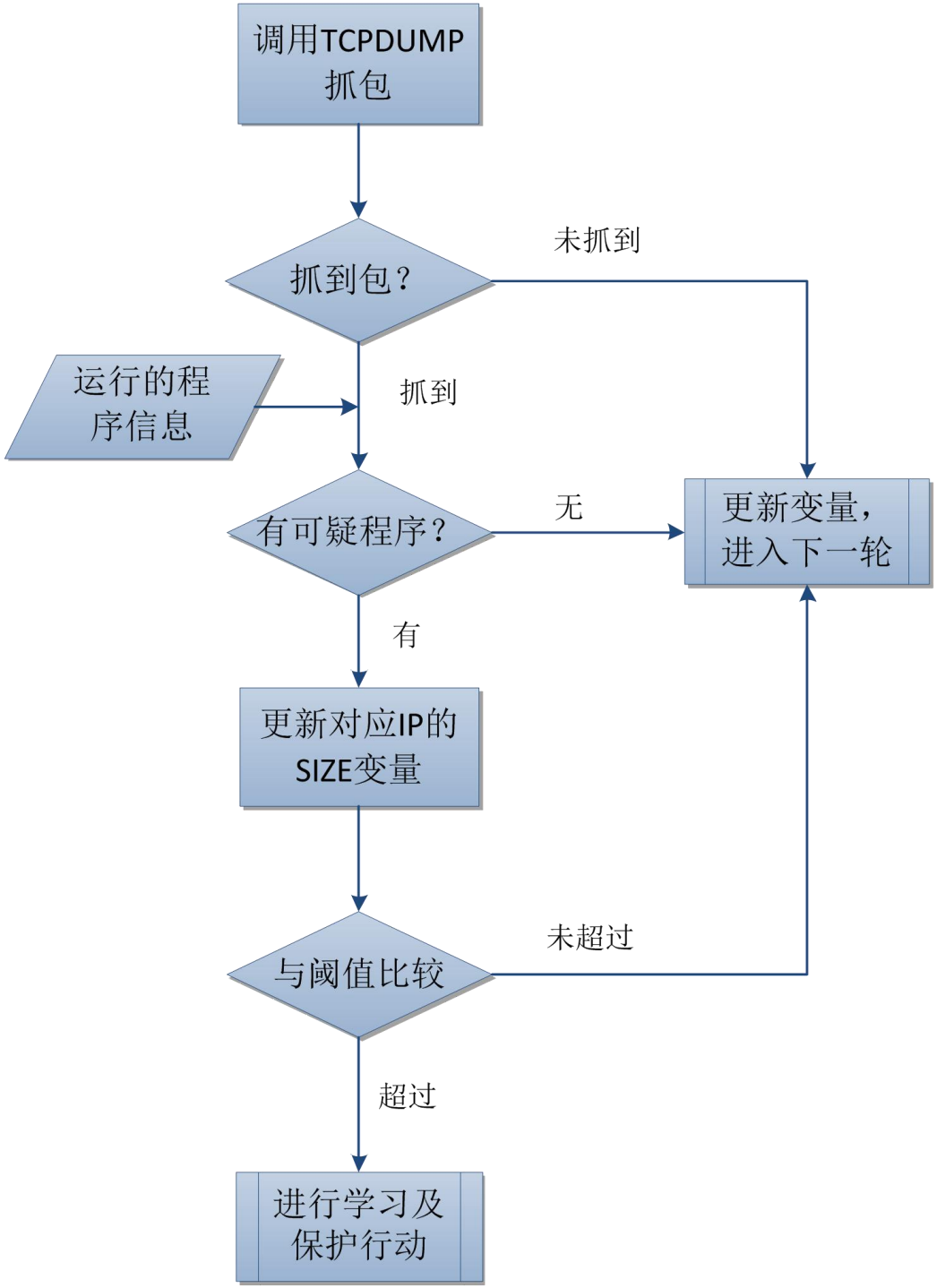
if trustless

then take action

}

next Δt

简化流程图


调用TCPDUMP抓包
抓到包？
未抓到
运行的程序信息
抓到
有可疑程序？
无
更新变量，进入下一轮
有
更新对应IP的SIZE变量
与阈值比较
未超过
超过
进行学习及保护行动

效果


23:14
隐私卫士防护工具
隐私保护
日志查看
阈值设置
隐私卫士保护您！


23:13
上传监听
已打开
软件列表
微信 正在运行
共有2项隐私权限
360手机助手
共有3项隐私权限
超级课程表
共有2项隐私权限
腾讯微博
共有1项隐私权限
WhatsApp 正在运行
共有1项隐私权限
QQ 正在运行
共有5项隐私权限
豌豆荚连接服务
共有3项隐私权限
豌豆荚
共有5项隐私权限
人人
共有4项隐私权限
百度云
共有3项隐私权限
Chrome


23:13
监控报告
生成日期 ：2014-08-01
生成时间 ：23:13:28
采取手段 ：关闭wifi
目标程序 ：微信
目的IP地址 ：120.204.200.66
当前阈值 ：650 B
上传包大小 ：659 B
保存报告
退出

总结

- 特色：与市场上的一些静态解决方案不同，该方案能够**在不影响用户正常使用的情况下进行动态监控并提供保护。**
- 创新一：将基于蠕虫检测的思想应用在手机隐私保护上。
- 创新二：将机器学习的思想应用在数据包大小和可疑应用的对应上。

Thank You !