黑灰产网络资产图谱构建与可视化（拟录用）

赵颖[1)], 付铄雯[1)], 赵鑫[1)], 徐雅琦[1)], 赵勇[1)], 陈运鹏[1)], 周芳芳[1)*], 黄鑫[2)], 李玉伟[2)], 陈卓[2)]

[1)] (中南大学计算机学院 长沙 410083)
[2)] (奇安信科技集团股份有限公司 北京 100015)
*(通信作者 Email 地址: zff@csu.edu.cn)

摘 要: 黄赌毒诈等黑灰产团伙的网络化运作严重破坏着网络生态和社会治安. 网络化运作依赖于掌握域名、IP 地址、安全证书等多种类型的网络资产. 由于公开的网站和域名等外围网络资产副本多且更换频繁, 查封核心网络资产, 比如: 重要 IP 地址和安全证书, 是目前打击网络黑灰产的主要手段之一. 但核心网络资产信息一般分散或隐藏在多个异构数据源中, 难以被获取和识别. 本文提出了一个黑灰产网络资产图谱构建方法, 从多源异构数据中广泛获取网络资产信息及关联关系, 并将其整合到点边双异质有向图中. 本文提出了一组黑灰产网络资产图可视化方法, 基于资产图拓扑特性改进了经典力导引布局算法和社区检测算法, 帮助用户观察和理解资产间复杂关联, 并快速识别核心资产及其影响范围. 本文还公布了一个大规模黑灰产网络资产图谱数据集, 梳理了数据集可支持的黑灰产治理需求, 展望了图谱分析面临的技术挑战, 旨在推动面向黑灰产治理的大数据分析技术的发展和创新.

关键词: 网络黑灰产; 网络资产; 图; 可视化; 公开数据集

中图法分类号: TP391.41 **DOI:**

Constructing and Visualizing Cyber Asset Graphs of Cybercrime Gangs

Zhao Ying[1)], Fu Shuowen[1)], Zhao Xin[1)], Xu Yaqi[1)], Zhao Yong[1)], Chen Yunpeng[1)], Zhou Fangfang[1)*], Huang Xin[2)], Li Yuwei[2)], and Chen Zhuo[2)]

[1)] (*School of Computer Science, Central South University, Changsha* 410083)
[2)] (*Qi An Xin Technology Group Co., Ltd, Beijing* 100015)

Abstract: The internet ecosystem is being severely disrupted by cybercrime gangs, such as online gambling and online drug trafficking. Various cyber assets, such as websites, domains, IPs, and SSL certificates, are fundamental to support cybercrime activities. Deactivating and banning cyber assets are widely-used manners to fight against cybercrimes. However, peripheral assets, such as websites and domains, are accessible but low-value targets for cybercrime fighting due to their easy replicability and replaceability. Core assets associated with numerous peripheral assets, are valuable targets but their information is generally hidden and scattered in heterogeneous online data sources. In this paper, we propose a cyber asset graph construction method for mining and integrating the information of and associations between cyber assets of cybercrime gangs from heterogeneous online data sources. We also propose a set of visualizations that present cyber asset graphs for asset association analysis and core asset identification. Furthermore, a large-scale cyber asset graph dataset is released to the public, aiming to promote the development and innovation of advanced data analysis technologies for cybercrime fighting.

Key words: Cybercrime; cyber asset; graph; visualization; open dataset

---

收稿日期: 2021-12-25; 修回日期: 2022-4-20. **基金项目**: 国家自然科学基金(61872388, 62072470). **赵颖**(1980—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为可视化与可视分析; **付铄雯**(2000—), 女, 硕士研究生, 主要研究方向为可视化与可视分析、数据挖掘. **赵鑫**(1997—), 女, 硕士研究生, 主要研究方向为可视化与可视分析. **周芳芳**(1980—), 通讯作者, 女, 博士, 教授, 博士生导师, 主要研究方向为可视化、虚拟现实、大数据技术. **黄鑫**(1992—), 男, 奇安信雷尔可视化平台部负责人, 主要从事数据可视化领域的技术研究和平台开发.

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[1-3]. “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[4], 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[1], 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.

网络黑灰产主要特点是链条化、团伙化、资产化、跨域化[2,4,5]. **链条化**是指产业的上中下游紧密配合完成非法牟利. 上游负责收集信息资源, 比如: 手机黑卡和网民隐私. 中游负责提供技术支持, 比如: 软硬件系统和网络环境. 下游负责收入变现, 比如: 非法支付和洗钱. **团伙化**是指整个业务链上多人分工明确, 各司其职. **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: 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. **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. 一是封堵外围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的线索.

然而, 监管部门在打击黑灰产核心网络资产时面临两个难题. **第一个难题是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, 从少量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. **第二个难题是缺乏直观的网络资产信息呈现手段.** 一个黑灰产团伙通常掌握成百上千个复杂关联的网络资产. 监管部门亟需对信息整合后的网络资产进行分析, 理解资产间复杂关联关系, 结合经验与场景决策需重点打击的核心资产和预估打击后的影响范围, 甚至找到真实世界中关联人员的相关信息.

针对第一个难题, 本文提出了一个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. 本文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.

针对第二个难题, 本文提出了一组黑灰产网络资产图可视化方法. 首先, 我们发现了黑灰产网络资产图谱具有全局稀疏、局部稠密、多簇多桥的拓扑特性. 然后, 我们改进了经典的 SE (Spring Embedder) 力导引布局算法以更快更好地呈现上述拓扑特性. 最后, 我们综合采用度中心性和随机游走中心性识别了图谱中的核心网络资产, 并改进了 LFM (Local Fitness Maximization) 社区检测算法来检测核心网络资产的影响范围. 这些方法可以帮助用户观察网络资产图的拓扑形态, 理解资产间的复杂关联关系, 找出需重点打击的核心网络资产和预估打击后的影响范围, 甚至发现真实世界中黑灰产团伙关联人的相关线索.

