

南京邮电大学

毕业设计（论文）

题 目 定向覆盖模糊测试工具的设计与实现

专 业 计算机科学与技术

学生姓名 雷尚远

班级学号 B190303 B19030334

指导老师 王子元

指导单位 计算机学院、软件学院、网络空间安全学院

日期： 2023 年 3 月 x 日至 2023 年 6 月 x 日

毕业设计（论文）原创性声明

本人郑重声明：所提交的毕业设计（论文），是本人在导师指导下，独立进行研究工作所取得的成果。除文中已注明引用的内容外，本毕业设计（论文）不包含任何其他个人或集体已经发表或撰写过的作品成果。对本研究做出过重要贡献的个人和集体，均已在文中以明确方式标明并表示了谢意。

论文作者签名：

日期： 年 月 日

摘 要

模糊测试（Fuzzing）是一种通过向目标系统提供非预期的输入并监视异常结果来发现软件安全漏洞的方法，是软件安全领域常用的方法之一。由于代码覆盖率与漏洞覆盖率密切相关，大多数模糊测试工具都是以代码覆盖率为导向。然而，由于大多数被覆盖测试的代码可能并不包含漏洞，这使得盲目地扩展代码覆盖率的方式在实际测试时效率较低。极端情况尤为如此。与盲目增加代码覆盖率的模糊测试不同，定向覆盖的灰盒模糊测试（DGF）将大部分时间用于检测特定目标区域（例如，易出错代码段）而不会浪费资源于不相关的部分。因此，DGF 特别适用于补丁测试、漏洞复现以及特殊漏洞检测等场景。目前，DGF 已成为一个快速发展的研究方向。基于一些先进的定向覆盖模糊测试工具的研究和相关调查，本文主要做了以下工作：

- (1) 基于现有的模糊测试工具框架 AFL（American Fuzzy Lop）以及 AFLGo 做了定向覆盖策略的设计和集成；
- (2) 实现了简单的定向覆盖的模糊测试命令行工具；
- (3) 针对相应的公开通用漏洞集（CVE）做了复现及定向实验对比测试。

此外本文亦通过分析工具设计以及实现过程中的局限性与不足，对于未来该方向的研究发展做出了一些展望。

关键词： 模糊测试；定向覆盖模糊测试；灰盒测试；软件安全

ABSTRACT

Fuzzing is a method of discovering software security vulnerabilities by providing unexpected inputs to a target system and monitoring for abnormal results. It is one of the commonly used methods in the field of software security. As code coverage is closely related to vulnerability coverage, most fuzz testing tools are guided by code coverage. However, blindly extending code coverage may be inefficient in practical testing since most of the covered code may not contain vulnerabilities, especially for corner cases. In contrast to blind code coverage-based fuzz testing, directed grey-box fuzzing (DGF) spends most of its time detecting specific target regions (such as error-prone code segments) rather than wasting resources on irrelevant parts. Thus, DGF is particularly suitable for scenarios such as patch testing, bug reproduction, and special bug detection. For now, DGF has become a fast-growing research area. Based on some advanced directed coverage fuzz testing tools and relevant investigations, this article mainly focuses on the following points of work:

- (1) Designed and integrated a directed coverage strategy based on the existing fuzzy testing tool framework AFL (American Fuzzy Lop) and AFLGo;
- (2) Implemented a simple command-line tool for directed fuzz testing;
- (3) conducted reproductions and directed experiments on corresponding public vulnerability databases (CVE) for comparative testing.

In addition, this article also provides some prospects for the future research and development of this direction by analyzing the limitations and deficiencies in the design and implementation process of the tool.

Keywords: Fuzzing; Directed Greybox Fuzzing; Greybox test; Software Security

目 录

第一章 绪论.....	1
1.1 背景分析.....	1
1.2 国内外研究现状.....	1
1.3 研究内容.....	3
1.4 论文结构.....	3
第二章 相关技术研究.....	5
2.1 模糊测试技术.....	5
2.1.1 基本概念定义.....	5
2.1.2 基本架构.....	6
2.1.3 模糊测试技术的分类.....	8
2.2 定向模糊测试技术.....	10
2.2.1 白盒定向模糊测试技术.....	10
2.2.2 灰盒定向模糊测试技术.....	11
2.3 研究动机.....	12
2.4 本章小结.....	12
第三章 定向模糊测试策略设计.....	13
3.1 AFLGo 架构研究.....	13
3.1.1 距离计算机制.....	14
3.1.2 能量调度机制.....	15
3.2 定向适应度指标设计.....	17
3.2.1 距离定义.....	17
3.2.2 可达目标函数集覆盖率.....	19
3.3 本章小结.....	21
第四章 基于 AFLGo 的定向模糊测试系统的实现.....	22
4.1 需求分析.....	22
4.2 架构设计.....	22
4.3 静态分析器的改进.....	22
4.4 定向模糊测试工具.....	22
4.5 本章小结.....	22
第五章 系统测试.....	23
5.1 系统测试概述.....	23
5.1.1 系统测试目标.....	23
5.1.2 系统测试环境.....	23
5.2 功能测试.....	23
5.3 实验评估.....	23

5.4 本章小结	23
第六章 总结与展望	24
6.1 总结	24
6.2 展望	24
结束语	25
致谢	26
参考文献	27

第一章 绪论

1.1 背景分析

“常用系统中可能会潜伏着严重的漏洞^[1]。”这一论述源自于模糊测试首次面世的论文。其揭示了一个事实，即随着软件技术的不断发展，软件安全问题就将日益成为愈发重视的议题。在当前的信息化时代，软件已经成为了人们生活、工作和娱乐的重要组成部分，这也意味着我们将面临着越来越多的安全威胁。因此，确保软件安全已经成为了一项非常重要的任务。

软件安全（Software Security）就是使软件在受到恶意攻击的情形下依然能够继续正确运行及确保软件被在授权范围内合法使用的思想。在当今社会，软件越来越普及，并被广泛应用于各个领域，包括电商、金融、医疗等。但是，由于软件的复杂性和开发过程中的缺陷，软件本身也存在着各种安全问题。这些问题可能导致信息泄露、数据损坏、远程攻击等，对个人、企业甚至整个社会造成巨大的损失。因此，保障软件安全显得尤为重要。

