2022 年第九届中国可视化与可视分析大会

数据可视分析挑战赛 挑战 1

（ChinaVis Data Challenge 2022 mini challenge 1）

赛题：黑灰产网络资产图谱可视分析

1. 背景介绍

网络黑灰产是指利用信息技术和网络技术，实施各类违法犯罪活动来谋取不正当利益的产业形态。目前，在互联网运行的内容秩序威胁型黑灰产是最常见网络黑灰产类型，它们以公开网站为载体来传播违法违规内容，开展网络诈骗、网络赌博、网络色情、违禁品交易等犯罪活动，严重侵害网络生态的健康发展，甚至威胁着网民生命财产安全。

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，因为网站复本多，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。

奇安信公司通过多种技术手段，在多个公开数据源中收集和整理黑灰产网络资产信息，形成了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。

2. 数据介绍

本次挑战赛提供的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：http://www.chinavis.org/2022/challenge.html

https://github.com/csuvis/CyberAssetGraphData

下面是这两个数据文件的字段说明。

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。

表 1. Node.csv 数据文件—字段说明

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 2. 节点类型说明

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 3. 黑灰产业务类型说明

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

```
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。

表 4. Link.csv 数据文件—字段说明

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 5. 边的名称说明

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

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。

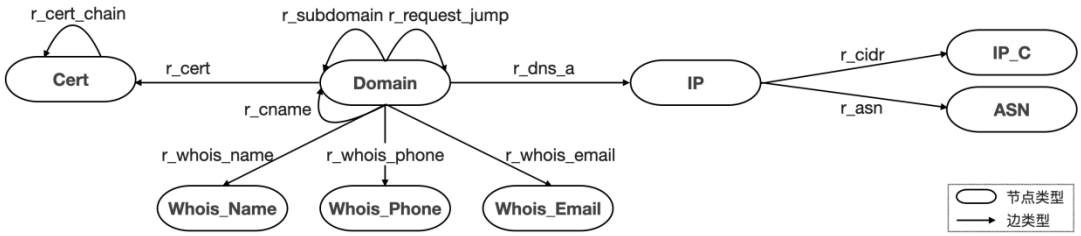

图 3. 黑灰产网络资产图谱抽象模型

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。

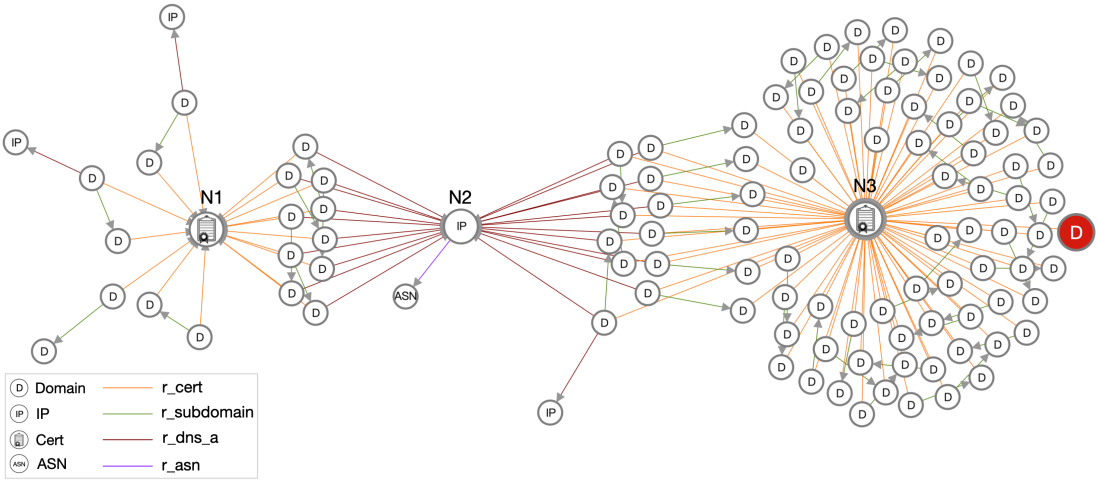

图 4. 某小型黑灰产团伙掌握的网络资产子图示例

3. 题目说明

挑战 1.1：请根据附录 1 所示的五个黑灰产团伙的网络资产线索，在黑灰产网络资产图谱数据集中分别挖掘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。（请将答案尽量控制在 2000 字、10 张图片、10 个表格内）

挑战 1.2：请在黑灰产网络资产图谱数据集中挖掘不少于五个网络资产子图（与挑战 1.1 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。（请将答案尽量控制在 2000 字、10 张图片、10 个表格内）

