分析不同 ip 是否属于同一个攻击者或攻击机构

目录

一．收集和分析数据

1. IP 日志
2. 部分终端恶意样本（样本名和终端指纹）
3. IP 绑定过域名（域名）
4. IP 解析过的域名
5. 360 威胁情报

分析后思路：若 ip 间存在相同恶意样本，终端指纹，绑定过同一域名，解析过同一域名，则是否可以认为不同 ip 存在某种联系。

方向：终端指纹以及绑定过相同域名具有最强关联性，相同恶意样本其次，主要针对正三个

方向的关系网进行可视化分析。解析域名数据量太大，笔者现有计算能力，所写算法难以进行可视化。

二．处理数据

所有数据处理前先做冗余数据检查，删除所有冗余数据。

1. 获取题目一中，在所有流量数据中存在 ip，从三个方向判断是否属于攻击 ip。
   a. 使用收集的攻击规则（正则）对每个流量数据日志进行处理，得到判断为攻击流量的记录，按 ip 分组统计每个 ip 使用的攻击类别（PHP 攻击，sql 注入等）以及攻击分数.
   b. 从 360 威胁情报.json 文件中提取数据，处理为 dataframe 格式。筛选出流量日志中 ip 的威胁情报，is_malicious 值为 TRUE 的 ip 即为有恶意行为的 ip。处理过程详见源码。
   c. 从部分恶意终端样本.csv 文件，按 ip 分组统计每个 ip 的恶意样本记录，筛选出流量日志中 ip 的恶意样本记录。处理过程详见源码

   由此得到攻击 ip 结果数据 result.csv（file_id,is_attacker,reason）
2. 处理 domain_info.csv。得到 ip 在某天绑定过某个域名记录。

   按 ip 分组处理每条记录，将绑定域名分割出来，得到（ip,date,domain,source_ip）。source_ip 是源 ip，ip 是域名情况中显示的 ip（暂不清楚两者详细关系，大多数情况下值相同）。以源 ip 为主。
3. 部分终端恶意样本.csv（Iip,md5,mid,date），可疑样本 md5 与文件名映射.json。筛选出攻击 ip 列表的恶意样本记录。分析数据后发现，md5 对应的恶意样本名是一个列表，需要关联两个文件，得到 ip 在某个时间投放的样本名。如果两个 ip 的恶意样本名相同，

则认为具有关联性。

4. dns.csv，筛选出攻击 ip 列表的解析域名记录。

   步骤 2-4 详见源码文件 get_ip_file.ipynb

5. 对恶意样本数据 ip_attack_file.csv，绑定域名数据 ip_domain.csv，域名解析数据 ip_log_dns_cate.csv 以此处理，得到 ip 一一对应的恶意样本文件名列表，终端指纹列表，绑定域名列表，解析域名列表。

6. 集成(步骤 5 所获取的)数据 ip_data_cluster.csv，得到每个 ip，流量攻击类别 reason，恶意样本记录数 attack_file_number，恶意样本名列表 attack_file_list，终端指纹列表 mid_list，绑定域名列表 bind_domain_list，解析域名列表 resolve_domain_list，解析域名类别列表 domain_category。

```
In [11]: ip_data_cluster.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 5879 entries, 0 to 5878
Data columns (total 8 columns):
ip                    5879 non-null object
reason                4236 non-null object
attack_file_number    5642 non-null float64
attack_file_list      5642 non-null object
mid_list              5642 non-null object
bind_domain_list      143 non-null object
resolve_domain_list   4625 non-null object
domain_category       4625 non-null object
dtypes: float64(1), object(7)
memory usage: 413.4+ KB
```