近年来，因为软件漏洞造成的损失案例屡见不鲜。2017 年，全球范围内爆发了 WannaCry 勒索病毒攻击事件，该攻击利用了微软 Windows 操作系统中的漏洞，并导致了数十亿美元的经济损失；2019 年，美国资讯技术服务公司 SolarWinds 遭受了一次大规模的网络攻击，该攻击利用了 SolarWinds Orion 平台软件中的漏洞，影响了包括美国联邦政府在内的许多组织和机构；2021 年 2 月，法国 LCL 银行的客户登录自己的银行应用程序时，看到的是别人的银行账户信息。原因是由于备份超级计算机系统（日本惠普公司制造）的程序存在缺陷，超级计算机系统出现了意外，其中存储（/LARGE0）中的某些数据被误删除；2021 年 12 月，知名日志框架 Log4j2 被爆出远程代码执行漏洞，影响了大量使用该框架的中间件和应用，给企业和用户带来了巨大的安全风险。

以上诸多例子可以说明，大多数的安全事件都是攻击者利用软件系统中的漏洞从而进行攻击引发的。因而可以帮助发现和修复安全漏洞的软件测试技术（Software testing）一直以来都是软件安全领域的一个重要议题。

软件测试可以通过模拟攻击者的行为来发现这些安全漏洞，并提供关于如何修复这些漏洞的信息。例如，黑盒测试可以探测应用程序中的安全问题，白盒测试可以评估应用程序的源代码中是否有漏洞，静态分析可以扫描源代码以发现潜在的安全问题，动态分析可以模拟攻击场景并检查应用程序的反应。

此外，软件测试还可以帮助确保应用程序在面对各种攻击时具有足够的鲁棒性和可靠性。它可以测试应用程序的身份验证和授权机制、加密技术、网络协议、输入输出数据验证等方面的功能，以确保应用程序满足安全需求。

1.2 国内外研究现状

软件安全信息系统和软件安全代码的有效安全项目往往依靠两种自动的安全测试：静态安全扫描测试和动态安全扫描测试。

在软件的开发期间，为了保证软件的安全性，通常会进行软件安全静态扫描。

这个过程是通过威胁建模和分析来完成的，其目的是对静态代码进行全面地扫描，以便及早地发现任何可能存在的安全漏洞。其是在不运行程序的情况下对软件进行测试和评估。静态分析可以检查代码、设计和文档等，以发现潜在的问题和错误，并确保软件符合某些标准或规范。由于本文主要探讨针对代码的漏洞审查，关于软件工程部分的文档、标准以及接口设计的测试技术在此不再赘述。利用数据流分析，符号执行以及污点分析等静态软件分析技术可以检查源代码中的错误和缺陷，包括语法错误、类型错误、内存泄漏、空指针引用等。与传统的动态测试相比，静态扫描可以更早地发现安全问题，因为它可以在代码尚未被编译或执行之前就进行检测。此外，静态扫描还可以减少测试成本，提高测试效率，并帮助开发团队更好地理解代码中的潜在安全风险。

软件安全动态扫描是一种对工作环境中实际运行的代码进行扫描的技术，它能够在代码运行时检测和分析可能存在的漏洞、缺陷和错误。与静态代码分析不同，动态扫描具有更强的准确性和实时性，因为它是在真实的环境中对代码进行测试和评估。通过使用动态扫描技术，开发人员和安全专家可以有效地识别并修复潜在的安全漏洞，从而保护软件系统免受攻击和破坏。此外，动态扫描还可以帮助企业遵循各种合规性标准和法规要求，确保其软件应用程序的安全性和稳定性。

模糊测试技术（Fuzzing）是动态安全扫描测试中重要的一种方式。而自从 1988 年模糊测试这一概念被提出后，这一方法一直在软件安全测试领域保持着较高的活跃度和关注度。在提出伊始，其主要用于测试操作系统。之后，随着软件技术的发展，模糊测试技术不断得到改进和推广，并应用于网络、移动设备等领域。目前，模糊测试技术已成为一种成熟的自动化测试技术，可以有效地检测软件中存在的漏洞和安全隐患。并且迄今为止其社区依然十分活跃，在 GitHub 上有超过 1000 个与模糊测试相关的仓库^[2]。为了防止被恶意攻击，许多商业软件公司，例如 Adobe, Cisco, Google, 和 Microsoft 都将模糊测试作为其雇员软件开发安全测试的必要环节。可以说，模糊测试是软件安全领域中一个经久不衰的热门议题。

而定向模糊测试（Directed Fuzzing）作为模糊测试的一个研究方向，主要关注重点区域（例如，易出错区域）并且将大部分的时间用于到达测试这些位置而不浪费资源在无关部分^[3]。从本源上讲，定向模糊测试工具早期的解决思路主要是基于利用程序分析和约束求解来生成检测不同程序执行路径输入符号的执行技术^[4-8]。然而，由于定向符号执行技术（DSE）依赖于大量的程序分析和约束求解，其受限于执行效率、兼容性和可扩展性的问题。

在 2017 年，Böhme 等人提出的 AFLGo^[9]引入了定向灰盒模糊测试（DGF）的概念。这是定向模糊测试的又一重要工作。其开创式地将位置的可达性问题转化为生成种子和其目标集之间距离的最小值问题。通过给更靠近目标集的种子更多的变异机会，它可以逐渐引导灰盒测试接近程序目标位置。与定向符号执行技术相比，DGF 有更好的可拓展性，并且在测试效率上有几个数量级的提升。例如，Böhme 等人可以在 20 分钟内重现 Heartbleed^[10]（CVE-2014-0160）漏洞，而定向符号执行工具 KATCH^[7]需要 24 小时以上^[9]。

