黑灰产网络资产图谱公开数据集说明

一、总体介绍

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。我们主要关注内容秩序威胁型黑灰产，它们以公开网站为载体来传播违法违规内容，比如：网络赌博、网络色情、违禁品交易。网络黑灰产主要特点是链条化、团伙化、资产化、跨域化。其中资产化是指黑灰产团伙掌握大量且复杂关联的多种网络资产，以支撑产业链运转，比如：网站运维需要域名和 IP 地址，业务支付需要安全证书。

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业务网站域名。普通网络资产是普通的不向网民直接公开的资产。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：某 IP 地址可能同时支持多个网站域名运行，又比如：同一团伙掌控的赌博业务和贩毒业务可能使用了同一数字安全证书。核心网络资产信息一般不直接向网民公开，并分散或隐藏在多个异构数据源中。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。原因来自三个方面：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。三是深度分析核心资产能挖掘多资产或多业务间的关联，有利于还原整个产业链，甚至发现虚拟网络世界背后的真实犯罪人员。

我们以外围网络资产为线索，采用多种技术手段，经过持续两年以上的数据采集和数据清洗，维护了一个大规模、高质量的黑灰产网络资产信息库。现在，我们将一部分数据以黑灰产网络资产图谱数据集形式公布出来。该数据集以点边双异质有向图为基本数据结构，节点表示网络资产实体，边表示网络资产间关联关系，节点规模 237 万，边规模 328 万。

我们期望通过数据集公开，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，探索黑灰产团伙在真实世界中的线索，为黑灰产网络资产打击策略提供辅助决策。我们还期望吸引广大科研人员关注黑灰产治理工作，推动面向黑灰产治理的大数据分析技术的发展，为营造健康的网络生态出谋献策。

二、数据介绍

黑灰产网络资产图谱数据集以 CSV 格式存储，压缩前 727MB，包括 **Node.csv** 和 **Link.csv** 数据文件。下面，我们分别介绍这 2 个数据文件的具体信息。

1、 Node.csv

Node.csv 数据文件大小为 231M，包括 237 万条数据记录，每一条数据记录对应一个节点信息，包括以下 4 个字段，如表 1 所示。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	名称经过了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 类节点有效)	String	['B']	共 10 类，见表 3

```
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

表 2. 节点类型说明

字段	说明	数量
Domain	网站域名	200 万
IP	网站的 IP 地址	20 万
Cert	网站用的 SSL 安全证书	13 万
Whois_Name	网站域名的注册人姓名	1.8 万
Whois_Phone	网站域名的注册人电话	0.2 万

Whois_Email	网站域名的注册人邮箱	0.4 万
IP_C	IP 的 C 段	0.6 万
ASN	IP 的自治域	0.03 万

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

2、 Link.csv

Link.csv 数据文件大小为 496M，包括 328 万条数据记录，每一条数据记录对应一条边，包括以下 3 个字段，如表 4 所示。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

图 2. Link.csv 数据样本示例

表 5. 边的名称说明

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