7. 处理 ip_data_cluster.csv。编写脚本，遍历每个 ip 具有关联性的“朋友”ip 列表。根据四个原则得到具有相同终端指纹 IP 列表，投放过相同恶意样本名 ip 列表，绑定过相同域名 ip 列表，解析过域名 ip 列表。

```
In [18]: ip_data_cluster.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5879 entries, 0 to 5878
Data columns (total 12 columns):
ip                          5879 non-null object
reason                      4236 non-null object
attack_file_number          5642 non-null float64
attack_file_list            5879 non-null object
mid_list                    5879 non-null object
bind_domain_list            5879 non-null object
resolve_domain_list         5879 non-null object
domain_category             4625 non-null object
mid_friends_list            5879 non-null object
attack_file_friends_list    5879 non-null object
bind_domain_friends_list    5879 non-null object
resolve_domain_friends_list 5879 non-null object
dtypes: float64(1), object(11)
memory usage: 597.1+ KB
```

三．社群发现和社区分析

1.数据集

由上述探索式数据分析得到，攻击 ip 相关数据。字段说明如下

流量攻击类别 reason，由 WAF 规则检测出来的流量中使用攻击类型，360 威胁情报恶意行为。

恶意样本记录数 attack_file_number，投放过恶意样本记录数

恶意样本名列表 attack_file_list，投放过的恶意样本名列表

终端指纹列表 mid_list，ip 的终端指纹列表

绑定域名列表 bind_domain_list，ip 绑定过的域名列表

解析域名列表 resolve_domain_list，ip 解析过的域名列表

解析域名类别列表 domain_category。Ip 解析过的域名类别（economic，education 等）

恶意样本名列表 attack_file_friends_list，与该 ip 投放过的相同恶意样本的 ip 列表

终端指纹列表 mid_friends_list，与该 ip 具有相同终端指纹的 ip 列表列表

绑定域名列表 bind_domain_ friends_list，与该 ip 绑定过的相同域名的 ip 列表

（解析域名列表 resolve_domain_ friends_list，与该 ip 解析过相同域名的 ip 列表，受笔者计算能力限制，置空）

```
In [18]: ip_data_cluster.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5879 entries, 0 to 5878
Data columns (total 12 columns):
ip                              5879 non-null object
reason                          4236 non-null object
attack_file_number              5642 non-null float64
attack_file_list                5879 non-null object
mid_list                        5879 non-null object
bind_domain_list                5879 non-null object
resolve_domain_list             5879 non-null object
domain_category                 4625 non-null object
mid_friends_list                5879 non-null object
attack_file_friends_list        5879 non-null object
bind_domain_friends_list        5879 non-null object
resolve_domain_friends_list     5879 non-null object
dtypes: float64(1), object(11)
memory usage: 597.1+ KB
```