此后，许多基于 AFLGo 的定向性改进工作被相继提出。在 2018 年，Hongxu

Chen 等人提出的 Hawkeye^[11] 指出了 AFLGo 的距离计算方式导致的能量分配偏向导致的在测试的路径选择上偏向于距离目标更近的路径（或最短路径），而这有可能漏掉一些存在于较长路径上潜在的漏洞，于是其使用了新的距离度量来辅助达到更精确的距离导引；在 2020 年 Wang 等人提出的 UAFL^[12] 开创了针对于特定漏洞行为的定向性导引，其利用目标行为次序而不是目标位置来查找释放后使用的漏洞，这种漏洞的内存操作一定要按照一定次序执行（例如，分配，使用，然后释放内存）才能触发；在 2021 年 Gwangmu Lee 等人提出的 CDGF^[13] 则指出 AFLGo 在路径选择中的不考虑执行路径导致的目标位置的顺序影响从而忽略满足特定行为引起崩溃条件的种子的缺陷，他们的解决思路不再是利用 AFLGo 的种子分配能量的方式来指引种子变异从而检测指定目标位置，而改为将其视作约束求解问题，通过设置一系列的约束来优先考虑选择符合要求的种子从而达到检测指定目标位置的目标；而在 2022 年由 Heqing Huang 等人提出的 BEACON^[14] 则从另一种方式进一步尝试提高定向模糊测试的速率：通过“剪枝”，即修建无效路径的方式。其结合轻量级的静态程序分析，来计算到达目标位置的抽象前提条件，并在运行时剪除那些不满足条件的路径，从而可以有效地提高效率，避免在无效或不可达的路径上浪费时间和资源。

1.3 研究内容

首先，针对于模糊测试技术，本文做了一个简单的梳理，对于目前模糊测试的基本主流技术框架做了分析和总结。自 1988 年这一技术的提出以来，相关技术研究百花齐放，技术快速迭代发展，直到近些年才有相关的总结研究和谱系分析^[2,15]，建立系统性的分析。本文在参考相关文献的基础上对模糊测试的技术框架做了一个梳理。

其次，针对于定向灰盒模糊测试的开篇之作 AFLGo，本文对其的技术架构做了一个详细的总结梳理。对于 AFLGo 的一些实现细节做了详细的分析和总结。除此之外，本文还探讨了 AFLGo 在实际应用中的优缺点以及相比于其他模糊测试工具所具备的特点。针对 AFLGo 的优点，我们详细阐述了其灵活性、高效性和可扩展性，同时也分析了其在某些情况下可能会出现的一些问题。此外，本文还介绍了 AFLGo 在不同场景下的应用，帮助读者更好地理解如何使用 AFLGo 进行测试。

最后，参考以 AFLGo, Hawkeye 为主的定向性模糊测试工具，针对于 AFLGo 的架构进行了改进和集成。本文主要从距离定义和目标函数集合覆盖率两个指标针对于定向策略做了设计和修改，并结合定向模糊测试工具 AFLGo 和 LLVM 的 Pass^[16] 技术做的静态分析的实现，实现了指标在 AFLGo 的集成。最终，本文在实现后通过实验成功复现 libxm2^[17] 的 CVE-2017-{9047,9048}，证实了集成的可行性和可靠性。

1.4 论文结构

本文共分六章，各个章节的内容如下：

第一章：主要介绍本文的研究背景、国内外研究现状、研究内容以及本文的

论文结构。

第二章：主要介绍本文涉及技术的详细定义及通用架构，包括模糊测试技术和定向模糊测试技术，以及阐述本文的研究动机。

第三章：介绍 AFLGo 的主要架构和适应度指标，以及针对于其不足设计出的定向模糊测试适应度指标。详细阐述相应指标的核心设计, 包括距离机制的修改和可达目标函数集覆盖率的补充。

第四章：结合目标场景对实现的定向模糊测试工具做了需求分析，详细介绍了基于 LLVM 的 Pass 技术实现的插桩以及静态分析过程。对于结构集成以及实现方式的缘由做了详细分析。

第五章：对于实现集成的工具进行了实验测试，并做了相应的测试评估，包括工具测试的可行性和性能评估。

第六章：总结了本文的工作并对于未来定向模糊测试技术的发展及研究方向做了展望。

第二章 相关技术研究

2.1 模糊测试技术

模糊测试（Fuzz Testing）是一种软件测试技术，其主要思想是通过向输入参数、文件、网络请求等随机或半随机注入无效、异常、边界数据，来检测目标系统在处理异常情况时的鲁棒性。模糊测试可以帮助发现那些未经预料的漏洞和错误，这些问题可能会导致应用程序崩溃、停止响应或者执行意外的操作。

在进行模糊测试时，测试人员通常需要编写一个模糊测试工具，该工具可以生成大量的随机测试用例，并将这些测试用例输入到应用程序中进行测试。模糊测试通常被认为是一种高效的测试技术，因为它可以在较短的时间内检测应用程序中的许多潜在问题。此外，由于测试用例是随机生成的，模糊测试可以找出一些没有被其他测试方法发现的漏洞和错误。

模糊测试的过程通常分为以下几个步骤：

- (1) 选择目标：选择需要测试的软件目标，如应用程序、库、操作系统等。
- (2) 寻找输入：确定需要对目标注入的输入类型和数据源，如输入参数、文件、网络请求等。
- (3) 创建模糊数据：使用随机或半随机的方式生成模糊数据，并将其注入到目标中。
- (4) 监控程序行为：监控被测试程序的运行行为，如崩溃、错误输出等。
- (5) 分析结果：对测试结果进行分析和报告，识别潜在的漏洞和安全问题，并反馈给开发人员进行修复。