挑战 1.3：请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。（请将答案尽量控制在 2000 字、5 张图片、3 个表格内）

对于挑战 1.1 和挑战 1.2，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、请按数据集的 csv 格式标准，用附件提供每个子图的详细节点和边信息，建议一个子图提供 1 个节点 csv 文件和 1 个边 csv 文件。

附录 1：网络资产线索

黑灰产团伙	节点 id	节点 name	节点类型
团伙 1（小型团伙）	Domain_c58c149eec59bb14b0c102a0f303d4c20366926b5c3206555d2937474124beb9	c58c149eec.com	Domain
	Domain_f3554b666038baffa5814c319d3053ee2c2eb30d31d0ef509a1a463386b69845	f3554b6660.com	Domain
团伙 2（中型团伙）	IP_400c19e584976ff2a35950659d4d148a3d146f1b71692468132b849b0eb8702c	156.241.xxx.xxx	IP
	Domain_b10f98a9b53806ccd3a5ee45676c7c09366545c5b12aa96955cde3953e7ad058	b10f98a9b5.com	Domain
团伙 3（中型团伙）	Domain_24acfd52f9ceb424d4a2643a832638ce1673b8689fa952d9010dd44949e6b1d9	24acfd52f9.com	Domain
	Domain_9c72287c3f9bb38cb0186acf37b7054442b75ac32324dfd245aed46a03026de1	9c72287c3f.com	Domain
	Domain_717aa5778731a1f4d6f0218dd3a27b114c839213b4af781427ac1e22dc9a7dea	717aa57787.com	Domain
	Domain_8748687a61811032f0ed1dcdb57e01efef9983a6d9c236b82997b07477e66177	8748687a61.com	Domain
	Whois_Phone_f4a84443fb72da27731660695dd00877e8ce25b264ec418504fface62cdcbbd7	+1.971xxxxx	Whois_Phone
团伙 4（大型团伙）	IP_7e730b193c2496fc908086e8c44fc2dbbf7766e599fabde86a4bcb6afdaad66e	23.82.xxx.xxx	IP
	Cert_6724539e5c0851f37dcf91b7ac85cb35fcd9f8ba4df0107332c308aa53d63bdb	6724539e5c	Cert
团伙 5（大型团伙）	Whois_Phone_fd0a3f6712ff520edae7e554cb6dfb4bdd2af1e4a97a39ed9357b31b6888b4af	+86.400xxxxx	Whois_Phone
	IP_21ce145cae6730a99300bf677b83bbe430cc0ec957047172e73659372f0031b8	3.234.xxx.xxx	IP
	Domain_7939d01c5b99c39d2a0f2b418f6060b917804e60c15309811ef4059257c0818a	7939d01c5b.com	Domain
