网络空间资源测绘: 概念与技术

郭 莉, 曹亚男*, 苏马婧, 尚燕敏, 朱宇佳, 张 鹏, 周 川

[1] 中国科学院信息工程研究所 信息内容安全技术国家工程实验室 北京 中国 100093
[2] 中国科学院大学 网络空间安全学院 北京 中国 100049

摘要 网络空间资源测绘是对各类网络空间资源及其属性进行探测、融合分析和绘制。本文给出网络空间资源测绘概念和技术体系, 主要包括: 网络空间资源测绘的相关概念, 网络空间测绘对象的分类体系和属性框架; 提出网络空间资源测绘的技术体系模型, 并从协同探测、融合分析和全息绘制三个层次探讨网络空间资源测绘的关键技术; 以网络资产评估、服务测绘两个应用场景为例, 对网络空间资源测绘技术体系的应用进行详细阐述。旨在为网络空间资源测绘理论和技术的研究和发展奠定基础。

关键词 网络空间资源测绘; 测绘技术体系; 网络测量; 探测; 分析; 绘制
中图法分类号 TP393.0 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2018.07.01

Cyberspace Resources Surveying and Mapping: The Concepts and Technologies

GUO Li, CAO Yanan*, SU Majing, SHANG Yanmin, ZHU Yujia, ZHANG Peng, ZHOU Chuan

[1]National Engineering Laboratory for Information Security Technologies, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Cyberspace resource surveying and mapping (CRSM) is the process of detecting, analyzing and visualizing of all kinds of cyberspace resources and their attributes. This paper mainly discusses the concepts and technical system of CRSM. In this paper, we firstly introduce the definition of related concepts of CRSM, and propose a classification system and an attribute graph of cyberspace resource. Then, we give a technical architecture of CRSM, and discuss the research approaches of synergistic detection, fusion analysis and all-round visualization, which are key technologies in CRSM. Finally, we describe two application acenarios of CRSM in the network asset valuation and service surveying and mapping field in detail. This paper aims to accelerate the development process of the theory and technologies in CRSM.

Key words cyberspace resources surveying and mapping; technical architecture of surveying and mapping; network measurement; detecting; analyzing; visualizing

1 引言

网络空间已经成为人类生产生活的“第二类生存空间”, 关系到经济、文化、科研、教育和社会生活的方方面面, 成为国家发展的重要基础。随着网络技术日新月异的发展, 网络空间中的资源种类越加丰富, 不仅包括传统的设备、逻辑拓扑等软硬件基础设施, 也包括网络用户、应用服务等动态多变的虚拟资源。传统的网络测量技术已不足以全面刻画网络空间的特性, 发展网络空间资源测绘技术刻不容缓。

网络空间资源测绘是对网络空间中的各类资源及其属性进行探测、融合分析和绘制。通过绘制网络空间资源全息地图, 全面描述和展示网络空间信息, 能够为各类应用(如网络资产评估、设备漏洞发现等)提供数据和技术支撑。因此, 研究网络空间资源测绘技术, 全面掌握网络空间特性及其资源分布, 对于推动国民经济和保障国家安全都具有十分重要的理论意义和应用价值。

近年来, 国内外相继出现了网络空间资源测绘的相关工作。美国是最早进行网络空间资源测绘的国家, 目前已形成了较为完整的网络空间探测基础设施和体系, 其中代表性的工作包括: 美国国防局

通讯作者: 曹亚男, 博士, 副研究员, caoyanan@iie.ac.cn。
本课题得到国家重点研发计划(No. 2016YFB0801300)资助
收稿日期: 2018-03-30; 修改日期: 2018-05-30; 定稿日期: 2018-06-19

的 X 计划[1], 美国国土局的 SHINE 计划[2], 美国国安局的藏宝图计划[3]等。 其中 X 计划以绘制网络空间地图、提高网络空间作战能力为目标, 探测网络逻辑拓扑和设备数据, 开发网络战场地图引擎和端到端网络战场地图感知平台, 研发单兵作战支持平台和集团作战支持平台。藏宝图计划以全网态势感知、侦察和攻击推演为目标, 对网络空间进行多层次的信息探测和数据分析, 形成大规模情报能力, 探测内容包括: BGP、AS 和 IP 地址空间信息。SHINE 计划主要针对美国本土网络安全态势感知, 建立美国本土网络空间关键基础设施信息数据库, 监测关键行业网络可达性及安全态势, 发现弱点设备和系统。在国内, 知道创宇公司的 Zoomeye[4]可对全球一些地区的路由设备、工业联网设备、物联网设备以及摄像头等基础设施进行探测。FOEYE[5]在全面网络资产测绘基础上, 重新定义了安全事件处理和漏洞扫描形式, 形成集资产探测管理、安全事件验证、智能统计分析、安全态势感知、持续安全监控为一体的全方位安全体系。

以上工作大多基于传统的网络测量技术, 仅对网络空间的基础设施和逻辑拓扑进行探测和分析, 并没有覆盖网络空间的全部资源。相关研究表明, 由于目前在学术界和业界尚未形成对于网络空间资源测绘相关概念的统一认知, 缺乏对网络空间资源测绘技术体系的顶层设计, 网络空间资源测绘技术仍处于起步阶段。因此, 本文主要围绕网络空间资源测绘的概念和技术体系进行研究和探讨, 旨在为网络空间资源测绘理论和技术的研究和发展奠定基础。

本文主要包含三部分内容: 首先介绍网络空间资源测绘的相关概念, 提出网络空间测绘对象的分类体系和属性框架; 然后, 提出网络空间资源测绘的技术体系模型, 并从“协同探测”、“融合分析”和“全息绘制”三个层次探讨网络空间资源测绘的关键技术和研究思路; 最后, 从资产评估、服务测绘两个方面, 对网络空间资源测绘的应用前景进行详细的阐述。

2 网络空间资源测绘相关概念

2.1 网络空间概念

网络空间的概念最早出现在 1984 年美国科幻小说《神经漫游者》中, 作者威廉·吉布森[6]将其描述为通过计算机设备进入的数据库空间。此后, 国内外学者对网络空间概念进行了研究和阐述, 基于不同应用需求及研究领域, 网络空间被赋予了不同的内涵和外延。总的来说, 已有定义可概括为以下三类: (1)强调网络空间的物质属性。Sterling B[7]认为网络空间依存于硬件、软件设备等物质基础, 是互联网(Internet)与万维网(World Wide Web)的近似概念。(2)强调网络空间的社会属性。Adams P C[8]与 Hillis K[9]认为网络空间是人基于互联网技术与社交行为结合产生的“空间感”, 并将网络空间看作是关于人在交流和再现的空间中对社会的感知。Wellman B 等[10]认为社会性的交互活动比技术内容更能体现网络空间的本质内涵。(3)强调网络空间中的操作和活动。Mayer M 和 Martino L[11]认为: 网络空间是创造、储存、调整、交换、共享、提取、使用和消除信息与分散的物质资源的全球动态领域。

从上述定义中可以看出, 网络空间具有物质属性(软硬件等基础设施)和社会属性(人的交互行为及其操作)。早期的定义从不同角度强调了网络空间中的某种组成要素, 但是均未全面、系统地对网络空间的要素进行概括和描述。

随着网络技术的发展, 网络空间的内涵和外延也在不断发生变化。方滨兴院士[12]进一步将网络空间的组成要素分为 4 种类型: 载体、信息、主体和操作。其中, 网络空间载体是网络空间的软硬件设施, 是提供信息通信的系统层面的集合; 网络空间信息是在网络空间中流转的数据内容, 包括人类用户及机器用户能够理解、识别和处理的信号状态; 网络空间主体是互联网用户, 包括传统互联网中的人类用户以及未来物联网中的机器和设备用户; 网络空间的操作是对信息的创造、存储、改变、使用、传输、展示等活动。

综合以上要素, 网络空间可被定义为“构建在信息通信技术基础设施之上的人造空间, 用以支撑人们在该空间中开展各类信息通信技术相关的活动。其中, 信息通信技术基础设施包括互联网、各种通信系统与电信网、各种传播系统与广电网、各种计算机系统、各类关键工业设施中的嵌入式处理器和控制器。信息通信技术活动包括人们对信息的创造、保存、改变、传输、使用、展示等操作过程, 及其所带来的对政治、 经济、文化、社会、军事等方面的影响”。其中, “载体”和“信息”在技术层面反映出“赛博”的属性, 而“主题”和“操作”是在社会层面反映出“空间”的属性, 从而形成网络空间。

网络空间可以划分为物理层、逻辑层和认知层。物理层的内涵是实体的空间位置信息、实体间的连接关系存在于物理世界, 可直接观察, 易于感知; 逻辑层则是由逻辑拓扑、业务流动和用户操作构成的复杂网络, 无法直接观察, 必须借助于工具进行感