模糊测试是一种简单有效的测试方法，可以在较短时间内发现大量的异常情况，并帮助开发人员提高软件质量和安全性。

2.1.1 基本概念定义

为了准确地讨论学界关注的模糊测试方向的问题以及梳理架构，我选择了采用 Manès 等人于论文^[2]中提出的基本概念的定义。

- **Fuzzing**: 模糊测试技术是指使用从超出被测试程序（PUT）的预期输入空间的输入空间（“模糊输入空间”）采样的输入来执行被测试程序。

事实上，模糊输入空间（fuzz input space）在定义中并不一定需要包含预期输入空间，其只需要包含预期输入空间所没有的输入即可；其次，因为在实践中大部分模糊测试几乎必然会选择多次迭代来实现测试，故而上述定义补充为“重复执行”依然很大程度上是准确的；最后，关于对于模糊输入空间的采样并不必然是随机的，这与算法设计的种子优先级排序以及采取的变异方式等实现方式有关，故定义并没有加上“随机采样”。

- **Fuzz Testing**: 模糊测试是指使用模糊测试技术（Fuzzing）来测试被测试程序是否违反正确性策略。

在诸多论文^[18-20]中依然可以见到另一种说法 — Fuzz Campaign(模糊测试活

动), 用以指明利用特定的模糊测试工具针对具体的被测试程序的一次实际测试活动, 但在本文中不再区分这一概念, 统一用模糊测试来做指代。

- **Fuzzer:** 模糊测试工具是对被测试程序执行模糊测试的程序。
- **Bug Oracle:** 错误检测器是一个程序, 可能是模糊测试工具的一部分, 它确定被测试程序的给定执行是否违反了特定的正确性策略。
执行模糊测试的目标当然是查找被测程序的漏洞。但是不同模糊测试工具设计的针对的软件正确性准则 (correctness policy) 并不相同。例如, 早期的模糊测试的主要的准则仅仅是生成的测试用例是否会导致被测程序的崩溃 (crash); 事实上, 以 SECFUZZ^[21] 为代表的许多测试工具就将模糊测试引入了包括网络密钥交换协议等非严格软件安全的领域。一般来讲, 任何可以通过实际执行的状况中观察到的策略, 都可以通过模糊测试来进行安全测试。故而所有确定执行表现是否违反特定的正确性策略的机制部分都可以被称为错误检测器。在本文中, 依然将错误检测器执行的软件正确性准则视为导致被测试软件崩溃, 其余准则不加以讨论。
- **Fuzz Configuration:** 模糊测试算法的配置是指所有能控制模糊测试算法执行的参数配置。
出于一般性, 在此处将所有能够影响到模糊测试算法执行的参数因子都统一称为算法配置。不同的模糊测试算法关注的参数各不相同。例如, 最简单的输入随机生成比特流输出至被测试程序的模糊测试工具的配置可能仅仅是被测试程序的状态空间, 而复杂的模糊测试算法的配置则可能涉及到算法执行时间, 随机输入的结构, 随机变异策略, 前置静态分析结果等等。而一般情况下, 最受大家关注的配置就是种子的状况。
- **Seed:** 种子是被测试程序的 (通常结构良好的) 输入, 一般用于通过修改它来生成测试用例。

2.1.2 基本架构

算法2-1是一般化的模糊测试算法。现如今的大部分模糊测试工具基本是以此流程架构设计。这个算法足够通用且具有一般性, 可以代表如今所有的模糊测试技术, 包括黑盒模糊测试, 白盒模糊测试以及黑盒模糊测试。

本算法也可以简单提炼为模糊测试算法的五个基本步骤, 即为: 预处理 (preprocessing)、调度 (schedule)、输入构造 (input generation)、输入评估 (input evaluation)、配置更新 (update)。下面将结合算法分别简单介绍各个阶段。

算法 2-1: 模糊测试算法

```

Input:  $\mathbb{C}$ ,  $t_{\text{limit}}$ 
Output:  $\mathbb{B}$  // a finite set of bugs
1  $\mathbb{B} \leftarrow \emptyset$ ;
2  $\mathbb{C} \leftarrow \text{Preprocess}(\mathbb{C})$ ;
3 while  $t_{\text{elapsed}} < t_{\text{limit}} \wedge \text{Continue}(\mathbb{C})$  do
4    $\text{conf} \leftarrow \text{Schedule}(\mathbb{C}, t_{\text{elapsed}}, t_{\text{limit}})$ ;
5    $\text{tcs} \leftarrow \text{InputGen}(\text{conf})$ ;
   //  $O_{\text{bug}}$  is embedded in a fuzzer
6    $\mathbb{B}', \text{execinfos} \leftarrow \text{InputEval}(\text{conf}, \text{tcs}, O_{\text{bug}})$ ;
7    $\mathbb{C} \leftarrow \text{ConfUpdate}(\mathbb{C}, \text{conf}, \text{execinfos})$ ;
8    $\mathbb{B} \leftarrow \mathbb{B} \cup \mathbb{B}'$ ;
9 return  $\mathbb{B}$ ;

```

预处理: 在算法2-1的第2行为针对程序输入的算法配置进行预处理。主要是用户通过提供一组模糊测试配置, 而后依据模糊测试算法进行处理从而返回一组可能经过修改的模糊测试配置。由于采取的模糊测试策略不同, 预处理的操作也就不同。例如接受可执行文件的黑盒测试一般不做处理, 而白盒测试和灰盒测试可能会依据算法的设置源代码或者二进制代码的插桩, 进行控制流图生成, 指针分析以及污点分析等软件分析流程。一些测试工具还会选择对用户提供的种子进行处理, 例如预筛选可能多余的种子, 对种子进行合并精简化等。此外, 一些针对内核或库文件进行很难直接运行程序进行测试的模糊测试工具可能还会准备加载相应的驱动程序用以提供测试。

调度: 在算法2-1的第4行为针对执行的算法配置进行调度设置。主要是通过更改或者选择特定的算法配置来决定接下来算法的行为。在此过程一般为处理模糊测试算法中的探索 (exploration) 与开发 (exploitation) 问题。探索问题是指在执行模糊测试中将此次分配的时间用以收集有关每个配置的更准确信息上, 以便为未来的决策提供信息; 而开发问题是指将分配的时间用以针对当前被认为会导致更有利结果的配置进行模糊测试。由于实际被测试的程序不同, 待测试的目标复杂度不同以及需要的精确信息的获取难度不同, 实际测试中这两阶段的分界线并不是那样的固定和明确。一般算法仅仅是通过设置策略来引导探索阶段收集特定所需的信息。