	Domain_587da0bac152713947db682a5443ef639e35f77a3b59e246e8a07c5eccae67e5	587da0bac1.com	Domain

附录 2：子图挖掘业务规则

挖掘网络资产子图需要深度结合应用场景和领域知识。在此，我们提供了一些基础业务规则以供参考。我们鼓励参赛选手引入更多业务规则及相关领域知识。

业务规则 1：建议主要在起始节点 3 跳关联内挖掘网络资产子图。

业务规则 2：建议参考边的关联强度挖掘网络资产子图。对于关联强度较弱的边指向的目标节点，不建议挖掘其 1 跳以外的节点；对于关联强度一般的边指向的目标节点，不建议挖掘其 2 跳以外节点。

业务规则 3：建议根根据实际场景丰富或过滤网络资产子图，比如：对于某些节点，允许加入与之关联的 3 跳外节点；又比如：当一个节点有成百上千个同类型邻居节点（关联关系类型也一样）时，可以适当过滤其邻居节点，以减少子图总体规模。

业务规则 4：我们推荐参赛者挖掘的网络资产子图规模如下：（1）小型黑灰产团伙的网络资产子图规模在 400 个节点、800 条边以内；（2）中型黑灰产团伙的网络资产子图规模在 800 个节点、1600 条边以内；（3）大型黑灰产团伙的网络资产子图规模在 3000 个节点、6000 条边以内。

附录 3：核心网络资产识别业务规则

我们提供了一些识别核心网络资产的基础业务规则以供参考，我们推荐参赛选手在实际应用过程中引入更多的业务规则及相关领域知识。

业务规则 1：如果某个网络资产 50%以上的邻边关联强度较弱，则该资产不被认为是核心网络资产。

业务规则 2：同时关联 2 个以上 IP 地址的 Domain 网络资产很大概率使用了内容分发网络。因此，Domain 网络资产所关联的多个 IP 地址不被认为是核心网络资产。

附录 4：关键链路识别业务规则

我们提供了一些识别关键链路的基础业务规则以供参考，我们推荐参赛选手在实际应用过程中引入更多的业务规则及相关领域知识。

业务规则 1：两个核心网络资产间长度大于 4 跳的路径不被认为是关键链路。

业务规则 2：两个核心网络资产间存在多条路径时，路径越短越重要。

业务规则 3：两个核心网络资产间路径的关联强度越强则越重要。

附录 5：黑灰产网络资产图谱构建方法

主要撰写人：中南大学 赵颖；奇安信科技集团 黄鑫 赵晋龙

1 引言

网络黑灰产是指网络世界中违法违规的产业形态，它们依托于网络技术和互联网环境，进行有组织、有目的、有分工的规模化违法违规活动，影响着网络生态的健康发展，甚至威胁着网民生命财产安全。“黑产”业务直接触犯法律，比如：黄赌毒枪业务、网络诈骗、黑客攻击等；“灰产”业务游走在法律边缘并为“黑产”提供辅助，比如：垃圾信息、恶意注册、虚假认证等。近年来，网络黑灰产呈现加速蔓延之势，2018 年《网络黑灰产治理研究报告》显示，当年国内超 7 亿网民受黑灰产影响，造成经济损失估算达 900 亿元，且网络诈骗案每年以 20%以上速度增长。2020 年《全球风险报告》指出，网络黑灰产的市场效益比肩世界第三大经济体，网络犯罪将是未来十年全球最大风险之一。

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。

黑灰产网络资产分为外围资产、普通资产和核心资产。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。普通网络资产是普通不直接向网民公开的资产。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。查证和封堵核心网络资产是目前打击黑灰产的主要手段之一。主要原因有三条。一是封堵外网资产效率低且被动滞后，因为网站复本多，存活周期短，域名更换频繁。二是封堵核心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。

