# 1 Trace

## 数据集信息

ta1-trace-e3-official-1.json/ta1-trace-e3-official-1.json

文本

描述已自动生成

程序中处理的事件类型有：EVENT\_OPEN、EVENT\_READ、EVENT\_WRITE、EVENT\_FORK、EVENT\_EXECUTE、EVENT\_CONNECT、EVENT\_SENDMSG、EVENT\_RECVMSG。

程序中处理的实体类型有：Subject、FileObject、NetFlowObject

## 事件信息

Trace数据集中，每一个事件的日志数据除了subject和object的uuid之外，不包含额外的信息，因此处理的时候先把遍历到的实体存下来，处理事件的时候根据uuid取找实体，提取信息后，组合成目标事件对象。

Trace中，例如EVENT\_RECVMSG、EVENT\_WRITE等事件，都有对SrcSinkObject和UnnamedPipeObject操作的事件记录，这两种实体，以及和这两种实体相关的事件，暂时没有提取和解析。

EVENT\_WRITE和EVENT\_READ两种事件都有对NetFlowObject的操作，定义的时候定义成网络事件（因为本质就是进程与网络对象的交互）。

## 程序逻辑

程序中遍历记录的时候遇到实体信息就存进Map，遇到事件信息就提取事件所对应的subject实体和object实体。

Subject实体的创建函数需要接收参数pid，但数据集中只有ppid，所以把ppid赋值给了pid；EVENT\_SENDMSG，EVENT\_CONNECT的数据流方向定义为OUT，EVENT\_RECVMSG的数据流方向定义为IN。

# 2 Theia

## 2.1 数据集信息

ta1-theia-e3-official-5m.json/ta1-theia-e3-official-5m.json

文本

描述已自动生成

程序中处理的事件类型有：EVENT\_OPEN、EVENT\_READ、EVENT\_WRITE、EVENT\_CLONE、EVENT\_EXECUTE、EVENT\_CONNECT、EVENT\_SENDTO、EVENT\_RECVFROM。

程序中处理的实体类型有：Subject、FileObject、NetFlowObject。

## 2.2 事件信息

少数Subject实体没有ppid信息，程序里把这些实体的ppid设置为0。

部分NetFlowObject的IP字段是NET\_LINK，程序里把NET\_LINK设置为-1.-1.-1.-1，这样做是因为发现如果不是合理的IP地址格式，会导致这一条实体记录无法被创建。

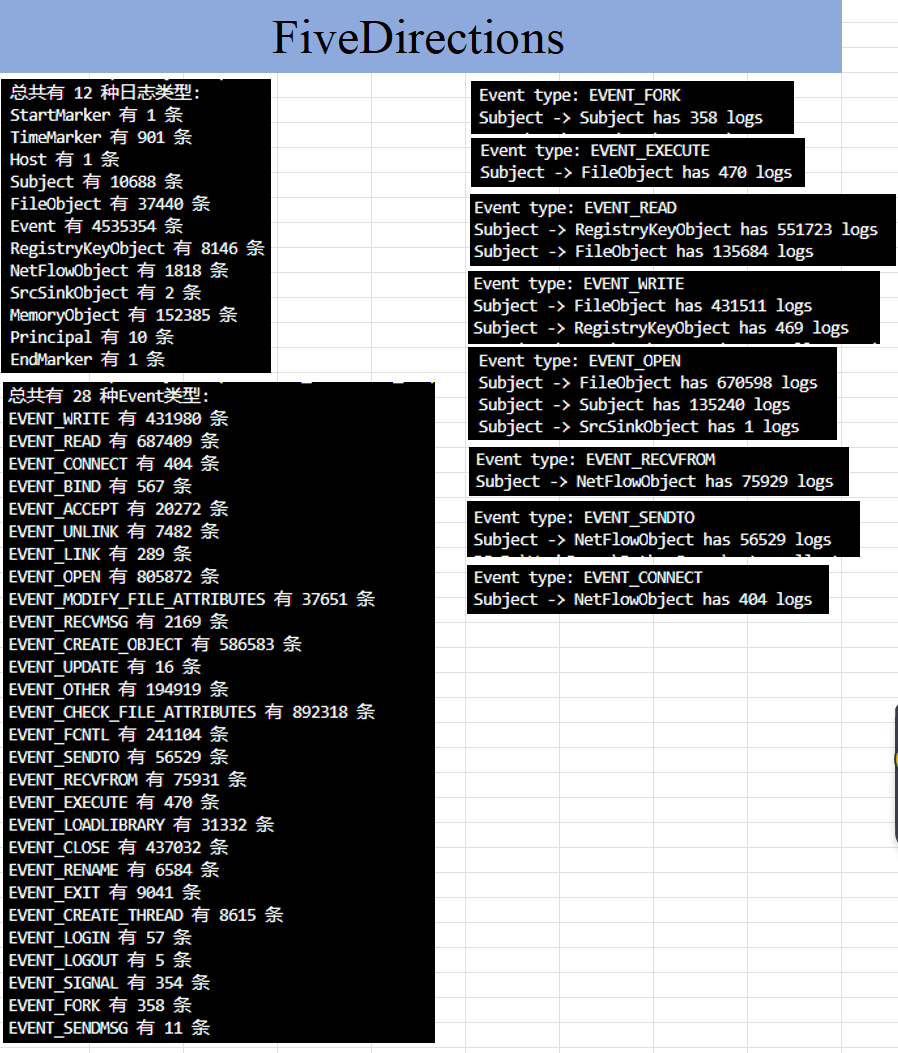
## 2.3 程序逻辑

程序中遍历记录的时候遇到实体信息就存进Map，遇到事件信息就提取事件所对应的subject实体和object实体。

把ppid缺失的Subject的ppid设置为0；把IP字段是NET\_LINK的设置为-1.-1.-1.-1。

# 3 FiveDirections

## 3.1 数据集信息



程序中处理的事件类型有：EVENT\_OPEN、EVENT\_READ、EVENT\_WRITE、EVENT\_FORK、EVENT\_EXECUTE、EVENT\_CONNECT、EVENT\_SENDTO、EVENT\_RECVFROM。

程序中处理的实体类型有：Subject、FileObject、NetFlowObject 、RegistryKeyObject。

## 3.2 事件信息

FiveDirections数据集中，事件Log本身包含有事件相关的信息，例如读文件的文件位置，对应的其文件实体Log本身不包含文件路径等信息。

NetFlowObject对象的IP地址有些是ipv6地址，目前没有处理，把它赋值成-1.-1.-1.-1，且ipv6中有一些是多播地址，IP地址中也存在NA，把NA也赋值成-1.-1.-1.-1；

数据集中包含很多对注册表的操作事件（特别是读取事件），目前把注册表事件当初文件一样处理了。

## 3.3 程序逻辑

程序中遍历记录的时候遇到实体信息就存进Map，遇到事件信息就提取事件所对应的subject实体和object实体。

用日志中的cid赋值给pid（因为很多Subject数据没有ppid字段）；把Subject数据中cmdline字段的值当作processName（因为Subject数据基本没有name字段），构建事件的时候需要的cmdline信息从EVENT数据中提取。