输入构造: 在算法2-1的第5行为针对选择的算法配置来生成相应的测试用例。这与采取的模糊测试技术有关, 一般根据此阶段的策略将模糊测试算法分成基于模型 (Model-based) 的算法和少模型 (Model-less) 的算法。由于不同的被测程序对于输入的格式要求不同, 故而对于测试用例的构造也有不同的策略。基于模型的模糊测试工具根据给定的模型生成测试用例, 给定模型描述了被测程序可能接受的输入或执行方式, 如准确描述输入格式的语法或较不准确的约束条件, 比

如标识文件类型的魔术值（magic values）。因此基于模型的算法又被称为生成式（Generation-based），通过输入模型来指导生成测试用例，常见于黑盒模糊测试。而少模型的算法则主要是依据算法伊始提供的种子来进行大量变异生成符合条件的测试用例。这种方式的出现是为了解决基于模型的生成输入效率不高的问题。除此以外，白盒测试技术可能会采取符号执行的方式来辅助生成测试用例。

输入评估：在算法2-1的第6行为针对生成的输入进行输入执行测试和进行种子评估。这一步会将生成的测试用例输入被测试程序来测试被测试程序。根据执行生成的执行信息和错误检测器来判断是否触发了被测试程序的异常行为，这一步将为下一步的配置更新提供反馈信息（对于黑盒测试，一般就是报告程序错误信息）。

配置更新：在算法2-1的第7行为依据上一步生成得到的执行信息和反馈信息对算法配置信息进行更新。这些变动将决定和引导接下来的模糊测试算法的行为。对于黑盒测试而言，这一步一般什么也不做。

2.1.3 模糊测试技术的分类

一般会将模糊测试技术分为黑盒模糊测试，白盒模糊测试以及灰盒模糊测试。与传统软件测试中的黑盒测试和白盒测试^[22]定义相似，对于模糊测试技术的分类是依靠模糊测试工具收集信息的数量来决定的^[2]。

模糊测试工具会有多种渠道收集信息，但是最主要的途径依然是通过程序的状态来获得。正如一般化算法2-1中所示，事实上只有只有算法的第2行和第7行会对算法的配置进行修改。这对应的即是算法的预处理部分和配置更新部分。也即，这两个步骤获取程序信息的数量也基本上决定了模糊测试工具的种类（或者讲，“颜色”）。

早期的模糊测试技术就是黑盒模糊测试^[23]（black-box fuzzing）。黑盒这一概念就来源于软件工程，在模糊测试的语境下就是并不获取被测试程序的内部信息，仅通过观察被测试程序的输入/输出行为来测试被测程序，将其内部视作一个黑盒。黑盒模糊测试不依赖于程序相关信息，通过生成大量随机或半随机的输入来检测程序的漏洞。大多数黑盒模糊测试也如前文所言，采取的是基于模型的输入生成策略，然后基于给定的模型或者结合提供的种子样例来产生随机输入。因此对应到算法2-1中的第2步与第7步，黑盒测试基本上并不会对模糊测试配置进行修改。但是有些黑盒测试可能会将合适的（如，引起了程序崩溃）测试用例再次加入种子池参与变异，故而不能完全依照变异策略来进行区分。常见的现代黑盒测试工具包括 Peach^[24]，Spike^[25]等。

黑盒模糊测试由于部署以及算法设计简单，故而在业内具有广泛地应用。但是黑盒模糊测试的缺陷在于由于其生成测试用例的随机性，导致其在测试实际情况下的程序效率较低。举个简单的例子，假设被测试的程序其中一条路径的执行条件为“if a == 34”，而四字节的 Int 型输入的可能性为 $2^{32} = 4,294,967,296$ 种，这意味着仅靠随机生成字节很难探测到该路径。这一特性使得黑盒测试在实际测试中只能探查程序的表面逻辑来暴露浅显的漏洞，而对于深层的逻辑错误探测就检

测效率较低。

与黑盒模糊测试对应的就是白盒模糊测试^[26]（white-box fuzzing）。白盒模糊测试就是通过分析被测试程序的内部以及测试时收集的信息来生成测试用例的模糊测试技术。白盒模糊测试一般会结合程序分析技术，例如符号执行，约束求解以及污点分析等来实现对被测试程序内部信息的全面获取，从而实现测试工具对被测试程序的覆盖率提升或者探测指定代码区域的测试目标。从理论上讲，白盒模糊测试根据获得的程序内部信息可以做到对所有执行路径的全面覆盖。而白盒模糊测试相对于黑盒模糊测试最大的区别就是其引入了重量级的程序分析部分来获取程序内部信息。结合到算法2-1讲就是算法的第2行预处理部分引入了程序分析技术来获取内部信息从而对算法配置进行修改，针对性的识别代码块来引导测试用例的生成^[26]。而自2008年Cadars等人在白盒模糊测试技术里引入了动态符号执行技术（Dynamic Symbolic Execution）^[4]后，白盒模糊测试可以通过结合预处理阶段的信息以及动态执行时收集的信息，可以生成一个约束系统，用于确定哪些路径可达，从而提高了模糊测试的覆盖率和效率。这对应到算法2-1中第7行的配置更新部分。常见的现代白盒模糊测试工具包括KLEE^[4]，SAGE^[27]等。

利用程序分析技术的白盒模糊测试可以提升对程序的覆盖率，从而实现对程序的全面的分析和检测。但是其在实际应用时，由于重量级的程序分析技术往往会耗费大量时间，以及在实际程序中随着规模的增加会出现“路径爆炸”的问题，使得白盒模糊测试技术难以达到理想的效果。并且由于其复杂的符号执行和约束求解过程，使得这一技术在使用时会十分低效。