综上所述, 本文主要有**三个贡献**: (1) 提出了

一个黑灰产网络资产图谱构建方法，能有效地从多源异构数据集中挖掘与整合网络资产信息及其关联关系. (2) 提出了一组黑灰产网络资产图可视化方法，能友好地呈现网络资产图拓扑结构，有效地识别图中的核心网络资产并检测其影响范围. (3) 公开了一个黑灰产网络资产图谱数据集，梳理了数据集可以支持的黑灰产治理需求，展望了图谱分析面临的技术挑战，旨在推动面向黑灰产治理的大数据分析技术的发展和创新.

2 相关工作

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[2,4]. 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[6]. 本文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[4,5,7]. 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[8]. 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[9]；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》[4]、《网络犯罪防范治理研究报告》[2]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[10-13].

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[14-15]. 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[16-19]. 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[15,20,21]. 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，我们用“图谱”来命名，而不是用“图”. 当前的黑灰产网络资产图谱可以看做简单知识图谱. 另外，我们将黑灰产网络资产图谱数据集中有意义的子图，称为(黑灰产)网络资产图. 它是一种异质信息图. 我们期望一个有意义的网络资产图能包含同一黑灰产团伙掌握的网络资产及其关联关系.

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[22-24]，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[14,20]，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[17,18].

2.3 图布局技术

图布局技术将难以理解的网络数据以拓扑图的形式展示出来，充分利用人的感知和认知能力，获取图中重要信息[25-26]. 本文采用了图布局技术来展示网络资产图，因此本小节讨论图布局相关工作.

图布局有三条技术路线：力导引布局[27-32]、约束布局[33-37]和降维布局[38-40]. 力导引布局算法利

用节点间力作用的物理模型实现图布局. 经典算法包括基于弹簧模型的 Spring 算法[27], 基于粒子模型的 FR 算法[28], 和结合弹簧模型和粒子模型的 SE 算法[29]. 另外, 许多算法对三大经典算法进行了局部改进, 比如: 改进 FR 算法局部布局不收敛的 GEM 算法, 改进 FR 算法布局效果的 ForceAtlas2 算法, 优化斥力计算复杂度的 Barnes Hut 算法[30-32]. 力导引布局容易理解, 易于实现, 全局和局部结构表现力好, 布局结果符合美学标准[41]. 力导引布局是目前科研和工程实践中的主流技术路线. 但力导引布局有两个缺点: (1)节点间力迭代计算耗时, (2)迭代结果可能无法全局最优.

约束图布局将图布局问题转化为最优化问题, 通过求解目标函数得到布局结果. KK 算法是经典代表[33], 其目标函数假设理想布局中节点对之间的距离为最短路径, 并通过 Newton-Raphson 方法求解. 该类算法优点是目标函数求解思路多, 比如: 共轭梯度下降、随机梯度下降、梯度投影[34-35]等, 并且可自定义或扩展优化项以满足用户特定布局需求, 比如: 限制连边长度[36]和子图布局形状[37]. 约束图布局有两个缺点: (1) 目标函数变量多, 求解复杂耗时, 难以达到全局最优. (2) 需要合理应用较多约束才能达到高质量布局效果, 使用难度大.

降维布局通过计算节点对在高维空间中的相似性, 利用降维技术把节点降维到二维平面中得到布局结果. 经典算法包括 HDE、Pivot MDS、DRGraph 等[38-40]. 该类算法的优点是布局速度快, 能应对大规模网络的高效布局需求. 但该类算法在降维过程中利用的信息较少, 导致局部结构表现力不好, 布局结果可读性差.

综合上述讨论和网络资产图拓扑特性(见 4.2 小节), 本文选用力导引布局为技术路线. 为解决力导引布局计算耗时问题, 已有工作分两种改进途径. (1) 加速渲染. 力导引图布局耗时步骤包括力迭代计算和计算结果渲染两部分. 尤其在节点数量达到万级时, 渲染耗时更明显. 采用 GPU 并行渲染可以明显提升渲染效率, 代表性方法是 NetV.JS[42]. (2) 加速力迭代计算. 节点间力迭代计算次数多, 会导致布局时间长. 目前, 普遍采用基于四叉树计算的 Barnes Hut 方法[30], 通过把邻近的节点分组并近似看为一个整体, 来近似计算节点之间的斥力, 从而加速节点间的力计算效率. 考虑到本文所用网络资产图案例规模主要在千数量级, 本文选用了加速力迭代计算途径来提高图布局效率.

3 网络资产图谱构建

根据黑灰产网络资产数量多、类型多、关联复杂的特点, 本文以构建网络资产图谱为总体思路, 来解决网络资产信息整合难题. 我们将某黑灰产团伙掌握的网络资产及其关联关系抽象为一个点边双异质图谱. 图谱中存在多类节点, 每类节点代表一种网络资产, 每个节点代表一个具体网络资产实体. 图谱中也存在多类边, 每类边代表网络资产间的一种关联关系, 每条边代表一个具体关系. 因此, 网络资产信息整合问题就转化为了网络资产图谱构建问题. 点边双异质图谱构建过程一般包括实体类型定义, 关系类型定义, 实体与关系提取三个步骤. 本章分别介绍这三个步骤.

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在线数据源中提取. 根据网络资产对黑灰产业务运作的影响程度, 我们将 8 种网络资产的重要程度划分为非常重要、重要和一般.

表 1 网络资产图谱实体类型说明

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.

