建立渗透测试型人才能力评估的综合评价模型

章 秀[1,2], 刘宝旭[1,2], 龚晓锐[1,2], 于冬松[1,2], 赵蓓蓓[1,2], 刘 媛[1,2]

[1]中国科学院信息工程研究所 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 网络安全人才的培养和选拔, 离不开一把衡量人才的“尺子”。以通用漏洞评分系统作为参考范例, 一个具备可操作性的评价模型, 不能只是一个抽象的思考模型, 而是应当包含准则、权重、量化取值方法、计算公式、得分和评级 6 个要素, 现有模型在这些要素上都有不同程度的缺失。因此, 本文以多轮问卷调查的形式, 综合运用了多种定性与定量评估方法, 建立起了具备以上 6 个要素的渗透测试型人才能力评估的综合评价模型, 取名为 *CEMoPT*。首先, 我们运用德尔菲法, 通过文献阅读归纳形成了评价准则结构和准则项定义; 然后, 采用层次分析法、熵权法和组合赋权法, 得到准则权重; 并设计了基于隶属度矩阵标注任务的方法以获得准则量化取值; 最后使用模糊综合评价法中相应的计算公式, 得到人才的得分和评级。我们设计了在线问卷, 招募了 72 名领域专家, 对 *CEMoPT* 的 6 个要素依次开展评议, 并严格遵循稳定性度量与共识性度量约束, 经历了最长 4 轮迭代。具体来说, *CEMoPT* 的评价准则包括 5 个基本度量组和 18 个准则项, 权重是主观权重和客观权重的组合赋权结果, 数学公式的核心元素是隶属度矩阵, 综合评级分为新手、学徒、高手、专家和大师 5 级。本文通过设计对比实验, 验证了 *CEMoPT* 的可靠性。模型建立过程严格遵循科学方法所要求的诸多度量和检验约束, 保证了 *CEMoPT* 的有效性。

关键词 渗透测试型人才; 综合评价模型; 德尔菲法; 层次分析法; 熵权法; 组合赋权法; 模糊综合评价法
中图法分类号 TP302.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2022.12.09

Establishing a Comprehensive Evaluation Model for the Competency Assessment of Pentesting Cybersecurity Talents

ZHANG Xiu[1,2], LIU Baoxu[1,2], GONG Xiaorui[1,2], YU Dongsong[1,2], ZHAO Beibei[1,2], LIU Yuan[1,2]

[1] 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 The cultivation and selection of talents are indispensable for a “ruler” to measure them. Taking the CVSS as an example, an evaluation model with high operability can not just be an abstract model of thinking. Furthermore, there are six essential elements: criterion, weight, a method to map the criterion with a corresponding numerical value, computational formula, score, and rating. Prior models lack those elements to varying degrees. Therefore, in the form of multiple rounds of questionnaire surveys, this paper uses several qualitative and quantitative evaluation methods to establish a comprehensive evaluation model with the above six essential elements for the competency assessment of pentesting cybersecurity talents, named *CEMoPT*. First, we summarized the criterion structure and definition by combining literature review with the Delphi method. Then, we applied the analytic hierarchy process, the entropy weight method, and the combination weighting method to obtain the weight of the criteria. Next, we designed a method of labeling tasks based on the membership matrix to map the criterion with a corresponding numerical value. Finally, the score and rating were calculated by taking advantage of the computational formula in the fuzzy comprehensive evaluation method. We designed an online questionnaire, recruited 72 subject matter experts, conducted reviews on the six essential elements of the *CEMoPT* in turn, and strictly followed the constraints of stability measure and consensus measure, and experienced a maximum of 4 rounds of iterations. Specific to *CEMoPT*, the criteria make up of 5 basic metric groups and 18 criterion items. The weight is a combination weight, which is a compromise between the subject weight and the object weight. The membership matrix is the core of the mathematical formula. Based on the score, the rating is divided into 5 levels, i.e., novice, apprentice, journeyman, expert, and master. The reliability of *CEMoPT* was verified by conducting a comparative experiment. To ensure the validity of *CEMoPT*, the research process strictly followed many constraints required by scientific methods.

通讯作者: 龚晓锐, 硕士, 正高级工程师, Email: gongxiaorui@iie.ac.cn。
本论文获得中国科学院网络测评技术重点实验室和网络安全防护技术北京市重点实验室资助。获得了北京市科委项目(No. Z191100007119010, No. Z181100002718002)课题资助。
收稿日期: 2020-08-22; 修改日期: 2020-12-26; 定稿日期: 2022-12-07

Key words pentesting cybersecurity talents; comprehensive evaluation model; the Delphi method; analytic hierarchy process; entropy weight method; combination weighting method; fuzzy comprehensive evaluation

1 引言

网络安全人才的培养和选拔, 离不开一把衡量人才的“尺子”。国内外在网络安全人才培养和能力评估方面的最新进展主要包括: 美国国家网络空间安全教育计划(National Initiative of Cybersecurity Education, NICE)发布的《NICE 网络安全人才队伍框架[1](NICE Cybersecurity Workforce Framework, NCWF)》; 国际计算机协会及相关机构成立的网络安全教育联合工作组(The Joint Task Force on Cybersecurity Education, JTF)发布的《网络安全专上学位课程指导[2](CSEC2017)》及后续《网络安全副学士学位课程指导[3](Cyber2yr2020)》; 英国国家网络安全中心(National Cyber Security Center, NCSC)发布的《网络安全知识体系[4-6](Cybersecurity body of knowledge, CyBoK)》; 中国网络空间安全人才教育联盟(The Cyberspace Security Talent Education Alliance of China, CEAC)发布的《网络空间安全工程技术人才培养体系指南[7-8]》; 以及美国能源部技术报告中提到的“胜任力盒子[9](Competency box)”; 美国诺瓦东南大学博士学位论文研究基于知识-技能-能力来衡量组织信息系统用户的网络安全胜任力[10](MyCyberKSAs); 360 网络安全大学提出的“知识-能力-素养模型[11]”; 还包括信息安全专业人员协会(Institute of Information Security Professionals, IISP)发布的《信息安全技能框架[12](Information Security Skills Framework, ISSF)》; 学术论文中提到的网络安全能力倾向和个人天赋研究框架[13-14](Cyber Aptitude and Talent Assessment, CATA)和人才队伍网络安全能力评估模型[15](Workforce Cyber Security Capability, WCSC)等共 10 项。

以通用漏洞评分系统[16](Common Vulnerability Scoring System, CVSS)作为参照范例, 一个具备可操作性的评价模型, 不能只是一个抽象的思考模型, 而是应当包含指标或准则、权重或系数、量化取值方法、计算公式、得分和评级等 6 个要素。CVSS 用于评估软件漏洞的严重性, 它包括基本、时间和环境 3 个度量组, 包括攻击向量、攻击复杂度、所需特权、用户交互、范围、机密性影响、完整性影响、可用性影响、利用代码成熟度、修复水平、报告可信度、发生改变的基本度量组、机密性要求、完整性要求、可用性要求等 15 个准则。然后, 它定义了每一个准则的可能取值、取值方法以及量化数值, 包括权重。最后, 它按照计算公式得出漏洞的得分, 并对应到 5 个评级: 无、低、中、高、严重。

参照 CVSS, 以上这 10 项最新研究进展在这 6 个要素上都有不同程度的缺失, 导致它们不具备很好的可操作性, 详见表 1。因此, 本研究综合运用了多种定性与定量评估方法, 建立起了具备以上 6 个要素的渗透测试型人才能力评估的综合评价模型(简称 *CEMoPT*)。

首先, 通过文献阅读[17](Literature review)和德尔菲法[18](The Delphi method)形成了评价准则结构和准则定义, 包括侦查探测、攻击准备、渗透控制、对抗博弈和取证分析等共 5 组, 包含开源情报收集、网络入口探测、威胁建模、Web 应用攻击、程序逆向分析、漏洞挖掘利用、密文解码破译、辅助攻击平台构建、初始据点获取、横向移动和域渗透、网络服务弱点利用、操作系统权限提升、应用软件产品破解、线下环境物理渗透、防护措施规避、技术反制、数据处理和取证、攻击事件分析等 18 项准则。然后, 使用层次分析法[19](Analytic Hierarchy Process, AHP)得到准则的主观权重, 通过熵权法[20](Entropy weight method)得到准则的客观权重, 并将二者应用博弈论组合赋权法[21](Combination weighting method)得到组合权重; 并设计了基于隶属度矩阵[22] (Membership matrix)标注任务的方法以获得准则量化取值; 最后, 使用模糊综合评价法[23](Fuzzy Comprehensive Evaluation, FCE)中相应的计算公式, 得到人才能力评估的得分与评级: 新手、学徒、高手、专家和大师。

建立这个综合评价模型需要克服以下挑战: (1)这是一个“以人为本[24](Human-centered)”的安全研究, 它不像是软件程序或者网络流量分析, 既没有现成的数据集, 也很难自主构建数据集; (2)研究对领域专家[25](Subject Matter Expert, SMEs)的依赖性很强, 招募到符合要求的专家并有效组织他们参与实验也是一件棘手难题; (3)研究是对抽象概念“能力”的度量, 模型的可靠性和有效性很难证明。

针对以上挑战, 研究围绕着评价模型的 6 个要素设计问卷调查[26](Questionnaire survey), 招募专家组成员 72 个, 经历 4 轮专家评议迭代。第一轮①为试点测试, 样本数为 24, 对试点测试中收集的有效专家反馈意见做合并修订。第二轮②为集中评议, 样本数为 48, 对专家群体反馈数据进行统计分析, 除“综

① 第一轮，试点测试，调查问卷在线链接地址：https://www.wjx.cn/jq/78432843

② 第二轮，集中评议，调查问卷在线链接地址：https://www.wjx.cn/jq/81471253

表 1 人才培养和评价模型的 6 个要素分析

Table 1 Analysis of six essential elements of prior models about cybersecurity talents

模型名称	抽象模型	准则	权重	量化取值方法	计算公式	评级	得分
NCWF		√					
CSEC2017、Cyber2yr2020		√					
CyBOK		√					
网络空间安全工程技术人才培养体系指南		√					
Competency box						√	
MyCyberKSA		√		√	√		√
知识-能力-素养模型	√						
ISSF						√	
CATA	√						
WCSC	√						
CVSS		√	√	√	√	√	√

合评级”之外, 其余议题均符合专家评议停止的检验条件。第三轮①是对“综合评级”的补充评议, 样本数为 45, 此轮评议通过停止检验条件。至此, 研究成功建立了综合评价模型。第四轮②是针对综合评价模型的可靠性和有效性分析, 样本数为 24。研究设计对比实验, 实验结果证明模型可靠性良好。研究用严谨的实验过程来间接保证综合评价模型有效性: 研究使用了多个定性或定量分析方法, 实验过程中的每一个流程, 都严格遵循该方法的科学流程与检验条件。研究还给出了模型的应用展望。

本文结构如下: 第二章介绍研究流程; 第三章介绍专家组; 第四章介绍准则; 第五章计算主观权重、客观权重和组合权重; 第六章构建隶属度矩阵, 得到得分和评级; 第七章评价模型可靠性和有效性; 第八章介绍模型应用展望; 第九章讨论专家重点关注的问题; 第十章介绍相关工作; 第十一章总结全文。

2 研究概述

2.1 术语定义

本研究针对的网络安全人才, 在学术界和工业界相关文献中对这类人才的称呼通常以下 6 种形式: 渗透测试工程师[27](Penetration Tester); 白帽子或黑帽子[28](Whitehat or Blackhat); 道德黑客[29](Ethical Hacker); 外部威胁行动者[30](External Threat Actor); 红队成员[31](Red-team member); 敌手模拟者[32] (Adversary Emulator)。他们具备“以攻击的形式来提高信息系统安全”的共同特征, 在本研究中将其统称为“渗透测试型人才(Pentesting Cybersecurity Talents)”。

本研究将“能力”定义为“胜任力(Competency)”, 它源于人力资源学科[33]: 一个人在特定的业务环境中、执行特定的任务, 所必须拥有的、可识别、可测量的, 知识、技能和能力(Knowledge, Skill, Abilities, KSAs), 以及其他与所开展工作相关的特征, 比如态度、行为和精力等。可见胜任力是一种综合 KSAs 三者的、更全面的描述, 在网络安全领域也有诸多应用[34-37]。在本文中遵从大众表达习惯仍简称为“能力”。

2.2 整体介绍

研究概述如图 1 所示, 按照流程、量化、共识、方法分为 4 个层次。

流程层 分为 3 个主要步骤: 确定评价准则、确定准则权重和确定综合评价。从结果看, 研究建立的综合评价模型最终是要给出综合评级和得分。

量化层 需要计算 3 个量化数值: 共识度[38](Consensus measure)、组合权重(Combination Weight, CW)和隶属度矩阵(Membership matrix)。其中组合权重又分为主观权重(Subject Weight, SW)和客观权重(Object Weight, OW)。主观权重的计算依赖层次分析法构造的成对比较矩阵[39](Pairwise comparison matrix), 客观权重的计算依赖专家自评数据的信息熵值。

共识层 包括对 6 个议题的专家评议: 准则结构、准则定义、准则重要性评价因素、准则量化取值、综合评级和得分分布。

方法层 用到了以下 4 种定性与定量评估方法: 德尔菲法, 目的在于组织专家有效沟通来形成领域共识; 层次分析法, 目的在于确定主观权重; 熵权法, 目的在于确定客观权重; 模糊综合评价法, 目的在于通过构造隶属度矩阵, 得到综合评级和得分。

① 第三轮，补充评议，调查问卷在线链接地址：https://www.wjx.cn/jq/86736893.aspx

② 第四轮，模型可靠性和有效性评价，调查问卷在线链接地址：https://www.wjx.cn/jq/86432988.aspx

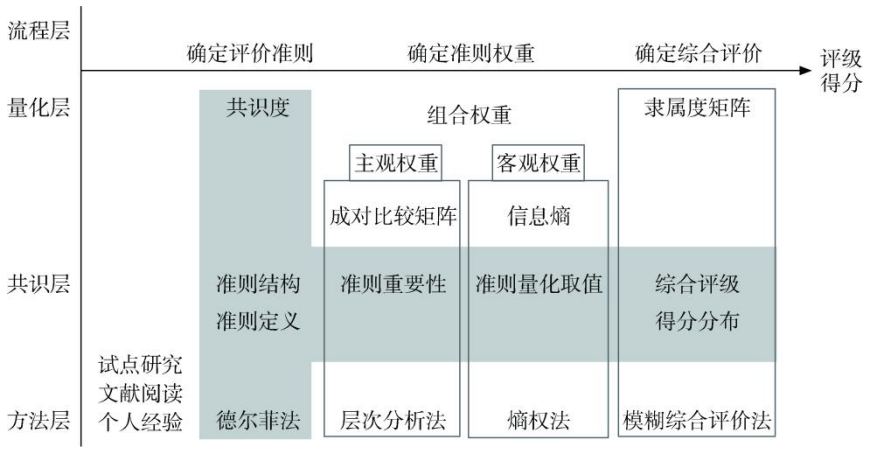

图 1　研究整体概述

Figure 1　Overview of the research

另外, 研究首先会基于个人经验、文献阅读、试点研究等方法得到一个综合评价模型的雏形, 这是研究的起点。研究通过散布有受控意见反馈的评议问卷的形式开展。

2.3　德尔菲法

德尔菲法是 20 世纪 50 年代由诺曼·达尔基[40](Norman Dalkey)提出。“德尔菲”起源于古希腊神话, 带有某种预见未来的隐喻。德菲尔法基于对集体智慧[41](Crowd wisdom)的信仰, 即专家群体的共识比某一个人的认知更可信。它的实施难点在于, 如何组织领域专家对某个具体的议题, 在不同个体认知的合作和冲突中寻求折中来达成领域共识。它在实施过程中非常依赖专家, 不仅需要专家专业性好还要求专家有足够的精力投入。

经典德尔菲过程包含以下 4 个特征[42]: (1)匿名性: 允许参与者自由表达自己的观点, 不需要和其他人保持一致, 也不受社会压力影响; (2)迭代: 允许参与者在每一个迭代轮次(Round)中完善自己的观点; (3)受控反馈(Controlled feedback): 组织者需要向参与者传达其他参与者的观点; (4)群体反馈的统计汇总(Statistical aggregation of group response): 对群体反馈数据进行统计学上的定量分析和解释。在这 4 个特征中最明显的特征是迭代, 按照经验 3 轮迭代就足够获得稳定的群体反馈, 后续迭代轮次往往变化很小。

德尔菲评议迭代终止[43]的检验条件, 即群体共识性的度量, 是德尔菲方法数据分析和解释的重要组成部分。这通常包含 2 个检验条件: (1)稳定性(Stability), 定义为专家群体反馈在统计上的一致性; (2)共识性, 定义为专家群体中多数成员达成的意见或立场。

参考文献[43], 本研究定义的德尔菲过程如图 2 所示。

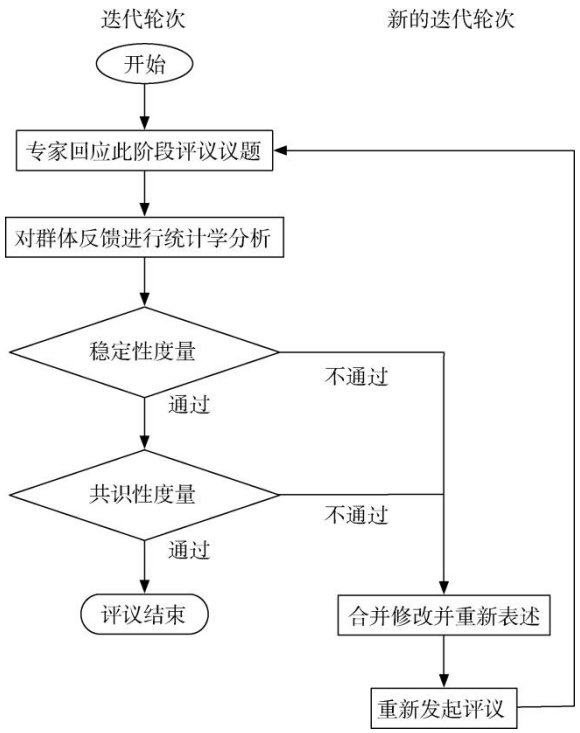

图 2　德尔菲评议过程

Figure 2　Procedure of the Delphi method

将这个过程形式化定义, 它是一个四元组: ***Delphi={Issue, Sample, Round, Consensus}***

其中: *Delphi*, 指一次专家评议; *Issue*, 指评议议题; *Sample*, 指有效专家样本数; *Round*, 指评议达成共识经历的迭代轮次数; *Consensus*, 指专家共识度。

2.3.1　稳定性度量

参照文献[44]用变异系数(Coefficient of Variation, $C.V$)来衡量群体反馈的稳定性。

$$C.V=\frac{S}{\bar{X}} \tag{1}$$

其中, S 表示数据标准差, $\bar{X}$ 表示数据均值。数值含义解读如下: $0<C.V\leqslant 0.5$, 稳定性好, 通过稳定性度量。$0.5<C.V\leqslant 0.8$, 稳定性一般, 可能需要进行额外的评议轮次, 依据共识性度量结果而定。$C.V>0.8$, 稳定性差, 不通过稳定性度量, 需要额外的评议轮次。

2.3.2 共识性度量

针对问卷设计的不同题型用到不同的共识性度量方法, 有以下 3 种:

(1) 描述性统计学分析(Descriptive statistics)通过统计处理可以简洁地用几个统计值来表示一组数据的集中趋势(Central tendency)或离散度(Dispersion), 比如, 均值、众数, 标准差。

(2) 针对 10 级净推荐值[45](Net Promoter Score, NPS)量表, 采用 NPS 值来度量共识。

$$NPS=\frac{\text{推荐者数}-\text{贬损者数}}{\text{总样本数}} \tag{2}$$

其中, 推荐者数是评分为 9～10 的样本数, 表示完全认可, 有强烈的推荐意愿; 贬损者数是评分为 1～6 的样本数, 表示不认可, 完全没有推荐意愿。另外, 评分为 7～8 的样本数, 表示基本认可但没有推荐意愿, 称为被动者数, 这部分数据很难表达明确的专家认知倾向。

大部分情况下, NPS 值在 0.05～0.1 之间徘徊; 如果 NPS 值在 0.5 以上则被认为不错; 如果 NPS 值在 0.7～0.8 之间则被认为极好。