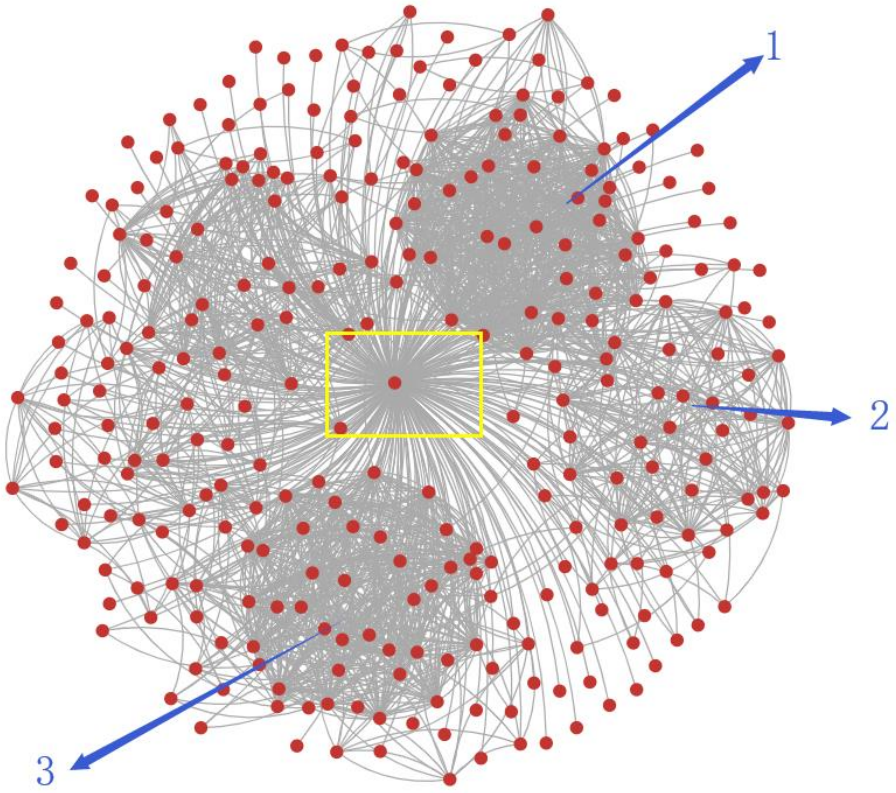
2. 可视化分析

使用无向图来构建力引导关系，虽然在安全领域的风控、反欺诈方向中使用有向图更为广泛一些，但目前检测出来关联关系是双向的，因此此处分析采用无向图。从终端指纹关系网，恶意样本关系，绑定域名关系建立可视化图。