然而，监管部门在打击黑灰产核心网络资产时普遍缺乏自动的网络资产信息整合手段。外围网络资产信息可以通过群众举报和网络搜索获得，但核心网络资产信息不直接向网民公开，并分散或隐藏在多个异构数据源中，比如：服务器 IP 地址存于域名解析数据库中，网站安全证书隐含在域名服务器资源请求返回内容中。监管部门亟需一种信息整合手段，从某个举报的非法网站域名为起点，广泛从多源数据中自动挖掘网络资产信息，并整合它们之间的关联关系。

本文介绍了黑灰产网络资产图谱构建方法。首先，我们定义了黑灰产常用的 8 类网络资产和 11 类网络资产间的关联关系。然后，我们构建了一个点边双异质的抽象图模型来描述网络资产类型及其关联关系类型。最后，我们综合使用了爬虫、检索、页面解析等技术手段，从 3 个外部数据源和 1 个内部数据源中挖掘与整合网络资产具体实体信息及其关联关系，最终形成黑灰产网络资产图谱数据集（下文中，我们亦简称为网络资产图谱）。我们公开了经过脱敏处理的黑灰产网络资产图谱数据集，该数据集包含 237 万个节点和 328 万条边。我们期望通过公布大规模、高质量的真实数据集，吸引更多科研人员关注黑灰产治理，推动面向黑灰产治理的大数据分析技术的发展和创新。

2 相关知识

2.1 黑灰产治理现状

网络黑灰产有四类，分别是内容秩序威胁型黑灰产、数据流量威胁型黑灰产、技术威胁型黑灰产和暗网。内容秩序威胁型黑灰产是最常见的黑灰产类型，主要以网站为载体来传播违法违规内容，以网络赌博、网络色情、违禁品交易最为猖獗。数据流量威胁型黑灰产通过流量劫持、恶意点击、刷单刷量、数据窃取等违法手段牟取不法利益。技术威胁类黑灰产为网络犯罪提供技术支持，比如：恶意注册、木马植入、钓鱼网站、恶意软件等。暗网是无法通过常规互联网搜索和访问的“不可见网”，充斥着大量违法犯罪交易，具有很强的匿名性、隐蔽性，是各类黑灰产的寄生平台。我们主要关注内容秩序威胁型黑灰产，该类黑灰产的业务需将网站暴露在公共网络中，因此，我们可以得到相关网站的域名，并以此为线

索，进一步在网络中挖掘相关网络资产信息。

我国一直积极关注网络黑灰产治理工作。近年来，国家司法机关相继推出多项指导政策和多部法律法规，如《民典法》、《网络安全法》、《国家网络空间安全战略》等，使得黑灰产治理有法可依。比如，2020 年全国网安部门联合发起“净网 2020”行动，重拳打击网络诈骗和网络赌博等违法犯罪活动。各大平台企业也群策群力，积极承担网络黑灰产治理责任，比如，2020 年抖音封禁 5 万多个涉黑灰产的账号；百度、阿里巴巴等企业联合发布了《网络黑灰产治理研究报告》、《网络犯罪防范治理研究报告》等。但黑灰产治理之路仍然任重道远，需要政府、企业、法律工作者、安全专家、学者等群策群力，加强跨界协同，推进技术攻坚，共同营造和谐的网络环境。

2.2 知识图谱构建

知识图谱能结构化地描述客观世界中的概念、实体及其关系，将信息表达成更接近人类认知的形式。知识图谱有强大的语义处理和互联组织能力，已经被广泛用于知识推理、智能推荐、自动问答、语义搜索等领域。知识图谱构建一般经过知识建模、知识获取、知识融合、知识存储等过程，涉及实体抽取、关系抽取、属性提取、实体消歧、知识合并等技术。本文黑灰产网络资产图谱的构建参考了知识图谱的构建过程和构建技术。

黑灰产网络资产图谱与知识图谱的相同点有三个方面。首先，网络资产图谱与知识图谱都用点边双异质图作为基本数据结构。然后，两者都有抽象概念层面的图模型和具体实体层面的图模型。知识图谱有本体层和实体层。网络资产图谱有网络资产类型和关系类型抽象图模型，也有具体网络资产实体层面的关系图。最后，两者的构建过程都基于对文本信息的分析和对实体与关系的抽取。因此，我们将整个数据集命名为“图谱”，当前数据集可以看做简单知识图谱；该数据集中任意子图，我们称之为网络资产图，是一种异质信息网络。