(3) 针对 5 级李克特量表[46](Likert Scale), 采用多数意见的平均百分比[47](Average Percent of Majority Opinions, APMO)来度量共识。本研究对其定义如下:

$$APMO=\frac{\text{同意的大多数}-\text{不同意的大多数}}{\text{总样本数}} \tag{3}$$

其中, 同意的大多数是评分为 4～5 的样本数, 表示认可; 不同意的大多数是评分为 1～2 的样本数, 表示不认可。通常选取 0.8 作为 $APMO$ 推荐截止率(Cut-off Rate)。

从公式(2)(3)可以看到, NPS 和 $APMO$ 值背后的含义是: 剔除位于中间值的、反映出专家对该议题的认知倾向模棱两可的样本数。

2.4 层次分析法

层次分析法是 20 世纪 70 年代由美国运筹学家托马斯·塞蒂[48](T.L.Satty)提出。它首先将多准则决策[49](Multi-Criteria Decision Making, MCDM)问题, 分解为目标(Object)、准则(Criterion)和方案(Alternative)等 3 个层次; 然后, 在准则层通过层次单排序或层次总排序求得准则的权向量; 最后, 在方案层将所有候选方案依次对每个准则进行单排序, 再用加权和的方法与准则权重递阶归并, 得到各候选方案对目标的优先权重, 优先权重最大者即为选出的最优方案。

层次分析法的核心特征在于: 不是将所有影响决策的因素放在一起进行比较, 而是将它们进行两两比较, 构造成对比较矩阵。因为本研究应用层次分析法是为了计算主观权重, 所以本研究只关注层次分析法中和准则层权向量计算相关的部分。

2.4.1 层次单排序

层次分析法在层次单排序的情况下权向量计算的流程如图 3 所示。

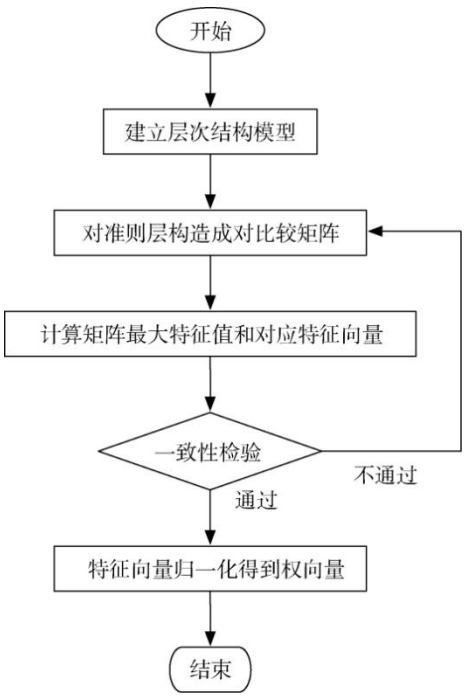

图 3 运用层次分析法计算权向量的流程

Figure 3 Procedure of calculating weight using AHP

主要步骤如下:

(1) 建立层次结构模型。在深入分析实际问题的基础上, 将影响问题决策的各个因素按照不同属性自上而下地分解成若干层次。最上层为目标层, 最下层通常为候选方案层, 中间为准则层。研究定义评价准则为:

$$C=\left\{C_i \middle| i\in[1,n]\right\} \tag{4}$$

其中, n 为准则数目。

(2) 对准则层构造成对比较阵。将准则项C_i和C_j两两比较, 构造成对比较矩阵。研究定义其为:

$$\boldsymbol{P}=\begin{bmatrix} p_{11} & \cdots & \cdots & \cdots & \cdots & p_{1n} \\ \vdots & \ddots & p_{ij} & p_{ik} & \vdots & \vdots \\ \vdots & p_{ji} & p_{jj} & p_{jk} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n1} & \cdots & \cdots & \cdots & \cdots & p_{nn} \end{bmatrix} \tag{5}$$

其中, $i,j,k\in[1,n]$,

$p_{ij}\in\{1,2,3,4,5,6,7,8,9,\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{5},\frac{1}{6},\frac{1}{7},\frac{1}{8},\frac{1}{9}\}$, 数字含义表示准则项$C_i$和$C_j$相比的重要性标度。

$\boldsymbol{P}$ 为正互反矩阵[50](Positive reciprocal matrix), 必须满足的基本约束有:

①对角线为 1, $p_{jj}=1$;

②对称位置的元素互为倒数, $p_{ij}=1/p_{ji}$;

基于专家是绝对理性的假设, 理论上 $\boldsymbol{P}$ 应该满足的一致性约束有:

③判断的传递性: $p_{ij}*p_{jk}=p_{ik}$。举例说明, 如果, p_{12}=3, 表示准则C_1和准测C_2相比, 重要性标度值为 3; p_{23}=2, 表示准则C_2和准测C_3相比, 重要性标度值为 2; 那么, 准则C_1和准测C_3相比, 重要性标度值理论上应该为 6, 即$p_{13}=p_{12}*p_{23}$。

p_{ij}的重要性标度取值方法为表 2 所示的成对比较标尺。

表 2 成对比较标尺

Table 2 Scale of pairwise comparison

标度值	含义
1	准则项C_i和准则项C_j相比一样重要
3	准则项C_i和准则项C_j相比稍微重要
5	准则项C_i和准则项C_j相比明显重要
7	准则项C_i和准则项C_j相比强烈重要
9	准则项C_i和准则项C_j相比极端重要
2,4,6,8	上述两相邻判断的中值
以上数值的倒数	准则项C_i和准则项C_j相比是χ, 则准则C_j和准则C_i相比是$\frac{1}{\chi}$

④记 $\boldsymbol{P}$ 的特征值向量为$\boldsymbol{\lambda}=[\lambda_1,\lambda_2,...,\lambda_n]^{\mathrm{T}}$, 记最大特征值为$\boldsymbol{\lambda}_{\max}$, 满足$\boldsymbol{\lambda}_{\max}=\lambda_1=n$[38], 其余特征值均为 0; 记$\boldsymbol{\lambda}_{\max}$ 对应的特征向量为 $\boldsymbol{U}=[u_1,u_2,...,u_n]^{\mathrm{T}}$, 满足 $PU=\boldsymbol{\lambda}_{\max}U=nU$。

然而, 由于客观事物的复杂性以及人们对事物进行判断比较时的模糊性, 在实际操作中需要 P 同时满足以上诸多限制是不可行的。退而求其次, 允许 P 存在一定程度的不一致性。一致性要求简化为: $\boldsymbol{\lambda}_{\max}$ 和 n 尽可能接近。

(3) 计算矩阵最大特征值和对应特征向量。该步骤就是常规的矩阵运算, 略过细节。

(4) 一致性检验。层次分析法利用一致性指标(Consistent Index, CI)、随机一致性指标(Random Index, RI)和一致性比率(Consistent Ratio, CR)来做一致性检验。CI 计算公式为:

$$CI=\frac{\boldsymbol{\lambda}_{\max}-n}{n-1} \tag{6}$$

RI 作为 CI 的参照是这样得到的: 假设阶数为 n, 随机构造满足基本约束①②的 $\boldsymbol{P}$, 这样得到的 $\boldsymbol{P}$ 大概率是不满足一致性约束的, 求 $\boldsymbol{P}$ 对应的$\boldsymbol{\lambda}_{\max}$记为一次采样。将这个过程重复很多次, 充分采样之后得到的$\boldsymbol{\lambda}_{\max}$的平均值, 即阶数为 n 时对应的 RI。RI 由查表获得, 如表 3 所示。

表 3 成对比较矩阵阶数为 n 时对应的 RI

Table 3 RI in line with the order of matrix P

阶数	1	2	3	4	5	6	7	8
RI	N/A	N/A	0.52	0.89	1.11	1.25	1.34	1.41

当 $n\leqslant 2$ 时, RI 没有意义, 表格中标记为 N/A(Not Applicable)表示此项不适用, 后文中与同样的情况时遵循此约定。

CR 的计算公式:

$$CR=\frac{CI}{RI} \tag{7}$$

如果 $CR<0.1$, 则认为 $\boldsymbol{P}$ 通过了一致性检验。否则需要重新构造 $\boldsymbol{P}$。

(5) 特征向量归一化之后得到权向量。因为层次分析法中成对比较矩阵的构建依赖于专家主观认知, 它是一种主观赋权法, 所以研究将层次分析法计算得到的权向量称为主观权重, 记为:

$$SW=[sw_1,sw_2,\cdots,sw_n] \tag{8}$$

其中, n 为准则数目。SW 和 C 一一对应表示 C 的主观权重, 是特征向量 $\boldsymbol{U}$ 的归一化, 计算公式为:

$$sw_i=\frac{u_i}{\sum_{i=1}^{i=n}u_i} \tag{9}$$

其中, $i\in[1,n]$, 满足: $sw_i\in[0,1]$, $\sum_{i=1}^{i=n}sw_i=1$。

2.4.2 层次总排序

层次单排序通常以 7 为分界线, 当准则项过多时, 即 $n>7$ 时, 就需要在层次单排序的基础上对准则项C_i进一步分解出子准则项(Sub Criterion), 记为:

$$SC=\left\{SC_{ij}\,\middle|\,i\in[1,n],j\in[1,m]\right\} \tag{10}$$

其中, n 表示准则层数目, m 为C_i对应的SC_{ij}的数目。

这种情况下, 需要对准则层 C 按照层次单排序方法构造 P, 计算 CI 做一致性检验并计算权向量

SW_i; 然后对每一个C_i对应的SC_{ij}, 按照层次单排序方法构造P_i, 计算CI_i做一致性检验并计算权向量SW_j。

在这些检验都通过之后, 对层次总排序做一致性检验, 综合一致性比率记为$CR^{'}$, 计算公式如下:

$$CR' = \frac{\sum_{i=1}^{i=n}(SW_i * CI_i)}{\sum_{i=1}^{i=n}(SW_i * RI_i)} \tag{11}$$

如果$CR^{'} < 0.1$, 则认为层次总排序通过一致性检验; 否则, 则需要对CI_i较大值对应的SC_{ij}重新构造P_i。

检验通过之后, 将各级权重对应相乘, 得到SC_{ij}对目标层的最终权重SW_{ij}, 计算公式如下:

$$SW_{ij}\prod_{i=1,j=1}^{i=n,j=m} SW_i * SW_j \tag{12}$$

满足: $SW_{ij} \in [0,1]$, $\sum_{i=1,j=1}^{i=n,j=m} SW_{ij} = 1$。

特别说明, 将准则进行分层只是为方便构造成对比较矩阵, 对权重计算没有实质影响。在后文中, 为简洁考虑, 准则定义仍然以公式(4)的形式来示例。

2.5 熵权法

熵是热力学中的一个物理概念, 用来度量系统的无序程度, 熵越大表示系统越无序。借鉴物理学中熵的概念, 香农提出了信息熵[51](Information Entropy), 用来描述数学上事件所包含的信息量的期望。借鉴信息熵的定义, 在熵权法中, 对于一个数据集, 如果准则对应的测量数据的信息熵越小, 说明测量值越离散, 所以该准则测量值对综合评价的影响越大, 准则权重就越高。反过来, 假设某准则对应的测量数据都相等, 说明该准则对应的测量值对综合评价完全没有影响, 即权重为 0, 表示该准则没有存在的意义, 需要被删除。

熵权法分为 4 个主要步骤[52], 如图 4 所示。

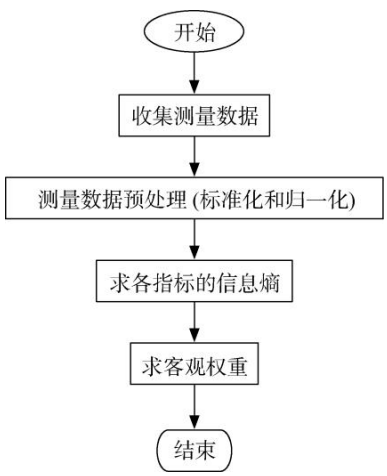

图 4 熵权重计算过程

Figure 4 Procedure of entropy weight method

(1) 收集测量数据。研究约定, 测量数据集包含 l 个样本, 即评估对象, 每个样本包括和准则 C 对应的测量值 n 个, χ_{ik}是准测项C_i在样本 k 采集到的测量数值。测量数据集记为:

$$X = \left\{ \chi_{ik} \middle| i \in [1,n], k \in [1,l] \right\} \tag{13}$$

(2) 测量数据预处理, 包括标准化和归一化。首先对数据做离差标准化处理, 使数据落到 0～1 之间。因为本研究全是正向指标(benefit-type), 即测量数值越高越好, 所以采用以下方式完成数据标准化处理。

$$\chi'_{ik} = \frac{\chi_{ik} - \mathrm{Min}(X_i)}{\mathrm{Max}(X_i) - \mathrm{Min}(X_i)} \tag{14}$$

其中, Max()是求最大值函数, Min()是求最小值函数。然后做归一化处理, 得到数据值的比例关系, 记为:

$$Y_{ik} = \frac{\chi_{ik'}}{\sum_{k=1}^{k=l}\chi_{ik'}} \tag{15}$$

(3) 求准则测量数据的信息熵。根据信息论中信息熵的定义, 准则C_i对应的测量值X_i的信息熵为:

$$E_i = -\frac{1}{Ln?(l)}\sum_{k=1}^{k=l} Y_{ik} Ln(Y_{ik}) \tag{16}$$

其中, $Ln(\,)$是求自然对数函数, 满足$0 \leqslant E_i \leqslant 1$。如果$Y_{ik} = 0$, 则定义$\lim_{Y_{ik}\to 0} Y_{ik}\, Ln(Y_{ik}) = 0$。

(4) 求准则的熵权重。因为熵权法依赖于对准则测量数据的信息熵计算, 它是一种客观赋权法, 所以本研究将熵权重称为客观权重, 记为:

$$OW = [ow_1, ow_2, \cdots, ow_n] \tag{17}$$

其中, n 为准则数目, OW 和 C 一一对应表示 C 的客观权重, ow_i的计算公式为:

$$ow_i = \frac{1 - E_i}{n - \sum_{i=1}^{n} E_i} \tag{18}$$

满足: $0 \leqslant ow_i \leqslant 1$, $\sum_{i=1}^{i=n} ow_i = 1$。

2.6 组合赋权法

主观赋权法依赖于专家主观认知, 客观赋权法依赖于测量数据。主观权重存在专家认知的主观偏差, 客观权重存在采样不充分带来的客观缺陷。因此, 对某个准则而言, 主观权重和客观权重可能存在冲突。组合赋权[23,53]就是在主观权重和客观权重之间寻求最佳均衡, 使得组合权重对二者都是最优方案。这非常符合博弈论(Game Theory)中的纳什均衡(Nash Equilibrium)的应用场景。

本研究定义组合权重为:

$$CW = [CW_1, CW_2, \cdots, CW_n] \tag{19}$$

CW 和 C 一一对应表示 C 的组合权重。最佳均衡系数为α_1和α_2, 则满足:

$$CW = \alpha_1 * SW + \alpha_2 * OW \tag{20}$$

组合赋权法的目的是: 求得α_1和α_2, 使得 CW 到 SW 的距离、CW 到 OW 的距离都能最小, 表示为:

$$\begin{aligned} &\mathrm{Min}\left\|CW - SW\right\|_2 \\ &\mathrm{Min}\left\|CW - OW\right\|_2 \end{aligned} \tag{21}$$

按照矩阵的微分性质, 取公式(20)和公式(21)的最优化一阶导数, 可转化为:

$$\begin{bmatrix} SW \cdot SW^{\mathrm{T}} & SW \cdot OW^{\mathrm{T}} \\ OW \cdot SW^{\mathrm{T}} & OW \cdot OW^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} SW \cdot SW^{\mathrm{T}} \\ OW \cdot OW^{\mathrm{T}} \end{bmatrix} \tag{22}$$

计算得到α_1和α_2, 然后做归一化处理:

$$\alpha_1' = \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{23}$$

其中, $i \in [1,2]$。将α'_1和α'_2代入公式(20)得到 CW。

2.7 模糊综合评价法

综合评价[53-54]是在对每一个评价准则做出单独评价的基础上, 进一步对所有评价准则给出一个整体评价。模糊综合评价法包括了定性评价和定量评价两个方面, 定性评价是给出综合评级, 定量评价是给出最终得分。方法的核心是通过设计隶属度函数[55-56]来确定隶属度矩阵, 也称作单因素评价矩阵。

模糊综合评价包括 6 个主要步骤, 如图 5 所示。

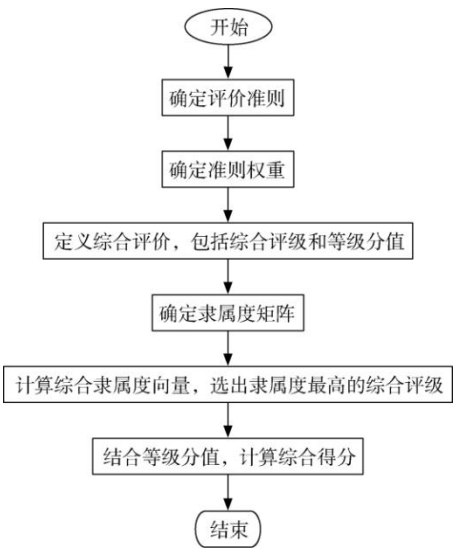

图 5 模糊综合评价法流程

Figure 5 Procedure of fuzzy comprehensive evaluation

(1) 确定评价准则。和公式(4)定义一致。

(2) 确定权准则权重。和公式(19)定义一致。

(3) 定义综合评价, 包括综合评级和等级分值, 通常分为 5 个等级。研究定义综合评级为:

$$V_{rating} = \left\{ Vr_t \middle| t \in [1,5] \right\} \tag{24}$$

定义等级分值为:

$$V_{scale} = [V_{s1}, V_{s2}, V_{s3}, V_{s4}, V_{s5},] \tag{25}$$

(4) 确定隶属度矩阵。研究定义为:

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & \cdots & r_{15} \\ \vdots & r_{it} & \vdots \\ r_{n1} & \cdots & r_{n5} \end{bmatrix} \tag{26}$$

其中, $i \in [1,n]$, n 为准则数目, $t \in [1,5]$。r_{it}表示评估对象 A 在评价准则C_i上对应的综合评级Vr_t上的隶属度。r_{it}可以通过模糊数学运算得出, 也可以通过对专家投票数据的统计百分比确定。

(5) 计算综合隶属度向量, 选出隶属度最高的综合评级。研究定义综合隶属度向量为:

$$\boldsymbol{B} = [b_1, b_2, \cdots, b_5] \tag{27}$$

综合隶属度向量的计算公式为:

$$\boldsymbol{B} = CW \times R \tag{28}$$

b_t表示评估对象 A 对应综合评级Vr_t上的综合隶属度。记 Max(b_t)为最大综合隶属度, 那么对应的Vr_t就是评估对象 A 的综合评级。

(6) 结合等级分值计算综合得分。研究定义评估对象 A 的综合得分为 $Score$, 计算公式为:

$$Score = V_{scale} \times B^{\mathrm{T}} \tag{29}$$

3 专家组信息

研究招募专家组成员 72 个, 以 4 轮问卷调查的形式开展专家评议。第一轮为试点测试, 样本数为 24, 对试点测试中收集的有效专家反馈意见做合并修订; 第二轮为集中评议, 样本数为 48, 对专家群体反馈数据进行统计分析, 除“综合评级”之外, 其余议题均符合专家评议停止的检验条件; 第三轮是对“综合评级”的补充评议, 样本数为 45, 此轮评议通过停止检验条件; 第四轮是针对综合评价模型的可靠性和有效性分析, 样本数为 24。

研究在评议问卷的第一部分, 对专家组成员的基本信息进行了多维度的统计, 包括性别、年龄、学历、从事网络安全的年限、工作单位性质、知识结构、技能类别和专业经验。通过图 6 可见男性比例很高; 通过图 7 可见年轻人比例很高。这些统计数据和公开的行业报告数据[57]一致。通过图 8 可见专家组的主要是博士研究生学历; 通过图 9 可见专家组人员接触网络安全的年限通常为 2～5 年; 通过图 10 可见专家组人员的工作单位多为科研院所。

研究将网络空间安全(Cybersecurity)细分为系统安全、移动安全、Web 安全、网络安全(Network security)、软件安全、应用安全、数据安全、密码学安全, 共8个类别, 图11统计了专家组人员的知识结