a. 终端指纹

终端指纹 ip 节点数：290

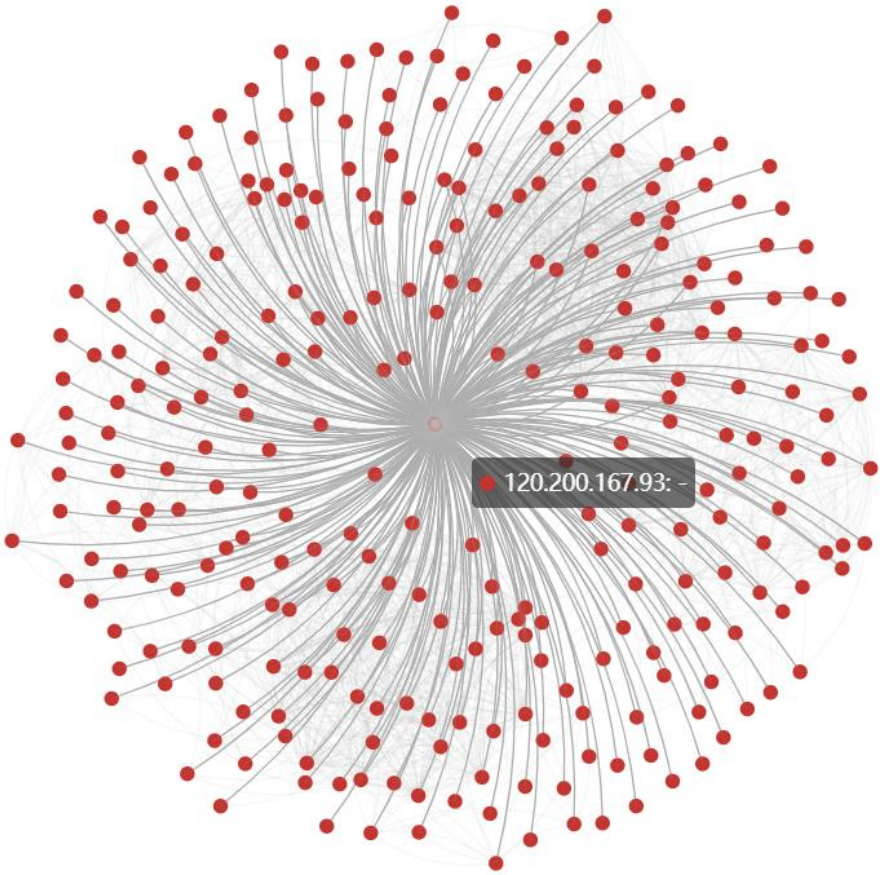
终端指纹 ip 关系数：1831



终端指纹力导图 3.1

可以很明显的发现最中心一点十分突出。以及环绕它的三个密集社区。

进一步查看它的关系网，非常直观。关联关系成伞形散开，几乎关联了大部分的 ip。



终端指纹力导图 3-1 中心点关系网

我们进一步探索其他三个密集社区。社区一：

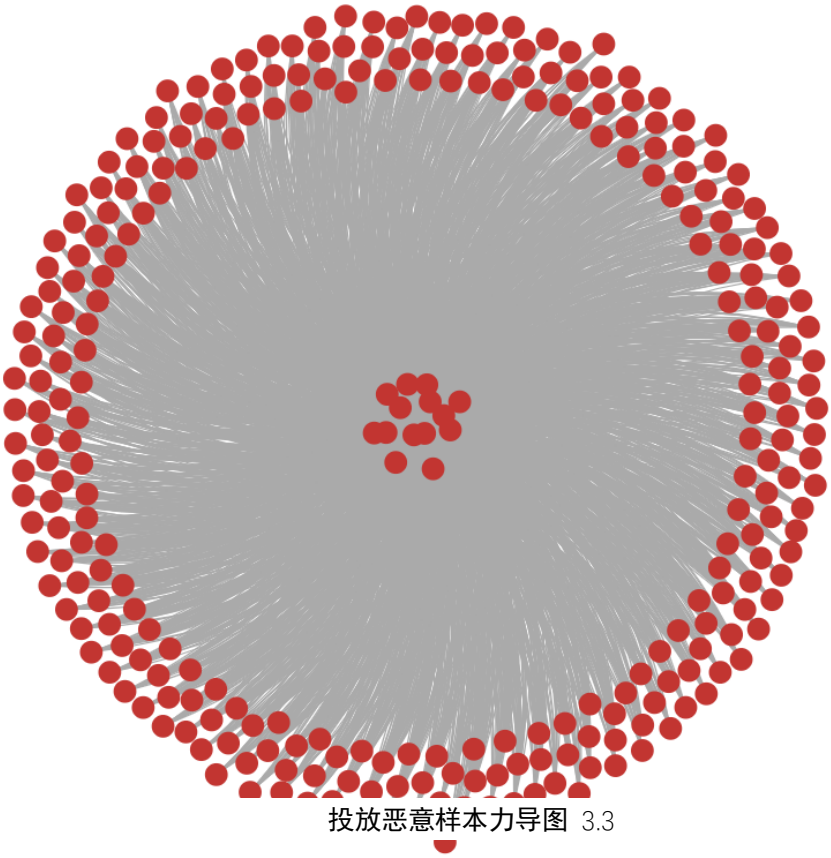
控制社区 1 的多个点都和中心点（ip:120.200.167.93）有关联关系。其他两个密集社区具有相同特性，此处不一一讲解