黑灰产网络资产图谱与知识图谱的不同点来自三个方面。首先，黑灰产网络资产图谱构建过程不涉及知识名词消岐等复杂环节，网络资产图谱的节点，比如：域名和 IP 地址等，大部分都有可唯一标识的信息。然后，知识图谱的本体层和实体层可以联合应用，但黑灰产网络资产图谱中的抽象图模型不能直接应用。最

后，当前的黑灰产网络资产图谱只有 8 类节点和 11 类边，对黑灰产网络运作知识覆盖不足，难以全面支撑知识推理、自动问答等知识图谱的主要应用方向。

3 网络资产图谱构建

黑灰产网络资产图谱构建的整体思路是：首先将黑灰产团伙掌握的网络资产及其关系抽象为一个点边双异质抽象图模型。图中存在多类节点和多类边，每类节点代表一种网络资产，每类边代表一种关系。然后具象化网络资产图谱，即，通过多种技术手段，从多个数据源中提取具体网络资产实体和关联关系的相关信息，得到黑灰产网络资产图谱数据集。数据集中每一个节点代表一个具体的网络资产，每条边代表某两个网络资产间具体的关系。

3.1 网络资产实体类型定义

网络资产实体类型是黑灰产业务运行需要用到的网络资产的分类抽象表达。图谱构建第一步是确定需要关注的网络资产类型，以及每个类型的重要性。经过多次专家研讨，本文确定了 8 种值得关注的网络资产实体类型，如表 1 所示，它们被大部分黑灰产使用，并且其具体实体信息可以通过技术手段从数据源中提取。

表 1 网络资产实体类型说明

序号	实体类型	说明	重要程度
1	Domain	网站域名	非常重要
2	IP	网站的 IP 地址	非常重要
3	Cert	网站用的 SSL 安全证书	非常重要
4	Whois_Name	网站域名的注册人姓名	重要
5	Whois_Phone	网站域名的注册人电话	重要
6	Whois_Email	网站域名的注册人邮箱	重要
7	IP_C	IP 的 C 段	一般
8	ASN	IP 的自治域	一般

非常重要的网络资产实体类型包括域名、IP 和安全证书，它们是黑灰产业务网络运营的基础。域名是网站便于记忆的网络地址。IP 是网站的实际网络地址。

安全证书在本文主要指 SSL 安全证书，它是网站运行的全球唯一安全执照，保障用户端和服务器间数据传输的安全性。

在注册黑灰产网站域名时需要提供注册人姓名、电话和邮箱。如果用真实个人信息注册，这三类信息将为监管部门提供网络黑灰产在真实世界中的关联人线索。因此，专家将三种 Whois 网络资产实体类型视为重要类型，它们不直接影响黑灰产业务运行，但能为黑灰产治理提供高价值信息。

IP 的 C 段和自治域是对黑灰产业务运维有益的网络资产实体类型。IP 的 C 段反映黑灰产团伙掌握了某 C 段内所有或大部分 IP 地址。IP 的自治域提供了黑灰产所掌握的 IP 地址所属国家的运营商、机构等信息，在 IP 地址初始化时分配。

3.2 网络资产关系类型定义

网络资产实体间关系建模是定义网络资产实体间可能存在哪些类型的关联关系，其关联强度（紧密程度）如何。经过多次专家研讨和初步数据分析，我们发现 8 种网络资产类型间可能存在 11 种关联关系。表 2 给出了这 11 种关系类型和其关联强度。图 1 给出了由 8 种网络资产类型和 11 种关系类型构成的抽象图谱模型。

表 2　网络资产图谱关系类型说明

序号	关系类型	说明	强度
1	r_cert	域名使用的安全证书	很强
2	r_subdomain	域名拥有的子域名	很强
3	r_request_jump	域名间跳转关系	很强
4	r_dns_a	域名对应的 IP 地址	很强
5	r_whois_name	域名的注册人姓名	强
6	r_whois_email	域名的注册人邮箱	强
7	r_whois_phone	域名的注册人电话	强
8	r_cert_chain	证书的证书链关系	一般
9	r_cname	域名对应的别名	一般
10	r_asn	IP 所属的自治域	弱
11	r_cidr	IP 所对应的 C 段	弱