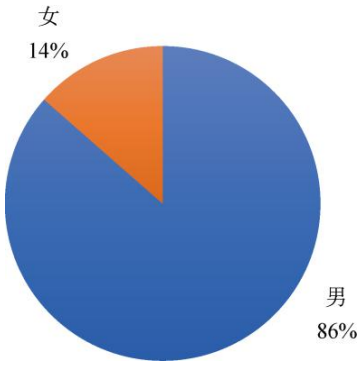

图 6 专家组性别信息

Figure 6 Gender of SMEs

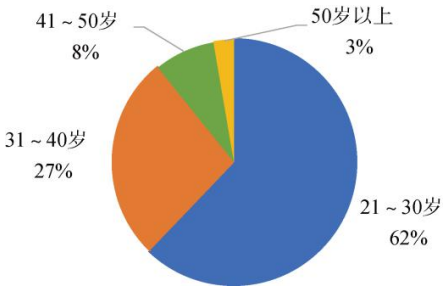

图 7 专家组年龄信息

Figure 7 Age of SMEs

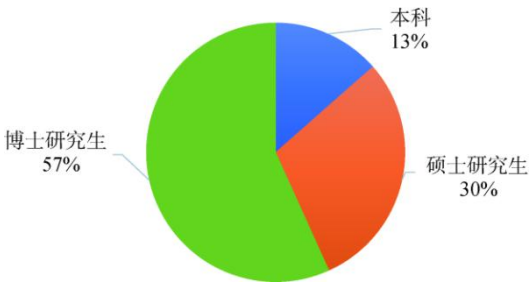

图 8 专家组学历信息

Figure 8 Education of SMEs

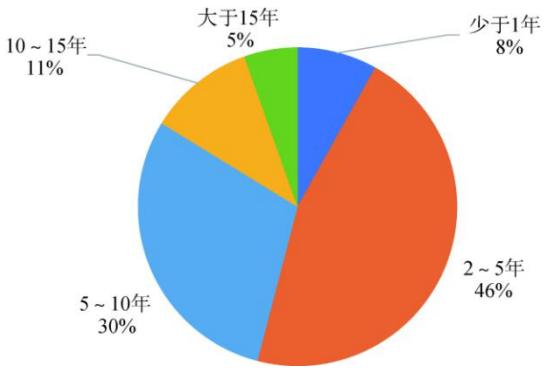

图 9 专家组从事网络安全领域的年限

Figure 9 Years engaged in cybersecurity of SMEs

构。研究列举了漏洞挖掘、漏洞利用、利用适配、程序逆向、安全编程、网站渗透、信息搜集、网络操作、渗透测试、密文破译和数字取证, 共 11 个安全相关的核心技能, 图 12 统计了专家组人员擅长的技能类别。研究枚举了能体现专家专业性的经验指标 13 项, 包括: 刷榜公开题库、参与安全竞赛、参与企业应急响应、发表学术论文、发表工业界会议报告、提交安全漏洞、拥有安全职业资格证书、拥有自己的知识共享平台、参与真实渗透测试项目、参与实践性质的安全教育项目、熟练使用渗透测试集成工具和贡献领域相关开源工具, 图 13 统计了专家组人员拥有的专业性经验。为了给专家以直观的认知, 同时避免不同专家认知差异导致反馈不一致的情况, 问卷中对每一个术语做了举例说明。

4 评价准则

4.1 准则结构

评价准则的结构和定义是文献阅读归纳的产物, 除了参考引言中提到的 10 项人才培养和评估最新研究进展之外, 还参考了大量的技术指导或描述模型, 主要有: 通用敌手、战术技术知识库[58](ATT&CK), 通用攻击模式枚举和分类[59](Common Attack Pattaern Enumeration and Classification, CAPEC), 网络杀伤链模型[60](Cyber Kill Chain), 统一杀伤链模型[61](Unified Kill Chain model), 渗透测试标准流程[62](Penetration Testing Execution Standard, PTES), 开放 Web 应用安全项目测试指导[63](Open Web Application Security Project Test Guide, OWASP-TG), Web 应用安全联合-威胁分类[64](Web Application Security Consortium Threat Classification, WASC-TC), 开源安全测试方法论[65](Open Source Security Testing Methodology Manual, OSSTMM), 信息系统安全评估框架[66](Information Systems Security Assessment Framework, ISSAF), 信息系统测试和评估技术指导[67](Technical Guide of Information Security Testing and Assessment)等。

参照层次分析法, 本研究按照“目标-准则-子准则-方案”分层定义了评价模型的整体结构, 如图 14 所示。

准则层分为共 5 组: 侦查探测、攻击准备、渗透控制、对抗博弈和取证分析。子准则层进一步细分展开, 共 18 项。按照章节 2.4 中的介绍, 本研究属于层次分析法中的层次总排序情况, 并且每一个层次单排序不超过推荐数目 7。

4.2 准则定义

网络攻防是一门实践性质的学科, 需要的是一种面向过程、面向经验的评价体系。为了使专家对准则定义有最直观的认知, 我们这里准则项的定义, 以“关键字驱动[68-69](Keyword-driven)”的方式来表述。准则定义如下:

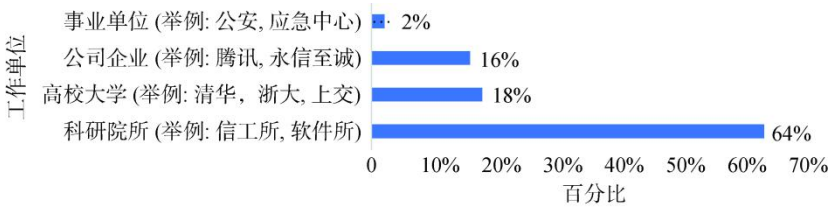

图 10　专家组的工作单位信息

Figure 10　Workplace of SMEs

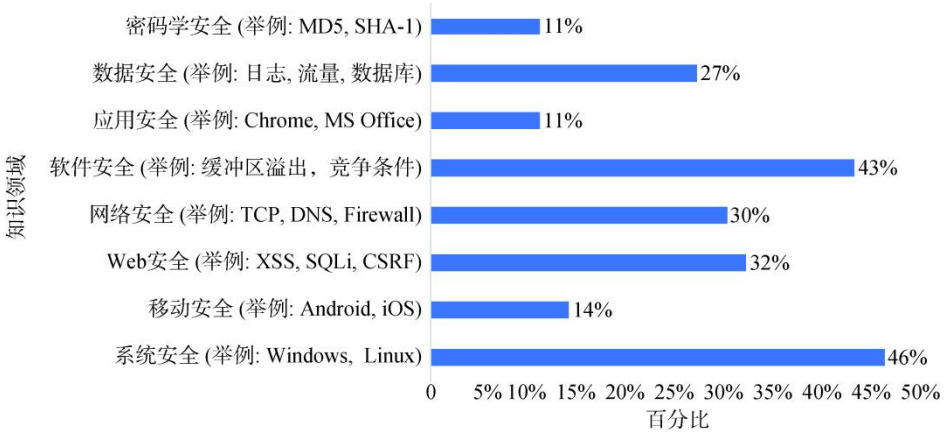

图 11　专家组了解的知识领域

Figure 11　Knowledge domain of SMEs

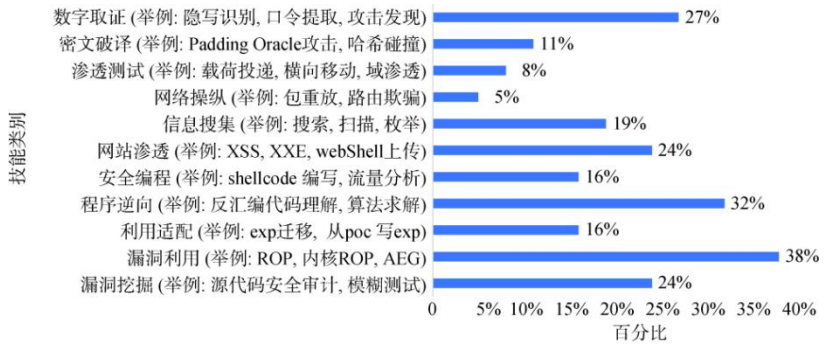

图 12　专家组擅长的技能类别

Figure 12　Skill category of SMEs

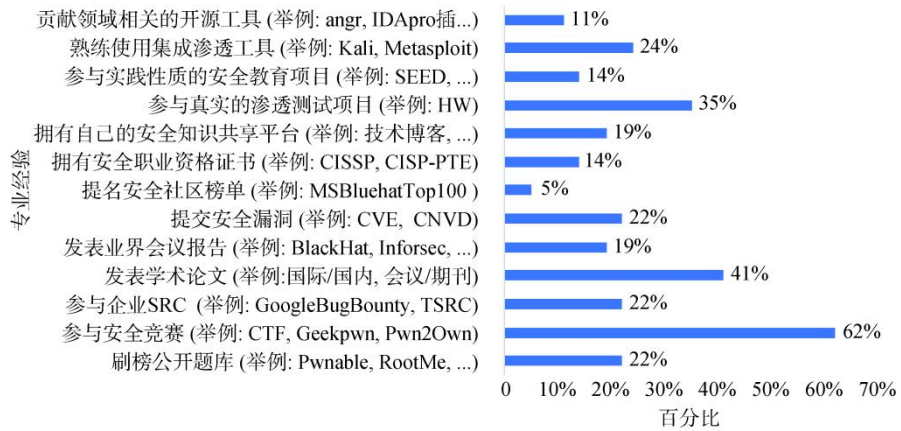

图 13　专家组拥有的专业经验

Figure 13　Professional experience of SMEs

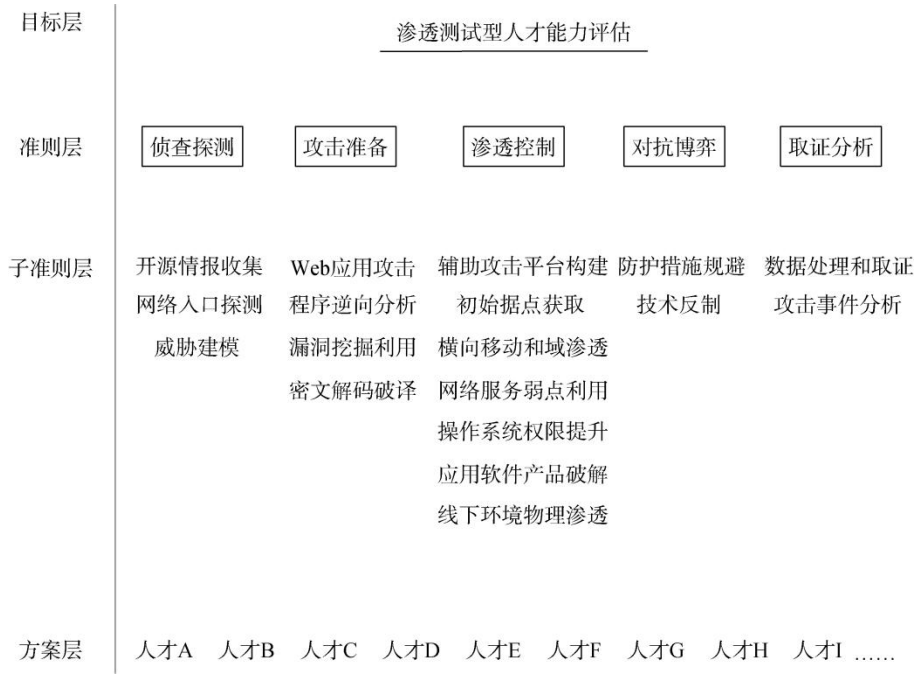

图 14 评价准则的整体结构

Figure 14 Structure of the criteria

(1) C_1: 侦查探测, 包括 3 个子准则

SC_{11}: 开源情报收集。通过公开来源情报库(Open-source intelligence, OSINT)/搜索引擎/社交网络/行业报告等途径收集数据, 并整理出组织结构/业务关系/外围信任链/人员特征等有效信息。

SC_{12}: 网络入口探测。通过网络扫描/指纹匹配/信息枚举/内容检查等途径, 探测目标的网络域名/IP 地址/开放端口/外部可访问服务等网络资产, 并积极寻找组件漏洞/系统口令等潜在的突破口。

SC_{13}: 威胁建模[70]。围绕着目标的网络资产和攻击向量, 通过绘制组织结构图/业务关系图/人物特征图/网络拓扑图/网站技术架构图等来帮助深刻理解目标信息系统, 从而有效规划攻击路径。

(2) C_2: 攻击准备, 包括 4 个子准则

SC_{21}: Web 应用攻击。遵循开放 OWASP 测试指导, 对目标 Web 应用进行输入验证/业务逻辑/服务配置/部署管理/授权和认证等测试, 并通过发起跨站脚本(Cross Site Script, XSS)/SQL 注入/XML 实体注入/反序列化漏洞利用等攻击, 达到上传 Web shell 或泄漏关键信息的目的。

SC_{22}: 程序逆向分析。通过反汇编、反编译二进制文件, 借助静态分析或动态调试工具, 对抗加壳/混淆等软件保护措施, 应对编译优化导致的信息缺失, 来理解程序逻辑与算法, 并通过求解出特定输入或修改特定代码来破解程序。

SC_{23}: 漏洞挖掘利用。通过代码审计/模糊测试等程序分析技术, 对软件中的缓冲区溢出/格式化字符串/竞争条件等漏洞进行挖掘, 并想办法突破 NX/ASLR/Stack Canary/RELRO/EIP/CFG 等安全缓解措施的限制, 利用 ROP/堆风水等技术, 劫持程序控制流。也包括以网络空间超级挑战赛[71-72](Cyber grand challenge, CGC)为代表的自动化漏洞利用技术(Automated Exploit Generation, AEG)。

SC_{24}: 密文解码破译。基于对哈希函数/数字签名/大数分解/离散对数/格计算等数学理论和 DES/AES/RSA/DSA/ECC 等现代密码协议的理解, 通过实施差分攻击/Padding Oracle 攻击/CFB 重放攻击/共模攻击/哈希长度扩展攻击等攻击方法, 破译密文。也包括凯撒/维吉尼亚编码等在内的古典加密解码。

(3) C_3: 渗透控制, 包含 7 个子准则

SC_{31}: 辅助攻击平台构建。构建支持 XSS 盲打/Reverse Shell 回连/验证码自动识别/动态域名等功能的辅助攻击平台, 并针对性地构建本地社工信息库/口令字典/Web shell/攻击载荷/1-day 漏洞利用脚本/0-day 漏洞利用脚本等资源储备库。

SC_{32}: 初始据点获取。通过网络钓鱼/水坑攻击/供应链污染等途径投递载荷, 或利用外部服务/应用漏洞, 或利用有效口令, 获取初始访问; 并通过定时任务/隐藏账户/开机自启等系统特性实现持久化, 最终通过植入远程访问工具(Remote Access Tool, RAT)/设置远程连接等途径建立起命令和控制(Command and Control, C&C)信道。

SC_{33}: 横向移动和域渗透。充分挖掘已渗透的内网主机上的攻击收益, 利用企业内部网络资产之间特殊的信任关系, 通过跳板/协议隧道等技术扩展内网访问权限, 通过有效口令/漏洞利用/会话劫持/哈希或票据传递等方法扩大内网控制范围, 最终接管内部关键信息管理系统或域控制器。

SC_{34}: 网络服务弱点利用。对常见的网络服务, 包括 Web 服务/框架/中间件(如 Apache/Ngnix/ThinkPHP/Weblogic/Struts2/Glassfish), 数据库(如 Redis/MySQL/PostgreSQL), 远程访问服务(如 SSH/RDP/VNC), 文件传输服务(如 FTP/Samba)等有丰富的渗透经验, 能够快速识别并利用其中不安全的默认配置/有缺陷的默认组件/弱口令/1-day 漏洞等。

SC_{35}: 操作系统权限提升。对某种主流操作系统, 如 Windows/Ubuntu/iOS/Android, 有内核漏洞挖掘或内核漏洞利用脚本调试经验, 能够利用内核漏洞或系统访问控制特性提升权限。在复杂情况下, 能够组合多个漏洞形成利用链, 来获取操作系统最高权限。

SC_{36}: 应用软件产品破解。对某款装机量显著的应用软件, 如浏览器/虚拟化程序/办公软件/邮件客户端/解压和归档程序等, 有丰富研究经验。能够执行浏览器沙盒逃逸、虚拟机逃逸等攻击, 或者能够基于某个 1-day 或 0-day 漏洞, 针对性构造可投递的攻击载荷。

SC_{37}: 线下环境物理渗透。对特定目标, 区别于通过网络的远程渗透方式, 而是通过近距离物理接触, 从目标的无线网络/门禁系统/打印机/投影仪/蓝牙设备等软硬件弱点进行突破, 包括执行社会工程学攻击。

(4) C_4: 对抗博弈, 包含 2 个子准则

SC_{41}: 防护措施规避。通过身份欺骗/日志清理/端口复用/时间戳修改/混淆命名等方法, 实现网络/主机/进程/文件等不同级别的隐藏, 从而有效规避 Web 应用防火墙/垃圾邮件自动过滤系统/入侵检测系统/蜜罐系统/安全审计系统/杀毒软件等防护系统的安全告警。

SC_{42}: 技术反制。在攻防多方参与的零和博弈场景中, 能够通过域名抢注/后门发现等技术有效接管他方建立的控制信道, 或者能够通过加入随机因素/数据反复擦除与重写/构建误导性入侵凭证等措施干扰他方进行取证溯源。

(5) C_5: 取证分析, 包括 2 个子准则

SC_{51}: 数据处理和取证。通过自主编程或使用专业工具, 对网络流量包/图片/归档文件/Pdf 或 Office 文档/视频或音频/磁盘镜像等原始数据进行处理, 识别文件格式, 提取嵌入子文件, 并从中寻找出口令/密钥/攻击行为等隐藏信息。

SC_{52}: 攻击事件分析。通过对磁盘镜像/内存镜像/网络流量/服务日志/录音录像/通话记录/恶意软件等情报进行分析, 能够将发生在不同时点、不同部位的攻击碎片串联起来, 还原攻击事件的来龙去脉。

4.3 准则评议

4.3.1 准则结构评议

从结构合理、逻辑清晰、粒度一致和内容完备等四个角度, 组织专家对准则结构进行评价。问卷题目设计为净推荐值 10 级量表题, 即评价分为 1～10 个候选等级, 1 表示完全不认可, 10 表示完全认可。参考图 2 所示专家评议流程, 发起第一轮专家评议, 收到有效反馈 24 份, 评议结果如图 15 所示。

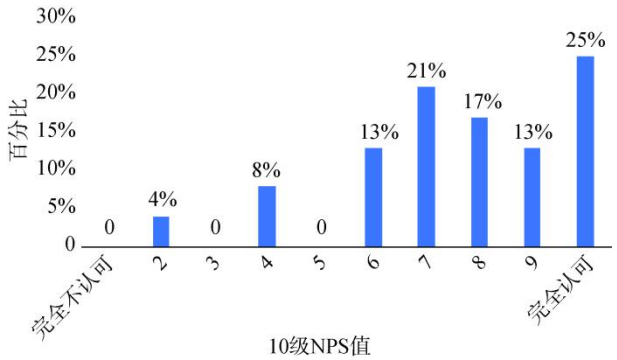

图 15 准则结构的净推荐值分布

Figure 15 Statistics of *NPS* value about criteria structure

对专家群体反馈进行统计分析, 数据均值为 7.58, 标准差为 2.12; 根据公式(1)求得 $C.V$ 值为 0.28, 小于 0.5, 通过稳定性度量; 根据公式(2)求得 NPS 值为 12.5%, 属于普遍水平, 但是参考众数为 10, 认为通过了共识性度量。专家评议过程结束, 评议轮次为 1。

4.3.2 准则定义评议

从(1)定义表述清晰无歧义; (2)定义符合常识无认知冲突; (3)该准则相比其他准则存在的必要性和合理性等三个角度, 组织专家组对准则定义进行评议。

该部分采用 5 级李克特量表来征集意见, 1 表示完全不认可, 5 表示完全认可。第一轮专家评议收集有效问卷数 24。按照章节 2.3 中介绍的方法, 对专家群体反馈进行统计分析, 详细数据见附录 1。采用 $C.V$ 来度量专家评议的稳定性, 所有准则的 $C.V$ 均小