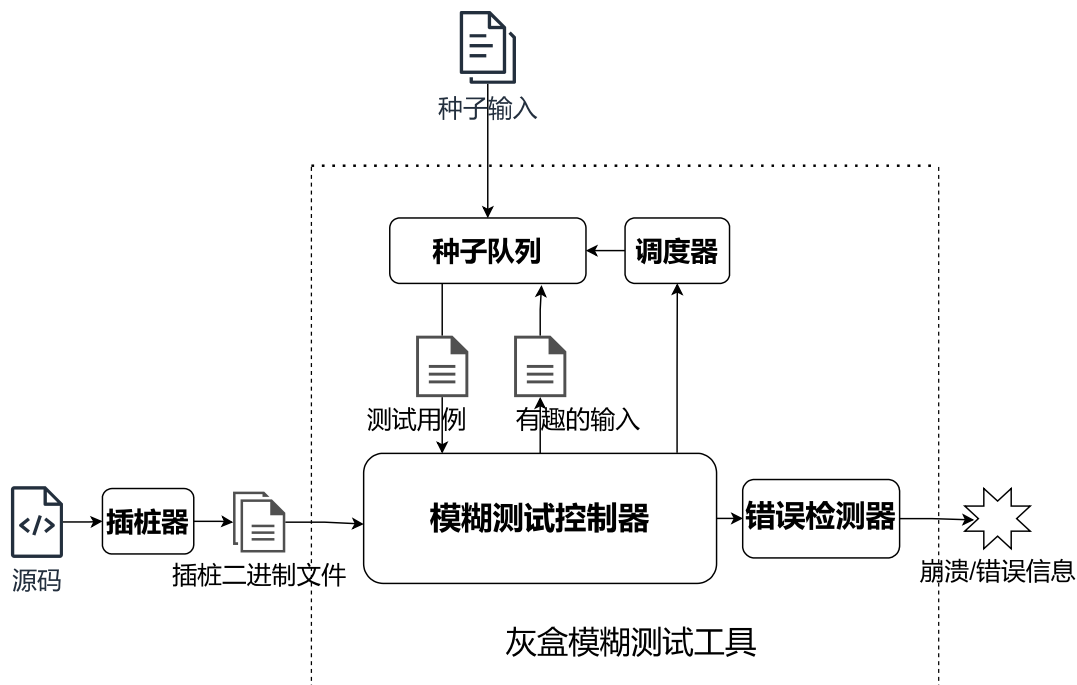


图 2.1 灰盒模糊测试架构

由于白盒模糊测试受限于符号执行自身的问题，灰盒模糊测试的方式越来越得到学界重视。这一技术方法介于上述两种技术之间。通常，灰盒模糊测试工具

可以获得被测试程序或其执行的一些内部信息。与白盒模糊测试工具不同，灰盒模糊器不推理被测试程序的完整语义；相反，其采用对被测试程序执行轻量级静态分析或收集有关其执行的动态信息，例如代码覆盖率，数据流等，以及配合良好的测试策略，来实现对被测试程序的整体测试，既综合了白盒模糊测试的优点，亦规避了其效率低下的缺点。

在 2013 年由 Google 发布的 AFL^[28](American Fuzzy Lop) 是灰盒模糊测试工具中最重要的一个成果。图2.1以其为例展示了一般的灰盒模糊测试工具的基本架构。AFL 是一款以覆盖率为导向的模糊测试工具，通过插桩的方法，采集输入数据对应的边覆盖率，作为模糊测试种子选取的衡量指标，通过设计适当的算法函数，较好的实现了灰盒模糊测试，达到了较高的代码覆盖率。而对应到算法2-1中，灰盒模糊测试主要在第 7 行利用收集到的执行信息（如代码覆盖率，种子发现的新路径等）来更新算法配置，从而指导测试用例的生成。虽然灰盒模糊测试一般也会在第 2 行的预处理阶段进行一些静态分析，但相对于白盒模糊测试而言，其所有的程序分析部分都会在此步或之前完成，是非常轻量级的程序分析技术。具体到 AFL 来讲，这一步仅仅是识别不同的路径并对程序进行插桩。

2.2 定向模糊测试技术

前文所述的包括 AFL 等在内的大部分模糊测试工具的主要是以覆盖率为导向（coverage-based）的模糊测试技术。直觉上来讲，模糊测试工具对程序的代码覆盖率越大，那么通过模糊测试可以探查到的程序漏洞一般来讲也会更多。故而无论是白盒模糊测试还是以 AFL 为代表的灰盒模糊测试都是以扩大代码覆盖率为测试用例生成的主要目标。

而在一些应用场景下，盲目地扩张代码覆盖率未必能取得很好的漏洞检测效果。例如，在对程序的补丁进行测试时，盲目扩大代码的覆盖率检测正常运行的部分是对资源的浪费。这启发了定向模糊测试（Directed Fuzzing）的思想，与其一味的扩大代码的覆盖率，不如重点检测特定的代码区域。

而现在意义上的定向模糊测试可以认为是将模糊测试生成的测试用例引导至指定的目标：既可以是指定的代码区域，即源码中的指定行数的代码区域或二进制文件中的虚拟地址；也可以是指定的特定行为或错误，比如 UAF（use-after-free）错误。当然，本文的设计是目标是定向覆盖特定区域，针对特定行为的定向模糊测试技术就不再讨论。

定向模糊测试技术除了上述的补丁测试（Patch Testing）的应用外，还可以用于崩溃重现（Crash Reproduction）、静态分析报告验证（Static Analysis Report Verification）以及数据流检测（Information Flow Detection）等。

2.2.1 白盒定向模糊测试技术

早期的定向模糊测试是由白盒模糊测试实现的。在 2011 年由 Ma 等人提出的定向符号执行^[6]（Directed symbolic execution，简称 DSE）技术是其中最具代表性的一项工作。

符号执行^[29]是一种自动化程序分析技术，它通过对程序输入进行符号化表示并应用约束求解来推导出程序的所有可能执行路径及其相关状态。在符号执行期间，程序中的每个变量都被表示为一个符号表达式，而不是一个具体值，这可以使分析器在不运行实际代码的情况下探索程序状态空间的各种可能性。