知; 认知层作为网络空间客观精神的外化, 承载着意识形态上层建筑, 无法直接观察, 只能根据其外在产物推测。

可见, 不同的网络空间层次包含了网络空间中的不同资源和资源的不同属性, 而同一类资源可能跨越不同的网络空间层次。例如, 硬件设备既包含位置属性(物理层)又包含设备间的逻辑拓扑关系(逻辑层); 网络用户既包含位置属性(物理层), 具有用户操作(逻辑层), 同时也具有意识形态(认知层)。

2.2 网络空间资源分类体系

广义的网络空间资源是网络空间中“载体”、“信息”、“主体”等各类要素的总和, 不仅覆盖通信基础设施、IP 网络、覆盖网络、应用支撑系统等互联网基础设施实体资源, 而且覆盖承载在实体设施之上的信息内容、用户等虚拟资源。传统的网络资源分类体系无法全面涵盖当前多种多样的网络空间资源, 因此我们提出了全新的网络空间资源分类体系和属性图谱。

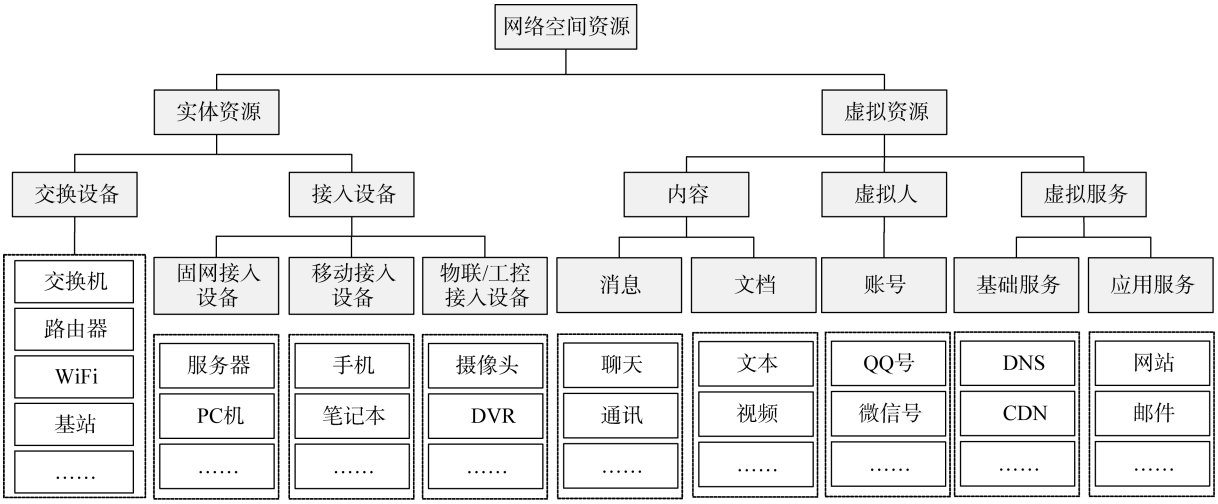

图 1 网络空间资源分类

Figure 1 Classification of Cyberspace Resources

网络空间绝大多数资源已有相应的分类命名, 如硬件资源(路由器、交换机、服务器等)和软件资源(视频、社交、电商网站等)。这些分类方法大多是基于行业角度的局部分类: 如计算机软件按照计算机体系结构的标准可分为系统软件和应用软件, 按照《软件产业统计制度修订说明》的标准则分为基础软件、中间件、应用软件等。总之, 这些分类方法并没有涵盖整个网络空间资源。相对于这些局部分类方法, 爱尔兰梅努斯大学的 Rob Kitchin 教授[13]在 2001 年将网络空间资源自下而上划分为网络实体、逻辑网络和赛博人, 初步对网络空间资源进行了全面划分。然而这种划分方法是从观察者的视觉划分网络空间, 并没有真正从网络空间视角去剖析和命名网络空间资源。

针对目前网络空间资源分类体系的缺失, 我们深入分析了各类网络空间资源的特点, 借鉴生物图谱、地学图谱、知识图谱等相关领域图谱的研究经验, 从理论与机理研究、分类方法探索等方面入手, 定义了全新的网络空间资源分类体系(如图 1 所示)。从物质形态和社会形态, 将网络空间资源为实体资源和虚拟资源。实体资源分为交换设备和接入设备: 其中交换设备包括交换机、路由器、Wifi、基站等; 接入设备包括移动接入设备(手机/笔记本等)和物联/工控接入设备(摄像头/DVR 等)。虚拟资源分为虚拟人、虚拟服务和虚拟内容: 其中虚拟人包括各类网络帐号; 虚拟服务包括基础服务(DNS/CDN 等)和应用服务(网站/邮件等); 虚拟内容则包括消息(聊天/通讯等)和文档(文本/视频等)。

2.3 网络空间资源属性图谱

网络空间资源属性是指网络空间资源具备的所有共同性质或独有性质的总和, 是资源必然的、基本的、不可分离的特性。以虚拟用户为例, 其共同属性涉及了用户名、性别、年龄等, 但不同的用户类别也可能具有其独有的属性。总之, 在网络空间资源测绘中, 资源属性塑造了资源的形态, 决定了资源的行为, 反映了资源间的关系。属性既直接从属于资源, 又能容纳和解释数据, 因此对资源属性进行描述有助于深入理解网络空间资源, 清晰认知网络空间。

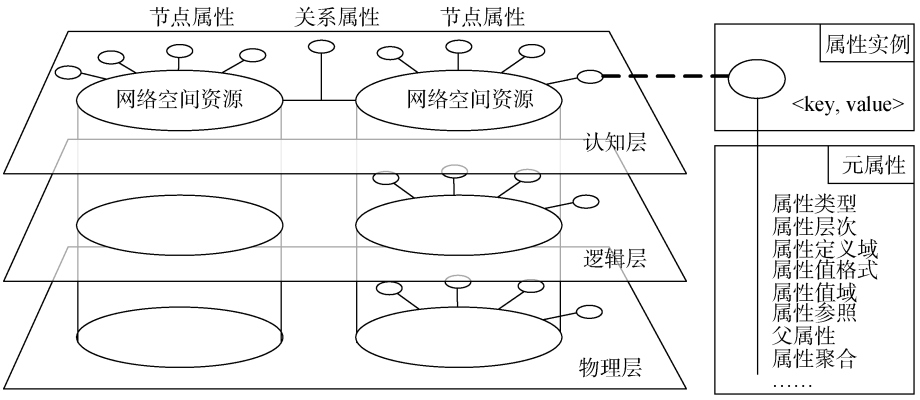

图 2　网络空间资源属性描述模型

Figure 2　Description Model of Cyberspace Resources

网络空间资源属性描述是指对网络空间资源的构成要素及属性进行统一组织和编码。国内外有关网络空间资源属性描述的典型研究有 Joseph W.Yoder 等人[14]提出的 AOM 自适应对象模型, 通过属性的复合化, 对对象内部进行属性细分。另外, Wille R[15]提出概念格来描述属性, 将概念格的每个节点表示为一个形式概念, 并将外延(概念所覆盖的实例)和内涵(覆盖实例的共同特征)嵌入其中。基于属性重要性和关系的知识发现方法也是一种属性描述方法, 该方法描述蕴含在事例中的属性。除此之外, 基于关联的多属性决策分析、知识表示、关系挖掘等都是属性描述的常用方法。

表 1　网络空间资源属性示例

Table 1　Samples of Cyberspace Resources Attributes

层次	典型属性
物理层	GPS 坐标、GPS 经度、GPS 纬度、城区、城市、省份、国家、大洲、尺寸、长度、宽度、高度、重量、功率、厂家、出厂日期、设备类别、型号、版本、端口数、占用端口数、连接介质、介质型号、介质速率
逻辑层	IP 地址 BIN、IP 地址 DEC、IP 地址点分、IP 地址、AS 域号、AS 名、运营商、网关、端口列表、端口信息、主机名、操作系统、OS 版本、资源类别、服务类别、域名、别名、连接类型、网段数量、连接速率
认知层	管理组织、管理者、责任者、服务类型、服务对象、服务主体、部署区域、调用方式、内容格式、内容语义、重要程度、依赖资源、活跃时间、连接类型、连接方式

然而, 已有的属性描述方法存在划分层次重叠问题。由于网络空间可以划分为物理层、逻辑层和认知层, 在对网络空间资源进行探测和感知时, 处于不同层次的资源会表现出不同的状态和行为, 即不同层次的资源属性各不相同。这种因层而异的特性产生了资源的全属性划分。因此, 我们将网络空间资源属性按物理层、逻辑层和认知层进行层级划分, 作为网络空间资源属性的标签; 属性可以依附于资源本体也可以依附于资源间的连接(如图 2 所示)。表 1 给出了网络空间资源在各个层次上的部分典型属性的示例。