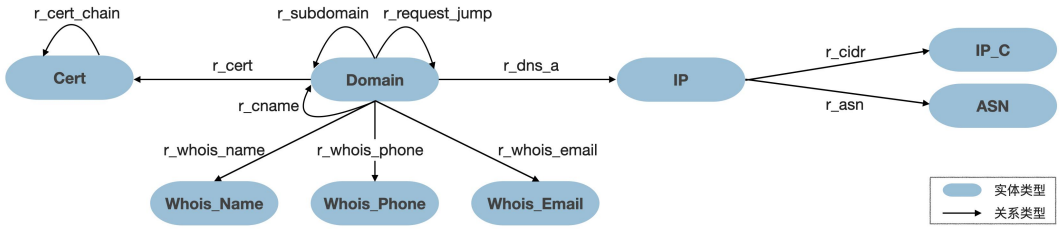

图 1. 黑灰产网络资产图谱抽象模型

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.

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.

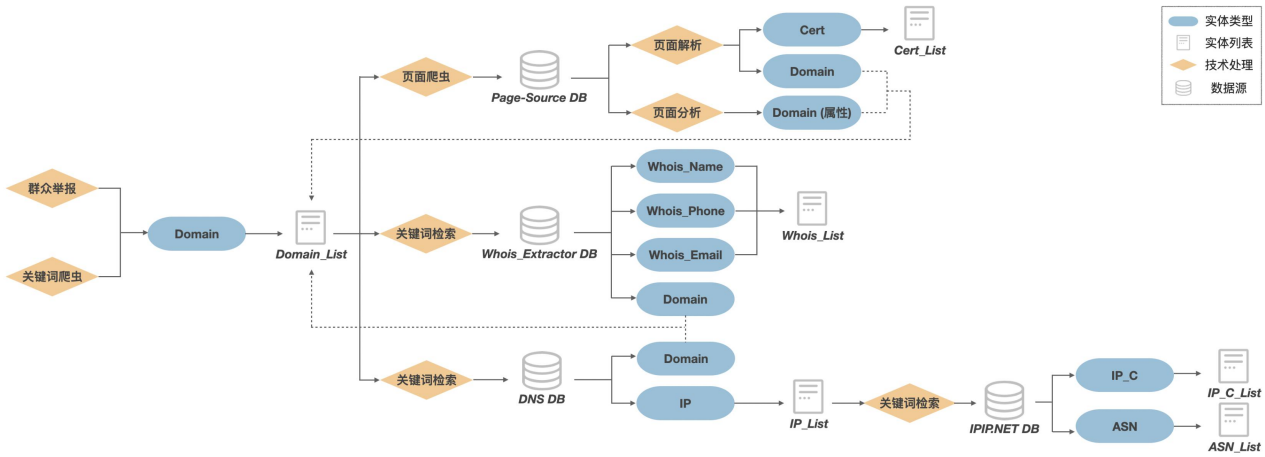

图 2. 黑灰产网络资产实体与关系信息提取流程图

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[43-44]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[44-45], 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[46-47], 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[48], 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* 中非黑灰产域名.

Whois_Extractor DB 是一个外部数据源[49]. 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[50]. 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[51]. 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, 我们能以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　网络资产图谱可视化

网络资产图谱中包含许多黑灰产团伙的网络资产及其关联关系. 我们将同一个黑灰产团伙掌握的网络资产及其关联关系称为一个网络资产图，是网络资产图谱中的一个有意义子图. 图 3 给出了一个真实的小型黑灰产网络资产图. 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.

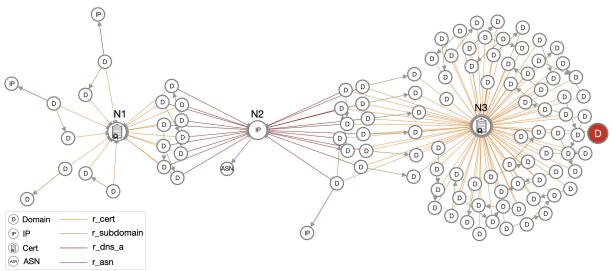

图 3. 小型黑灰产网络资产图示意图.

为了更好地利用网络资产图，监管部门亟需一系列可视化方法直观呈现，以便于观察网络资产拓扑形态，理解资产间的复杂关联关系，结合经验与场景来决策需重点打击的核心资产和预估打击后的影响范围，甚至找到真实世界中嫌疑人员的相关信息. 本章将给出一个初步的网络资产图可视化解决方案.

4.1 网络资产图可视化需求

经过与黑灰产治理专家研讨，网络资产图可视化的需求主要体现在下面三个方面:

R1: 网络资产图需要一种可视化展示方法,来直观呈现网络资产间的关联关系，帮助用户观察总体拓扑形态和理解复杂关联.

R2: 网络资产图可视化需要帮助用户快速识别核心资产与关键链路，因为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(见第一章). 核心资产是图中连接了大量资产的核心节点. 关键链路是连接图中多个核心资产的边.

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.

R3: 网络资产图可视化需要呈现核心资产的影响范围，以便于用户评估打击效果.

4.2 网络资产图拓扑布局方法

图布局技术是对图进行拓扑结构呈现的核心技术(R1). 现有主流图布局技术路线包括力导引布局[27-32]、约束布局[33-37]和降维布局[38-40]. 我们对这三种技术路线下的十多种具体布局算法进行了实验比较，根据网络资产图本身拓扑特性和可视化需求，最后选择了 SE (Spring Embedder) 力导引布局算法[29]，布局效果如图 4(a)和图 5(a)所示. 该算法能较好地呈现网络资产图的整体拓扑，也能较为清晰地展示局部结构信息.

通过观察图的拓扑结构，我们总结了网络资产图的 3 个常见拓扑特性.