对于程序的定向覆盖的需求，Ma 等人将这一位置可达性问题转换为迭代约束满足问题，也即经典的约束求解问题。通过指定待探测代码区域，定向符号执行技术利用程序分析来分析所有可到达指定区域的路径，而后通过约束求解生成所有符合执行相应路径的测试用例以来达到对指定区域的覆盖和测试。Ma 等人提出了两种符号执行方式：最短距离符号执行（Shortest-distance symbolic execution，简称 SDSE）和调用链后向符号执行（Call-chain-backward symbolic execution，简称 CCBSE）。SDSE 是通过过程间控制流图，从起点出发计算可以到达指定区域的最短距离从而得到合适的路径条件；而 CCBSE 则是通过从指定区域自后向前执行符号执行，沿着调用链向前查询所有可能到达起点的路径条件。在得到路径条件后，再通过设置路径约束条件以求解满足条件的生成测试用例。而后测试用例即可依据设定的路径达到指定代码区域，从而实现对特定区域的测试，也即定向覆盖模糊测试。

除此以外，无论是结合污点分析技术的 BuzzFuzz^[5]，还是结合动态符号执行的 KLEE^[4]，都无外乎是在符号执行引擎中实现，采用探索可行路径的状态空间的方式来实现定向覆盖。

但是这一方式依然具有白盒模糊测试的弊病，那就是对路径条件的探索和种子生成条件的指引的有效性是以模糊测试的效率为代价的。当实际被测试程序具有很高的复杂度和规模时，DSE 采取的重量级程序分析将面临严重的路径爆炸问题，这也使得实际执行效果并不理想。

2.2.2 灰盒定向模糊测试技术

由 Böhme 等人于 2017 年提出的 AFLGo^[9]引入了定向灰盒模糊测试（Directed Grey-box Fuzzing，简称 DGF）的概念。这一工具也正是本文重点研究的架构。相较于繁重的约束求解来解决定向的位置可达性问题，AFLGo 采取从另一角度来实现本目标。Böhme 等人先前于 2016 年做出的工作^[30]将模糊测试的路径执行过程建模为马尔科夫链，这为 AFLGo 的设计提供了思路。

事实上我们知道，一个程序中大多数执行路径是不会被经常执行的，而经常执行的路径往往只占有所有路径中的小部分。在 AFL 的输入构造环节中种子的能量分配却是平均分配的，这使得那些较少被执行的路径获取到的种子能量因为相对不足而需要更多的时间才能探索暴露出漏洞。对此，Böhme 等人在他们的工作 AFLFast^[30]中给出的解决办法是将模型模糊测试程序建模为对马尔可夫链接状态空间的遍历过程，将种子变异导致的程序执行路径转移概率，视为马尔可夫链上的状态转移概率。然后通过制定相应的能量分配策略，使得马尔可夫链状态空间的遍历过程，更倾向于低访问频率的区域，使得能量分配更加合理。通过针对 AFLFast 和 AFL 的对比实验表明，这样的能量分配策略使得在同样的时间中，AFLFast 有更

多的机会去接近低访问频率的区域从而能够发现比 AFL 多指数级的崩溃^[30]。

于是，针对于定向覆盖的问题，AFLGo 就将其从位置可达性问题转换为能量分配的优化问题。AFLGo 舍弃了繁重的路径条件的求解，而是转而在预处理阶段计算每个基本块与目标位置的距离并将其插桩进入二进制文件，再通过运行时评估每个种子执行路径相对于目标区域的平均距离从而利用种子的能量调度机制来影响测试用例的生成从而达到定向覆盖的目标。这一机制还会在章节3.1详细介绍。

现在 DGF 仍然是模糊测试领域中非常先进的技术。其结合了定向模糊测试和灰盒模糊测试的思想，可以在不需要完全了解被测系统内部工作原理的情况下，检测出潜在的漏洞和错误。其将大部分的能量分配给重点的代码区域，而不浪费资源于其他无意义的部分，可以更快更高效地测试代码。同时，相比于白盒定向模糊测试方法，定向灰盒模糊测试可以避免过多地关注程序内部的实现细节，从而节省测试用例设计和编写的时间成本。

2.3 研究动机

自从定向灰盒模糊测试技术被提出来以来，就有许多相关的研究工作被提出。虽然 AFLGo 作为定向灰盒模糊测试方向开山之作在该方向的使用表现卓有成效，但是其策略在一些方面依然存在着一定的缺陷。后续又涌现出了一批相关的工作，例如 Hawkeye^[11]，CDGF^[13]，BEACON^[14]等。

这些工作指出了 AFLGo 的诸多不足，例如 Chen Hongxu 等人认为 AFLGo 在能量分配上偏向较短的路径，这会导致漏洞的遗漏^[11]；Gwangmu Lee 等人则认为 AFLGo 在路径上的优化问题处理方式会遗漏目标位置的顺序影响以及不考虑导致崩溃产生的数据条件，从而忽略满足崩溃条件的种子^[13]；Heqing Huang 则指出 AFLGo 保留的过多无效路径仍然会影响模糊测试执行的路径以及耗费资源^[14]。

这些工作中 Chen Hongxu 等人的 Hawkeye^[11]对 AFLGo 的思路沿用最为整体，改进措施较为详细。但是他们的工作是在其自研的模糊测试框架 FOT^[31]上实现，且该项工具并未开源。这使得对于其成果即灰盒模糊测试的性能提升较难验证。于是我的定向算法指标设计主要参考了他们指出的问题，并尝试在开源的 AFLGo 框架上实现集成指标，以期达到 AFLGo 的改进。

2.4 本章小结

本章首先详细介绍了模糊测试的一般算法框架，并介绍了模糊测试技术的分类，分别详细区别了黑盒模糊测试，白盒模糊测试以及灰盒模糊测试的不同，一并介绍了不同技术的特点和相关工具。而后介绍了本文重点关注的定向模糊测试技术，介绍了技术发展初期的定向白盒模糊测试技术原理和相关工具，现在主要流行的定向灰盒模糊测试技术与典型工作 AFLGo。最后，介绍了本文的研究动机和主要参考工作。