2.4　网络空间资源测绘

“测绘”的概念最初源于地理测绘学, 是指“对自然地理要素或者地表人工设施的形状、大小、空间位置及其属性等进行测定、采集和绘制”。地理测绘的概念包括了“测量”和“制图”两项主要内容。

相对于地理测绘的目标是描述和标注地理位置, 最初的网络空间测绘的概念主要是“采用一些技术方法, 来探测全球互联网空间上的节点分布情况和网络关系索引, 构建全球互联网图谱的一种方法”。已有的网络空间测绘系统大多采用主动或被动探测的方法, 来绘制网络空间中的设备画像。例如, 知道创宇公司的 Zoomeye[3]可对全球一些地区的路由设备、工业联网设备、物联网设备等基础设施进行探测; Shodan[16]采用搜索引擎技术, 可以让用户使用各种过滤器查找连接到互联网的特定类型的设备; Caida[17]在全球范围内分布大量测量探针, 发现网络拓扑和节点设备; ANT 实验室开展的 AMITÉ、MR-Net 项目[18]通过探测互联网资源的使用现状、跟踪拓扑和流量的变化趋势, 并在网络地图上标注出相关信息以帮助研究者更好地改善网络的安全性和提高防御能力。其他类似的系统还包括 Foeye[5]、Gperf[19]等。

上述概念和系统仅强调了对网络基础设施和网络拓扑的探测和绘制, 测绘的对象不全面, 方法也局限于传统的网络测量技术。由于网络空间资源涵

盖了基础设施、数据资源、虚拟用户等网络空间要素, 网络空间资源测绘的涵义需要覆盖所有的网络空间资源类型。网络空间资源测绘是网络测量的发展和延伸, 与网络测量有着许多共通之处, 网络测量的数据获取和分析建模的技术均可用于网络空间资源测绘中, 但网络空间资源测绘与网络测量存在一些不同之处, 主要包括:

(1) 目标不同: 两者均是为了更好地了解、认识和优化网络, 网络测量的目的是对网络拓扑和网络性能进行度量以及对网络的规律进行建模; 但网络空间资源测绘的目的在于全面掌握网络空间资源的属性和状态, 绘制网络空间资源全息地图。

(2) 方法不同: 网络测量的过程主要是数据获取和分析建模, 存在一系列的测量方法、平台和技术; 网络空间资源测绘除了“探测”、“分析”外, 还有“绘制”的过程。“分析”除了包括对单个对象进行建模, 还包括对多类对象的融合分析、从社会空间向物理空间的映射等。测绘方法主要采用“协同探测”、“融合分析”、“迭代演进”、“全息绘制”等技术思路。

(3) 对象和范围不同: 网络测量的对象主要是网络拓扑和网络性能等, 测量范围根据需要确定, 可以是局部网络, 可以是某一类型的全网络, 也可以是跨多个网络; 而网络空间资源测绘的测量内容包括实体资源和虚拟资源, 包括网络空间的全部要素, 测量的范围为整个网络空间。

(4) 结果和应用不同: 网络测量的结果是一系列的模型、规律、特征等, 各类测量之间具有较强的独立性, 每种测量结果通常可以直接应用于对网络的模拟和性能优化; 网络空间资源测绘的结果是一个网络空间资源全息地图, 是多类型测量结果的融合, 因而可以更广泛地支持资产风险评估、网络性能评价、病毒主动预警、攻击轨迹刻画等。

通过上述分析, 我们将网络空间资源测绘定义为“对网络空间中的各类虚实资源及其属性进行探测、分析和绘制的全过程”。具体内容包括: 通过网络探测、采集或挖掘等技术, 获取网络交换设备、接入设备等实体资源以及信息内容、用户和服务等虚拟资源及其网络属性; 通过设计有效的定位算法和关联分析方法, 将实体资源映射到地理空间, 将虚拟资源映射到社会空间, 并将探测结果和映射结果进行可视化展现; 将网络空间、地理空间和社会空间进行相互映射, 将虚拟、动态的网络空间资源绘制成一份动态、实时、可靠的网络空间地图。

3 网络空间资源测绘技术体系

网络空间资源测绘是对网络空间进行探测、分析和绘制的过程。与传统的网络测量方法相比, 网络空间资源测绘技术体系, 除了包含传统的数据获取和分析建模方法外, 还具备“协同探测”、“迭代演进”、“融合分析”和“全息绘制”技术特点。下面, 我们对网络空间资源测绘技术体系和其中的关键技术点进行阐述。

3.1 网络空间资源测绘技术体系

网络空间资源测绘体系是一个“探测(Detecting)、分析(Analyzing)、绘制(Visualizing)、应用(Applying)”的循环过程(DAVA Loop), 对各种网络空间资源进行协同探测, 获取探测数据, 对这些数据进行融合分析和多域映射, 形成网络空间资源知识库; 在此基础上, 通过多域叠加和综合绘制来构建网络空间资源全息地图; 最后, 根据不同的场景目标按需应用这一全息地图, 通过迭代演进使得测绘能力不断提升。如图 3 所示, 其中:

探测(Detecting): 精确全面地获取网络空间实体资源和虚拟资源测量数据的过程, 测量的内容包括资源及其属性和数字化活动, 对实体资源测量又包含实体测量、IP 测量、拓扑测量等, 对虚拟资源的测量又包含用户测量、服务测量等。网络空间资源测量方法应该满足以下四方面的需求: “稳定”、“准确”、“全面”、“可重复”。稳定性要求网络空间资源轻微变化不会导致测量方法失效, 准确性要求测量结果能够准确反映网络空间资源的真实情况, 全面性要求测量方法和结果能够尽可能全面地获取和覆盖被测资源各种参数数据, 可重复则是在相同测量条件下, 多次测量结果应是一致的。

分析(Analyzing): 从测量结果中提取资源及其属性, 并进行分析建模和关联映射的过程, 实现对网络空间资源高精度全景画像和追踪定位。分析的内容包括对实体资源和虚拟资源的属性提取、关联和画像, 以及向物理空间和社会空间的关联映射。网络空间资源分析需要解决复杂属性解析、缺失属性填充、多表征归一、跨域映射等一系列关键问题, 分析的结果是形成一系列网络空间资源知识库。

绘制(Visualizing): 基于测量结果和分析结果, 将多维的网络空间资源及其关联关系投影到一个低维的可视化空间, 构建网络空间资源的分层次、可变粒度的网络地图, 实现对多变量时变型网络资源的可视化过程。绘制需要对数量巨大、多源异构的信息数据, 进行时间、空间、类型等一体化组织, 基于

统一的时空基准数据模型和资源标识, 对数据进行有效关联组织和可视化表达, 对网络空间资源的分布、状态、发展趋势等进行全方位动态展示。

应用(Applying): 根据网络空间资源全息地图, 面向不同的综合业务, 应用不同的层次数据。如地理地图一样, 网络空间资源全息地图既可以独立使用, 也可以与其他资源、状态、外部信息、知识图谱等叠加。网络空间资源测绘的结果可以应用于改进网络部署、提升网络性能、评估网络安全态势、主动预警和防御网络攻击等场景

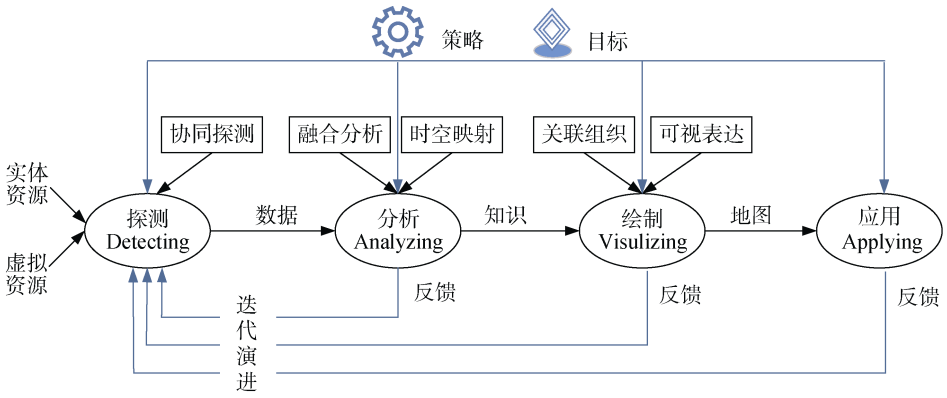

图 3 网络空间资源测绘技术体系

Figure 3 Technical Architecture of Cyberspace Resources Surveying and Mapping (DAVA Loop)

3.2 协同探测

在传统的网络测量中, 探测方法有多种分类标准。按探测方式, 可以分为主动探测和被动探测; 按探测点的多少, 可以分为单点探测和多点探测; 按探测内容, 可分为拓扑探测、性能探测和流量探测等; 按探测点所在层次, 可分为网络层探测和应用层探测; 按探测者是否主动配合, 分为协作式和非协作式; 按探测所采用的协议, 分为基于 BGP 协议、基于 TCP/IP 协议、基于 SNMP 协议以及基于应用层协议(如 HTTP)等。每种探测方法都有相应的优点和缺点, 许多相关工作[20-26]已经对此进行了详细综述。