全局稀疏，局部稠密. 网络资产图属于稀疏图，因为边数远小于有相同节点数的完全图的边数[52]. 图谱中有一些局部区域存在较多连边，如图 5(a)的左中部区域，呈现局部稠密特点.

簇结构多且类型丰富. 簇结构是由一组节点同时关联到一个或多个中心节点而形成. 网络资产图的簇结构多，而且存在单中心簇，多中心簇，单级簇，多级簇等多种类型. 以图 4(b)为例，该图存在多个明显的簇结构，左侧簇是一个单中心多级簇，它从一个证书节点作为中心向外辐射，辐射有两级，第一级是直接与该证书节点相连的大量域名节点，第二级是与这些域名节点相连的 IP 节点. 右上簇由多个域名节点同时连接到三个中心节点构成，三个中心节点分别是 Whois_Name, Whois_Phone 或 Whois_Email 类型节点，反映了同一个注册人注册了多个域名，该注册人可能是黑灰产团伙在真实世界中的关联人.

桥结构多且类型丰富. 桥结构是由关联不同簇结构的节点和连边形成. 网络资产图的簇结构多, 因此簇间关联也较为复杂, 存在两个簇中心直接关联或间接(多跳)关联, 两个簇间单桥关联或多桥关联等多种关联形式. 图 5(b)中有明显的多跳关联和多桥关联的桥结构.

为了更好地呈现上述拓扑特性, 我们对 Spring Embedder 算法进行了 3 个方面的优化. 新布局算法命名为 SE-BH-CAG (Spring Embedder-Barnes Hut for Cyber Asset Graph). 下面我们具体介绍优化方法.

(1) 基于近似计算节点斥力的布局速度优化. SE 算法通过迭代计算节点间的引力、斥力, 使所有节点达到受力平衡来完成布局. Barnes Hut 算法提出了近似计算节点间斥力的方法[30], 通过把邻近的节点分组并近似看为一个整体, 来构造四叉树进行节点间斥力的近似计算. 因此, 我们在 Spring Embedder 算法中应用了上述思想, 来减少节点的斥力计算复杂度, 提高布局速度.

(2) 基于四叉树信息复用技术的布局速度优化. Barnes Hut 近似计算节点斥力思想的应用, 会使 SE 算法在每次布局迭代时利用四叉树来近似计算本次迭代的所有节点间的斥力. Robert Gove 提出了四叉树复用技术[53], 即, 不需要在每次布局迭代都全部重新构建四叉树, 特别是在迭代布局后期, 两次迭代间节点位置变化不大. 因此, 我们把上述四叉树复用技术应用在 Spring Embedder 算法中, 来节省重新构造四叉树的时间, 以此来提高布局速度.

(3) 优化簇内边和桥接边的弹簧力. 网络资产图具有多簇多桥的特点, 当簇较大较多时, 图布局后难免出现簇拥挤现象. 我们提出增加簇内边拉力, 增加桥接边斥力, 使得簇结构内部的节点和边布局更加紧凑, 桥结构中的节点和边布局更为稀疏, 有效减少簇拥挤现象, 获得更为视觉舒适的布局效果. 簇结构和桥结构的识别方法见下个小节.

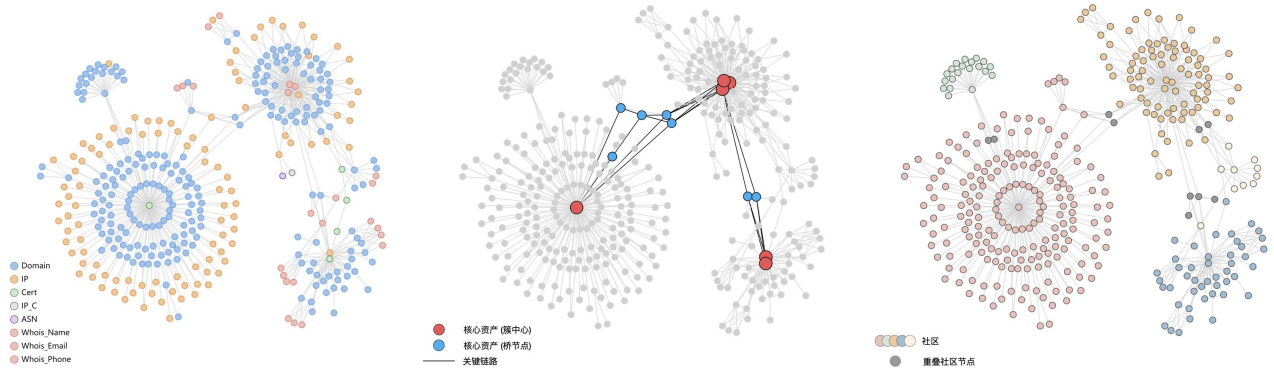

图 4. 一个具有 368 个节点和 643 条边的中型黑灰产网络资产图案例. (a)网络资产图布局效果图; (b)网络资产图中核心资产与关键链路识别结果; (c)网络资产图社区检测结果.