很强的关系类型有四种，它们反映了黑灰产核心网络资产间的直接关联，能有效还原黑灰产核心网络资产链条。(1) r_cert 表示域名和证书之间的关联，比如：某域名对应的网站使用了某安全证书。(2) r_subdomain 表示域名对应的子域名，即下级域名。(3) r_dns_a 反映了域名对应的 IP 地址。一个域名被访问时需要通过 DNS 服务解析为具体 IP 地址。在 DNS 解析中一个域名对应的 IP 地址被称为 DNS A 记录，所以该关系命名为 r_dns_a。(4) r_request_jump 表示两个域名间存在自动跳转关系，即打开某域名对应网站时会自动跳转到另外一个域名对应的网站。这是黑灰产团伙常用的隐藏与引流策略，自动跳转到的目标网站可能才是真实业务网站。

强度一般的关系类型有两种。(1) r_cert_chain 表示安全证书和上级安全证书或证书签发机构之间的关联。SSL 安全证书采用层级化管理体系，比如：某证书签发机构掌握了顶级证书，该机构可以给申请者派发下级证书。r_cert_chain 关联强度一般，因其无法直接证明黑灰产团伙与证书签发机构间存在利益联系。(2) r_cname 是域名和域名别名之间的关联。在 DNS 解析中，一个域名对应的别名被称为 CNAME 记录，所以该关系命名为 r_cname。为了便于域名变更和 IP 地址变更，可以在 DNS 服务中为多个域名设置同一个别名，再为别名设置 IP 地址。但如果域名使用了内容分发服务，域名与其别名间就失去直接关联。因此关联强度一般。

三个与域名注册人相关的关系类型属于强关联，因为这三种关系能为监管部门提供网络黑灰产在真实世界中的关联人线索。r_asn 和 r_cidr 属于弱关联，因为自治域和 IP 段的覆盖范围广，无法提供确切的黑灰产业务链信息。

3.3 网络资产实体与关系提取

实体类型定义和关系类型定义实现了黑灰产网络资产图谱的抽象建模，如图 1 所示。在抽象模型的指导下，我们需要将网络资产图谱具象化，即通过技术手段从数据源中提取具体网络资产实体和关系的相关信息。但具象化过程面临两个挑战。(1)监管部门初始时只有少量群众举报的黑灰产网站域名信息，需要通过一定技术手段并借助外部数据源，提取其它类型实体和关系的相关信息，称之为信息富化。(2)不同类型的网络资产信息分散在不同的数据源中，数据源异构性导致

单一技术手段无法完成信息富化，需要组合多种技术手段对实体和关系信息进行挖掘与整合。

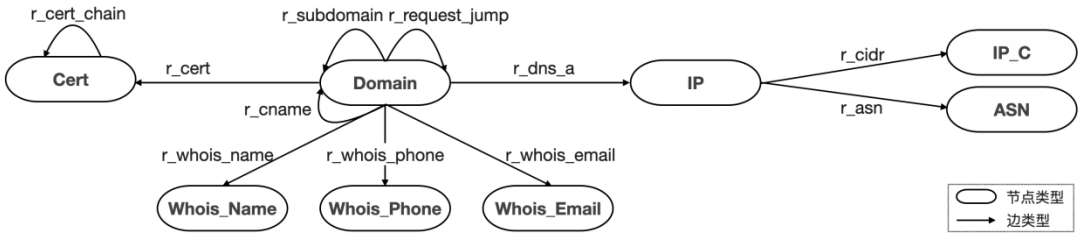

图 1 黑灰产网络资产图谱抽象模型

我们深入分析了各个可用数据源的特点，对不同数据源设计了针对性技术手段，形成了一套完整的网络资产实体和关系提取流程。如图 2 所示，该流程利用了 5 种技术手段，从 1 个内部和 3 个外部数据源中提取实体和关系。下面具体介绍。

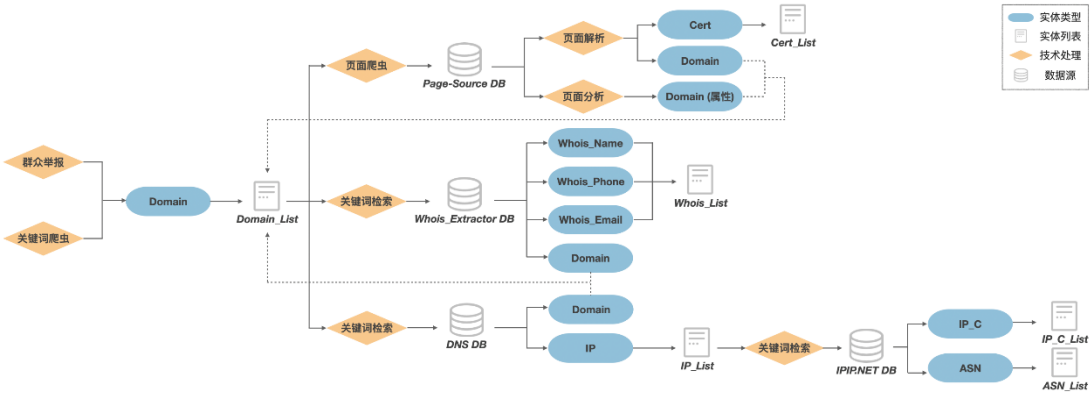

图 2 黑灰产网络资产实体与关系信息提取流程图

域名是信息富化的源头。监管部门通过群众举报可获得一些域名，但数量有限。为了扩充黑灰产域名，我们定义了近百个黑灰产网站高频关键词，比如棋牌、娱乐平台、直播等，然后采用关键词爬虫技术在搜索引擎中自动爬取相关域名并将获得的 Domain 实体存储在 *Domain_List* 中。

Page-Source DB 是由黑灰产治理技术团队维护的内部数据源。我们采用页面爬虫技术，根据 *Domain_List* 中的域名，向域名服务器发送请求，爬取对应页面的源代码。然后把域名服务器返回的内容与爬取的页面源代码存储在 *Page-Source DB* 中。我们研发了页面解析技术和页面分析技术来分别处理存储数据。页面解析技术基于正则表达式分析域名服务器的返回内容，从中挖掘出域名使用