与传统的网络测量方法不同, 为实现对大规模复杂多样的网络空间资源进行探测, 对网络空间资源的探测需要采用“协同探测”的方法, 即主被动协同、多点协同、协作协同。

1) 主被动协同

主动探测根据探测需要构造特定的探测包并向网络注入探测流量, 通过接收探测包流经网络时各网络参与者或探测目标的响应来获得探测结果。被动探测是在网络中选择一组探测节点(关键网络设置或者主机)并在其上部署探针进行监听, 通过收集经过该节点的流量或发往该节点的请求来获得探测结果。主被动协同探测是通过主动方式进行资源探测并扩散被动探测节点, 通过被动方式获取信息, 再针对这些信息通过主动方式进一步在网络空间探测, 进而获得更多更全面的结果。

主被动协同探测可以利用主动探测和被动探测的优点, 弥补不足, 从而获得更为准确全面的探测结果。主动探测的优点在于能够有针对性地对目标进行探测, 即使这些目标不主动产生流量或者流量不经过探测节点, 探测程序部署位置也较为灵活。但是, 主动探测也具有一定的局限性, 当目标对探测行为进行检测并主动屏蔽时, 则无法获得准确的探测结果, 主动探测也无法准确获得一些网络节点之间的通联关系, 此外, 主动探测会产生新的探测流量, 因而会增加网络负载, 并且主动探测不可避免地会对网络当前状态产生干扰, 使探测结果产生一定的偏差。被动探测由于不需要发送探测数据包, 或者仅在收到用户请求时才发送必要的响应数据, 对网络影响较少, 探测结果更为准确, 也可以获得一些主动探测无法获得的数据。但是被动探测也有一定的局限性, 如探测实现复杂度较高, 尤其是骨干链路, 高速网络流量的获取和分析都是探测需要解决的难点问题, 探测范围和结果准确度依赖于探测程序的部署位置、处理能力等。在设计和实施主被动协同探测时应尽可能减少对网络的影响, 提高探测的结果的准确性, 同时还应注意通过协同方式尽量避免流量捕获和解析等带来隐私和安全问题。

2) 多点协同

网络空间资源的分布广泛、种类复杂多样, 在研究初期、资源有限或对某一范围进行探测时可采用

单点探测, 但单个探测点的能力有限, 大规模大范围的网络探测通常需要分布式部署多个探测节点, 综合多点的信息以获得更全面准确的探测结果。多点协同探测在探测基础设施中表现的是探测网络组织结构是多地多点部署, 在探针方面表现的是多探针分布式部署、探测任务并发执行、多种探针联动等。多点协同探测需要解决探测点部署位置、探测任务调度、探测网络通信等问题。

为了对这些广泛分布的异构探测节点进行统一组织, 还需构建一套网络空间资源协同探测的基础设施, 解决探测网络组织模型、探测任务调度策略、平台运维管理等问题, 支持对互联网、物联网、工控网等网络空间中各类实体资源和虚拟资源探测。探测平台既可以兼容现有的探测平台, 如 Archipelago[23]、DIMES[27]、iPlane[28]、RIPE Atlas[29]、SamKnows[30]、BISmark[31-33]等, 也可以独立建设。探测平台的网络接入位置包括网络边缘和网络核心, 网络接入方式包括固网接入和移动接入, 平台的组成既有通用服务器也有专用的硬件设备, 平台主要以分布式部署为主, 设备来源既可以有众筹的方式也可以独立维护运行。

3) 协作协同

协作协同包括两方面: 探测参与者的协作和探测实施者的协作。探测参与者的协作是指探测目标或管理者对探测知情或主动配合, 是一种协作式探测, 相较于非协作探测, 能够更容易获得准确的探测结果。被动探测在部署探测点时通常需要相应的管理者协助, 但一些探测目标和网络管理者出于隐私和安全的角度均更倾向于不配合网络探测, 虽然采用主动探测的方式能够不依赖于探测参与者主动配合, 但网络空间资源测绘试图建立一种网络开发者、管理者、用户的合作模式, 既保护用户隐私不被侵犯, 又能够增加对网络空间资源的认知和理解, 从而更好地创建、管理、使用网络空间资源。探测实施者的协作, 是指在构建探测平台或实施探测时, 各组织机构合作, 平台共享、资源共享、能力共享、结果共享。

3.3 融合分析

网络空间资源种类和属性丰富, 表现形式多样, 资源数量巨大、关系复杂, 这对数据分析和处理的及时性、准确性和可靠性提出了新的要求。目前, 单源少量数据的分析已无法满足需求, 需要采用更有效的分析方法, 实现对大规模流式数据、多源数据的深度融合分析。总的来说, 网络空间资源分析方法呈现以下三方面的发展趋势:

1) 从小数据离线分析到大数据在线分析

网络空间的数据规模巨大。以虚拟资源为例, 网络空间中存在亿级服务资源[34], 数十亿级虚拟用户[35], ZB 级内容和数据。同时, 由于网络空间资源数据具有很强的实时性和动态性, 这就要求分析技术能够对大规模流式数据进行快速、准确的处理。在实时流数据处理中, 由于数据持续不断更新, 无法一次性存储在数据库中, 因此需要使用概要提取[36, 37]技术, 在内存中保留处理过的概要数据代表以前的历史数据, 同时采用增量方式对概要数据进行更新。以良好的概要数据结构设计为基础, 各类基础流数据分析技术包括流数据聚类[38]、分类[39]、时序预测[40], 事件流预测[41]等。在数据处理平台方面, 可采用 MapReduce[42]、Hadoop[43]和 Spark[44]等分布式处理模式, 或使用 Spark Streaming、Storm、Flink 为代表的流处理大数据系统。

2) 从浅层分析方法到深度分析模型

浅层模型在有限的样本和计算单元的条件下, 对于复杂问题的泛化能力有限, 不能很好地解决多类网络空间资源的分析和计算问题。深度学习相对于浅层学习而言, 能够通过学习一种深层的非线性网络结构, 实现复杂函数逼近, 为实现“端到端”的网络空间资源分析系统提供了契机。

随着深度学习技术的日益成熟, 深度学习模型也初步应用于网络空间资源测绘领域: 在实体设备数据分析方面, 相关研究通过对 GSM 信号进行分析, 确定设备的指纹特征, 同时利用 BP 神经网络对信号的指纹特征进行分类,以识别发射设备的类型; 在虚拟用户画像建模处理方面, 采用卷积神经网络方法实现对虚拟用户缺失属性推断和用户兴趣挖掘[45, 46], 采用用户嵌入表示学习方法实现跨社交网络账号关联[47, 48]; 在虚拟用户行为分析方面, 采用长短时记忆网络实现用户的异常行为检测和隐私对抗[49-51]; 在服务分析方面, 采用自编码器和多种深度神经网络对网站指纹进行构建和识别[52]; 在内容分析方面, 采用循环神经网络实现恶意软件检测[53, 54]; 采用深度神经网络与生成对抗训练、强化学习等融合框架实现文本、视频等多媒体内容分析[55, 56]。在以上应用领域中, 深度分析模型所实现的效果均超过了目前最好的浅层模型。

3) 从单一对象和数据源分析到融合分析

针对单一数据源的简单分析无法囊括不断扩大的网络空间资源对象和复杂的数据类型, 需要采用数据融合技术, 对多源信息进行综合处理, 实现对网络空间资源及其属性的精准刻画。数据融合主要

是针对各类可用数据形式化表达的信息融合, 数据关联质量与效果的优劣关系到系统对融合结果的处理。因此, 为了实现资源精准分析, 如何利用各类网络空间资源独有的数据特征和资源之间的相互关系, 进行数据的交叉验证和关联映射, 将成为实现资源精准分析的关键。

1) 基于多类资源融合的交叉验证

对网络空间不同类型的资源属性进行交叉分析验证, 可实现资源特定维度属性的填充, 并发现数据之间的不一致性。在实体资源交叉验证方面, 可利用多种信息源(如 Web 服务器地标信息、互联网机构主页、黄页等), 挖掘实体设备地标, 综合各类信息源对地标的可信度进行评估。在实体和虚拟资源交叉验证方面, 一方面, 利用实体资源定位技术, 获取实体设备 IP 的对应地理位置; 另一方面, 从流量或文本等数据源中挖掘操作该实体的虚拟用户的位置, 可实现实体资源和虚拟用户位置的一致性验证。在虚拟资源交叉验证方面, 可利用异常用户对虚拟服务、虚拟内容的访问行为分析, 实现对异常服务、异常内容的发现; 与此同时, 利用异常服务和异常内容的用户访问日志, 再发现新的异常用户。这是一个迭代分析、协同训练的过程。

