DATA130012h.01 数据可视化部分 Project

黑灰产网络资产图谱可视分析

一、背景介绍

网络黑灰产具有链条化、团伙化、资产化和跨域化等特点。链条化是指黑灰产形成了环环相扣的上、中、下游产业链，共同配合完成非法牟利。资产化是指黑灰产团伙掌握大量且关联复杂的多种网络资产，以支撑产业链的网络化运转，比如：上游信息盗取需要木马和钓鱼网站，中游业务网站运维需要域名和 IP 地址；下游支付需要安全证书。跨域化是指黑灰产团伙为躲避追查，将一部分网络资产和成员布置在境外。

分析黑灰产团伙掌握的网络资产是打击黑灰产的重要切入点。网络资产可以分为外围网络资产、普通网络资产和核心网络资产。外围网络资产主要是向网民直接公开的黑灰产业务网站域名。核心网络资产是关系到许多外围网络资产运行或关联多个业务线的网络资产，比如：同时支持多个网站域名运行的某IP地址，又比如：同一黑灰产团伙掌控的赌博业务网站和违禁品交易业务网站共同使用的数字安全证书。核心网络资产信息一般不直接向网民公开，部分核心网络资产信息隐藏在多种公开数据源中。普通网络资产介于其它两类网络资产之间。

查证和封堵黑灰产团伙掌握的核心网络资产是目前打击黑灰产的主要手段之一。原因来自三个方面：一是封堵外围网络资产效率低且被动滞后，因为网站副本多，存活周期短，域名更换频繁。二是封堵核心网络资产可以让许多非法网站失效或陷入安全风险，造成高额恢复成本。三是深度分析核心网络资产有利于发现关联多资产或多业务的关键链路，还原多个业务线之间的联系，甚至发现真实世界中控制黑灰产的嫌疑人。

现有一个黑灰产网络资产图谱数据集（已脱敏）。该数据集以点边双异质有向图为数据结构，节点为网络资产，边为网络资产间关联关系。该数据集包含8类网络资产和11类资产关联关系，共有237万个节点，328万条边。假设你是网络黑灰产治理人员，请设计一套可视分析方案，从数据集中找到一些由同一黑灰产团伙掌握的网络资产子图，并识别子图中的核心资产与关键链路，将结果用图表形式呈现出来。

二、数据介绍

黑灰产网络资产图谱数据集（已脱敏）以CSV格式存储，压缩前722MB，包括 **Node.csv** 和 **Link.csv** 两个数据文件。

下载地址：https://pan.baidu.com/s/1m56y1h7qSGofwAiCE-Dv-g 提取码: bfqm

2.1 Node.csv

Node.csv数据文件大小为229M，包括237万条数据记录，每一条数据记录一个节点，包括表1所示的4个字段。图1展示了Node.csv的数据样本。

字段	说明	类型	示例	说明
id	节点id	String	Domain_0d9f06a82e90193f68e72e53acd55e23c74afb0e3589608627e423c64d19f6db	唯一标识节点
name	节点名称	String	0d9f06a82e.com	经过了MD5加密和无效化脱敏处理
type	节点类型	String	Domain	共8类，见表2
industry	黑灰产业务类型(只对Domain类型节点有效)	String	['B']	共10类，见表3

表1. Node.csv 数据文件—字段说明

字段	说明	数量	重要程度
Domain	网站域名	200万	非常重要
IP	网站的IP地址	20万	非常重要
Cert	网站用的SSL安全证书	13万	非常重要
Whois_Name	网站域名的注册人姓名	1.8万	重要
Whois_Phone	网站域名的注册人电话	0.2万	重要
Whois_Email	网站域名的注册人邮箱	0.4万	重要
IP_C	IP的C段	0.6万	一般
ASN	IP的自治域	0.03万	一般

表2. 节点类型说明

industry字段值	黑灰产业务类型	说明
A	涉黄	该域名的网站涉及色情传播
B	涉赌	该域名的网站涉及网络传播
C	诈骗	该域名的网站涉及网络诈骗，如仿冒著名网站
D	涉毒	该域名的网站涉及毒品交易
E	涉枪	该域名的网站涉及枪支交易
F	黑客	该域名的网站是嵌入恶意信息的黑客网站，如嵌入木马的钓鱼网站
G	非法交易平台	该域名的网站涉及非法交易，如个人信息买卖
H	非法支付平台	该域名的网站是非法支付平台
I	其他	其他黑灰产业务网站