于 0.5, 说明稳定性好。采用 $APMO$ 来度量专家评议的共识度, 有 9 项准则的 $APMO$ 值大于 80%, 满足推荐截止率; 但是有 8 项准则的 $APMO$ 值小于 80%, 而且最低的数值为 58%, 远低于推荐截止率。因此, 此轮专家评议未通过共识性度量, 需要进行新一轮的专家评议。

依据第一轮专家反馈意见, 研究重新定义了大部分准则, 并增加了一个新的准则项(SC_{37}: 线下环境物理渗透)。再次采用 5 级李克特量表的形式开展专家评议, 此轮收集到 46 个有效样本, 对专家群体反馈进行统计学分析。采用 $C.V$ 来度量专家评议的稳定性, 所有准则的 $C.V$ 均小于 0.5, 说明稳定性好。采用 $AMPO$ 来度量专家评议的共识度, 有 13 项准则的 $APMO$ 值大于 80%, 满足推荐截止率; 有 5 项准则的 $APMO$ 值虽然小于 80%, 但是和推荐截止率非常接近。故而认为, 此轮专家评议通过共识性度量, 评议结束, 评议轮次为 2。准则定义两轮专家评议的 $APMO$ 值, 统计信息如图 16 所示。

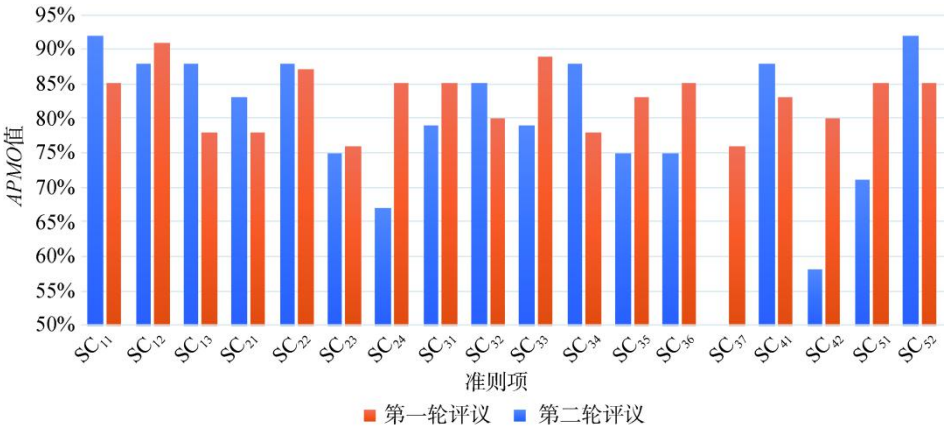

图 16 准则定义两轮专家评议的 $APMO$ 值

Figure 16 $APMO$ of two Delphi rounds about criteria

5 准则权重

5.1 主观权重

回顾章节 2.4, 本研究采用层次分析法来获得主观权重, 即通过专家对准则进行两两比较, 构造成对比较矩阵 $\boldsymbol{P}$ 来求得权重 SW。成对比较标记量尺如表 2 所示, 将准则进行成对比较时是按照“重要性”来标度。但是, 在第一轮专家评议中, 研究发现“重要性”本身也是一个模糊的概念, 不同专家对“重要性”的主观认知不一样。因此, 在第二轮专家评议中, 研究首选度量专家对“重要性”的共识, 然后基于这个共识构造成对比较矩阵 $\boldsymbol{P}$。

5.1.1 准则重要性

通过对第一轮专家评议反馈的分析, 研究归纳了影响专家对“重要性”评价的因素, 有以下 5 个:

(1) 入门门槛高低。学习该准则所描述的技能是否依赖复杂的理论知识, 是否容易获得练习环境。通常情况下, 入门门槛越高, 此类人才越可能稀缺, 准则重要性越高。

(2) 学习曲线是否陡峭。学习该准则所描述的技能, 学习瓶颈出现的早晚, 突破学习瓶颈需要投入精力的多少。通常情况下, 学习曲线越陡峭, 此类人才越可能稀缺, 准则重要性越高。

(3) 社区筛选度。该准则所描述的技能, 对应的学习天花板的高低; 在该技能上持续投入能否积累出安全社区的身份标识[73](Identity Capital)。通常情况下, 社区筛选度越好, 准则重要性越高。

(4) 人才市场的竞争力。从就业角度考虑市场供需, 该准则所描述的技能是否为安全行业研究热点, 拥有该准则所描述的技能是否能够在求职时为自己争取有竞争力的薪酬。通常情况下, 人才市场竞争力越好, 准则重要程度越高。

(5) 是否为复杂攻击链构造的必要因素。参考高级持续性威胁模型和网络杀伤链模型, 考虑该准则描述的技能对于构造复杂攻击链来说是否必要。可参照该准则描述的技能, 在已披露的网络空间攻击事件分析报告中是否出现、是否为关键步骤来度量。通常, 准则和复杂攻击链的联系越密切, 准则重要程度越高。

本研究以排序题的形式收集专家对这 5 个重要性影响因素的综合排序, 表格中的 1～5 表示排序位置, 排序数字越小表示越重要。对应权数是排序位置的逆序, 如排序位置为 1, 对应权数为 5。数据信息如表 4 所示。

表 4 准则重要性影响因素排序

Table 4 Ranking of factors about criteria importance

重要性影响因素	排序					平均综合得分	综合排序
	1	2	3	4	5		
入门门槛高低	12	7	5	13	11	2.92	3
学习曲线是否陡峭	1	10	10	13	14	2.21	5
社区的筛选度	11	15	13	4	5	3.48	1
人才市场的竞争力	11	10	17	7	3	3.40	2
是否为复杂攻击链构造的必要因素	13	6	3	11	15	2.81	4

研究收集样本数 48 个, 评价因素的平均综合得分计算公式如下:

$$\text{平均综合得分}=\frac{\sum \text{频数*权数}}{\text{样本数}} \tag{30}$$

按照平均综合得分的大小, 专家群体对 5 个重要性评价因素的综合排序情况, 按照重要性依次递减的情况是: 社区筛选度, 人才市场的竞争力, 入门门槛高低, 是否为复杂攻击链构造的必要因素, 学习曲线是否陡峭。

5.1.2 层次单排序权重

研究以矩阵填空的题目形式, 依据章节 5.1.1 中得到的专家群组对准则重要性影响因素排序的共识, 按照表 2 所示的标度方法, 组织专家对准则层成对比较矩阵进行标注。实验中该部分收回有效问卷 32 份, 对矩阵中每一个成对比较标度, 选取众数作为最终标度。准则层次单排序结果如表 5 所示。

表 5 准则层成对比较矩阵及一致性检验

Table 5 Pairwise comparison matrix of C_i and consistency check

	C_1	C_2	C_3	C_4	C_5	n	λ_{max}	CI	RI	CR	一致性检验	λ_{max} 对应特征向量 $\boldsymbol{U}$	权向量(SW_i)
C_1	1	1/3	1/5	3	2							0.2238	0.1262
C_2	3	1	1/3	4	3							0.4375	0.2466
C_3	5	3	1	5	4	5	5.22	0.06	1.11	0.05	通过	0.8498	0.4790
C_4	1/3	1/4	1/5	1	1/2							0.1028	0.0580
C_5	1/2	1/3	1/4	2	1							0.1602	0.0903

按照章节 2.4 中介绍的层次分析法, 我们对准则 C 进行层次单排序。首先, 构造成对比较矩阵 $\boldsymbol{P}$, 然后计算λ_{max}为 5.22, 依据公式(6)求得 CI 为 0.06, 查表 3 可得 n 为 5 时 RI 参考值为 1.11, 依据公式(7)求得 CR 为 0.05, 满足 CR<0.1, 通过一致性检验。

然后对λ_{max} 对应的特征向量 $\boldsymbol{U}$ 做归一化处理, 得到准则C_i对应权向量, 数据见表 5 最后一列, 对应于表 6 第一部分中的一级权重。可见准则层按照重要性依次递减排序是: 渗透控制, 攻击准备, 侦查探测, 取证分析, 对抗博弈。而且, 权重数据的相对离散, 可见层次分析法具备很好的去平均化。

5.1.3 层次总排序权重

用同样的方法分别对准则C_i对应的子准则SC_{ij}, 依次进行层次单排序, 构造成对比较矩阵 $\boldsymbol{P}_i$, 计算 CI_i, 做一致性检验, 如表 7 所示, 详细计算数据见附录 2。

可见本研究构造的成对比较矩阵 $\boldsymbol{P}_i$ 的CR_i值远小于界限值, 表明矩阵 $\boldsymbol{P}_i$一致性良好。一致性检验通过之后, 对λ_{max_i}对应的特征向量U_i做归一化, 得到准则SC_{ij}对应权重SW_j, 数据见表 6 第一部分中的二级权重。

之后按照公式(11)计算组合一致性比率CR'为 0.05, 结果小于 0.1, 层次总排序通过一致性检验。按照公式(12)将各级权重对应相乘得到主观权重SW_{ij}, 数据见表 6 第一部分中的主观权重。

5.2 客观权重

因为考虑到专家主观权重可能存在偏差, 所以通过测量数据得到的熵值分析来对其进行修正。此阶段, 我们通过专家自评来得到测量数据。为了使不同专家在自评的时候, 能够遵循同一个标准, 需要首先约定自评等级(Level, 记为 L), 即准则量化取值方法; 然后, 开展专家自评获得测试数据集; 最后, 对自评数据进行熵值分析来获得客观权重。

5.2.1 准则量化取值

本研究对评价准则量化取值方法做如下约定。自评等级分为 5 级, 对应量化数值为 1～5, 每级定义如下:

表 6　多层次组合权重计算

Table 6　Multi-hierarchy combination weight calculating

一级准则	一级权重	二级准则	二级权重	主观权重 (SW)	信息熵	客观权重 (OW)	组合权重 (CW)
C_1	0.1262	SC_{11}	0.1095	0.0138	0.9574	0.0189	0.0158
		SC_{12}	0.5816	0.0733	0.9332	0.0296	0.0557
		SC_{13}	0.3090	0.0390	0.8927	0.0476	0.0425
C_2	0.2466	SC_{21}	0.4117	0.1015	0.8649	0.0599	0.0848
		SC_{22}	0.1800	0.0443	0.8984	0.0450	0.0446
		SC_{23}	0.3098	0.0761	0.8770	0.0545	0.0674
		SC_{24}	0.0984	0.0243	0.8115	0.0835	0.0481
C_3	0.4790	SC_{31}	0.1191	0.0570	0.9189	0.0359	0.0485
		SC_{32}	0.2935	0.1406	0.8693	0.0579	0.1074
		SC_{33}	0.2342	0.1122	0.8395	0.0711	0.0957
		SC_{34}	0.0867	0.0415	0.8763	0.0548	0.0468
		SC_{35}	0.1702	0.0815	0.8716	0.0569	0.0716
		SC_{36}	0.0549	0.0263	0.8810	0.0528	0.0369
		SC_{37}	0.0414	0.0198	0.7077	0.1296	0.0639
C_4	0.0580	SC_{41}	0.7500	0.0435	0.8450	0.0687	0.0536
		SC_{42}	0.2500	0.0145	0.8072	0.0854	0.0430
C_5	0.0903	SC_{51}	0.3333	0.0301	0.9459	0.0240	0.0277
		SC_{52}	0.6667	0.0602	0.9463	0.0238	0.0456

表 7　子准则一致性检验

Table 7　Consistency check of SC_{ij}

	n	λ_{max}	CI_i	RI_i	CR_i	层次单排序一致性检验
P_1	3	3.00	0	0.52	0	通过
P_2	4	4.10	0.03	0.89	0.04	通过
P_3	7	7.41	0.07	1.36	0.05	通过
P_4	2	2	0	N/A	N/A	N/A
P_5	2	2	0	N/A	N/A	N/A
CR'						0.05
层次总排序一致性检验						通过

L_1: 了解该领域概念/术语/原理/流程/技术/操作/工具, 能够无障碍阅读/沟通/自主学习;

L_2: 掌握了该领域经典问题的求解模式, 能够独立解决大部分教科书式的问题;

L_3: 形成了自己在该领域相对完善知识和技能体系, 能够对复杂的问题进行分析/调试/求解;

L_4: 建立了自己在该领域独到的技能树/代码库/工具集, 能够快速高效地学习或开发新的问题模式;

L_5: 积累出自己在该领域的深厚经验, 能够解决实际场景中的真实问题或对社区有创造性产出。

研究采用 5 级李克特量表的形式, 按照章节 2.3 介绍的德尔菲流程, 开展专家评议。专家评议数据如图 17 所示。

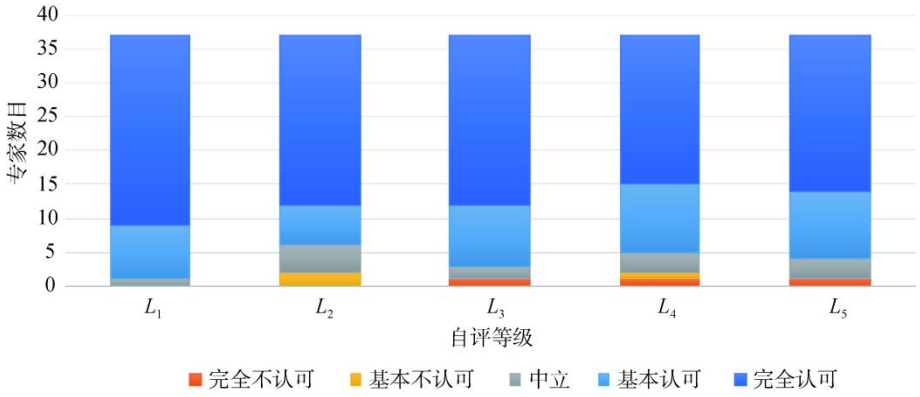

图 17　准则量化取值评议数据

Figure 17　Statistic about numerical value of the criteria

这一部分经历了 2 轮评议, 第一轮专家评议收集到有效样本数 24 个, 未通过共识性度量。第二轮专家评议收集有效样本数 37 个, 可见评议结果为“基本认可”和“完全认可”所占比例十分高, 数据分析表明L_1～L_5的 $APMO$ 值都通过了推荐截止率, 通过共识性度量, 评议终止。因为该部分评议方法和数据分析, 与章节 4.2 中对准则定义的评议非常接近, 故在文中省略, 详细评议数据分析见附录 3。

5.2.2 自评数据集

基于 5.2.1 中专家对自评等级形成的共识, 研究通过组织专家自评来得到准则测量数据。研究包含准则 18 项, 收回有效反馈 32 份。因为测量数据集过多, 所以在文中略去, 详细数据见附录 4。

5.2.3 熵权重

在获得测量数据集之后, 根据章节 2.5 中介绍, 对每一个子准则SC_{ij}对应的测量数据, 首先依据公式(14)完成离差标准化处理, 依据公式(15)完成归一化处理; 然后依照公式(16)计算准则测量数据的信息熵 E_i; 最后依照公式(17)计算客观权重 OW。测量数据得到的信息熵和客观权重见表 6 的第二部分。

5.3 组合权重

依据章节 2.6 中介绍的组合赋权法, 将已经获得的主观权重 SW 和客观权重 OW 按照公式(22)求出最佳均衡系数$\alpha_1 = 0.70$ 和$\alpha_2 = 0.44$, 然后按照公式(23)归一化之后得到$\alpha'_1 = 0.61$ 和$\alpha'_2 = 0.39$, 将其带入公式(20)得到组合权重, 见表 6 的第三部分。将主观权重、客观权重和组合权重, 展示在折线图里面如图 18 所示。

观察数据可以发现:

(1) 主观权重中最高的两项: 初始据点获取, 横向移动和域渗透。它们都是一种战术(Tactic), 而不是一种具体的技术(Technique)或流程(Procedure)。这表达出来专家的主观认知倾向: 不管使用什么样的技术或者进行什么样的流程, 能达成渗透控制的目标最重要。

(2) 客观权重中最高的两项: 线下环境物理渗透和技术反制。这是意料之外的结果, 但是深入思考之后可以发现客观权重数据呈现出一种倾向: 越是小众的评价准则, 客观权重越大。这表明某个评价准则对应领域的安全人才比较少, 但是这少部分在这个领域留下来的人才, 都是在该领域深耕细作之后的佼佼者。故而小众领域的准则, 测量数值区分度更大, 客观权重更大。

(3) 组合权重位于主观权重和客观权重之间, 是一种折中。研究得到的组合权重, 按照权重从高到底排序展示, 如图 19 所示。

6 综合评价

6.1 综合评级

回顾章节 2.1 中对本研究的“人才”和“能力”两个术语的定义, 依据认知系统工程科学家罗伯特霍·夫曼(Robert R. Hoffman)的研究[74-75], 以及在知名游戏(如我的世界和上古卷轴)中描述玩家水平的通用约定, 本研究定义评级为:

V_{rating}=[新手，学徒，高手，专家，大师] (31)

每一个评级的定义如下:

新手(Novice): 处于从安全社区中获取知识的最初阶段, 需要跟着解题报告(Writeup)学习典型问题的求解模式, 或跟着指导教程练习使用专业工具。

学徒(Apprentice): 可使用成熟工具, 或借鉴已有的代码框架/攻击载荷/漏洞利用脚本, 或对已有工具/脚本进行适配性修改, 来对某个精心构造的脆弱点(Vulnerable-by-design)进行渗透, 比如参与安全竞赛并开始有解题贡献。

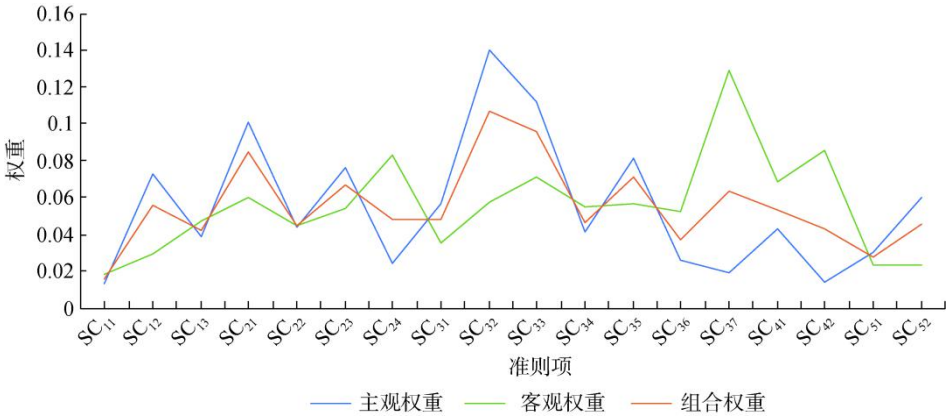

图 18 基于主观权重和客观权重计算组合权重

Figure 18 CWcalaulated by combining SW with OW

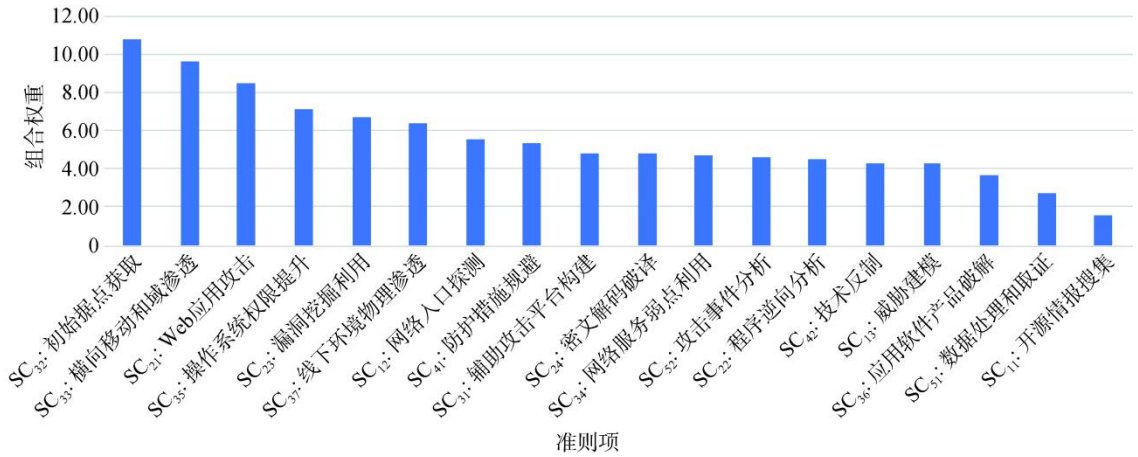

图 19　以降序方式展示的准则组合权重

Figure 19　*CW* of criteria displayed in descending order

高手(Journeyman): 通过对典型问题求解方法和技术的训练, 能够转变为解决真实世界中的具体问题, 比如从反复训练在安全竞赛中求解浏览器沙箱逃逸题目, 到提交 Chrome 浏览器沙箱逃逸漏洞拿到谷歌漏洞赏金(Bug bounty)。