2) 基于多源数据融合的关联映射

由于网络空间资源的隐匿性特点, 存在“人物多身份、服务多镜像、内容多副本”的特点。通过对不同数据源下网络空间资源的多维属性、关系和行为特征进行融合分析, 能够实现多源数据中同一资源对象的关联映射。在人物身份识别方面, 可利用虚拟用户关系网络结构、用户属性和用户行为特征, 学习用户的嵌入表示和网络的嵌入表示, 实现跨网络多账号用户之间的关联映射, 将虚拟用户映射到社会空间, 识别用户的真实身份[57]。在同源服务识别方面, 首先利用基于视觉信息的网页分割算法定义网页的语义结构, 每一个节点代表一个语义块, 每个语义块都有一个 DOC 值描述其内部内容的关联性; 然后, 对网站进行分块, 利用向量相似度度量(EMD)进行网页相似度检测, 将具有相似主页的网站看作是同源网站[58, 59]。在相似内容检测方面, 通过计算词级、短语级、句子级的嵌入表示, 在嵌入向量空间计算文本之间的相似度, 从语义层面解决各个层次的文本相似性检测问题[60, 61]; 将两张图片看成一张双通道的图像, 采用 Siamese 网络对两张图片的特征向量构造相似度损失函数, 通过对网络进行训练, 能够判别任意两张图像是否匹配[62, 63]。

3.4 全息绘制

网络空间资源绘制技术是在网络空间资源探测和分析的基础上, 将多维的网络空间资源及其关联关系投影到一个低维可视化空间, 旨在构建网络空间资源的层次化、可变粒度的网络地图, 实现对多变量时变型网络资源的可视化呈现[64]。

在绘制理论方面, 美国科学家 Marc A.Smith 和爱尔兰 Rob Kitchin 教授先后提出了网络空间的社区、地图和映射的理论[65], 介绍了如何利用地理学、制图学、计算机通信、信息可视化等领域的研究成果建立网络空间可视化的方法。武汉大学艾廷华教授在 2016 年提出了网络空间资源表达的符号可视化理论以及应用数学法则进行测量的方法[66]。

在绘制技术方面, 国内外研究团队虽然提出了一些准则, 如网络空间地理学图像的电信网络分析方法、网络空间景观制图的若干法则[67]、拓扑可视化[68]等。然而, 已有研究主要基于地理空间对实体设备和拓扑关系的可视化, 如何对高维、动态的虚拟资源进行绘制, 如何将网络空间中的多类资源投影到地理空间进行绘制缺乏系统和成熟的技术思路。因此, 需要综合可视化、图形学、数据挖掘理论与方法, 研究新的网络空间资源测绘理论模型和可视化方法。在绘制全息地图的需求下, “叠加绘制”和“时空建模”将成为网络空间资源绘制的研究重点。

1) 网络空间多类资源叠加绘制

针对网络空间资源数据的时间特性和空间特性, 结合数据间的关联性, 建立数据可视化关联模型, 动态维护关联关系, 并根据网络空间资源所涉及多种属性进行资源聚类, 计算节点的分布位置, 实现在虚拟空间和现实空间上的数据关联、数据流向、数据分布展示; 针对网络资源探测的大规模多种类别数据信息, 结合空间、地点、时间等多个维度, 打破地图上数据信息展示单一的格局, 对不同数据进行多方面、多层次的处理, 并将不同维度数据进行合理的叠加处理, 将组织关系、物理位置、网络行为、通联日志等数据以网络拓扑视图、地理分布视图以及时间维度等多种方式展现于全息地图之上, 实现数据处理全过程的灵活操控。

2) 地理空间与网络空间时空建模

研究虚拟网络空间与实体地理空间中对象时空变化机理和模型、统一时空基准建立与维持、时空坐标系转换、时空动态语义和时空演化模型等, 构建虚拟网络空间和实体地理空间统一描述的时空坐标体系, 为网络空间资源的测绘建立时空基准框架,

实现虚拟网络空间和实体地理空间资源中多源异构、海量高维、动态变化的时空数据的融合与统一管理、描述与应用。针对网络空间资源属性的动态时变特性, 研究网络空间资源的多维度、跨尺度、实时动态绘制技术, 实现虚拟网络空间与真实地理空间场景的一体化真三维绘制与动态切换、网络空间地图要素符号自动综合、网络空间三维景观动态交互等。

4 测绘技术应用示例

在第三节中, 我们给出了网络空间资源测绘的DAVA Loop 通用体系结构。下面我们以区域资产评估和服务测绘为例, 阐述 DAVA Loop 在具体应用中的工作流程, 并说明其有效性。

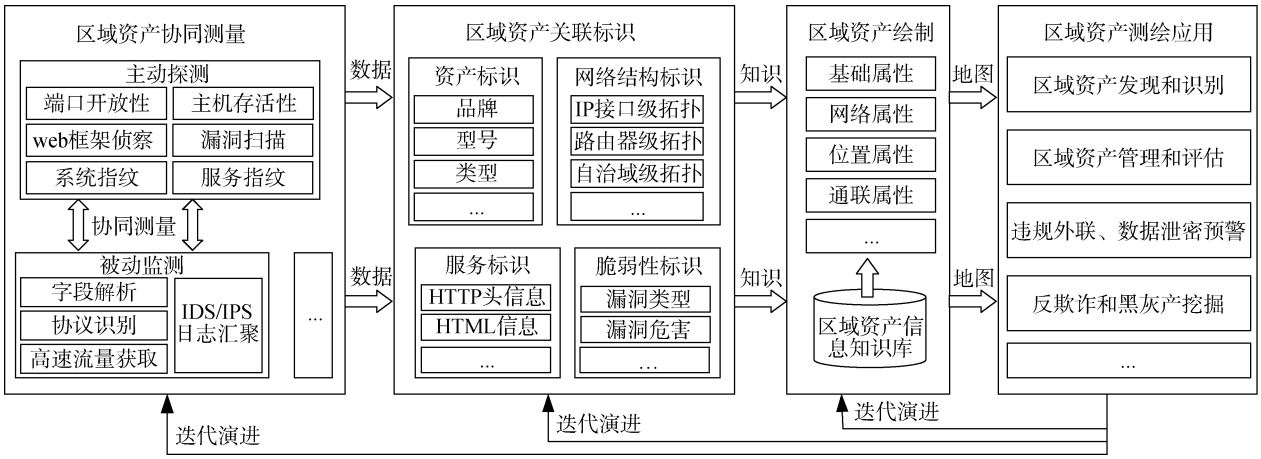

图 4 基于测绘体系的区域资产评估框架

Figure 4 The Technical Architecture of Network Assets Evaluation

4.1 区域资产发现、识别和风控

区域资产是对组织具有价值的信息或资源, 是安全策略保护的对象。威胁、脆弱性以及风险都是针对资产而客观存在的。准确掌握某一资产的攻击面是防御和攻击的关键。通过网络空间测绘技术, 对某个区域的网络资产进行识别和控制, 可以更好地保护个人和组织的数据, 防范已经存在和可能存在的风险。

为了获得精准的区域资产信息, 基于 DAVA Loop 的资产测绘对区域内的网络资产进行协同测量, 通过对数据进行融合分析和多域映射, 形成区域资产信息知识库; 在此基础上, 通过多域叠加和综合绘制来构建网络空间资产全息地图; 最后, 根据区域资产发现和识别场景、漏洞引起的业务风险监测场景、预警违规外联、数据泄密隐患场景、资产管理评估场景和网络空间反欺诈场景等目标, 应用这一面向资产的全息地图, 通过迭代演进使得资产测绘能力不断提升。

1) 区域资产协同探测

区域资产协同测量主要分为主动探测和被动监测。其中, 主动探测主要包括设备端口开放性检测、主机存活性检测、WEB 框架侦察、系统指纹发现、服务指纹发现和漏洞扫描等; 被动解析子模块主要有高速流量获取、协议识别和字段解析、IDS/IPS 日志汇聚等功能。

主动探测在探测前需要测试网络范围, 除了穷举网络地址, 也可使用 Dmitry、Scapy 等专用工具查询目标网络中的 IP 地址和域名信息, 测试网络范围; 此外, 反扫描设备的存在以及地址空间过大等问题, 使得主动探测在确定扫描范围存在盲区, 且无法获得实时的资产状态。因此本文的资产测绘使用了主被动协同测量的框架。一方面主动探测采用传统的网络范围测试手段, 另一方面, 由被动流量中过滤得到的 IP、端口和域名等信息, 作为主动探测的目标输入。被动流量过滤得到的 IP、端口和域名等信息包括本区域内部的资产和与本区域相关联的资产信息组成。