表3. 黑灰产业务类型说明

```
Domain_0586b66338e82edf74a0a7d65d1e5835a86647b2e3781e5718c6330e0aca3617,0586b66338.com,Domain,['B']
Cert_fb7076fed16346aeb065c7d6f984ddff37b8dd4b35d2bd1a07f30ef7b819b03d,fb7076fed1,Cert,[]
IP_37f7ed5739b43757ff23c712ae4d60d16615c59c0818bf5f2c91514c9c695845,5.180.xxx.xxx,IP,[]
IP_44e642e648fa555970bfd01596dc1b67e65b357e469479b4105fed2758339462,156.245.xxx.xxx,IP,[]
Cert_5dd7cba66d526fbaaa23b4f2c375f2a10cf4cc9e927682e9602f423a9ae96d38,5dd7cba66d,Cert,[]
Whois_Name_da9834465d7bf75b26f00e78a2412c55a9bb160ab439ee4c0e7742c507a6ac78,lixxxxxxi,Whois_Name,[]
Whois_Email_e3ed53e22963da2784dc9aad7a83c123790617384f67d719fa31fa1c1872a417,sbiqqxxxxx@xxx.xxx,Whois_Email,[]
Whois_Phone_b9383e2d6af1ab1d9f4648f2b7bd348fb875f829124662f2ff4b510af4b66b89,+86.870xxxxx,Whois_Phone,[]
IP_C_80052b75991b23fad5ef78809203fc4e0f4af613c2414f51eba45772149a9625,156.245.xxx.0/24,IP_C,[]
Domain_a7eb1ab42b77f5806e61efe29fefa61bb58686f00f241c1753e7f399448e90f7,a7eb1ab42b.com,Domain,[]
ASN_894a39aa8f6405a82567c5c1832fd3a6b110552c2fe84eafa929a3e603fc4387,AS_894a39aa8f,ASN,[]
```

图1. Node.csv数据样本示例

2.2 Link.csv

Link.csv数据文件大小为493M，包括328万条数据记录，每一条数据记录对应一条边，包括表4所示的3个字段。图2展示了Link.csv的数据样本。

字段	说明	类型	示例	说明
relation	边类型	String	r_dns_a	共11类，见表5
source	源节点	String	IP_37f7ed5739b43757ff23c712ae4d60d16615c59c0818bf5f2c91514c9c695845	源节点的id字段值
target	目标节点	String	Domain_2d3bbcec29453b6f56fb85ea28e8e5ea5fc5f5562e0f896b6b52b113a6cc1e44	目标节点的id字段值

表4. Link.csv数据文件—字段说明

relation字段	说明	数量	关联强度
r_cert	域名使用的安全证书	23万	很强
r_subdomain	域名拥有的子域名	45万	很强
r_request_jump	域名间跳转关系	0.06万	很强
r_dns_a	域名对应的IP地址	205万	很强
r_whois_name	域名的注册人姓名	10万	较强
r_whois_email	域名的注册人邮箱	2.8万	较强
r_whois_phone	域名的注册人电话	1.9万	较强
r_cert_chain	证书的证书链关系	1.5万	一般
r_cname	域名对应的别名	13万	一般
r_asn	IP所属的自治域	6.9万	较弱
r_cidr	IP所对应的C段	17万	较弱