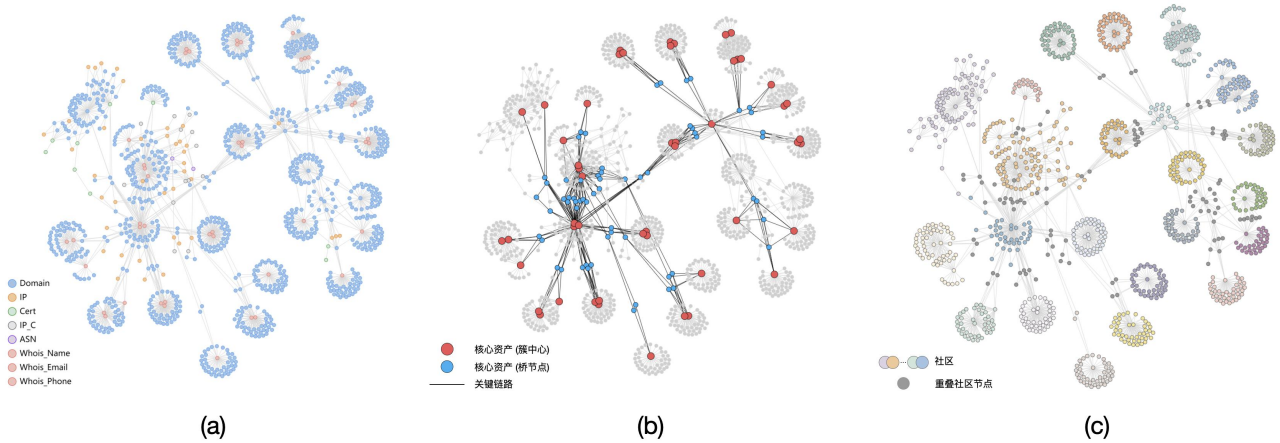

图 5. 一个具有 835 个节点和 1233 条边的大型黑灰产网络资产图案例. (a)网络资产图布局效果图; (b)网络资产图中核心资产与关键链路识别结果; (c)网络资产图社区检测结果.

4.3 网络资产图核心资产识别方法

根据 R2 和上个小节总结的拓扑特性, 网络资

产图的核心资产可以等价为簇结构的中心节点, 简称簇中心; 网络资产图的关键链路可以等价为桥结构上的节点和边, 简称为桥节点和桥接边. 因此, R2 需求就转化为了识别图中的簇中心和桥节点.

中心性是常用来衡量节点重要性的方法[54-61]. 本文仍然沿用了这一思路. 我们实验了度中心性、最短路径介数中心性、随机游走中心性、桥接中心性、通信中心性、准最短路径中心性等多种中心性度量指标[56-61]. 我们发现度中心性与簇中心在图拓扑层面的特征相匹配; 随机游走中心性与簇中心和桥节点在图拓扑层面的特征相匹配, 因为可以衡量节点对图连通性的影响力. 因此, 我们组合使用了度中心性和随机游走中心性来识别簇中心和桥节点. 我们还结合了两条业务规则, 过滤不重要的簇中心和桥节点: (1) 如果簇中心一半以上的邻边是“一般”或“弱”关联强度, 那么该簇中心不能作为核心资产. (2) 如果桥节点通过“一般”或“弱”关联强度的边与其他簇中心或桥节点连接, 那么该桥节点不能作为关键链路中的节点.

我们的方法用来识别核心资产和关键链路的效果如图 4(b)和图 5(b)所示. 以图 5(b)为例, 该网络资产图包含数量较多的簇结构, 簇之间由少量的桥节点与桥接边产生关联. 本文的方法找到了所有簇结构, 并将绝大部分簇结构的中心节点识别为核心资产. 但图右侧有两个簇结构的中心节点并未被识别为核心资产, 因为这两个簇的中心节点是域名类型节点, 与它直接关联的边是 r_cname 类型的边, 其关联强度一般. 该图中的核心资产(簇中心)主要是 Whois 类型、安全证书类型和域名类型的节点, 封堵这些核心资产可以使大量黑灰产网站面临安全风险甚至直接瘫痪. 该图中关键链路主要由域名和 IP 类型节点, r_request_jump、r_subdomain 和 r_dns_a 类型边组成. 封堵关键链路上的网络资产, 可以斩断资产簇之间的联系, 对黑灰产团伙的正常业务运转造成影响与干扰.

4.4 网络资产图核心资产影响范围检测

社区是由图中一组连接紧密的节点组成, 同时这些节点与社区外部的节点连接稀疏[62-63]. 社区检测是用来检测图中节点聚集现象的一种重要方法[63-64]. 根据黑灰产资产图的拓扑特性, 图中每个簇结构可以类比为一个社区. 因此, 我们可以将核心资产影响范围检测需求(R3)转换为: 利用已经获取的簇中心信息, 对网络资产图进行社区检测.

社区检测算法有很多, 比如: GN、Louvain、WalkTrap、LFM、Mod_m、Link 等[65-70]. 我们选择了 LFM 社区划分算法作为基本算法. LFM 的基本思想是从种子节点出发, 以贪心策略从局部向外扩展来检测社区. 该思想启发我们将已识别得到的簇中心作为种子节点. 基于 LFM 算法并结合核心资产识别方法, 我们设计了一个检测核心资产作用范围的社区检测算法, 命名为 LFM-CAG (LFM for Cyber Asset Graph). 该算法分为 3 个步骤. (1) 以簇中心为种子节点进行广度优先扩展, 直至遇到社区边界节点(桥节点)停止. 此时, 种子节点以及通过种子节点遍历到的所有节点, 被归纳为同一个社区. (2) 基于加权投票原则[71]为桥节点分配社区. 首先, 我们根据 4.3 节的核心资产识别结果, 将桥节点的一度邻居节点分为簇中心节点、桥节点和普通节点 3 类, 并分别为这三类节点设置了由高到低的不同权重, 即节点投票时持有的票数. 然后, 我们计算桥节点所有一度邻居节点的社区投票结果. 比如, 簇中心节点的权重是 5, 投票时, 该簇中心节点就为它所属的社区投 5 票; 最后, 我们把得票数最多的社区作为桥节点的社区. (3) 对于无法通过加权投票分配社区的桥节点, 即, 有多个社区获得了相同的票数, 且票数最多, 那么我们就将桥节点的社区作为票数最多的多个社区, 桥节点就被视为同时属于多个社区的重叠节点. 经过以上 3 个步骤, 就得到了以簇中心为起点的社区划分结果. 图 4(c)和图 5(c)展示了 LFM-CAG 算法结果. 图中不同颜色代表不同社区, 灰色节点表示多个社区间的重叠节点.