2) 区域资产关联标识

通过主动探测得到的资产标识与被动监测得到的资产标识具有明显的差异。一方面, 主动探测具有资产标识细节信息更加明确, 更适合局域网资产(如打印机、DVR、NVR 等)的探测发现等特点; 另一方面, 类似防火墙的网络隔离使得主动探测需要设计针对复杂网络的绕过技术, 实现代价陡增。与之对应的被动监测得到的资产标识具有实时性高、监测代价小的优点, 但是被动流量种类和属性丰富, 协议多样, 数量

巨大、关系复杂, 同时具有加密和云化特点。

在设计区域资产关联时需要充分考虑两种测量方式的优劣, 针对数据分析和处理的及时性、准确性和可靠性, 建立精准资产识别框架, 采用数据融合技术, 从网络结构、设备指纹、服务指纹、浏览器指纹等细节处对多源信息进行综合处理, 实现对区域资产及其属性的精准刻画。

在主被动测量数据下, 利用各类不同类型资产的行为特点、品牌特点、功能特点、网络结构特点等, 发现独有的数据特征和资源之间的相互关系, 进行数据的交叉验证和关联映射, 形成对多维度多属性的通用高精度识别框架; 针对不同类型、品牌训练不同参数, 通过优化的机器学习算法进行设备的分层识别; 利用多维度关联验证方法, 对设备型号等综合属性进行验证过滤, 剔除不精确的判断, 提升识别精度。

3) 区域资产绘制

以可视化的方式呈现区域资产信息, 包括基础属性、网络属性、位置属性、通联属性等。例如基础属性包括设备名称、设备型号、设备类型、生产厂家、基本参数(CPU、内存等)、出厂日期、外型参数等; 网络属性包括设备 IP、使用协议、开放端口、运行系统信息(系统类别、版本号)、发现时间、漏洞信息、banner 信息、页面特征等; 位置属性包括时区、国家、区域代码、所在地、经纬度、地址等; 通联属性包括相关组件号、关系类型、映射信息等。区域资产地图可与其他资源、状态、外部信息、知识图谱等叠加, 按需展示关注点。

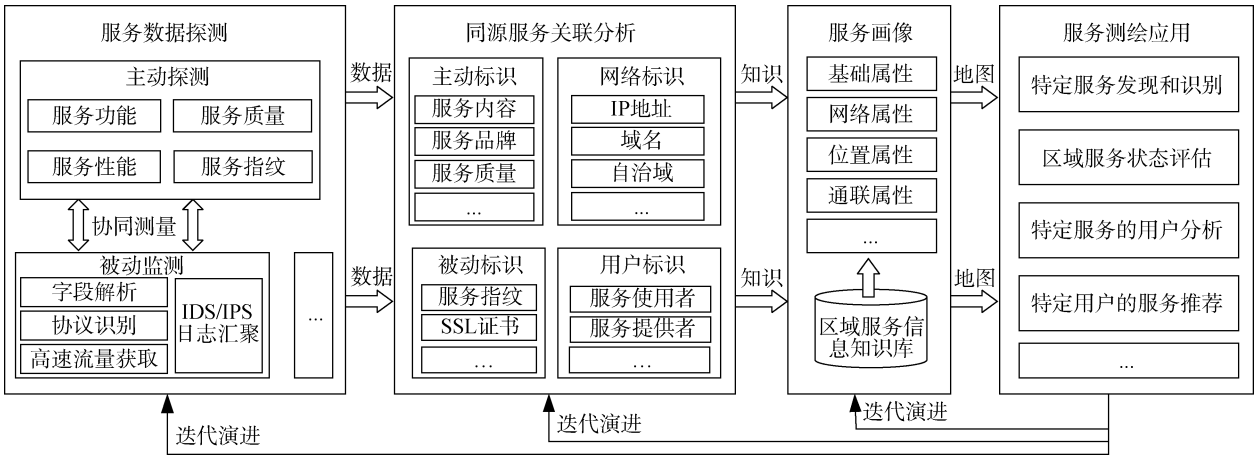

图 5 基于测绘体系的虚拟服务评估框架

Figure 5 The Technical Architecture of Virtualization Service Evaluation

4) 区域资产测绘应用

根据资产测绘地图, 面向不同的业务场景, 应用不同的层次数据, 提供决策支持, 并通过迭代演进使得资产测绘能力不断提升。典型应用场景如下:

(1) 区域资产发现和识别场景: 通过主动扫描、流量监控等多种资产采集方式, 自动获取资产和开启的服务。对关注区域内的资产进行分析与统计, 便于安全监管部门, 对区域内重要信息系统资产、物联网设备资产等了解和掌控。

(2) 漏洞引起的业务风险监测场景: 根据系统预置常见、热门漏洞, 根据资产组件特征判断漏洞影响资产数量; 基于对资产和漏洞的统计分析, 对资产的数量、分布、组件应用以及漏洞、威胁资产进行深入的态势感知及告警, 实现资产、漏洞的安全监测。

(3) 预警违规外联、数据泄密隐患场景: 自动发现管理网内同时连接内网和互联网的设备, 上报设备内网 IP 地址、互联网出口 IP 地址、外联时间及访问的网址。自动发现内网环境下连通互联网的风险设备点, 上报设备信息, 及时预警。

(4) 资产管理评估场景: 监控资产设备使用规范性; 发现企业内部的违规性行为; 精准推送资产漏洞并结合特制漏洞专扫、弱口令专扫等, 实现区域内网络资产合规性、违规性、存活性、脆弱性的综合检测与评估。

(5) 网络空间反欺诈场景: 绘制网络空间上设备的网络节点和网络连接关系图, 给各设备的画像; 结合风控理论, 在反欺诈和黑灰产发现的实践中, 为电商、支付、在线信贷等行业提供精准的身份识别。

4.2 服务测绘

网络空间服务是指网络空间软件设施中的各种泛在应用, 其中最典型的一种应用就是网站。网络空

间服务包含内在的属性和外延的关系, 其中属性又可以分为功能属性和非功能属性, 功能属性包括服务 IP 地址, 服务内容, 服务提供者, 服务协议, 服务状态, 服务使用者等。非功能属性包括服务性能(吞吐量和延迟), 服务可靠性, 服务价格等。关系主要包括服务之间的关系和服务与使用者之间的关系。

网络空间服务具有隐匿化, 小众化和加密化的特点, 网络空间服务测绘的目标就是利用主被动协同探测和智能分析手段, 发现动态、时变、隐匿的服务属性和关系, 通过“地图”的方式进行可视化展示, 以支撑网络空间安全的各种应用。基于 DAVA Loop 循环的服务测绘体系结构包括以下方面:

1) 服务数据探测

主被动协同测量主要分为主动探测和被动监测。其中, 主动探测基于服务指纹构造特定的探测请求, 通过接收探测包流经网络时各网络参与者或探测目标的响应来获得测量结果; 通过主动探测, 可探测得到服务功能、服务性能、服务质量、服务指纹等。被动监测基于服务指纹在网络流量中进行匹配, 然后对匹配命中结果进行过滤和统计; 通过被动监测, 可获取高速流量、进行协议识别和字段解析、IDS/IPS 日志汇聚等。

2) 同源服务关联分析

在主被动测量数据下, 利用同源服务的内容、品牌、结构相似的特点, 发现独有的数据特征和服务之间的相互关系, 进行数据的交叉验证和关联映射, 形成对同源服务的高精度识别框架。这里的同源服务是指服务内容相同或者相似的服务。

关联分析模块首先利用极大似然估计法和多重插补方法对测量结果进行真值判定并且对缺失属性值进行填充。接着对同源服务进行聚类。最后对服务关系进行挖掘, 最终生成包括服务实体, 服务关系, 服务属性和服务类别的服务知识库。

3) 服务知识库可视化

该模块采用文本可视化、网络可视化和时空可视化对服务知识库进行展示, 并且通过网络空间坐标系, 以“地图”的形式呈现特定区域的服务画像。

服务画像包括基础属性、网络属性、位置属性、通联属性等。例如基础属性包括企业法人、主办单位名称、网站备案/许可证号、网站名称、网站首页网址等; 网络属性包括网站 IP、网站域名、AS 号码等; 位置属性包括国家、区域代码、所在地、经纬度、地址等; 通联属性包括外链网址, 服务使用者 IP 等。服务信息在时空范围内可与其他资源信息等叠加, 按需展示关注点。

4) 服务测绘应用

根据服务测绘地图, 面向不同的业务场景, 应用不同的层次数据, 提供决策支持, 并通过迭代演进使得服务测绘能力不断提升。典型应用场景如下:

(1) 特定服务发现和识别场景: 通过主动扫描、流量监控等多种探测方式, 获取特定服务的信息。对关注区域内的特定服务进行分析与统计, 便于安全监管部门。

(2) 区域服务状态评估场景: 绘制网络空间上服务影响范围状态图, 在网络攻防的实践中, 为网络靶场等应用提供精准的攻击效果评估。

(3) 特定服务的用户分析场景: 绘制网络空间上服务和用户的连接关系图, 对特定服务的用户以及潜在用户进行群体分析。

(4) 特定用户的服务推荐场景: 绘制网络空间上服务和服务的连接关系图, 对特定用户进行服务推荐。

5 总结和展望

面向新时代网络空间技术的应用需求, 传统的网络测量技术逐渐发展为对网络空间各类资源的发现、分析和展示(即网络空间资源测绘)技术。总体而言, 网络空间资源测绘的相关研究还处于起步阶段, 缺乏对网络空间资源测绘概念的统一认知, 尚未建立一套完整的网络空间资源理论模型和技术体系。

本文首先介绍了网络空间、网络空间资源及网络空间资源测绘等概念的内涵和外延, 提出了网络空间资源的分类体系和属性图谱; 然后,提出了网络空间资源测绘的通用技术体系(DAVA Loop), 并对“探测”、“分析”、“绘制”等关键技术的研究思路和相关工作进行了综述和探讨; 最后, 基于该技术体系的基本框架, 以网络资产评估和服务测绘为具体应用场景, 分析了技术体系的有效性。网络空间资源测绘领域存在大量的理论和技术问题, 需要我们进一步探讨和研究。在未来工作中, 我们将围绕“协同探测”、“融合分析”和“全息绘制”开展更加深入和具体的研究工作。

致谢 本课题得到国家重点研发计划“网络空间测绘”项目(2016YFB0801300)资助。

参考文献

[1] Nakashima, Ellen, With Plan X, Pentagon seeks to spread US military might to cyberspace, Washington Post, 30, 2012

[2] Grant Tim, On the military geography of cyberspace, Leading Issues in Cyber Warfare and Security: Cyber Warfare and Security,

2:119, 2015

[3] O. Rashid, I. Mullins, P. Coulton and R. Edwards, “Extending cyberspace: location based games using cellular phones,” Computers in Entertainment, vol. 4, no. 1, 2006.

[4] ZoomEye, https://www.zoomeye.org/.

[5] Foeye, http://www.baimaohui.net/foeye#gnys.

[6] G. William, “Neuromancer,” Translated by Limin Lei and Chu’an Wen (in Chinese), Shanghai: Shanghai Scientific & Technological Education Publishing House, pp. 7-9, 1999.

[7] B. Sterling, “The Geography of the Internet Infrastructure: A Simulation Approach Based on the Barabasi-Albert Model,” Evolutionary Economic Geography, vol. 63, no. 1, pp. 19-37, 2010.

[8] P. C. Adams, “A taxonomy for communication geography,” Progress in Human Geography, vol. 35, no. 1, pp. 37-57, 2011.

[9] K. Hillis, “Digital sensations: Space, identity, and embodiment in virtual reality,” London: Routledge, pp. 55-59, 1999.

[10] B. Wellman, J. Salaff, D. Dimitrova, L. Garton, M. Gulia and C. Haythornthwaite, “Computer networks as social networks: collaborative work, telework, and virtual community,” Knowledge & Communities, vol. 22, no. 22, pp. 179-207, 2000.

[11] M. Mayer and L. Martino, “Towards geographies of cyberspace,” Progress in Human Geography, vol. 22, no. 2, pp. 385-406, 1998.

[12] B.X. Fang, “Define cybersecurity,” (in Chinese), Journal of Network and Information security. Vol. 4, no. 1, pp. 1-5.
(方滨兴,“定义网络空间安全”, *网络与信息安全学报*, (4): 1-5。)

[13] Dodge and M. Kitchin, “Mapping cyberspace. International Encyclopedia of Human Geography,” no. 1, pp. 356–367, 2000.

[14] J. W. Yoder and R. Johnson, “The Adaptive Object-Model Architectural Style,” Software Architecture-IFIP International Federation for Information Processing, 2002.

[15] R. Wille, “Restructuring lattice theory: An approach based on hierarchies of concepts,” In: Rival I ed.Dordrecht: Reidel, pp. 445-470, 1982.

[16] Shodan. https://www.shodan.io/.

[17] C. Shannon, D. Moore, K. Keys, M. Fomenkov, B. Huffaker and K. Claffy, “The internet measurement data catalog,” ACM SIGCOMM Computer Communication Review, vol. 35, no. 5, pp. 97-100, 2005.

[18] A. Dainotti, K. Benson, A. King, M. Kallitsis, M. Kallitsis and E. Glatz, et al., “Estimating internet address space usage through passive measurements,” ACM SIGCOMM Computer Communication Review, vol. 44, no. 1, pp. 42-49, 2013.

[19] J.P. Wu, “Current state and future of China education and research network”, New Technology of Library and Information Service, vol. 6, pp. 9-11, 1996.
(吴建平, 中国教育和科研计算机网 CERNET 的现状和发展, *现代图书情报技术*, 1996, 6: 9-11。)

[20] R. Motamedi, R. Rejaie and W. Willinger, “A survey of techniques for Internet topology discovery,” IEEE Communications Surveys & Tutorials, vol. 17, no. 2, pp. 1044-1065, 2015.

[21] B. Donnet and T. Friedman, “Internet topology discovery: a survey,” IEEE Communications Surveys & Tutorials, vol. 9, no. 4, pp. 56-59, 2007.

[22] K. Levchenko, A. Dhamdhere, B. Huffaker, K. Claffy, M. Allman and V. Paxson, “Packetlab: a universal measurement endpoint interface,” in Internet Measurement Conference(IMC’17), pp. 254-260, 2017.

[23] K. Claffy, Y. Hyun, K. Keys and M. Fomenkov, “Internet Mapping: From Art to Science,” in Cybersecurity Applications & Technology Conference for Homeland Security(CATCH’09), pp. 205-211, 2009.

[24] Y.H. Niu, X.H. Ren and J.P. Bi, “A survey of Internet network measurement methods,” Computer application and software, vol. 23, no. 7, pp. 11-13, 2006.
(牛燕华, 任新华, 毕经平, “Internet 网络测量方式综述”, *计算机应用与软件*, 2006, 23(7): 11-13。)

[25] Z.P. Cai, F. Liu, W.T. Zhao, X.H. Liu and J.P. Yin, “Deployment Model and Optimization Algorithm of Network Measurement,” Journal of software, vol. 19, no. 2, pp. 419-431, 2008.
(蔡志平, 刘芳, 赵文涛, 刘湘辉, 殷建平, “网络测量部署模型及其优化算法”, *软件学报*, 2008, 19(2):419-431。)

[26] H.L. Zhang, B.X. Fang, M.Z Hu, Y. Jiang,C.Y. Zhan and S.F. Zhang, “Survey and Analysis of Internet,” Journal of software, vol. 14, no.1, pp. 110-116, 2003.
(张宏莉, 方滨兴, 胡铭曾, 姜誉, 詹春艳, 张树峰, “Internet 测量与分析综述”, *软件学报*, 2003, 14(1):110-116。)

[27] Y. Shavitt and E. Shir, “DIMES: Let the Internet Measure Itself,” SIGCOMM Comput. Commun. Rev., vol. 35, no. 5, pp. 71–74, Oct. 2005.

[28] H. V. Madhyastha et al., “iPlane: An information plane for distributed services,” in Proc. 7th Symp. OSDI 2006, pp. 367–380.

[29] “RIPE Atlas”, RIPE NCC, https://atlas.ripe.net/, 2018.

[30] “SamKnows”, https://samknows.com/global-platform, 2018.

[31] S. Sundaresan, S. Burnett, N. Feamster, and W. de Donato, “Bismark: A Testbed for Deploying Measurements and Applications in Broadband Access Networks,” in Proc. USENIX ATC, 2014, pp. 383–394.

[32] V. Bajpai and J. Schönwälder, “A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts,” IEEE Communications Surveys & Tutorials, vol. 17, no. 3, pp. 1313-1341, 2015.

[33] U. Goel, M. P. Witie, K. C. Claffy and A. Le, “Survey of End-to-End Mobile Network Measurement Testbeds, Tools, and Services,” IEEE Communications Surveys & Tutorials, vol. 18, no.

1, pp. 105-123, 2016.

[34] Netcraft report: http://searchdns.netcraft.com/

[35] “Gigital in 2018 in oceania,” We Are Social and Hootsuite, Jan.2018.

[36] H. V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosala, K. C. Sevcik and T. Suel, “Optimal histograms with quality guarantees,” In VLDB, vol. 98, pp. 24-27, Aug. 1998.

[37] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan and M. Strauss, “Surfing wavelets on streams: One-pass summaries for approximate aggregate queries,” In VLDB, vol. 1, pp. 79-88, Sept. 2001.