专家(Expert): 对网络安全的某个子领域有深入研究, 在该领域同行中享有一定的声誉, 比如, 在智能化软件分析领域提出一套自动化漏洞利用生成技术, 在移动安全领域实现特定安卓设备组合漏洞攻击等。

大师(Master): 基于对网络安全全貌的深刻认知, 具备前瞻力和全局观, 逐步在特定安全团队中表现出来领导能力, 能够团队挑战高精尖任务, 比如构造复杂攻击链渗透真实信息系统, 或溯源分析真实攻击事件。

分析以上综合评级定义, 是一个典型的从浅到深、从点到面的过程。本研究以 5 级李克特量表的形式, 按照图 2 所示的专家评议流程开展专家评议。第一轮专家评议的有效样本数为 24, 未通过共识性度量, 专家反馈意见的核心是“定义一直围绕着构造复杂攻击链, 过于片面”; 第二轮专家评议的有效样本数为 37, 未通过共识性度量, 专家反馈意见的核心是“从人才在安全社区中的知识获取和知识输出行为表现来进行定义, 非常抽象”; 第三轮专家评议的有效样本数为 45, 通过共识性度量, 评议终止, 评议数据如图 20 所示。详细数据见附录 5。

6.2　得分分布

本研究定义和综合评级对应的等级分值为:

$$V_{scale}=[20，40，60，80，100] \tag{32}$$

参照概率论和统计学, 本研究将渗透测试型人才的得分分布定义为一个低限是 0、众数是 40、上限是 100 的连续概率分布, 即三角分布[76-77](Triangular Distribution)。

该分布的概率密度函数(Probability Density Function, PDF)和累积分布函数(Cumulative Distribution Function, CDF)如图 21 所示。

概率密度函数计算公式为:

$$P(x)=\begin{cases}\dfrac{x}{2000} & (0\leqslant x\leqslant 40)\\ \dfrac{100-x}{3000} & (40<x\leqslant 100)\end{cases} \tag{33}$$

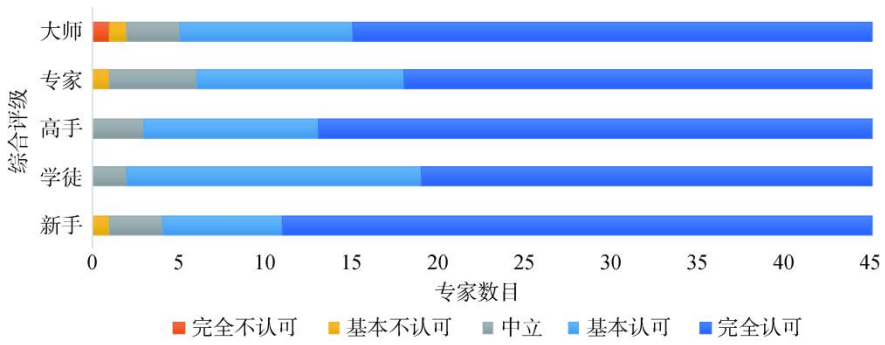

图 20　第三轮综合评级的评议数据

Figure 20　Statistics of V_{rating} *in* the third Delphi round

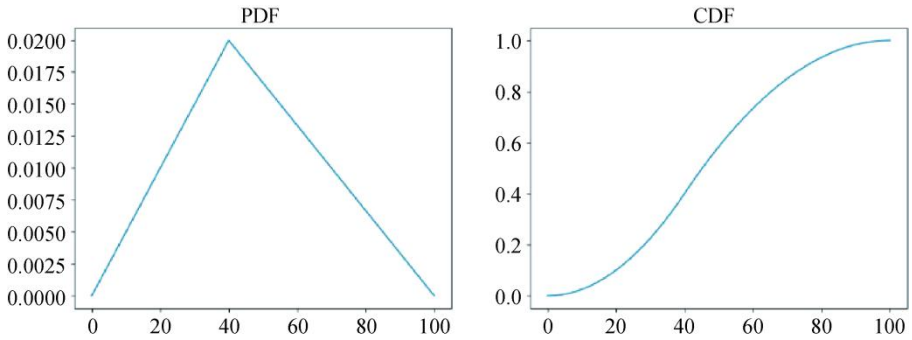

图 21　得分的概率密度函数和累计分布函数曲线

Figure 21　PDF and CDF curve of *Score*

其图形曲线表示如图 21 中(a)图所示。

该分布的累积分布函数计算公式为:

$$D(x)=\begin{cases}\dfrac{x^2}{4000} & (0\leqslant x\leqslant 40)\\ 1-\dfrac{(100-x)^2}{6000} & (40<x\leqslant 100)\end{cases} \tag{34}$$

其图形曲线表示如图 21 中(b)图所示。

依据图 21 所示的得分概率曲线, 得到对应的渗透测试型人才分布模型如图 22 所示。

图片左侧与$\boldsymbol{V}_{rating}$、V_{scale}一致, 表示评级与等级分值; 图片右侧标识两个评级之间的区域比例; 图片下侧标识 5 个评级自下而上的累积比例。可见, 各评级之间的区域比占不是平均分配的, 以学徒到高手最多, 占比接近 1/3, 符合直观认知; 评级的累计比例也不是线性增长的, 与图 21 中的(b)曲线一致。

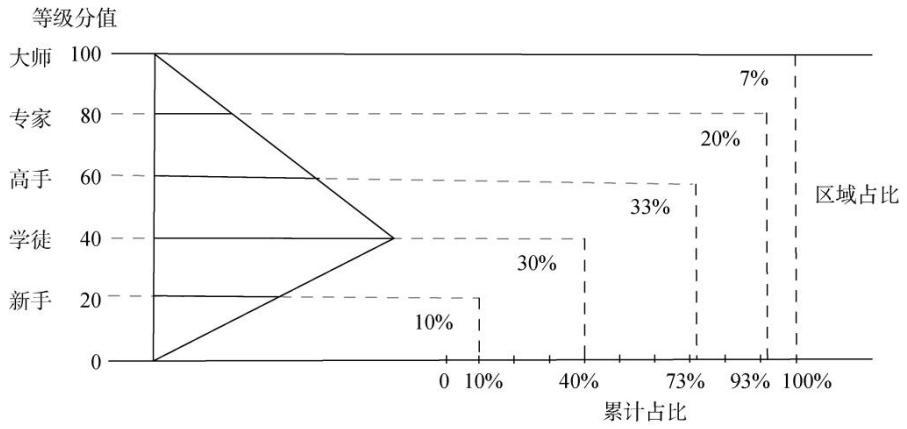

图 22　渗透测试型人才分布模型

Figure 22　Distribution of pentesting cybersecurity talents

渗透测试型人才分布模型评议数据如图 23 所示。

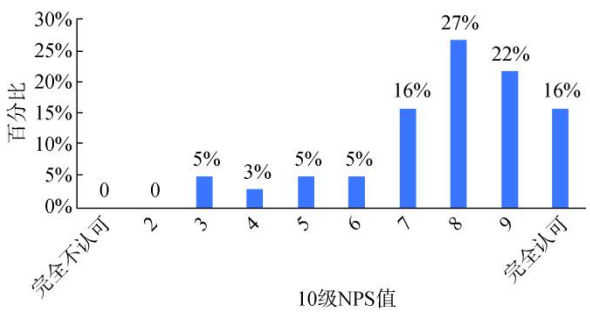

图 23　人才分布模型专家评议数据

Figure 23　Statistics of the distribution about talents

研究采用 10 级净推荐值量表来组织专家评议。第一轮专家评议, 样本数为 24, *NPS* 值为 12.5%, 结果良好, 但是专家反馈“脚本小子”“实践者”“研究员”这样的评语十分突兀。故研究借鉴认知心理学, 修改了综合评级名称。第二轮专家评议, 有效样本数为 37, *NPS* 值达到了 18.92%, 相比第一轮有显著提高。

6.3　隶属度矩阵

回顾研究, 目前我们已经获得了评价准则 SC、组合权重 CW、综合评级V_{rating}和评级量尺V_{scale}。根据章节 2.6, 隶属度矩阵 $\boldsymbol{R}$ 是映射评估对象 A 到综合评价 V 的核心。

本研究以安全团队 *NeSE*®作为专家组，随机挑选出 7 名成员作为评估对象，记为:

$$A=\{A_l|l\in[1,7]\} \tag{35}$$

组织专家对评估对象A_l在准则SC_{ij}上最适合的描述性评语V_{rating}进行投票；通过投票数据的百分比统计得到对应的隶属度r_{it}。以评估对象A_1为例，求得隶属度矩阵 $\boldsymbol{R}$ 如表 8 所示。

评估对象A_1在该准则SC_{ij}上的最大隶属度，在表 8 中以字体加粗来标记。其他评估对象的详细投票数据见附录 6。

表 8　评估对象(A_1)的隶属度矩阵和综合评价结果

Table 8　Membership matrix and comprehensive evaluation result of the alternative (A_1)

afang	新手	学徒	高手	专家	大师
开源情报搜集	0	0.29	**0.43**	0.09	0.19
网络入口探测	0.05	0.29	**0.52**	0	0.14
威胁建模	0.1	0.2	**0.35**	0.25	0.1
Web 应用攻击	0.05	0.24	**0.48**	0.23	0
程序逆向分析	0	0.04	0.22	**0.48**	0.26
漏洞挖掘利用	0	0	0	0.22	**0.78**
密文解码破译	0.22	**0.36**	0.32	0.1	0
辅助攻击平台构建	0	0.1	**0.57**	0.28	0.05
初始据点获取	0.05	0.18	**0.45**	0.18	0.14
横向移动和域渗透	0.09	0.23	**0.45**	0.09	0.14
网络服务弱点利用	0	0.29	**0.33**	0.24	0.14
操作系统权限提升	0	0	0.26	0.30	**0.44**
应用软件产品破解	0	0.23	0.27	0.18	**0.32**
线下环境物理渗透	0.05	0.27	**0.36**	0.18	0.14
防护措施规避	0	0.05	0.32	**0.45**	0.18
技术反制	0	0.14	**0.38**	0.33	0.15
数据处理和取证	0.15	0.15	**0.4**	0.2	0.1
攻击事件分析	0.09	0.05	**0.38**	0.29	0.19
有效样本数			23		
综合隶属度	0.05	0.17	**0.36**	0.22	0.20
评级			**高手**		
得分			**67.00**		

以最大隶属度值所在的等级，表示评估对象A_l在该准则SC_{ij}上的对应等级。如果出现两个等级的隶属度相等，且均为最大的情况，则取这两个等级的中间值为对应等级。将评估对象A_l在该准则SC_{ij}上的对应等级绘制成雷达图，表示评估对象的技能偏好，选取具备代表性的数据A_1,A_2,A_3，如图 24 所示。

隶属度矩阵 $\boldsymbol{R}$，理应通过动手实践型考核来获得，这才符合本研究针对渗透测试型人才能力评估的初衷。但是此处为了先完整地介绍评估模型，暂时采用问卷形式的专家投票方法来获得隶属度矩阵 R。在下一章，会介绍如何通过动手实践型考核来获得隶属度矩阵 $\boldsymbol{R}$。

6.4　评级和得分

根据章节 2.7 中的公式(28)，用组合权重 CW 和隶属度矩阵 $\boldsymbol{R}$ 相乘，得到综合隶属度向量 $\boldsymbol{B}$，数据记

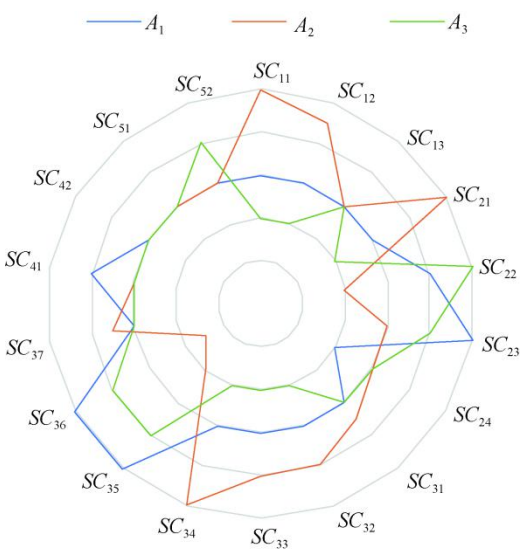

图 24　评估对象(A_1,A_2,A_3)的技能倾向

Figure 24　Skill aptitude of alternatives (A_1,A_2,A_3)

录在表 8 中。其中b_3最大, 故而评估对象A_1的最适合的综合评级为Vr_3, 即“高手”。

根据章节 2.7 中的公式(29), 用等级量尺V_{scale}和综合隶属度向量的转置B^{T}相乘, 得到评估对象A_1的综合评价得分$Score$为 67.00。用同样的方法, 求得其他评估对象的评级和得分数据, 按照从高到低排序如表 9 所示。

表 9 评估对象的综合评价得分排名

Table 9 Ranking by comprehensive evaluation

编号	网络 ID	有效样本数	评级	得分	备注
A_1	afang	23	高手	67.00	-
A_2	rebirth	21	专家	66.05	-
A_3	Ender	22	学徒	55.04	-
A_4	d3adf1sh	17	学徒	54.99	-
A_5	giglf	16	学徒	51.87	-
A_6	Mote	15	学徒	49.22	-
~~A_7~~	~~m0nshaw~~	9	~~学徒~~	~~51.94~~	样本太少数据失真

该部分研究收回有效问卷 24 份, 为了保证反馈信息的可靠性, 要求专家只对自己了解的人进行投票, 故每一个评估对象的有效样本数不一样。评估对象A_7因为是新加入 *NeSE* 团队的, 对其了解的团队成员并不多, 这导致了评议数据的有效样本数很少, 数据失真。

评估对象A_2在多数准则上表现良好, 所以依据综合隶属度向量 B, 其中b_4以微弱的优势最大, 故而评估对象A_2的最适合的综合评级为Vr_4, 即“专家”; 但是, 评估对象A_1凭借着大师级的漏洞挖掘利用、操作系统权限提升和应用软件产品破解能力, 在总分上微弱取胜。两者其实很难分出伯仲。可见, 综合评价的评级和得分可以描述评估对象不同的维度: 评级描述评估对象在多数准则上的主要表现; 得分描述评估对象在全部准则上的整体实力。

7 评价模型分析

7.1 可靠性分析

在定性与定量评估中, 可靠性[78](Reliability)是指使用测量工具(Instrument)获得的测量结果, 是否总是一致。包括 3 种类型的一致性度量: (1)重测信度(Test-Retest Reliability), 指间隔一段时间之后重复同样的实验, 得到的实验数据和上一次实验数据相比是否一致。(2)内部一致性(Internal Consistency), 指交叉项目之间, 专家对各个项目的反馈所传达出来的信息或认知倾向, 是否具备一致性。通常用折半相关系数(split-half correlation)来度量, 有 *Pearson*[79]相关系数, *Spearman Brown*[79]系数和 *Cronbach's* α[79]系数。(3)评估者间的可靠性(Interrater Reliability), 指不同的评估者使用同一个工具测量同一个对象, 获得的测量结果是否一致。

通过以上介绍可见, 可靠性主要是对结果的评估。在本研究中, 考虑研究的具体特征和可操作性设计对比实验: 按照两种方式, 即使用综合评价模型进行排序和不使用综合评价模型直接排序, 比较两种排名顺序是否存在冲突, 来评价模型的可靠性。第一种方式, 使用综合评价模型排序如表 9 所示; 第二种方式, 不使用综合评价模型直接排名, 实验中组织 *NeSE* 团队成员, 以排序题的形式对评估对象A_l直接排名, 并按照公式 5-1 处理数据, 得分和排序展示如图 25 所示。

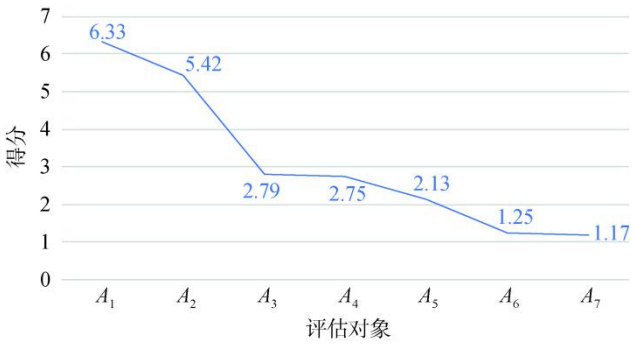

图 25 评估对象的专家投票排名

Figure 25 Ranking directly by voting of SMEs

将两种排序结果进行对比, 可以发现: 除去失真的评估对象A_7, 其余排序一致, 证明模型可靠性良好。

7.2 有效性分析

在定性与定量评估中, 有效性[78](Validity)指测量值接近预期真实值的程度。因为预期真实值是不可知的, 所以有效性的度量指标只是使测量值更可信的间接证据。包括 4 种有效性度量指标: (1)表面有效性(Face validity), 理解为充分性, 某个评价准则出现在评价体系中是合理的, 不是无关的。(2)内容有效性(Content validity), 理解为完备性, 评价体系所包含的评价准则覆盖了要评估对象的方方面面, 没有缺漏。(3)效标关联有效性(Criterion validity), 理解为测量结果应该和支撑该测量结果的效标联系密切, 即准则权重大。(4)效标判别有效性(Discriminant validity), 测量结果应该和不支持该测量结果的效标联系不密切, 即准则权重小。

通过以上介绍可见, 有效性主要是对准则和权

重的评估。本研究在设计问卷的时候, 遵循了问卷设计准则是: BRUSO[80], 即简要(Brief)、相关(Relevant)、明确(Unambiguous)、具体(Specific)、客观(Objective)。在实验过程中, 每一个步骤都严格遵循科学的方法流程与检验条件, 如表 10 所示, 严谨的实验过程间接保证了 *CEMoPT* 的有效性。

表 10　综合评价模型建立过程总结

Table 10　Summary of the process for establishing the comprehensive evaluation model

	评议议题	专家样本数	评议轮次	检验条件
定性分析	准则结构	24	1	$C.V < 0.5$ $NPS > 0.1$ $APMO > 0.8$
	准则定义	24,37	2	
	准则重要性	48	1	
	准则量化取值	24,37	2	
	综合评级	24,37,45	3	
	得分分布	24,37	2	
定量分析	主观权重	32	1	$CR < 0.1$
	客观权重	32	1	N/A
	综合得分	24	1	

8　评价模型应用展望

8.1　评估任务标注

在章节 6.3 中, 通过分析专家组对某一个评估对象的投票数据来求得隶属度矩阵。这样做方便完整地介绍模型。然而, 在评价模型的实际应用中, 对渗透测试型人才的考察需要动手实践才能考察, 这就需要评估任务(Task)。评估任务是一个有明确考察意图的、考察结果可观测或可度量的动手实践环境。

评价模型在实际应用中需要解决的问题是: 如何使用“评估对象求解评估任务”的方式, 模拟“专家投票”的方式来获得隶属度矩阵, 这个过程如图 26 所示。

首先, 运用综合评价模型以隶属度矩阵的形式来标注评估任务, 这样每一个评估任务都有自己独特的属性标签; 然后, 评估对象求解评估任务, 如果达到预期的考核指标(比如获得旗帜信息、获得操作系统最高管理权限等)就认为求解成功, 则对该评估对象用该评估任务的隶属度矩阵模拟一次有效的专家投票。如果评估对象未能成功求解评估任务, 就按照无效的专家投票处理。

评估任务的标注属性除了隶属度矩阵相关的属性之外, 还有几个其他属性: 测量项、工具或命令, 环境或组件, 这是为了给出更详细的任务描述, 丰富评估报告内容; 前置任务和后续任务, 这是描述评估任务之间的依赖关系, 这种依赖关系将一系列看似独立的评估任务关联起来, “连点成线(Connect the dots)”, 组成一个有实际意义的攻击或渗透场景。

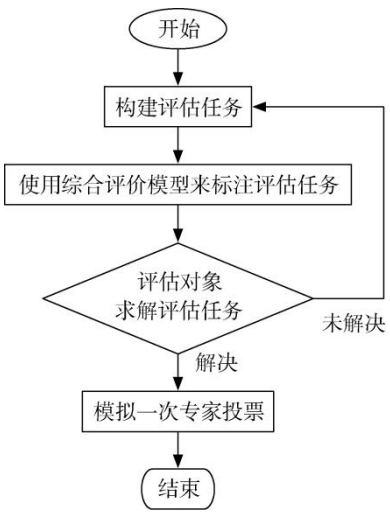

图 26　使用任务求解模拟专家投票获得隶属度矩阵

Figure 26　Using task-solving simulating SME-voting to get membership matrix

研究在附录 7 中给出了一个任务标注示例, 该示例是通过对解题报告[81]进行开放编码[82](Open-coding)得到。

研究初步构建了评估任务集, 评估任务的形式包括文件、容器、虚拟机、物理设备和复杂网络场景。文件形式的任务适合离线分析、不需要和环境交互的情况, 比如程序逆向分析、数据处理和取证。容器形式的任务适合轻量交互的情况, 比如 Web 应用攻击、网络服务弱点利用。虚拟机形式的任务适合操作系统级交互的情况, 比如操作系统权限提升。物理设备适合打印机、无线路由器这种典型物理设备层的交互需求, 比如线下环境物理渗透。复杂网络场景适合网络级的复杂交互, 比如网络入口探测、初始据点获取、横向移动和域渗透等。

不同于其他 4 种形式, 复杂网络场景作为一种特殊的类别, 在形式上通常是若干个以上 4 种形式的任务组合, 但是复杂网络场景的意义在于, 把这些看似独立的任务关联成一个完整的、具有现实意义的业务场景攻防模拟。场景的设计通常包含多个主机节点、多个网络连通或隔离域、多种网络服务、多个或多种类型的漏洞, 十分复杂。场景的实现通常需要基于网络靶场平台。

8.2　评估实验设计

评估实验的设计在于如何有效组织标注好的评估任务, 如何有效地组织评估对象参与评估实验, 来最大程度地测量到评估对象的能力上限?