5 数据公开

基于黑灰产网络资产图谱构建技术, 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①. 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. 该数据集还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 721M. 每个节点的字段信息

① https://github.com/csuvis/CyberAssetGraphData

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[72-79]. 本章从下面几个方面来梳理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, 比如: 3 跳内能相互关联的核心资产可以看做归属同一黑灰产团伙. 对于关联强度较弱的边指向的目标节点, 其 1 跳邻居以外的节点可以不归入同一黑灰产团伙; 对于关联强度一般的边指向的目标节点,其 2 跳以外节点可以不归入同一黑灰产团伙. 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. 本文并未探讨子图挖掘技术.

(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[25-26].

(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. 本文提出了一个初步地自动化识别方案. 但识别精度、速度和泛化能力还有一定提升空间. 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[77-79].

(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.

7 结论

本文提出了一个黑灰产网络资产图谱自动构建方法, 以少量黑灰产网站域名为线索, 以点边双异质有向图为基本数据结构, 自动从多源异构数据中挖掘和整合多类网络资产信息及关联关系. 本文还提出了一组适用于网络资产图的可视化方法, 包括图布局方法、核心资产识别方法、核心资产影响范围检测算法, 帮助人们观察图拓扑结构, 理解资产间的复杂关联, 快速识别核心资产, 预估打击核心资产后的影响范围. 本文提出的可视化方法可以应用到社交网络、数据资产网络、蛋白质相互作用网络等, 这些网络的拓扑结构特性与网络资产图类似. 另外, 本文公开了一个大规模黑灰产网络资产图谱数据集, 总结了该数据集可支撑的黑灰产治理需求, 展望了可能遇到的技术挑战, 旨在促进面向黑灰产治理的大数据分析技术的创新和发展.

参考文献(References):

[1] Li M. Critical Human Risk Assessment Thinking and Governance Approaches - An Overview of the Global Risk Report 2006-2021[J]. Disaster Reduction in China, 2021, 5(1): 8-15(in Chinese)
(李明. 人类重大风险评估思路与治理方法-2006-2021 年《全球风险报告》综述[J]. 中国减灾, 2021, 5(1): 8-15)

[2] Security Insider. 2019 Cybercrime Prevention and Governance Study Report[EB/OL]. [2019-12-18]. https://www.secrss.com/articles/16003

[3] Zhao Y, Fan X, Zhou F, *et al*. A Survey on Network Security Data Visualization[J]. Journal of Computer-Aided Design & Computer Graphics, 2014, 26(5): 687-697(in Chinese)
(赵颖，樊晓平，周芳芳，等. 网络安全数据可视化综述[J]. 计算机辅助设计与图形学学报, 2014, 26(5): 687-697)

[4] Yu H. Modality and Regulation of the Underground Industry Chain of Cybercrime[J]. Journal of National Prosecutors College, 2021, 1(1): 41-51(in Chinese)
(喻海松. 网络犯罪黑灰产业链的样态与规制[J]. 国家检察官学院学报, 2021, 1(1): 41-54)

[5] Zheng Z. Dualistic Governance of Legal Governance and Technical Governance in the Network Society[J]. China Legal Science, 2018, 0(2): 108-130(in Chinese)
(郑智航. 网络社会法律治理与技术治理的二元共治[J]. 中国法学, 2018, 0(2): 108-130)

[6] Hatta M. Deep Web, Dark Web, Dark Net: A Taxonomy of “Hidden” Internet[J]. Annals of Business Administrative Science, 2020, 19(6): 277-92

[7] Yu H. Research on the Judicial Application of New Types of Cybercrime[J]. China Journal of Applied Jurisprudence, 2019, 6(1): 150-165(in Chinese)
(喻海松. 新型信息网络犯罪司法适用探微[J]. 中国应用法学, 2019, 6(1): 150-165)

[8] Chinese government website. The Ministry of Public Security's “Internet Purification 2020” campaign is in full operation[EB/OL].http://www.gov.cn/xinwen/2021-03/08/content_5591521.htm

[9] TikTok Security Center. TikTok Security Annual Report[EB/OL].https://aweme.snssdk.com/magic/eco/runtime/release/605ab2e1754aee032238b540?appType=douyin

[10] Zhao Y, Yang K, Chen S, *et al*. A Benchmark for Visual Analysis of Insider Threat Detection[J]. SCIENCE CHINA Information Sciences, 2022, 65(9): 199102:1-199102:3

[11] Zhao J, Zhang J. The underground industry needs a multi-pronged approach to governance[J]. China Information Security, 2017, 12(1): 73-74
(赵军，张建肖. 网络黑灰产治理须多管齐下[J]. 中国信息安全, 2017, 12(1): 73-74)

[12] Zhou F, Huang W, Zhao Y, *et al*. ENTVis: A Visual Analytic Tool for Entropy-Based Network Traffic Anomaly Detection[J]. IEEE Computer Graphics and Applications, 2015, 35(6): 42-50

[13] Shi R, Yang M, Zhao Y, *et al*. A Matrix-Based Visualization System for Network Traffic Forensics[J]. IEEE Systems Journal, 2016, 10(4): 1350-1360

[14] Ji S, Pan S, Cambria E, *et al*. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications[J/OL]. IEEE Transactions on Neural Networks and Learning Systems. 1-21[2021-4-26]. https://doi.org/10.1109/TNNLS.2021.3070843

[15] Hogan A, Blomqvist E, Cochez M, *et al*. Knowledge graphs[J]. ACM Computing Surveys, 2021, 54(4):71:1-71:37