结论：得到多层级关系网，从 ip:120.200.167.93 进行多级攻击扩散

b. 投放恶意样本关系

恶意样本 ip 节点数：290

恶意样本 ip 关系数：4036



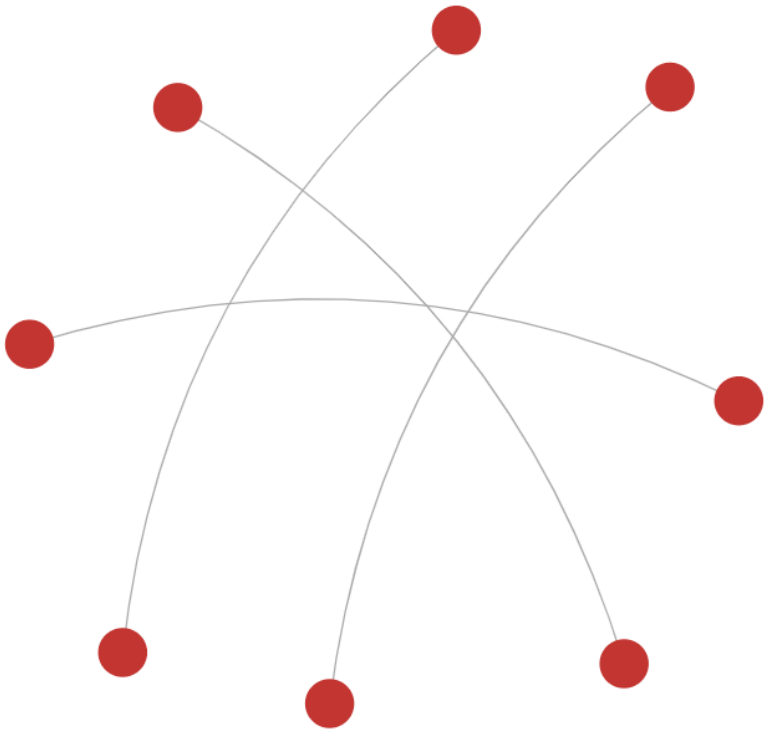
投放恶意样本力导图 3.3

根据 ip 投放过的恶意样本名得到的关联关系，具有更强烈的直观性，中心数十个 ip 为控制中心。

c．绑定域名关系

绑定域名 ip 节点数：8

绑定域名 ip 关系数：4



绑定域名力导图 3.4

可能是攻击域名绑定难度比较大，关系网非常小，互相具有关联关系的 ip，几乎可以认为属于同一个攻击者或者攻击机构，分析数据可以发现有关联性的 ip 值大部分具有相似性。

```
bind_domain_links
```

```
]: [{'source': '111.14.211.67', 'target': '111.14.211.68'},
    {'source': '123.185.222.82', 'target': '123.185.222.117'},
    {'source': '218.24.4.109', 'target': '218.24.4.100'},
    {'source': '218.25.140.72', 'target': '211.140.196.90'}]
```

3.结合多种关联关系分析

（1）获取恶意样本控制中心的 ip 列表。

（2）获取终端指纹关系中，源 ip 或者目标 ip 属于恶意样本中心 ip 列表的关系。并且筛选出与终端指纹中心点（120.200.167.93）的关系。

两表对比如下，虽然 ip：120.200.167.93 不属于恶意样本呢控制中心 ip 列表中，但是从终端指纹关系网看，它控制了所有恶意样本中心的 ip。

	source	target_number
0	218.58.75.140	289
1	223.100.159.142	289
2	36.102.222.104	286
3	59.45.61.162	289
4	59.46.137.3	289
5	59.46.196.254	289
6	59.46.196.90	289
7	59.46.212.196	289
8	59.46.219.83	289
9	59.46.227.90	289
10	59.47.37.164	289
11	59.47.37.197	288
12	59.47.37.198	284
13	59.47.37.83	288

恶意样本中心 ip 列表 3.5

	source	target	target_number
120.200.167.93	218.58.75.140		289.0
120.200.167.93	223.100.159.142		289.0
120.200.167.93	36.102.222.104		286.0
120.200.167.93	59.45.61.162		289.0
120.200.167.93	59.46.137.3		289.0
120.200.167.93	59.46.196.254		289.0
120.200.167.93	59.46.196.90		289.0
120.200.167.93	59.46.212.196		289.0
120.200.167.93	59.46.219.83		289.0
120.200.167.93	59.46.227.90		289.0
120.200.167.93	59.47.37.164		289.0
120.200.167.93	59.47.37.197		288.0
120.200.167.93	59.47.37.198		284.0
120.200.167.93	59.47.37.83		288.0

120.200.167.93 与恶意样本中心关系对比 1

四．源码复现说明

由于大多数是进行数据分析，所有代码都在 jupyter notebook 下编写。

环境：anaconda 3（64bit）python 3

第三方库：from pyecharts import Graph

源码说明：

1. 处理数据（2-5 步骤）源码位于 get_ip_file.ipynb，所有代码已进行函数归类，每个函数都有相应注释。需要将所用基础数据文件路径进行更换。
2. 代码生成的所有文件都将存放于相对路径 ip_file 文件夹下，ip_file 存放有第一题的结果文件 result.csv。
3. 处理数据（5-7）源码位于 analysisIpRelation.ipynb，获得各个 ip 间关系的脚本计算量十分大，运行时间很长。
4. 可视化分析位于 visi_analysis.ipynb