针对第一个问题, 研究在初步标注评估任务的过程中总结出了 4 种类型的评估任务。(1)元技能评估, 这类任务通常是相对简单的设定、相对明确的考察点, 考察基本的知识、原理、方法、技术、流程、工具使用等。(2)应用技能评估, 这类任务通常是针对网络信息系统中已披露漏洞, 包括漏洞识别、使用成熟漏洞利用模块发起利用、从概念验证代码(Proof-of-Concept, PoC)编写利用脚本(Exploit, EXP)和利用脚本在不同平台下的适配。(3)单靶机控守, 这类任务在安全社区中被称为“Boot2root”型靶机, 是复杂攻击链中的一个单点突破缩影。主要针对以下 4 个评价准则的考察: 网络入口探测、Web 应用攻击、网络服务弱点利用、操作系统权限提升。(4)攻击链构造, 这一类任务致力于模拟真实世界中的复杂网络攻击, 被网络安全培训、网络靶场攻防演练、网络安全竞赛等诸多领域作为首要宗旨。

针对第二个问题, 研究分析了网络安全领域不同的竞赛形式或人才评估实验设计形式, 提出了与评估任务类型对应的 4 种有效的评估实验开展形式。(1)夺旗赛(Capture the flag, CTF), 形式简单直接, 适合元技能类的任务考察; (2)网络寻宝[83](Network treasure hunting), 类似于网络空间的捉迷藏, 重点考察对目标的寻找和识别, 适合应用技能类任务, 此类任务可能识别漏洞比利用漏洞更需要花费精力, 更需要对嘈杂上下文环境的歧义容忍(Ambiguity tolerance), 更需要经验积累带来的敏锐直觉; (3)网络占山为王[84-85](Network King of the Hill, NetKotH), 此种考察形式的重点是考察控制权的获得和维持, 适合单靶机控守类任务考察; (4)网络探索和漏洞利用[86-87](Network Explore and Vulnerability Exploit, *Explore-Exploit*), 适合攻击链构造类任务, 因为网络探索和漏洞利用正是复杂攻击链构造的两个核心步骤。此类任务的设计和实现非常难, 需要设计者像导演一样, 对评估过程的每一个核心步骤、每一个关键分支都安排清楚, 需要实现者掌握网络部署、系统配置、漏洞环境复现、漏洞利用触发条件等都一系列实践技能。文献[87]研究了在有限资源条件下设计与实现真实度高的网络探索和漏洞利评估实验要解决的 3 个关键问题: 建模、设计与实现。该攻击链构造评估实验以内网渗透赛的形式开展, 实验中最长的渗透路径包含 4 个跳板机, 组合利用了 3 个漏洞和 1 个服务, 可作为后续开展此类评估实验的范例。

9 讨论

在专家评议过程中收集到的专家对研究方法和议题的诸多反馈, 该部分对专家普遍关注的问题进行回答。

问题 1: 为什么不按照软件安全、系统安全、网络安全这种典型安全领域类别的分类方式构建准则?

这种安全领域类别确实是典型网络安全学科课程的划分方式, 但是, 本研究中我们期望建立的综合评价模型是面向操作和面向经验的。考虑的重点是, 不同的准则之间能够构成一个完整的杀伤链, 而不是独立的准则考察。一方面, 这更符合渗透测试型人才了考察的初衷, 因为真实攻击过程往往是复杂的, 是多个领域知识或技能的综合运用; 另一方面, 在评估实验设计的时候, 许多准则往往也没办法单独考察, 比如开源情报搜集, 失去了后续应用场景就很难考察情报的价值。

问题 2: 使用层次分析法求主观权重的时候, 填写成对比较矩阵 $\boldsymbol{P}$ 十分艰难, 尤其是准则SC_{3j}有 7 个子准则, 要填写一个 7 维方阵, 需要进行 21 次成对比较。既然这样, 为什么主观权重不通过直接分配比重来求得?

直接分配比重看起来是一种简单可行的方法, 研究一开始也采用过这个方法来获得主观权重, 但是实验结果发现: 通过直接比重划分得到的准则项权重, 和它本身的重要性无关, 而是和它所在层级的准则数量直接相关, 这显然是非常不合理的。在附录 8 中给出了直接比重划分得到的权重数据, 一级准则项SC_i有 5 个, 使用直接比重划分一级准则权重在 20%左右。二级准则SC_{3j}有 7 个, 使用直接比重划分二级准则权重在 14%左右, 这样求得SC_{3j}的组合权重在 3%左右。作为对比, 二级准则SC_{4j}有 2 个, 使用直接比重划分二级权重在 50%左右, 这样求得SC_{4j}的组合权重在 10%左右。这是因为比重划分约束了每一个层级的总比重为 100%, 而人在划分比重的时候倾向于平均赋权; 而且, 实验发现随着样本数越多, 这种平均化的趋势越明显。

对比表 7 中通过层次分析法得到的主观权重相比, 可以发现层次分析法具备很好的去平均化能力, 因为两两比较的时候只需要关注两个比较对象的相对重要性。

问题 3: 层次分析法也可以对方案层, 即评估对象做综合评价, 为什么研究不用层次分析法解决综合评价的问题? 而只用层次分析法求权重?

层次分析法的核心是构造成对比较矩阵, 求得矩阵行元素的优先权重向量。如章节 2.4.1 介绍, 对评价准则构造成对比较矩阵, 可以得到评价准则的

权重, 是层次分析法中最通用的基础步骤。

使用层次分析法做综合评价有两种模式, 第一种模式是相对测量(Relative measurement), 这个模式是在方案层, 将所有评估对象依次对每一个评价准则构造成对比较矩阵, 得到每个评估对象在每个评价准则上的优先权重。第二种模式是绝对测量(Absolute measurement), 这个模式是在准则层, 对所有评价准则依次对每个评价准则的每个可能取值(如优、良、中、差等)构造成对比较矩阵, 得到每个评价准则的每个可能取值的相对优先权重。暂且不讨论这两种模式的优劣, 以上两种模式存在一个共同的问题: 一方面, 它们过度依赖成对比较矩阵, 以相对测量为例, 假设评估对象数目为 l, 评价准则数目为 n, 综合评价需要构造 n 次 $l \times l$ 维的矩阵, 在每个矩阵中进行$\frac{n(n-1)}{2}$次成对比较; 另一方面, 它们都没有给出一种普适的可信的成对比较矩阵构造方法, 目前这种仅依赖主观赋值加一致性检验约束的构造方法, 如果反复使用太多次是不够客观和可信的。

以上, 在本研究的应用场景下, 层次分析法适合求主观权重, 而不适合做综合评价。

问题 4: 使用研究建立的综合评价模型开展人才能力评估, 和安全竞赛、职业资格认证的联系或区别是什么?

安全竞赛确实是当前网络安全人才培养和选拔的有效手段。本研究评价准则的建立和定义也参考了当前安全竞赛所考察的技能维度。但是一方面, 安全竞赛是社区性质的活动, 不同竞赛的题目类别、考察点、出题风格和题目难度缺乏有效统一的组织; 另一方面, 竞赛往往是以团队性质参与的, 所以很难通过竞赛对个人的能力进行准确的综合评价。

网络安全领域的职业资格认证也是十分有效的人才筛选方法, 本研究评价准则的定义也参考了网络安全领域相关的职业资格认证考试白皮书。但是, 职业资格认证是人才市场的准入门槛, 是进入某个领域、胜任某个岗位需求的基线, 它很难触及到评估对象的能力上限。

本研究建立的综合评价模型, 试图建立一种人才能力评估的通用量尺, 这是本研究解决的问题。在模型应用展望中, 研究后续将基于这个量尺构建一个统一的结构化组织的评估任务集; 并且, 研究会根据不同的任务形式与不同的评估层次, 开展不同形式的评估实验, 试图触及到评估对象的能力上限, 这是后续研究正在解决的问题。

问题 5: 为什么要通过标注评估任务的方式, 来获得关于评估对象的测量信息, 而不是收集评估对象的操作过程而开展评估呢?

目前能够在评估过程中采集到的数据非常有限, 通常只是网络流量数据, 很难从主机层拿到详细的数据。这是普遍存在的技术局限性, 也是这个原因导致了当前安全竞赛和职业资格认证考试都是结果导向的判定。但是, 和当前安全竞赛和职业测评不同的是, 本研究对评估任务的标注更加详细, 除了综合评价模型相关的属性之外, 还有测量项、工具或命令、环境或组件等更加详细的信息。这样在评估结束的时候, 除了给出评估对象在每个评价准则上的隶属度、综合评级和得分之外, 还能对评估对象的行为过程、技能偏好等进行细致的画像。

10 相关工作

以人为本的安全研究已逐渐成为网络安全领域顶级学术会议中的热点研究。不同于传统安全研究, 此类研究需要有效组织相关人员参与其中, 通过收集来自真实用户的真实数据来理解和分析特定的安全问题。这类研究主要使用半结构化采访、问卷调查、过程观察、资料分析、开放编码、人在环中、对比实验等方法。研究主要意义在于解释现象、给出建议、验证假设、度量偏好等。在此, 以美国马里兰大学教授米歇尔·妈祖雷克(Michelle Mazurek)团队的几项最新研究为例进行介绍: 文献[88]分析了逆向工程师做程序逆向的过程, 给出了对逆向工具的交互设计建议; 文献[89]通过分析 BiBiFi[90]竞赛的数据, 理解软件安全错误是怎么被软件开发者引入; 文献[91]比较了黑客和软件测试人员的漏洞发现过程; 文献[92]分析在数字世界里面, 人们对不同来源的安全建议是如何选择和接受的等。

德尔菲法在安全研究中的应用主要适用于研究议题涉及的知识面太广, 超出了个人认知范畴, 因此需要利用专家群体智慧辅助决策的情况。文献[10]使用德尔菲法通过 5 轮专家评议, 从 KSAs 三个角度出发, 制定了衡量信息系统用户是否具备访问组织网络权限的胜任力量表。文献[93]组织领域专家通过 3 轮德尔菲过程, 挑选出在用户购买物联网设备之前需要告知他们的涉及安全与隐私的标签, 并研究如何将这些标签依据告知的必要性按照主标签和二级标签分层展示。

11 总结

本研究综合运用了多个定性和定量分析方法, 包括德尔菲法、层析分析法、熵权法、组合赋权法和模糊综合评价法, 建立了渗透测试型人才能力评

估的综合评价模型 *CEMoPT*。研究招募 72 名领域专家, 通过 4 轮问卷调查的形式开展专家评议。研究建立的评价模型, 包括准则、权重、量化取值方法、计算公式、得分和评级等 6 个要素, 具备很好的可操作性。模型可靠性评价和有效性评价良好, 可作为渗透测试型人才能力评估的“尺子”。同时, 研究给出了基于隶属度矩阵的评估任务标注方法, 用任务求解来模拟专家投票的隶属度矩阵获得方法, 以及后续评估实验设计和实现展望。

致 谢 本论文获得中国科学院网络测评技术重点实验室和网络安全防护技术北京市重点实验室资助。获得了北京市科委项目(No.Z191100007119010, No.Z181100002718002)课题资助。感谢专家组成员对本研究的支持。感谢 *NeSE* 战队成员的大力支持。

参考文献

[1] Dan S. A Guide to the National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework (2.0)[M]. Auerbach Publications, 2016.

[2] Curriculum guidelines for post-secondary degree programs in cybersecurity (CSEC2017). *A Report in the Computing Curricula Series Joint Task Force on Cybersecurity Education*. 2017.

[3] Cybersecurity Curricular Guidance for Associate-Degree Programs (Cyber2yr2020), *A Report endorsed by Association for Computing Machinery* (*ACM*) *and Committee for Computing Education in Community Colleges (CCECC)*, 2020.

[4] Rashid A, Danezis G, Chivers H, et al. Scoping the Cyber Security Body of Knowledge[J]. *IEEE Security & Privacy*, 2018, 16(3): 96-102.

[5] Hallett J, Larson R, Rashid A. Mirror, mirror, on the wall: What are we teaching them all? Characterising the focus of cybersecurity curricular frameworks[C]. *2018 {USENIX} Workshop on Advances in Security Education*. 2018.

[6] Shoemaker D, Kohnke A, Sigler K. The Cybersecurity Body of Knowledge: The ACM/IEEE/AIS/IFIP Recommendations for a Complete Curriculum in Cybersecurity[M]. CRC Press, 2020.

[7] 中国网络空间安全人才教育联盟, 网络空间安全工程技术人才培养体系指南, 2019.

[8] 中国网络空间安全人才教育联盟, 网络空间安全工程技术人才培养体系指南 2.0, 2020.

[9] O'Neil L R, Assante M, Tobey D. Smart grid cybersecurity: Job performance model report[R]. *Pacific Northwest National Lab.* (*PNNL*), Richland, WA (United States), 2012.

[10] Nilsen R. Measuring Cybersecurity Competency: An Exploratory Investigation of the Cybersecurity Knowledge, Skills, and Abilities Necessary for Organizational Network Access Privileges [Ph.D. dissertation]. Nova Southeastern University, 2017.

[11] 360 网络安全大学,网络安全人才能力发展白皮书, 2020.

[12] Institute of Information Security Professionals (IISP), Information Security SKills Framework (ISSF), 2010.

[13] Campbell S G, O’Rourke P, Bunting M F. Identifying Dimensions of Cyber Aptitude[J]. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2015, 59(1): 721-725.

[14] Campbell S G, Saner L D, Bunting M F. Characterizing cybersecurity jobs: Applying the cyber aptitude and talent assessment framework[C]. *The Symposium and Bootcamp on the Science of Security*, 2016: 25-27.

[15] Ani U D, He H M, Tiwari A. Human Factor Security: Evaluating the Cybersecurity Capacity of the Industrial Workforce[J]. *Journal of Systems and Information Technology*, 2019, 21(1): 2-35.

[16] Hanford S. Common vulnerability scoring system, v3.1 [C]. *Technical report, Forum of Incident Response and Security Teams*. 2019.

[17] Rowley J, Slack F. Conducting a Literature Review[J]. *Management Research News*, 2004, 27(6): 31-39.

[18] Skulmoski G J, Hartman F T, Krahn J. The Delphi Method for Graduate Research[J]. *Journal of Information Technology Education*: *Research*, 2007, 6: 1-21.

[19] Saaty T L. Decision Making with the Analytic Hierarchy Process[J]. *International Journal of Services Sciences*, 2008, 1(1): 83.

[20] Zhu Y X, Tian D Z, Yan F. Effectiveness of Entropy Weight Method in Decision-Making[J]. *Mathematical Problems in Engineering*, 2020, 2020: 1-5.

[21] Xu Z S, Da Q L. Study on Method of Combination Weighting[J]. *Chinese Journal of Management Science*, 2002, 10(2): 84-87.
(徐泽水, 达庆利. 多属性决策的组合赋权方法研究[J]. *中国管理科学*, 2002, 10(2): 84-87.)

[22] Dombi J. Membership Function as an Evaluation[J]. *Fuzzy Sets and Systems*, 1990, 35(1): 1-21.

[23] Lai C G, Chen X H, Chen X Y, et al. A Fuzzy Comprehensive Evaluation Model for Flood Risk Based on the Combination Weight of Game Theory[J]. *Natural Hazards*, 2015, 77(2): 1243-1259.

[24] Mancuso V F, Strang A J, Funke G J, et al. Human Factors of Cyber Attacks[J]. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2014, 58(1): 437-441.

[25] Hopkins P, Unger M. What is a 'subject-matter expert'? [J]. *Journal of Pipeline Engineering*, 2017, 16(4).

[26] Converse J M, Presser S. Survey questions: handcrafting the standardized questionnaire[M]. Beverly Hills: Sage Publications, 1986.

[27] Konstantinos X, Iain S, Huw R, et al. Penetration Testing and Vulnerability Assessments: A Professional Approach[C]. *Paper presented at the International Cyber Resilience conference, Security Research Institute Conferences*, 2010.

[28] Matthews J, Goerzen M. Black Hat Trolling, White Hat Trolling, and Hacking the Attention Landscape[C]. *Proceedings of The 2019 World Wide Web Conference*, 2019: 523-528.

[29] Caldwell T. Ethical Hackers: Putting on the White Hat[J]. *Network Security*, 2011, 2011(7): 10-13.

[30] Horejsi J, Lunghi D, Pernet C, et al. earth akhlut: exploring the tools, tactics, and procedures of an advanced threat actor operating a large infrastructure [J], 2020.

[31] Brangetto P, Caliskan E, Rõigas H. Cyber Red Teaming-

Organisational, technical and legal implications in a military context[J]. *NATO CCD CoE*, 2015.

[32] MITRE ATT&CK Evaluation, Methodology Overview: Adversary Emulation, URL: https://attackevals.mitre.org/adversary-emulation.htmlPenetration.

[33] Chuck and Allen, Competencies 1.0 (Measurable Characteristics), *HR-XML*, Recommendation 2001-Oct-16.

[34] Vybornov A, Miloslavskaya N, Tolstoy A. Designing Competency Models for Cybersecurity Professionals for the Banking Sector[M]. Information Security Education. Information Security in Action. Cham: Springer International Publishing, 2020: 81-95.

[35] Yaokumah W. Cyber Security Competency Model Based on Learning Theories and Learning Continuum Hierarchy[M]. Global Cyber Security Labor Shortage and International Business Risk. IGI Global, 2019: 94-110.

[36] Prifti L, Knigge M, Kienegger H, et al. A Competency Model for Industrie 4.0[J]. 2017.

[37] Behrens S G, Alberts C, Ruefle R. Competency lifecycle roadmap: toward Performance readiness[R]. CMU CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 2012.

[38] von der Gracht Heiko A. Consensus Measurement in Delphi Studies[J]. *Technological Forecasting and Social Change*, 2012, 79(8): 1525-1536.

[39] Kou G, Ergu D J, Chen Y, et al. Pairwise Comparison Matrix in Multiple Criteria Decision Making[J]. *Technological and Economic Development of Economy*, 2016, 22(5): 738-765.

[40] Dalkey N, Helmer O. An Experimental Application of the DELPHI Method to the Use of Experts[J]. *Management Science*, 1963, 9(3): 458-467.

[41] Suriowiecki J. The Wisdom of Crowds: Why the many are Smarter than the few and how Collective Wisdom Shapes Business, Economies, Societies and Nations[J]. *Choice Reviews Online*, 2004, 42(3): 42-1645.

[42] Rowe G, Wright G. The Delphi Technique as a Forecasting Tool: Issues and Analysis[J]. *International Journal of Forecasting*, 1999, 15(4): 353-375.