表5. 边的名称说明

```
r_dns_a,IP_bc3271fb9ecbb1a888cfcad82529e43432b54b3e4b0606db1b63f7b878e98e37,Domain_3c12294d75e586455f55489ef861e8973795e98c93e0b1fdf768305551fa21d6
r_subdomain,Domain_149bebae336db20900cd0be3f423b1744f0757ba2456d6ab4b985099364ffb73,Domain_3c12294d75e586455f55489ef861e8973795e98c93e0b1fdf768305551fa21d6
r_whois_name,Domain_3c12294d75e586455f55489ef861e8973795e98c93e0b1fdf768305551fa21d6,Whois_Name_af9c8790603b2045d997ea7062e2fd93c931560ae48932b95f20085663878464
r_whois_email,Domain_3c12294d75e586455f55489ef861e8973795e98c93e0b1fdf768305551fa21d6,Whois_Email_2e7c374df8dfbeb2a499b2686e7a448539e49a3ea9bd97ece8de39d1f1a45856
r_whois_phone,Domain_3c12294d75e586455f55489ef861e8973795e98c93e0b1fdf768305551fa21d6,Whois_Phone_4939081cd8c3df7854212ca0855ddcf12a4a1ae4b7eba4c6dbdae8ae2507a03b
r_asn,IP_88ca9d074a27a2212f56cbf588a44e4e8c7e3b331d4cd76fe6d45971788e6ad0,ASN_894a39aa8f6405a82567c5c1832fd3a6b110552c2fe84eafa929a3e603fc4387
r_subdomain,Domain_9fc9d03394e206e849fc84bb181e8f5e375b80abf8267235841dfe828a350e4a,Domain_5f8cde6da8765c697ccd110e56de9fc1190060db64edd11116711cad643e917e
r_dns_a,Domain_9fc9d03394e206e849fc84bb181e8f5e375b80abf8267235841dfe828a350e4a,IP_bc3271fb9ecbb1a888cfcad82529e43432b64b3e4b0606db1b63f7b878e98e37
r_cidr,IP_bc3271fb9ecbb1a888cfcad82529e43432b64b3e4b0606db1b63f7b878e98e37,IP_CIDR_ac0bb4a963926bcdc47fbef02b55d7991da54af99f75c413afeb238108af90c4
r_dns_a,Domain_149bebae336db20900cd0be3f423b1744f0757ba2456d6ab4b985099364ffb73,IP_bc3271fb9ecbb1a888cfcad82529e43432b64b3e4b0606db1b63f7b878e98e37
r_dns_a,Domain_3c12294d75e586455f55489ef861e8973795e98c93e0b1fdf768305551fa21d6,IP_bc3271fb9ecbb1a888cfcad82529e43432b64b3e4b0606db1b63f7b878e98e37
```

图2. Link.csv 数据样本示例

2.3 黑灰产网络资产图谱模型

黑灰产网络资产图谱数据集中包括8种类型的节点和11种类型的边，图3给出了黑灰产网络资产图谱抽象模型，说明了各类型节点间的可能关联关系类型。

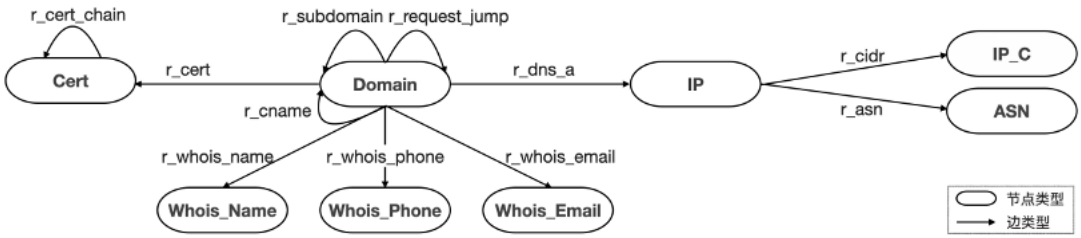

图3. 黑灰产网络资产图谱抽象模型

图4给出了以“5.180.xxx.xxx”IP地址为线索（图4中红色节点），在黑灰产网络资产图谱数据集中挖掘到的小型黑灰产团伙的网络资产子图。图中的N1、N3是安全证书节点，绝大部分域名关联到这两个安全证书。N2是IP节点，许多域名关联到这个IP地址。这些现象反映了许多域名（业务网站）共同使用了这两个安全证书，并且一部分网站部署在了同一个IP地址（服务器）上。另外，这些域名对应的网站大部分都是涉赌、涉黄、涉枪、游戏私服类网站。综上，该子图中的网络资产可能由同一个黑灰产团伙掌握，该黑灰产团伙同时开展了多项非法业务，其核心网络资产是N1、N2和N3，这三个核心网络资产之间的通路是网络业务的关键链路。