[16] Lao N, Mitchell T, Cohen W W. Random Walk Inference and Learning in a Large Scale Knowledge Base[C] //Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. USA: Association for Computational Linguistics, 2011: 529-539

[17] Wang X, He X, Cao Y, *et al*. KGAT: Knowledge Graph Attention Network for Recommendation[C] //Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). New York: ACM, 2019: 950-958

[18] Banerjee P, Baral C. Self-Supervised Knowledge Triplet Learning for Zero-Shot Question Answering[C] //Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). USA: Association for Computational Linguistics, 2020: 151-162

[19] Kang C, Vadrevu S, Zhang R, *et al*. Ranking related entities for web search queries[C] //Proceedings of the 20th International Conference Companion on World Wide Web. New York: ACM, 2011: 67-68

[20] Yang J, Xu B, Hu J, *et al*. Accurate and Efficient Method for Constructing Domain Knowledge Graph[J]. Journal of Software, 2018, 29(10): 2931-2947(in Chinese)
(杨玉基，许斌，胡家威，等. 一种准确而高效的领域知识图谱构建方法[J]. 软件学报, 2018, 29(10): 2931-2947)

[21] Liu Q, Li Y, Duan H, *et al*. Knowledge Graph Construction Techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582(in Chinese)
(刘峤，李杨，段宏，等.知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600)

[22] Kpcke H, Rahm E. Frameworks for entity matching: A comparison[J]. Data & Knowledge Engineering, 2010, 69(2): 197-210

[23] Rijula K, Susmija R, Sourangshu B, *et al*. Task-Specific Representation Learning for Web-Scale Entity Disambiguation[C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans:AAAI press, 2018: 5812-5819

[24] Phan M C, Sun A, Tay Y, *et al*. NeuPL: Attention-Based Semantic Matching and Pair-Linking for Entity Disambiguation[C] //Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: ACM, 2017: 1667-1676

[25] Zhao Y, Shi J, Liu J, *et al*. Evaluating Effects of Background Stories on Graph Perception[J/OL]. IEEE Transactions on Visualization and Computer Graphics: 1-16[2021-08-26]. https://doi.org/10.1109/TVCG.2021.3107297

[26] Zhao Y, Jiang H, Chen Q, *et al*. Preserving Minority Structures in Graph Sampling[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1698-1708

[27] Eades P A. A Heuristic for Graph Drawing[J]. Congressus Numerantium, 1984, 42(42): 149-160

[28] Fruchterman T, Reingold E M. Graph drawing by force-directed placement[J]. Software Practice & Experience, 2010, 21(11): 1129-1164

[29] Kobourov S G. Spring Embedders and Force Directed Graph Drawing Algorithms[D/P]. 2012, Arizona: University of Arizona.

[30] Barnes J, Hut P. A hierarchical O(N log N) force-calculation algorithm[J]. Nature, 1986, 324(6096): 446-449

[31] Mathieu J, Tommaso V, Sebastien H, *et al*. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software[J]. Plos One, 2014, 9(6): e98679-e98679

[32] Frick A, Ludwig A, Mehldau H. A Fast Adaptive Layout Algorithm for Undirected Graphs[C] //Proceedings of DIMACS International Workshop, Princeton: Graph Drawing, 1994: 388-403

[33] Kamada T, Kawai S. An algorithm for drawing general undirected graphs[J]. Information Processing Letters, 1989, 31(1): 7-15

[34] Zheng J X, Pawar S, Goodman D F M. Graph Drawing by Stochastic Gradient Descent[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(9): 2738-2748

[35] Dwyer T, Koren Y, Marriott K. Constrained graph layout by stress majorization and gradient projection[J]. Discrete Mathematics, 2009, 309(7): 1895-1908

[36] Yuan X, Che L, Hu Y, *et al*. Intelligent graph layout using many users' input[J]. IEEE transactions on visualization and computer graphics, 2012, 18(12): 2699-2708.

[37] Wang Y, Wang Y, Sun Y, *et al*. Revisiting stress majorization as a unified framework for interactive constrained graph visualization[J]. IEEE transactions on visualization and computer graphics, 2017, 24(1): 489-499

[38] Harel D, Koren Y. Graph Drawing by High-Dimensional Embedding[C] //Proceedings of 10th International Symposium on Graph Drawing. Heidelberg: Springer, 2002: 207-219