的安全证书，并将获得的 Cert 实体与 r_cert 关系、r_cert_chain 关系存入 *Cert_List* 中。页面解析技术还可以提取页面源代码中基于嵌入 JS 的自动跳转域名，并将获得的 domain 实体和 r_request_jump 关系存入 *Domain_List* 中，以实现网络资产信息的迭代挖掘。页面分析技术采用训练好的词向量与文本分类模型，比如：FastText，对页面进行分类，并将分类结果作为域名属性存储在 *Domain_List* 中，主要类别包括：涉黄、涉赌、诈骗、涉毒、涉枪等。页面分析技术还可以帮忙剔除 *Domain_List* 中的非黑灰产域名。

Whois_Extractor DB 是一个外部数据源。它提供域名的公开注册信息，比如：域名注册人姓名、注册人电话、注册人邮箱、注册机构等。我们采用关键词检索技术，从 *Whois_Extractor DB* 库中检索 *Domain_List* 中每个域名的 Whois_Name，Whois_Phone 和 Whois_Email 信息，并将获取到的 Whois 实体和 Whois 关系存入 *Whois_List* 中。另外，域名注册时可同时注册一个域名的下级域名（子域名），因此在 *Whois_Extractor DB* 中检索某域名时，还能提取到 Domain（子域名）实体和 r_subdomain 关系，需存入 *Domain_List* 中，以实现网络资产信息的迭代挖掘。

DNS DB 是一个外部数据源。它提供域名解析信息，主要包括：域名对应的 IP 地址和别名。我们采用关键词检索技术，从 *DNS DB* 中提取 *Domain_List* 中每个域名的 IP 地址和域名别名信息，将 IP 实体和 *r_dns_a* 关系存入 *IP_List* 中，将域名（别名）实体和 *r_cname* 关系存入 *Domain_List* 中。

IPIP.NET DB 是一个外部数据源。它提供 IP 地址画像。我们采用关键词检索技术，从 *IPIP.NET DB* 中提取 *Domain_List* 中每个域名的 IP_C 段信息和自治域信息，并将获得的 IP_C 实体和 r_cidr 关系存入 *IP_C_List*，ASN 实体和 r_asn 关系存入 *ASN_List*。

通过上述网络资产实体和关系的提取，我们能从群众举报和关键词爬虫获取到的少量黑灰产网站域名为入口，通过一系列信息富化过程，将大量网络资产实体及其关系整合到一张图谱中，为后续查封非法网络资产和调查黑灰产团伙提供网络作战地图。

4　网络资产图示例

使用上述网络资产图谱构建方法，能在整个数据集中，以一个域名为起点，获取到由成百上千个复杂关联的网络资产构成的点边双异质有向图，我们称之为网络资产图。图 3 给出了一个真实的小型黑灰产团伙的网络资产图，它是整个数据集的一个子图。图中红色节点是起点域名。图中有 111 个节点，其中 Domain 类节点 104 个，IP 类节点 4 个，Cert 类节点 2 个，ASN 类节点 1 个。图中有 181 条边，r_cert 类边 103 条，r_subdomain 类边 50 条，r_dns_a 类边 27 条，r_asn 类边 1 条。我们逐一查看了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，反映了该网络资产图背后的黑灰产团伙是一个运作多种非法业务的复合型团伙。

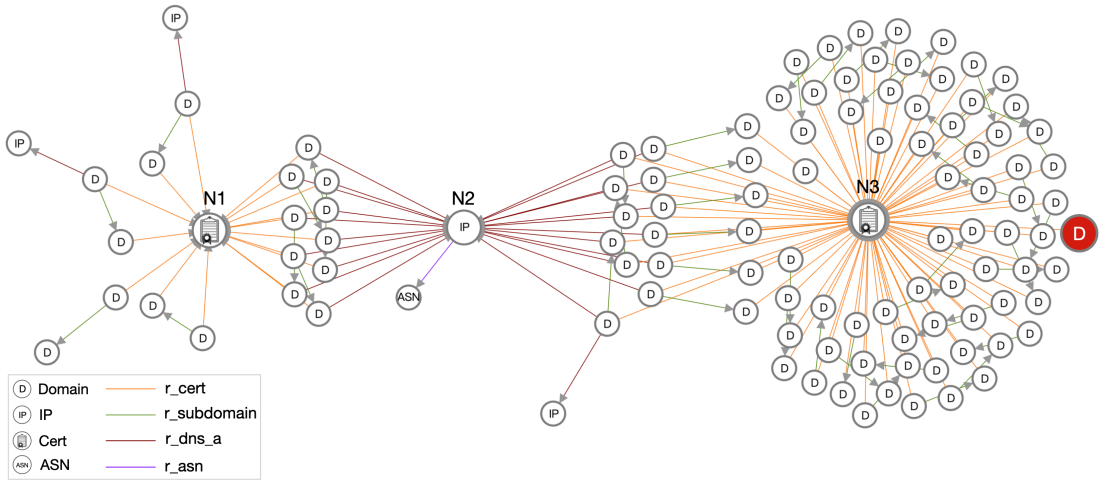

图 3　某小型黑灰产团伙掌握的网络资产及其关联关系示例