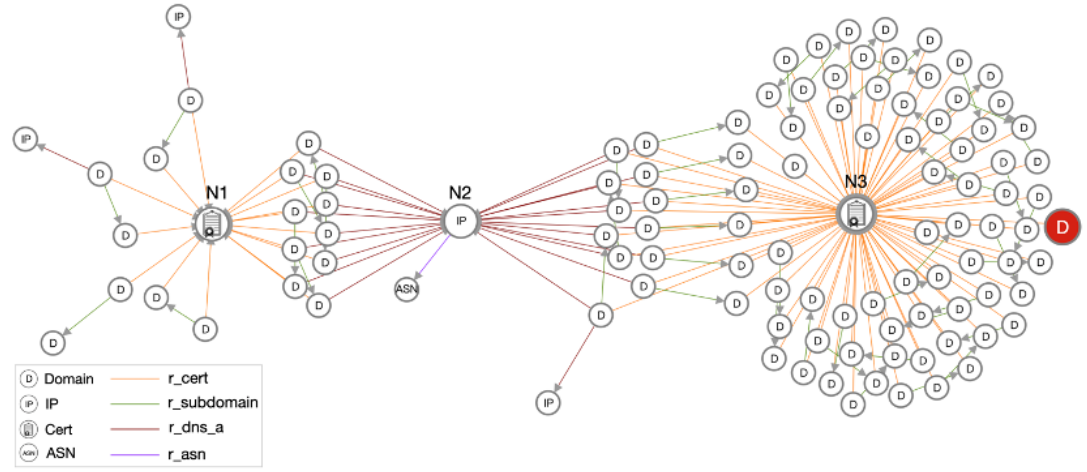

图4. 某小型黑灰产团伙掌握的网络资产子图示例

三、题目说明

任务 1.1：请设计并实现一个完整的可视分析系统，包含**多个视图和视图交互**，对其功能和分析流程进行描述。简述采用的可视分析方法，比如：子图挖掘方法、核心网络资产识别方法、关键链路识别方法、图可视化方法、图交互分析方法等。（30 分）

任务 1.2：请使用你实现的系统，根据附录1所示的五个黑灰产团伙的网络资产线索，选择**至少三个线索**，在黑灰产网络资产图谱数据集中挖掘其对应的网络资产子图（一个子图期望是由同一个黑灰产团伙掌握的网络资产及其关联关系）；识别每个子图中的核心网络资产和关键链路；用图表的形式呈现结果并简要分析每个黑灰产团伙网络运作机制。（40 分）

任务 1.3：请使用你实现的系统，在黑灰产网络资产图谱数据集中挖掘**至少两个**网络资产子图（与任务 1.2 不同的子图）；识别每个子图中的核心网络资产和关键链路；用图表的形式呈现结果并简要分析每个子图对应的黑灰产团伙的网络运作机制。（40 分）

Note：对于任务 1.2 和任务 1.3，答题内容建议包含下述内容：

1、每个子图的节点与边的总数量和分类型数量的统计列表；

2、每个子图的核心网络资产与关键链路列表；

3、每个子图的图拓扑结构可视化结果，建议包含核心网络资产与关键链路信息；

4、简要描述你通过可视化系统发现该子图的逻辑。

四、作业说明

- 2 - 3名同学一组，请同学们自行组队。
- 提交的作业文件（zip 格式，命名为“PJ_成员一_成员二_成员三.zip”）。须包括:
    1. 可运行的可视分析系统；
    2. 视频，包含系统介绍和一个子图挖掘案例的描述（5 分钟左右）；
    3. 作业报告（pdf 格式）；
- 作业文件提交截止日期: 2022 年 5 月 26 日晚 23:59 分
- Presentation 日期: 2022年 5 月 26 日
- 请勿抄袭他人代码，一经发现 0 分处理。
- 该题目改自 ChinaVis2022 中国可视化与可视分析大会数据可视化竞赛（http://chinavis.org/2022/challenge.html）赛道一，鼓励同学们参与比赛，按照竞赛要求提交结果~