[38] P. Domingos and G. Hulten, “Mining high-speed data streams,” In: Ramakrishnan R, Stolfo S, Pregibon D, eds. Proc. of the 6th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press (KDD), pp. 71-80, 2000.

[39] P. Domingos, G. Hulten and L. Spencer, “Mining time-changing data streams,” In: Provost F, Srikant R, eds. Proc. of the 7th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM Press (KDD), pp. 97-106, 2001.

[40] P. A. Dinda, “Design, implementation, and performance of an extensible toolkit for resource prediction in distributed systems,” IEEE Transactions on Parallel and Distributed Systems, vol. 17, no. 2, pp. 160-173, 2006.

[41] A. Bifet, G. Holmes, R. Kirkby and B. Pfahringer, “Moa: Massive online analysis. Journal of Machine Learning Research,” Journal of Machine Learning Research, vol. 11, no. 2, pp. 1601-1604, 2010.

[42] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

[43] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans and B. Saha, “Apache hadoop yarn: Yet another resource negotiator,” In Proceedings of the 4th annual Symposium on Cloud Computing. ACM. pp. 1-16, Oct. 2013.

[44] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, ... and A. Ghodsi, “Apache spark: a unified engine for big data processing,” Communications of the ACM, vol. 59, no. 11, pp. 56-65, 2016.

[45] Y.N. Cao, S. Wang, X.X. Li, C. Cao, Y.B. Liu and J.L. Tan, “Inferring Social Network User's Interest Based on Convolutional Neural Network,” International Conference on Neural Information Processing (ICONIP), pp. 657-666, 2017.

[46] X.X Li, Y.N. Cao, Y.M. Shang, Y.B. Liu, J.L. Tan and L. Guo, “Inferring User Profiles in Online Social Networks Based on Convolutional Neural Network,” International Conference on Knowledge Science, Engineering and Management (KSEM), pp. 274-286, 2017.

[47] J.W. Zhang, C.Y. Xia, C.W. Zhang, L.M. Cui, Y.J. Fu and S. Yu. Philip, “BL-MNE: Emerging Heterogeneous Social Network Embedding through Broad Learning with Aligned Autoencoder,” In: Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), pp. 605-614, 2017.

[48] H.W Wang, J. Wang, J.L. Wang, M. Zhao, W.N. Zhang, F.Z. Zhang, X. Xie and M.Y. Guo, “GraphGAN: Graph Representation Learning with Generative Adversarial Nets.” In: Proceedings of the 2018 Association for the Advancement of Artificial Intelligence (AAAI), 2018.

[49] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols and S. Robinson, “Deep Learning for Un-supervised Insider Threat Detection in Structured Cybersecurity Data Streams,” arXiv pre-print arXiv: 1710.00811, 2017.

[50] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi and M. Ghogho, “Deep Learning Approach for Network Intrusion Detection in Software Defined Networking,” In Wireless Networks and Mobile Communications (WINCOM), pp. 258-263, 2016.

[51] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias and K. Li, “AI2: Training a big data machine to defend,” In Big Data Security on Cloud (CBD), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on IDS, pp. 49-54, 2016.

[52] V. Rimmer, D. Preuveneers, M. Juarez, T. V. Goethem and W. Joosen, “Automated Website Fingerprinting through Deep Learning,” Network and Distributed System Security Symposium, 2018.

[53] R. Pascanu, J. W. Stokes, H. Sanossian, M. Marinescu and A. Thomas, “Malware classification with recurrent networks,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE. (ICASSP), pp. 1916-1920, Apr. 2015.

[54] B. Athiwaratkun and J.W. Stokes, “Malware classification with LSTM and GRU language models and a character-level CNN,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE. (ICASSP), pp. 2482-2486, Mar. 2017.

[55] E. Denton, S. Chintala, A. Szlam and R. Fergus, “Deep generative image models using a laplacian pyramid of adversarial networks,” pp. 1486-1494, 2015.

[56] R. Paulus, C. Xiong and R. Socher, “A deep reinforced model for abstractive summarization,” arXiv preprint arXiv: 1705.04304, 2017.

[57] T. Man, H.W. Shen, S.H. Liu, X.L. Jin and X.Q. Cheng, “Predict Anchor Links across Social Networks via an Embedding Approach,” Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), pp. 1823-1829, 2016.

[58] Z.P. Chen, P. Zhang and Q.Y. Liu, “ProxyDetector: A Guided Approach to Finding Web Proxies,” 2017 IEEE 42nd Conference on Local Computer Networks (LCN). IEEE, EI indexed: 20180304661670 (CCF-C), 2017.

[59] X.Q. Zhou, P. Zhang, C.Y. Huang, Z.P. Chen, Y. Sun and Q.Y. Liu,

"Phishing web page discovery method based on similarity of page layout", Journal of communication (s1), pp. 116-124, 2016.
(邹学强, 张鹏, 黄彩云, 陈志鹏, 孙永, 刘庆云, "基于页面布局相似性的钓鱼网页发现方法", *通信学报(s1)*, 2016, 116-124。)

[60] H. He, J. Wieting, K. Gimpel, J. Rao and J. Lin, "UMD-TTIC-UW at SemEval-2016 Task 1: Attention-Based Multi-Perspective Convolutional Neural Networks for Textual Similarity Measurement," International Workshop on Semantic Evaluation & Tutorials, pp. 1103-1108, 2016.

[61] T. Kenter and M. D. Rijke, "Short Text Similarity with Word Embeddings," ACM International on Conference on Information and Knowledge Management & Tutorials, pp. 1411-1420, 2015.

[62] J. Bromley, I. Guyon, Y. Lecun and R. Shah, "Signature verification using a "Siamese" time delay neural network," International Conference on Neural Information Processing Systems (NIPS), Morgan Kaufmann Publishers Inc, vol. 7, pp. 737-744, 1993.

[63] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," Computer Vision and Pattern Recognition IEEE. (CVPR), pp. 4353-4361, 2015.

[64] B.X. Fang, P. Zou and S.B. Zhu, "Research on Cyberspace Sovereignty," *Engineering Sciences*, vol. 18, no.6, pp.1-7, 2016.
(方滨兴, 邹鹏, 朱诗兵, "网络空间主权研究", *中国工程科学*, 2016, 18(6) :1-7。)

[65] M. A. Smith, S. M. Drucker, R. Kraut and B. Wellman, "Counting on community in cyberspace," In Proc. CHI'99 Extended Abstracts on Human Factors in Computing Systems, 1999, 87-88.

[66] T.H. AI, M.J. Zhou and Y.J. Chen, "The LOD Representation and TreeMap Visualization of Attribute Information in Thematic Mapping," Acta Geodaetica et Cartographica Sinica, vol. 42, no. 3, pp. 453-460, 2013.
(艾廷华, 周梦杰, 陈亚婕, "专题地图属性信息的 LOD 表达与 TreeMap 可视化", *测绘学报*, 2013, 42(3): 453-460。)

[67] Z.W. Sun, Z. Lu and Y. Wang, "The Geography of Cyberspace Review and Prospect," Advances in earth science, vol. 22, no. 10, pp. 1005-1011, 2007.
(孙中伟, 路紫, 王杨, "网络信息空间的地理学研究回顾与展望", *地球科学进展*, 2007, 22(10) :1005-1011。)

[68] K. Taşdemir and E. Merényi, "Exploiting data topology in visualization and clustering of self-organizing maps," IEEE Transactions on Neural Networks, vol. 20, no. 4, pp. 549-62, 2009.

郭莉 于 1994 年在湘潭大学获得硕士学位。现任中国科学院信息工程研究所副所长、正研级高工。研究领域为网络空间安全。Email: guoli@iie.ac.cn

朱宇佳 于 2012 年在北京邮电大学获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为网络空间安全。研究兴趣包括: 域名安全、网络行为测量。Email: zhuyujia@iie.ac.cn

曹亚男 于 2012 年在中国科学院计算技术研究所获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为知识计算与文本挖掘。研究兴趣包括: 知识抽取、深度学习、网络用户分析。Email: caoyanan@iie.ac.cn.

张鹏 于 2013 年在中国科学院计算技术研究所获得博士学位。现在中国科学院信息工程研究所副研究员, 研究领域为网络安全和大数据处理和挖掘。Email: pengzhang@iie.ac.cn

苏马婧 于 2013 年在哈尔滨工业大学信息安全专业获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为网络安全、网络测量、网络用户行为分析。Email: sumajing@iie.ac.cn

周川 于 2013 年在中国科学院数学与系统科学研究院概率论与数理统计专业获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为数据挖掘。研究兴趣包括: 社交网络分析、图挖掘、异常检测。Email: zhouchuan@iie.ac.cn

尚燕敏 于 2015 年 1 月在中国科学院计算技术研究所信息安全专业获得博士学位。现在中国科学院信息工程研究所从事研究工作。研究领域为社会网络分析。Email: shangyanmin@iie.ac.cn