我们以图 3 为例来解释核心资产与关键链路。图 3 中的 N1 和 N3 是安全证书节点，绝大部分域名节点关联到这两个安全证书。N2 是 IP 节点，许多域名节点关联到 N2 节点。这些现象反映了该黑灰产团伙掌握的域名（业务网站）共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。因此，图中 N1、N2 和 N3 节点很可能是该黑灰产团伙的核心网络资产，它们之间的多条关系路径都是关键链路。封堵 N1 和 N3 节点对应的安全证书，可以让关联网站面临被监听、篡改、流量劫持、数据泄露等问题，封堵 N3 节点对应的 IP 地址，可以让关联网站难以被访问，使得该团伙需要付出较多的时间和经济成本来恢复业务。

5 数据公开

基于网络资产图谱构建技术，我们经过持续两年多的数据采集和数据清洗，形成了一个大规模、高质量的黑灰产网络资产信息库。本文公开一个黑灰产网络资产图谱数据集。该数据集能帮助黑灰产治理从业人员了解黑灰产的网络运维模式，掌握黑灰产团伙的网络资产分布，制定黑灰产网络资产打击策略，探索黑灰产团伙在真实世界中的线索。我们也希望能吸引广大科研人员关注黑灰产治理，推动面向黑灰产治理的大数据分析技术的发展。

本次公开的黑灰产网络资产图谱数据集由 237 万个节点和 328 万条边组成，以 CSV 格式的文件存储，未压缩时大小共计 722M。每个节点的字段信息有：节点 ID；节点名称，比如：域名字符串或 IP 地址；节点类型（见表 1），对域名类型节点还提供黑灰产业务属性，比如：涉黄、涉赌、诈骗、涉毒、涉枪等。每条边的字段信息有：边类型（见表 2），源节点 ID 和目标节点 ID，边方向由源指向目标节点。

该图谱数据集完全来源于真实世界，为防止泄露黑灰产网站的真实信息，我们对每个节点的名称字段进行了脱敏处理。对于域名、证书和自治域类型的节点，我们在节点名称上做了加密处理。对于 IP 和 IP_C 段类型节点，我们对具体 IP 地址进行了无效化处理。对于 Whois 类型节点，我们只保留了节点名称的少数字符，并使其它字符无效化。

6 未来挑战

黑灰产网络资产信息的挖掘、整合和分析的需求还在不断深化。机器学习、数据挖掘、可视分析、人机交互等大数据分析技术在黑灰产治理中将发挥越来越重要的作用。我们梳理了黑灰产网络资产分析的相关需求和技术挑战。

（1）丰富数据源和拓展网络资产类型。本文通过 3 个外部数据源与 1 个内部数据源获取了 8 类网络资产。无论是数据源还是网络资产类型，未来都有进一步丰富和拓展的空间，力争更全面地还原黑灰产团伙掌握网络资产的形态、数量和关联关系。

（2）网络资产图谱的子图挖掘。当黑灰产网络资产图谱数据集较大时，子图

挖掘是黑灰产网络资产图谱分析的重要需求，需要综合运用图挖掘技术，并深度结合业务知识。子图挖掘是指：给出任意一个或几个网络资产节点信息，比如：域名或 IP 地址，在大规模图谱数据集中找到一个子图，该子图期望是由同一个黑灰产团伙掌握或运作的网络资产的全貌。

（3）大型网络资产图的可视化与交互探索。从本文公开数据集的经验来看，中小规模黑灰产团伙掌握的网络资产为数百个，大型黑灰产团伙掌握的网络资产数量为 2-3 千，超大型黑灰产团伙掌握的网络资产数量可达 1-2 万。作为分析网络资产图的总体拓扑结构和局部关联模式的主要技术手段，图可视化与交互探索技术必然面临布局速度、渲染速度、交互体验流畅性等挑战。

（4）智能识别核心网络资产。查封核心网络资产是目前打击黑灰产的主要手段之一。但前提条件是快速准确地识别核心网络资产及其影响范围。另外，核心网络资产识别无标准答案，业务知识和专业经验依赖性强，人机混合智能应是主要发展方向。

（5）动态构建多人协同作战环境。黑灰产治理可以类比为一个多人协同的动态作战过程，既有线上信息情报收集、整合与分析，也有线下实地调查取证。整个过程随时都可能出现新资产，新关联和新线索，还可能遭遇黑灰产团伙的对抗。监管部门需要一个多人协同的动态作战环境，使得安全专家和执法人员可以在网络空间与真实世界融合的作战地图上，借助 VR/AR 等多模态人机交互技术，灵活使用各类大数据分析技术，协同感知作战态势变化，共同制定和及时调整战略战术。