[43] Dajani J S, Sincoff M Z, Talley W K. Stability and Agreement Criteria for the Termination of Delphi Studies[J]. *Technological Forecasting and Social Change*, 1979, 13(1): 83-90.

[44] English J M, Kernan G L. The Prediction of Air Travel and Aircraft Technology to the Year 2000 Using the Delphi Method[J]. *Transportation Research*, 1976, 10(1): 1-8.

[45] Grisaffe D B. Questions about the ultimate question: conceptual considerations in evaluating Reichheld's net promoter score (NPS)[J]. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 2007, 20: 36.

[46] Joshi A, Kale S, Chandel S, et al. Likert Scale: Explored and Explained[J]. *British Journal of Applied Science & Technology*, 2015, 7(4): 396-403.

[47] Kapoor, Peter. Systems approach to documentary maritime fraud [Ph.D. dissertation], University of plymouth, 1987.

[48] Saaty T L. The analytic hierarchy process (AHP)[J]. *The Journal of the Operational Research Society*, 1980, 41(11): 1073-1076.

[49] Triantaphyllou E. Multi-Criteria Decision Making Methods[M]. Applied Optimization. Boston, MA: Springer US, 2000: 5-21.

[50] Shiraishi S, Obata T, Daigo M. Properties of a Positive Reciprocal Matrix and Their Application to Ahp[J]. *Journal of the Operations Research Society of Japan*, 1998, 41(3): 404-414.

[51] Carter T. An introduction to information theory and entropy[J]. *Complex systems summer school*, Santa Fe, 2007.

[52] Delgado A, Romero I. Environmental Conflict Analysis Using an Integrated Grey Clustering and Entropy-Weight Method: A Case Study of a Mining Project in Peru[J]. *Environmental Modelling & Software*, 2016, 77: 108-121.

[53] Shan C J, Dong Z C, FAN K, et al. Application of combination weighting method to weight calculation in river health evaluation [J]. *Journal of Hohai University (Natural Sciences)*, 2012, 40(6): 622-628.

[54] Chen J F, Hsieh H N, Do Q H. Evaluating Teaching Performance Based on Fuzzy AHP and Comprehensive Evaluation Approach[J]. *Applied Soft Computing*, 2015, 28: 100-108.

[55] Bai S M, Chen S M. Evaluating Students' Learning Achievement Using Fuzzy Membership Functions and Fuzzy Rules[J]. *Expert Systems With Applications*, 2008, 34(1): 399-410.

[56] Chameau J L, Santamarina J C. Membership Functions I: Comparing Methods of Measurement[J]. *International Journal of Approximate Reasoning*, 1987, 1(3): 287-301.

[57] 智联招聘和奇安信, 网络安全人才市场状况研究报告, 2019.

[58] MITRE C. Adversary Tactic Technique & Common Knowledge (ATT&CK), 2020, URL: https://attack.mitre.org/.

[59] MITRE C. Common Attack Pattern Enumeration and Classification (CAPEC). 2014. URL: http://capec.mitre.org.

[60] Bahrami P N, Dehghantanha A, Dargahi T, et al. Cyber Kill Chain-Based Taxonomy of Advanced Persistent Threat Actors: Analogy of Tactics, Techniques, and Procedures[J]. *Journal of Information Processing Systems*, 2019, 15(4): 865-889.

[61] Pols P, van den Berg J. The Unified Kill Chain[J]. *CSA Thesis, Hague*, 2017: 1-104.

[62] Nickerson C, Kennedy D, Smith E, et al. Penetration testing execution standard[J]. 2014. URL: http://www.pentest-standard.org/index.php/Main_Page.

[63] Meucci M, Muller A. Owasp testing guide v4[J]. *OWASP Foundation*, 2008.

[64] Auger R, Barnett R. Web Application Security Consortium: Threat Classification Version 1.0[J]. *Web Application Security Consortium* (www. webappsec. org), 2004.

[65] Herzog P. OSSTMM 3-The Open-Source Security Testing Methodology Manual: Contemporary Security Testing and Analysis[J]. *ISECOM-Institute for Security and Open Methodologies*, 2010.

[66] Rathore B, Brunner M, Dilaj M, et al. Information systems security assessment framework (ISSAF) draft 0.2. 1[J]. 2009.

[67] Scarfone K, Souppaya M, Cody A, et al. Technical guide to information security testing and assessment[J]. *NIST Special Publication*, 2008, 800(115): 2-25.

[68] Wu P, Sismanis Y, Reinwald B. Towards Keyword-Driven Analytical Processing[C]. *The 2007 ACM SIGMOD international conference on Management of data*, 2007: 617-628.

[69] Zhang Y, Xu F F, Li S, et al. HiGitClass: keyword-driven hierarchical classification of GitHub repositories[C]. *2019 IEEE International Conference on Data Mining*, 2020: 876-885.

[70] Bodeau D J, McCollum C D, Fox D B. Cyber threat modeling: survey, assessment, and representative framework[R]. *MITRE CORP MCLEAN VA MCLEAN*, 2018.

[71] Song J, Alves-Foss J. The DARPA Cyber Grand Challenge: A Competitor's Perspective[J]. *IEEE Security & Privacy*, 2015, 13(6): 72-76.

[72] Song J, Alves-Foss J. The DARPA Cyber Grand Challenge: A Competitor's Perspective, Part 2[J]. *IEEE Security & Privacy*, 2016, 14(1): 76-81.

[73] CÔTÉ J E. Sociological Perspectives on Identity Formation: The Culture-Identity Link and Identity Capital[J]. *Journal of Adolescence*, 1996, 19(5): 417-428.

[74] Hoffman R R. How can Expertise be Defined? Implications of Research from Cognitive Psychology[M]. Exploring Expertise. London: Palgrave Macmillan UK, 1998: 81-100.

[75] Ericsson K A. The Cambridge handbook of expertise and expert performance[M]. 2nd ed.

[76] Johnson D. The Triangular Distribution as a Proxy for the Beta Distribution in Risk Analysis[J]. *Journal of the Royal Statistical Society*: *Series D* (*the Statistician*), 1997, 46(3): 387-398.

[77] Weisstein E W. Triangular distribution, from mathworld–a wolfram web resource[J]. URL: https://mathworld.wolfram.com/TriangularDistribution.html.

[78] Price P C, Jhangiani R S, Chiang I C A. Reliability and validity of measurement[J]. *Research methods in psychology*, 2015.

[79] Eisinga R, Grotenhuis M T, Pelzer B. The Reliability of a Two-Item Scale: Pearson, Cronbach, or Spearman-Brown? [J]. *International Journal of Public Health*, 2013, 58(4): 637-642.

[80] Peterson R A. Rating scales[J]. *Constructing effective questionnaires*, 2000: 61-81.

[81] Workthrough of Cengbox, Gamebox in Vulnhub, URL: https://www.hackingarticles.in/cengbox-1-vulnhub-walkthrough/.

[82] Khandkar S H. Open coding[J]. *University of Calgary*, 2009, 23: 2009.

[83] Childers N, Boe B, Cavallaro L, et al. Organizing Large Scale Hacking Competitions[M]. Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010: 132-152.

[84] NetKotH (Network King of the Hill), URL: https://netkoth.org/

[85] Bock K, Hughey G, Levin D. King of the hill: A novel cybersecurity competition for teaching penetration testing[C]. *2018 {USENIX} Workshop on Advances in Security Education*. 2018.

[86] Zhang X, Liu B X, Gong X R, et al. State-of-the-Art: Security Competition in Talent Education[M]. Information Security and Cryptology. Cham: Springer International Publishing, 2018: 461-481.

[87] Zhang X, Liu B X, Gong X R, et al. Explore-Exploit: A Security Competition Modeling the Real-World Network Penetration Scenario[J]. *Journal of Cyber Security*, 2020, 5(4): 55-71.
(章秀, 刘宝旭, 龚晓锐, 等. Explore-Exploit: 一种模拟真实网络渗透场景的安全竞赛[J]. *信息安全学报*, 2020, 5(4): 55-71.)

[88] Votipka D, Rabin S M, Micinski K, et al. An Observational Investigation of Reverse Engineers' Processes[EB/OL]. 2019: arXiv: 1912.00317. https://arxiv.org/abs/1912.00317.

[89] Votipka D, Fulton K R, Parker J, et al. Understanding Security Mistakes Developers Make: Qualitative Analysis from Build It, Break It, Fix it[C]. *The 29th USENIX Conference on Security Symposium*, 2020: 109-126.

[90] Parker J, Hicks M, Ruef A, et al. Build It, Break It, Fix It: Contesting Secure Development[J]. *ACM Transactions on Privacy and Security*, 2020, 23(2): 10.

[91] Votipka D, Stevens R, Redmiles E, et al. Hackers *vs*. testers: A comparison of software vulnerability discovery processes[C]. *2018 IEEE Symposium on Security and Privacy*, 2018: 374-391.

[92] Redmiles E M, Malone A R, Mazurek M L, et al. I think they're trying to tell me something: Advice sources and selection for digital security[C]. *2016 IEEE Symposium on Security and Privacy*, 2016: 272-288.

[93] Emami-Naeini P, Agarwal Y, Cranor L F, et al. Ask the Experts: What should be on an IoT Privacy and Security Label? [EB/OL]. 2020: arXiv: 2002.04631. https://arxiv.org/abs/2002.0463.

章秀 于 2013 年在华中科技大学获得学士学位。现在中国科学院信息工程研究所攻读博士学位。研究领域为人才能力评估。研究兴趣包括安全教育、攻防场景等。Email: zhangxiu@iie.ac.cn

刘宝旭 于 2002 年在中国科学院研究生院获得博士学位。现任中国科学院信息工程研究所研究员。第六研究室主任。研究领域为网络安全攻防对抗、网络安全评测技术等。Email: liubaoxu@iie.ac.cn

龚晓锐 于 2014 年在北京大学获得硕士学位。现任中国科学院信息工程研究所正高级工程师。研究领域为攻防对抗。研究兴趣包括移动互联网安全、网络安全攻防对抗等。Email: gongxiaorui@iie.ac.cn

于冬松 于 2015 年在中国科学技术大学获得学士学位。现在中国科学院信息工程研究所攻读博士学位。研究领域为安卓系统安全。研究兴趣包括内核漏洞利用、系统访问控制等。Email: yudongsong@iie.ac

赵蓓蓓 于 2020 年在中国矿业大学获得学士学位。现在中国科学院信息工程研究所攻读博士学位。Email: zhaobeibei@iie.ac.cn

刘媛 于 2015 年在清华大学获得软件工程硕士学位。现任中国科学院信息工程研究所助理研究员。研究领域为攻防对抗。研究兴趣包括深度学习在二进制逆向方面的应用、软件测评等。Email: liuyuan@iie.ac.

附录 1: 准则定义的两轮评议数据

表 11 准则定义的第一轮专家评议数据分析

Table 11 Data analysis of the definition of criteria in the first Delphi round

准则项	1 完全不认可	2 基本不认可	3 中立	4 基本认可	5 完全认可	众数	均值	标准差	*C.V*	*APMO*
SC_{11}: 开源情报收集	0	0	2	7	15	5	4.54	0.64	0.14	92%
SC_{12}: 网络入口探测	0	1	1	8	14	5	4.46	0.76	0.17	88%
SC_{13}: 威胁建模	0	0	3	11	10	4	4.29	0.68	0.16	88%
SC_{21}: Web 应用攻击	0	1	2	10	11	5	4.29	0.79	0.18	83%
SC_{22}: 程序逆向分析	0	0	3	7	14	5	4.46	0.71	0.16	88%
SC_{23}: 漏洞挖掘利用	0	2	2	8	12	5	4.25	0.92	0.22	75%
SC_{24}: 密文解码破译	0	3	2	8	11	5	4.13	1.01	0.24	67%
SC_{31}: 辅助攻击平台构建	0	0	5	9	10	5	4.21	0.76	0.18	79%
SC_{32}: 初始据点获取	0	0	3	10	11	5	4.33	0.69	0.16	85%
SC_{33}: 横向移动和域渗透	0	2	1	6	15	5	4.42	0.91	0.21	79%
SC_{34}: 网络服务弱点利用	0	1	1	10	12	5	4.38	0.75	0.17	88%
SC_{35}: 操作系统权限提升	0	1	4	8	11	5	4.21	0.87	0.21	75%
SC_{36}: 应用软件产品破解	0	1	4	6	13	5	4.29	0.89	0.21	75%
SC_{41}: 防护措施规避	0	0	3	6	15	5	4.5	0.71	0.16	88%
SC_{42}: 技术反制	1	2	4	7	10	5	3.96	1.14	0.29	58%
SC_{51}: 数据处理和取证	0	2	3	4	15	5	4.33	0.97	0.22	71%
SC_{52}: 攻击事件分析	0	0	2	9	13	5	4.46	0.64	0.14	92%
有效样本数:				24						

表 12 准则定义的第二轮专家评议数据分析

Table 12 Data analysis of the definition of criteria in the second Delphi round

准则项	1 完全不认可	2 基本不认可	3 中立	4 基本认可	5 完全认可	众数	均值	标准差	*C.V*	*APMO*
SC_{11}: 开源情报收集	0	1	5	12	28	5	4.46	0.77	0.17	85%
SC_{12}: 网络入口探测	0	0	4	11	31	5	4.59	0.65	0.14	91%
SC_{13}: 威胁建模	1	1	6	15	23	5	4.26	0.92	0.22	78%
SC_{21}: Web 应用攻击	0	2	6	16	22	5	4.26	0.85	0.20	78%
SC_{22}: 程序逆向分析	0	1	4	16	25	5	4.41	0.74	0.17	87%
SC_{23}: 漏洞挖掘利用	1	1	7	13	24	5	4.26	0.94	0.22	76%
SC_{24}: 密文解码破译	0	1	5	12	28	5	4.46	0.77	0.17	85%
SC_{31}: 辅助攻击平台构建	0	0	7	11	28	5	4.46	0.74	0.16	85%
SC_{32}: 初始据点获取	0	1	7	11	27	5	4.39	0.82	0.19	80%
SC_{33}: 横向移动和域渗透	0	1	3	15	27	5	4.48	0.71	0.16	89%
SC_{34}: 网络服务弱点利用	1	2	4	11	28	5	4.37	0.96	0.22	78%

续表

准则项	1 完全不认可	2 基本不认可	3 中立	4 基本认可	5 完全认可	众数	均值	标准差	$C.V$	$APMO$
SC_{35}: 操作系统权限提升	0	2	4	14	26	5	4.39	0.82	0.19	83%
SC_{36}: 应用软件产品破解	0	1	5	10	30	5	4.50	0.77	0.17	85%
SC_{37}: 线下环境物理渗透	0	3	5	12	26	5	4.33	0.91	0.21	76%
SC_{41}: 防护措施规避	0	2	4	13	27	5	4.41	0.82	0.19	83%
SC_{42}: 技术反制	0	2	5	13	26	5	4.37	0.84	0.19	80%
SC_{51}: 数据处理和取证	0	1	5	14	26	5	4.41	0.77	0.17	85%
SC_{52}: 攻击事件分析	0	0	7	10	29	5	4.48	0.74	0.17	85%
有效样本数				46						

附录 2: 层次分析法中成对比较矩阵数据

表 13　侦查探测对应子准则的成对比较矩阵及一致性检验

Table 13　Pairwise comparison matrix of SC_{1j} and consistency check

	SC_{11}	SC_{12}	SC_{13}	n	λ_{max}	CI	RI	CR	一致性检验	λ_{max} 对应特征向量 $\boldsymbol{U}$	权向量(SW_{1j})
SC_{11}	1	1/5	1/3							–0.1640	**0.1095**
SC_{12}	5	1	2	3	3.00	0	0.52	0	通过	–0.8711	**0.5816**
SC_{13}	3	1/2	1							–0.4629	**0.3090**

表 14　攻击准备对应子准则的成对比较矩阵及一致性检验

Table 14　Pairwise comparison matrix of SC_{2j} and consistency check

	SC_{21}	SC_{22}	SC_{23}	SC_{24}	n	λ_{max}	CI	RI	CR	一致性检验	λ_{max} 对应特征向量 $\boldsymbol{U}$	权向量(SW_{2j})
SC_{21}	1	2	2	3							0.7423	**0.4117**
SC_{22}	1/2	1	1/2	2							0.3246	**0.1800**
SC_{23}	1/2	2	1	4	4	4.10	0.03	0.89	0.04	通过	0.5866	**0.3098**
SC_{24}	1/3	1/2	1/4	1							0.1774	**0.0984**

表 15　渗透控制对应子准则的成对比较矩阵及一致性检验

Table 15　Pairwise comparison matrix of SC_{3j} and consistency check

	SC_{31}	SC_{32}	SC_{33}	SC_{34}	SC_{35}	SC_{36}	SC_{37}	n	λ_{max}	CI	RI	CR	一致性检验	λ_{max} 对应特征向量 $\boldsymbol{U}$	权向量(SW_{3j})
SC_{31}	1	1/3	1/3	2	1/2	3	4							–0.2688	**0.1191**
SC_{32}	3	1	2	4	2	3	5							–0.6624	**0.2935**
SC_{33}	3	1/2	1	3	2	5	3							–0.5284	**0.2342**
SC_{34}	1/2	1/4	1/3	1	1/3	3	3	7	7.41	0.07	1.36	0.05	通过	–0.1956	**0.0867**
SC_{35}	2	1/2	1/2	3	1	3	3							–0.3840	**0.1702**
SC_{36}	1/3	1/3	1/5	1/3	1/3	1	2							–0.1240	**0.0549**
SC_{37}	1/4	1/5	1/3	1/3	1/4	1/2	1							–0.0933	**0.0414**

表 16　对抗博弈对应子准则的成对比较矩阵及一致性检验

Table 16　Pairwise comparison matrix of SC_{4j} and consistency check

	SC_{41}	SC_{42}	n	λ_{max}	CI	RI	CR	一致性检验	λ_{max} 对应特征向量 $\boldsymbol{U}$	权向量(SW_{4j})
SC_{41}	1	3	2	2.00	0	N/A	N/A	N/A	0.9487	**0.75**
SC_{42}	1/3	1							0.3162	**0.25**

表 17　取证分析对应子准则的成对比较矩及一致性检验

Table 17　Pairwise comparison matrix of SC_{5j} and consistency check

	SC_{51}	SC_{52}	n	λ_{max}	CI	RI	CR	一致性检验	λ_{max} 对应特征向量 $\boldsymbol{U}$	权向量(SW_{5j})
SC_{51}	1	1/2	2	2.00	0	N/A	N/A	N/A	0.4472	**0.3333**
SC_{52}	2	1							0.8944	**0.6667**

附录 3: 准则量化取值评议数据

表 18　准则量化取值的第一轮专家评议数据分析

Table 18　Statistics of the numerical value of criteria in the first Delphi round

自评等级	量化取值	1 完全不认可	2 基本不认可	3 中立	4 基本认可	5 完全认可	众数	均值	标准差	$C.V$	$APMO$
L_1	1	0	2	2	10	10	4,5	4.17	0.90	0.22	**75%**
L_2	2	0	2	5	11	6	4	3.88	0.88	0.23	**63%**
L_3	3	1	1	6	8	8	4,5	3.88	1.05	0.27	**58%**
L_4	4	1	1	9	3	10	5	3.83	1.14	0.29	**46%**
L_5	5	1	1	9	2	11	5	3.88	1.17	0.30	**46%**
有效样本数					**24**						

表 19　准则量化取值第二轮专家评议数据分析

Table 19　Statistics of the numerical value of criteria in the second Delphi round

自评等级	量化取值	1 完全不认可	2 基本不认可	3 中立	4 基本认可	5 完全认可	众数	均值	标准差	$C.V$	$APMO$
L_1	1	0	0	1	8	28	5	4.73	0.50	0.11	**97%**
L_2	2	0	2	4	6	25	5	4.46	0.89	0.20	**78%**
L_3	3	1	0	2	9	25	4	4.54	0.83	0.18	**89%**
L_4	4	1	1	3	10	22	5	4.38	0.94	0.21	**81%**
L_5	5	1	0	3	10	23	5	4.46	0.86	0.19	**87%**
有效样本数					**37**						