[39] Brandes U, Pich C. Eigensolver methods for progressive

multi-dimensional scaling of large data[C] //Proceedings of International Conference on Graph Drawing. Heidelberg: Springer, 2006: 42-53
[40] Zhu M, Chen W, Hu Y, *et al*. DRGraph: An Efficient Graph Layout Algorithm for Large-scale Graphs by Dimensionality Reduction[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1666-1676
[41] Purchase H C. Metrics for Graph Drawing Aesthetics[J]. Journal of Visual Languages & Computing, 2002, 13(5): 501-516
[42] Han D, Pan J, Zhao X, *et al*. NetV. js: A web-based library for high-efficiency visualization of large-scale graphs and networks[J]. Visual Informatics, 2021, 5(1): 61-66
[43] Baroni, M, Bernardini, S, Ferraresi, A, *et al*. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora[J]. Lang Resources & Evaluation, 2009, 43(3): 209-226
[44] Zhou L, Lin L. Survey on the research of focused crawling technique[J]. Computer Applications, 2005, 0(9): 1965-1969(in Chinese)
(周立柱, 林玲. 聚焦爬虫技术研究综述[J]. 计算机应用, 2005, 0(9): 1965-1969)
[45] Suchomel V, J Pomikálek. Efficient Web Crawling for Large Text Corpora[C] //Proceedings of the 7th Web as Corpus Work-shop, Stroudsburg: Association for Computational Linguistics, 2012: 39-43
[46] Thompson K. Programming techniques: Regular expression search algorithm[J]. Communications of the ACM, 1968, 11(6): 419-422
[47] Chapman C, Stolee K T. Exploring regular expression usage and context in Python[C] //Proceedings of the 25th International Symposium on Software Testing and Analysis, New York: Association for Computing Machinery, 2016: 282-193
[48] Joulin A, Grave E, Bojanowski P, *et al*. Bag of Tricks for Efficient Text Classification[C] //Proceedings of the 15th Conference of the European Chapter. Valencia: Association for Computational Linguistics, 2017: 427-431
[49] Bigdata Services of Domain's Whois [EB/OL]. https://www.whoisextractor.in/
[50] DNS Toolkit for Python [EB/OL]. https://www.dnspython.org/
[51] IPIP.NET [EB/OL]. https://www.ipip.net/
[52] Fiedler M, Sekanina M. Problem 25, in theory of graphs and its applications[M]. Smolenice: Publishing House of the Czechoslovak Academy of Sciences, 1964: 1-13.
[53] Gove R . It Pays to Be Lazy: Reusing Force Approximations to Compute Better Graph Layouts Faster[C] //Proceedings of 11th Forum Media Technology. St. Pölten: CEUR, 2018: 43-51
[54] Lü L, Chen D, Ren X, *et al*. Vital nodes identification in complex networks[J]. Physics Reports, 2016, 650: 1(1)-63
[55] Liu J, Ren Zhuoming, Wang Binhong. A survey of node importance ranking in complex networks[J]. Chinese Journal of Physics, 2013, 62(17): 9-18(in Chinese)
(刘建国, 任卓明, 郭强, 等. 复杂网络中节点重要性排序的研究进展[J]. 物理学报, 2013, 62(17): 9-18)
[56] Grofman B, Owen G. A game theoretic approach to measuring degree of centrality in social networks[J]. Social Networks, 1982, 4(3): 213-224
[57] Freeman L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40(1): 35-41
[58] Brandes U, Fleischer D. Centrality measures based on current flow[C] //Proceedings of the 22nd annual conference on Theoretical Aspects of Computer Science (STACS'05). Berlin: Springer-Verlag, 2005: 533-544
[59] Hwang W, Cho Y, Zhang A, *et al*. Bridging centrality: Identifying bridging nodes in scale free networks[C] //Proceedings of the 12nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '06). New York: ACM, 2006: 20-23
[60] Estrada E, Higham D J, Hatano N. Communicability betweenness in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(5): 764-774
[61] Medeiros D, Campista M E M, Mitton N, *et al*. The power of quasi-shortest paths:ρ-geodesic betweenness centrality[J]. IEEE Transactions on Network Science and Engineering, 2017, 4(3): 187-200
[62] Cheng X, Shen H. Community structures for complex networks[J]. Journal of Systems Science and Complexity, 2011, 8(1): 57-70
(程学旗, 沈华伟. 复杂网络的社区结构[J]. 复杂系统与复杂性学, 2011, 8(01): 57-70)
[63] Coscia M, Giannotti F, Pedreschi D. A classification for community discovery methods in complex networks[J]. Statistical Analysis and Data Mining, 2011, 4(5): 512-546
[64] Santo F, Darko H. Community detection in networks: A user guide[J]. Physics Reports, 2016, 659(1): 1-44
[65] Newman M, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 26113:1-26113:15
[66] Blondel V D, Guillaume J L, Lambiotte R, *et al*. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10(1): 10008-10016
[67] Libraries M. Computing communities in large networks using random walks[C] //Proceedings of the 20th international conference on Computer and Information Sciences (ISCIS'05), Berlin: Springer-Verlag, 2005: 284-293
[68] Lancichinetti A, Fortunato S, J Kertész. Detecting the over-lapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 1-20.
[69] Luo F, Wang J Z, Promislow E. Exploring local community structures in large networks[C] //Proceedings of the 5th IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong: IEEE Computer Society, 2006: 233-239
[70] Ahn Y Y, Bagrow, J, Lehmann S. Link communities reveal multiscale complexity in networks[J]. Nature, 2010, 446(7307): 761-764
[71] Zhang J, Chen D, Dong Q, *et al*. Identifying a set of influential spreaders in complex networks[J]. Scientific Reports, 2016, 6(1): 27823-27830
[72] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444
[73] Shi Y, Zhao Y, Zhou F, *et al*. A Novel Radial Visualization of Intrusion Detection Alerts[J]. IEEE Computer Graphics and Applications, 2018, 38(6): 83-95
[74] Zhao Y, Zhao X, Chen S, *et al*. An Indoor Crowd Movement Trajectory Benchmark Dataset[J]. IEEE Transactions on Reliability, 2021, 70(4): 1368-1380
[75] Zhou Z, Shi C, Shen X, *et al*. Context-aware Sampling of Large Networks via Graph Representation Learning[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1709-1719
[76] Zhao Y, Luo F, Chen M, *et al*. Evaluating Multi-Dimensional Visualizations for Understanding Fuzzy Clusters[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 12-21
[77] Zhou F, Lin X, Liu C, *et al*. A survey of visualization for smart manufacturing[J]. Journal of Visualization, 2019, 22(2): 419-435
[78] Zhao Y, Luo X, Lin X, *et al*. Visual Analytics for Electromagnetic Situation Awareness in Radio Monitoring and Management[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 590-600
[79] Zhong Z, Zhao Y, Shi R, *et al*. A User-Centered Multi-Space Collaborative Visual Analysis for Cyber Security[J]. Chinese Journal of Electronics, 2018, 27(5): 910-919