附录 4: 专家在每一个评价准则上的自评数据

表 20　专家在每一个评价准则上的自评数据

Table 20　Data of self-assessment in line with each criterion of SMEs

	SC_{11}	SC_{12}	SC_{13}	SC_{21}	SC_{22}	SC_{23}	SC_{24}	SC_{31}	SC_{32}	SC_{33}	SC_{34}	SC_{35}	SC_{36}	SC_{37}	SC_{41}	SC_{42}	SC_{51}	SC_{52}
SME_1	4	3	3	2	1	1	1	3	3	1	1	1	1	1	1	1	4	4
SME_2	2	3	1	2	3	2	4	2	3	2	3	3	3	2	3	3	2	1
SME_3	3	2	2	2	3	4	2	3	2	3	2	4	3	2	3	2	2	3
SME_4	4	5	3	4	1	1	1	4	3	3	4	3	1	2	3	1	2	3
SME_5	4	4	3	4	3	3	3	3	4	4	4	3	3	2	4	4	4	5
SME_6	2	3	2	4	1	1	2	3	3	2	3	2	1	1	1	1	2	2
SME_7	2	2	2	1	5	5	1	2	1	1	1	3	4	1	1	2	3	4
SME_8	2	1	1	1	3	2	1	4	1	1	1	1	1	1	1	1	2	1
SME_9	4	3	2	2	1	1	1	2	2	2	2	1	2	1	1	1	4	3
SME_{10}	2	2	3	1	4	1	4	1	1	2	1	2	4	1	3	1	3	5
SME_{11}	3	3	2	2	4	4	4	3	2	2	3	2	4	2	2	2	4	3
SME_{12}	3	2	1	2	5	3	2	3	2	1	2	2	2	1	2	2	4	4

续表

	SC_{11}	SC_{12}	SC_{13}	SC_{21}	SC_{22}	SC_{23}	SC_{24}	SC_{31}	SC_{32}	SC_{33}	SC_{34}	SC_{35}	SC_{36}	SC_{37}	SC_{41}	SC_{42}	SC_{51}	SC_{52}
SME_{13}	1	1	1	1	3	3	1	2	1	1	2	3	2	1	3	1	1	2
SME_{14}	4	3	3	1	2	1	1	1	1	1	1	1	1	1	1	1	4	4
SME_{15}	3	3	3	2	5	5	5	3	3	3	3	3	5	3	5	5	5	5
SME_{16}	3	4	4	5	3	5	2	4	4	4	5	4	4	4	4	4	3	4
SME_{17}	4	4	3	4	2	2	2	5	3	2	3	2	2	2	4	2	2	4
SME_{18}	3	3	4	2	4	5	2	3	2	2	3	4	4	1	3	3	3	4
SME_{19}	2	2	1	1	3	3	1	3	2	2	2	3	3	1	1	1	1	1
SME_{20}	3	1	1	1	3	3	1	1	1	1	1	1	3	1	1	1	1	2
SME_{21}	2	3	3	3	2	2	2	3	2	2	3	3	2	1	3	3	3	3
SME_{22}	3	3	3	2	4	4	3	4	3	3	4	3	4	4	3	3	2	2
SME_{23}	3	2	4	1	1	4	1	3	1	1	1	1	1	1	1	1	5	4
SME_{24}	2	2	1	3	4	2	3	1	2	1	2	1	2	1	1	1	3	5
SME_{25}	4	3	3	2	1	1	1	4	1	1	1	1	1	1	1	1	4	4
SME_{26}	3	2	2	2	1	1	1	2	1	1	2	1	1	1	2	2	3	2
SME_{27}	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	2
SME_{28}	5	5	3	5	4	3	2	1	3	4	5	5	3	1	5	3	5	5
SME_{29}	4	4	5	4	3	3	3	5	4	3	3	3	3	4	3	3	4	4
SME_{30}	4	3	3	3	4	5	2	3	3	3	3	3	3	3	3	3	3	3
SME_{31}	4	4	4	4	3	4	4	4	4	3	4	4	3	3	3	4	4	4
SME_{32}	4	5	5	3	4	4	2	4	3	3	4	3	4	2	4	3	4	5

附录 5: 综合评级的评议数据

表 21　综合评级的第一轮专家评议数据分析

Table 21　Data analysis of V_{rating} in the first Delphi round

综合评级	1 完全不认可	2 基本不认可	3 中立	4 基本认可	5 完全认可	众数	均值	标准差	$C.V$	$APMO$
新手	0	2	4	9	9	4,5	4.04	0.93	0.23	**67%**
脚本小子	1	0	6	9	8	4	3.96	0.98	0.25	**67%**
安全实践者	0	1	2	9	12	5	4.33	0.80	0.18	**83%**
安全研究员	1	0	4	6	13	5	4.25	1.01	0.24	**75%**
黑客	1	2	3	1	17	5	4.29	1.21	0.28	**63%**
有效样本数	**24**									

表 22　综合评级的第二轮专家评议数据分析

Table 22　Data analysis of V_{rating} in the second Delphi round

综合评级	1 完全不认可	2 基本不认可	3 中立	4 基本认可	5 完全认可	众数	均值	标准差	$C.V$	$APMO$
新手	1	1	4	10	21	5	4.32	0.96	0.22	**78%**
学徒	0	0	3	14	20	5	4.46	0.64	0.14	**92%**
老手	0	0	1	15	21	5	4.54	0.55	0.12	**97%**
专家	0	0	6	13	18	5	4.32	0.74	0.17	**84%**
大师	1	3	2	12	19	5	4.22	1.04	0.25	**73%**
有效样本数	**37**									

表 23　综合评级的第三轮专家评议数据分析

Table 23　Data analysis of V_{rating} in the third Delphi round

综合评级	1 完全不认可	2 基本不认可	3 中立	4 基本认可	5 完全认可	众数	均值	标准差	*C.V*	*APMO*
新手	0	1	3	7	34	5	4.64	0.70	0.15	**84%**
学徒	0	0	2	17	26	5	4.53	0.58	0.13	**96%**
高手	0	0	3	10	32	5	4.64	0.60	0.13	**93%**
专家	0	1	5	12	27	5	4.44	0.78	0.18	**84%**
大师	1	1	3	10	30	5	4.49	0.88	0.20	**84%**
有效样本数				**45**						

附录 6: 基于专家投票获得的评估对象的隶属度矩阵和综合评价结果数据

表 24　评估对象(A_7)的隶属度矩阵和综合评价结果

Table 24　Membership matrix and comprehensive evaluation result of the alternative (A_7)

m0nshaw	新手	学徒	高手	专家	大师
开源情报搜集	0	**0.67**	0.33	0	0
网络入口探测	0.11	**0.78**	0.11	0	0
威胁建模	0.22	0.33	**0.45**	0	0
Web 应用攻击	0.2	**0.8**	0	0	0
程序逆向分析	0	0	0.27	**0.64**	0.09
漏洞挖掘利用	0	0	0.27	**0.55**	0.18
密文解码破译	0.09	0.36	**0.55**	0	0
辅助攻击平台构建	0.1	0.3	**0.4**	0.2	0
初始据点获取	0.11	**0.67**	0.22	0	0
横向移动和域渗透	0.11	**0.78**	0	0.11	0
网络服务弱点利用	0	**0.56**	0.33	0.11	0
操作系统权限提升	0	0.18	**0.64**	0.09	0.09
应用软件产品破解	0	0.09	0.27	**0.55**	0.09
线下环境物理渗透	0.11	**0.45**	0.33	0.11	0
防护措施规避	0.1	0.1	**0.8**	0	0
技术反制	0.1	0.2	**0.6**	0.1	0
数据处理和取证	0.1	0.1	**0.8**	0	0
攻击事件分析	0.1	0.2	**0.6**	0.1	0
有效样本数			9		
综合隶属度	0.09	**0.41**	0.34	0.13	0.03
评级			学徒		
得分			51.94		

表 25 评估对象(A_3)的隶属度矩阵和综合评价结果

Table 25 Membership matrix and comprehensive evaluation result of the alternative (A_3)

Ender	新手	学徒	高手	专家	大师
开源情报搜集	0.05	**0.65**	0.1	0.1	0.1
网络入口探测	0.11	**0.79**	0.05	0.05	0
威胁建模	0.16	0.26	**0.47**	0.11	0
Web 应用攻击	0.21	**0.74**	0.05	0	0
程序逆向分析	0	0.05	0.09	0.27	**0.59**
漏洞挖掘利用	0	0.09	0.45	**0.46**	0
密文解码破译	0.05	0.14	**0.48**	0.24	0.09
辅助攻击平台构建	0.05	0.35	**0.55**	0.05	0
初始据点获取	0.16	**0.68**	0.16	0	0
横向移动和域渗透	0.32	**0.58**	0.1	0	0
网络服务弱点利用	0.11	**0.68**	0.21	0	0
操作系统权限提升	0	0.05	0.11	**0.84**	0
应用软件产品破解	0	0.05	0.09	**0.73**	0.14
线下环境物理渗透	0.14	0.19	**0.52**	0.15	0
防护措施规避	0.05	0.3	**0.5**	0.1	0.05
技术反制	0.05	0.2	**0.6**	0.1	0.05
数据处理和取证	0.05	0.25	**0.45**	0.2	0.05
攻击事件分析	0	0.15	0.35	**0.5**	0
有效样本数			22		
综合隶属度	0.09	**0.37**	0.29	0.21	0.04
评级			学徒		
得分			55.04		

表 26 评估对象(A_2)的隶属度矩阵和综合评价结果

Table 26 Membership matrix and comprehensive evaluation result of the alternative (A_2)

rebirth	新手	学徒	高手	专家	大师
开源情报搜集	0	0	0.33	0.33	**0.34**
网络入口探测	0	0	0.14	**0.43**	**0.43**
威胁建模	0.05	0.05	**0.4**	0.35	0.15
Web 应用攻击	0	0	0.1	0.24	0.6
程序逆向分析	0.2	**0.6**	0.2	0	0
漏洞挖掘利用	0.19	0.14	**0.33**	0.24	0.1
密文解码破译	0.15	0.3	**0.45**	0.1	0
辅助攻击平台构建	0	0.15	**0.4**	**0.4**	0.05
初始据点获取	0	0.05	0.25	**0.4**	0.3
横向移动和域渗透	0	0	0.32	**0.47**	0.21
网络服务弱点利用	0	0	0.26	0.32	**0.42**
操作系统权限提升	0.21	**0.42**	0.16	0.16	0.05
应用软件产品破解	**0.37**	**0.37**	0.16	0.1	0
线下环境物理渗透	0.21	0.16	**0.26**	**0.26**	0.11
防护措施规避	0.05	0.21	**0.48**	0.26	0
技术反制	0.05	0.26	**0.42**	0.27	0
数据处理和取证	0.11	0.31	**0.32**	0.26	0
攻击事件分析	0.1	0.16	**0.37**	0.32	0.05
有效样本数			21		
综合隶属度	0.09	0.16	0.28	**0.29**	0.18
评级			专家		
得分			66.05		

表 27 评估对象(A_6)的隶属度矩阵和综合评价结果

Table 27 Membership matrix and comprehensive evaluation result of the alternative (A_6)

Mote	新手	学徒	高手	专家	大师
开源情报搜集	0	0.33	0.33	**0.34**	0
网络入口探测	0	0.07	**0.57**	0.36	0
威胁建模	0.07	**0.43**	0.21	0.29	0
Web 应用攻击	0	0	**0.33**	0.6	0.07
程序逆向分析	0.4	**0.47**	0.13	0	0
漏洞挖掘利用	0.33	**0.53**	0.14	0	0
密文解码破译	0.27	**0.53**	0.2	0	0
辅助攻击平台构建	0	0.36	**0.5**	0.14	0
初始据点获取	0.07	0.21	**0.5**	0.22	0
横向移动和域渗透	0	0.29	**0.43**	0.28	0
网络服务弱点利用	0	0.21	0.22	**0.57**	0
操作系统权限提升	0.29	**0.43**	0.21	0.07	0
应用软件产品破解	**0.54**	0.38	0.08	0	0
线下环境物理渗透	0.23	**0.38**	0.23	0.08	0.08
防护措施规避	0.14	0.36	**0.43**	0.07	0
技术反制	0.07	0.43	**0.5**	0	0
数据处理和取证	0.14	**0.43**	0.36	0.07	0
攻击事件分析	0.07	0.36	**0.43**	0.14	0
有效样本数			15		
综合隶属度	0.18	**0.34**	0.33	0.14	0.01
评级			学徒		
得分			49.22		

表 28 评估对象(A_4)的隶属度矩阵和综合评价结果

Table 28 Membership matrix and comprehensive evaluation result of the alternative (A_4)

d3adf1sh	新手	学徒	高手	专家	大师
开源情报搜集	0.07	**0.36**	0.28	0.14	0.14
网络入口探测	0.21	**0.58**	0.14	0	0.07
威胁建模	0.14	**0.36**	0.29	0.21	0
Web 应用攻击	0.29	**0.71**	0	0	0
程序逆向分析	0	0	0.12	**0.47**	0.41
漏洞挖掘利用	0	0	0.23	**0.59**	0.18
密文解码破译	0.12	0.38	**0.44**	0.06	0
辅助攻击平台构建	0.2	0.07	**0.6**	0.13	0
初始据点获取	0.13	**0.47**	0.33	0.07	0
横向移动和域渗透	0.27	**0.53**	0.13	0.07	0
网络服务弱点利用	0.06	**0.5**	0.25	0.13	0.06
操作系统权限提升	0.06	0.19	**0.38**	0.37	0
应用软件产品破解	0	0	0.2	**0.6**	0.2
线下环境物理渗透	0.07	0.2	**0.4**	0.27	0.06
防护措施规避	0	0.2	**0.53**	0.27	0
技术反制	0	0.2	**0.67**	0.13	0
数据处理和取证	0.07	0.2	**0.6**	0.07	0.06
攻击事件分析	0.07	0.27	**0.47**	0.13	0.06
有效样本数			17		
综合隶属度	0.12	**0.32**	0.31	0.20	0.05
评级			学徒		
得分			54.99		

表 29 评估对象(A_5)的隶属度矩阵和综合评价结果

Table 29 Membership matrix and comprehensive evaluation result of the alternative (A_5)

giglf	新手	学徒	高手	专家	大师
开源情报搜集	0.15	**0.31**	**0.31**	0.08	0.15
网络入口探测	0.15	**0.69**	0.08	0	0.08
威胁建模	0.15	**0.47**	0.23	0.15	0
Web 应用攻击	0.31	**0.69**	0	0	0
程序逆向分析	0	0	0.18	**0.63**	0.19
漏洞挖掘利用	0	0.19	0.31	**0.5**	0
密文解码破译	0.13	0.27	**0.53**	0.07	0
辅助攻击平台构建	0.07	0.21	**0.5**	0.22	0
初始据点获取	0.29	**0.43**	0.14	0.14	0
横向移动和域渗透	0.36	**0.5**	0	0.14	0
网络服务弱点利用	0.14	**0.43**	0.21	0.22	0
操作系统权限提升	0.07	0.2	**0.47**	0.2	0.06
应用软件产品破解	0	0	0.36	**0.5**	0.14
线下环境物理渗透	0.14	**0.43**	0.29	0.14	0
防护措施规避	0	**0.5**	0.29	0.21	0
技术反制	0	0.36	**0.43**	0.21	0
数据处理和取证	0.07	0.21	**0.43**	0.14	0.15
攻击事件分析	0.08	0.21	**0.5**	0.21	0
有效样本数			16		
综合隶属度	0.14	**0.37**	0.26	0.20	0.03
评级			学徒		
得分			51.87		

表 30 评估对象的技能倾向

Table 30 Skill aptitude of all alternatives

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
SC_{11}	3	5	2	2	2.5	3	2
SC_{12}	3	4.5	2	2	2	3	2
SC_{13}	3	3	3	2	2	3	3
SC_{21}	3	5	2	2	2	4	2
SC_{22}	4	2	5	4	4	2	4
SC_{23}	5	3	4	4	4	2	4
SC_{24}	2	3	3	3	3	1	3
SC_{31}	3	3.5	3	3	3	3	3
SC_{32}	3	4	2	2	2	4	2
SC_{33}	3	4	2	2	2	3	2
SC_{34}	3	5	2	2	2	4	2
SC_{35}	5	2	4	3	3	2	3
SC_{36}	5	1.5	4	4	4	1	4
SC_{37}	3	3.5	3	3	2	3	2
SC_{41}	4	3	3	3	2	3	3
SC_{42}	3	3	3	3	3	3	3
SC_{51}	3	3	3	3	3	3	3
SC_{52}	3	3	4	3	3	3	3

附录 7: 评估任务标注示例

表 31 评估任务标注示例

Table 31 Example of task labeled

	新手	学徒	高手	专家	大师	测量项	工具/命令	环境/组件
开源情报搜集	1							
网络入口探测		2				M_1: 活跃主机发现 M_2: 端口和服务识别	netdiscover nmap	
威胁建模	1							
Web 应用攻击	1	2				M_3: 爆破管理页面 M_4: SQli 注入攻击获取口令(web 登录口令和 ubuntu 普通用户口令) M_5: 任意文件上传攻击	dirb sqlmap	apache
程序逆向分析	1							
漏洞挖掘利用	1							
密文解码破译	1							
辅助攻击平台构建	1							
初始据点获取		2				M_7: 上传 php-reverse-shell M_8: 提升至交互式 shell M_9: 利用口令从 www-data 提升至 ubuntu 用户	nc	
横向移动和域渗透	1							
网络服务弱点利用	1							
操作系统权限提升			3			M_{10}: 植入文件权限检查和隐藏进程枚举工具 pspy64s M_{11}: 枚举主机进程权限和隐藏进程, 发现 UID=0 的可读可写权限的运行脚本 md5check.py M_{12}: 使用 msfconsole 的 web_delivery 模块生成 payload M_{13}: 将 payload 写入 md5check.py M_{14}: 获得 root 权限的 meterpreter session	wget metasploit	ubuntu
应用软件产品破解	1							
线下环境物理渗透	1							
防护措施规避	1					M_6: 修改文件扩展名绕过文件上传限制		
技术反制	1							
数据处理和取证	1							
攻击事件分析	1							
任务形式	☐文件 ☐容器 ☑虚拟机 ☐物理设备 ☐复杂网络场景							
评估层次	☐元技能评估 ☐应用技能评估 ☑单靶机控守 ☐攻击链构造							
评估实验形式	☐夺旗赛 ☐网络寻宝赛 ☑网络占山为王 ☐网络探索和漏洞利用							
综合评级	☐新手 ☑学徒 ☐高手 ☐专家 ☐大师							
有效代码行数	☑小于 100 行 ☐100-1000 行 ☐大于 1000 行							
关键点	M_{11}: 枚举主机进程权限和隐藏进程, 发现 UID=0 的可读可写脚本 md5check.py							
前置任务	无							
后置任务	无							

附录 8: 通过直接分层比重划分得到的主观权重数据

说明: 通过专家直接分配比重得到的主观权重非常接近平均化参考值, 平均化参考值和每一级准则数目成负相关, 因此此种方法求得的主观权重不可信, 从而间接佐证了层次分析法的有效性。

表 32　通过直接分层划分比重得到的主观权重

Table 32　Multi-hierarchy *SW* calculating by SMEs who tend to divide the weight equally

一级准则	一级权重(%)	二级准则	二级权重 (%)	主观权重 (%)	主观权重 平均化参考值 (%)	二级准则 数目	一级权重 平均化参考值 (%)	一级准则数目
C_1	17.42	SC_{11}	30.54	**5.32**	**7**	3	20	5
		SC_{12}	33.3	**5.80**	**7**			
		SC_{13}	36.16	**6.30**	**7**			
C_2	21.56	SC_{21}	28.31	**6.10**	**5**	4	20	
		SC_{22}	24.19	**5.22**	**5**			
		SC_{23}	28.25	**6.09**	**5**			
		SC_{24}	19.25	**4.15**	**5**			
C_3	25.28	SC_{31}	11.27	**2.85**	**3**	7	20	
		SC_{32}	16.59	**4.19**	**3**			
		SC_{33}	19.08	**4.82**	**3**			
		SC_{34}	15.95	**4.03**	**3**			
		SC_{35}	15.22	**3.85**	**3**			
		SC_{36}	11.54	**2.92**	**3**			
		SC_{37}	10.35	**2.62**	**3**			
C_4	20.08	SC_{41}	56.78	**11.40**	**10**	2	20	
		SC_{42}	43.22	**8.68**	**10**			
C_5	15.67	SC_{51}	44.73	**7.02**	**10**	2	20	
		SC_{52}	55.27	**8.66**	**10**			