第三章 定向模糊测试策略设计

3.1 AFLGo 架构研究

算法 3-1: 定向灰盒模糊测试算法

```

Input:  $\mathbb{S}$  // a finite set of seeds
Input:  $\mathbb{T}$  // a finite set of targer sits
Output:  $\mathbb{S}'$  // a finite set of buggy seeds
1  $\mathbb{S}' \leftarrow \emptyset$ 
2  $SeedQueue \leftarrow \mathbb{S}$ 
3  $Graphs \leftarrow GraphExt(Code)$ 
4  $BBdistance \leftarrow DisCalcu(\mathbb{T}, Graphs)$ 
5 while  $!signal \wedge t_{elapsed} < t_{limit}$  do
6    $s \leftarrow Dequeue(SeedQueue)$ 
7    $trace \leftarrow Execution(s)$ 
8    $distance \leftarrow SeedDis(trace, BBdistance)$ 
9    $e \leftarrow AssinEnergy(s, t_{elapsed}, distance)$ 
10  for  $i \leftarrow 1$  to  $e$  do
11     $s' \leftarrow Mutation(s)$ 
12    if  $s'$  crashes then  $\mathbb{S}' \leftarrow \mathbb{S}' \cup s'$ 
13    if  $IsIntersting(s')$  then  $Enqueue(s', SeedQueue)$ 
14 return  $\mathbb{S}'$ 

```

本文以 AFLGo 为主要研究对象，且主要改进及算法实现均在 AFLGo 上实现，故本节将详细介绍 AFLGo 的基本架构。

算法3-1展现了 AFLGo 的基本执行流程。可以看到与算法2-1的基本结构相同。AFLGo 采取的算法配置包括：种子队列，种子分配能量以及种子执行路径的距离。而 AFLGo 主要的修改可以分为在预处理阶段阶段和运行阶段两个阶段。

在算法3-1的第 3 行和第 4 行是算法的预处理阶段。在这一阶段，AFLGo 通过先进行一次插桩编译计算出被测试程序的调用图（Call Graph，简称 CG）和每个函数对应的控制流图（Control-flow Graph，简称 CFG），这一步是算法3-1的第 3 行实现。再通过利用提供的目标（即源码中对应的代码行号）识别出目标代码所在的函数（Target Function），在 CG 中利用迪杰斯特拉最短路径算法得到每个函数到达目标函数的最短距离。而后结合对应每个函数的 CFG 计算函数中的基本块到目标代码所在基本块（Target Basic Block）的距离。这一步是在算法3-1的第 4 行实现。此后，AFLGo 会再执行一次插桩将代码距离植入每一个基本块。

在算法3-1的第 7、8 和 9 行是算法的算法配置更新阶段。针对种子池中的所有输入种子，AFLGo 都会将其执行一遍，并收集执行过程中种子轨迹的平均距离。

这一步由算法的第 7、8 行完成。而后，算法会依据种子的距离以及算法已执行的时间，向种子分配能量。注意，在 AFLGo 中将种子的能量定义为由种子生成的测试用例数量。通过利用模拟退火的优化方式对种子分配能量，来引导测试用例更多的倾向于覆盖指定区域。

3.1.1 距离计算机制

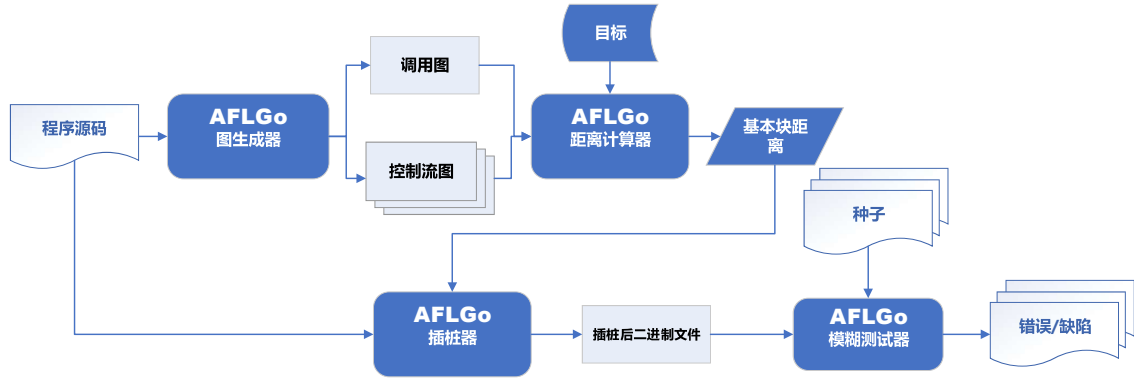


图 3.1 AFLGo 基本框架

图3.1展示了 AFLGo 的基本框架。与图2.1所展示的灰盒模糊测试的基本框架相比，AFLGo 做出较大改变的就是预处理的插桩部分。AFLGo 利用 LLVM 的 pass^[16]对程序的源码实现了编译器级别的分析和插桩。利用 AFLGo 的图生成器，AFLGo 将得到程序的 CG 和 CFG，这将用于参与计算 AFLGo 的函数级距离和基本块级目标距离。此外，可以根据提供的源码行号指定的目标得到目标代码所在的目标函数集 T_f 和目标基本块集 T_b 。而后，根据 CG 就可以计算函数级目标距离。

函数级距离 $d_f(n, n')$ 定义为调用图 CG 中函数 n 和 n' 之间最短路径上的边数。而函数 n 与目标函数 T_f 之间的函数级目标距离 $d_f(n, T_f)$ 则被定义为 n 与任何可达目标函数 $t_f \in T_f$ 之间函数距离的调和平均值：

$$d_f(n, T_f) = \begin{cases} \text{undefined}, & \text{if } R(n, T_f) = \emptyset \\ \left[\sum_{t_f \in R(n, T_f)} d_f(n, t_f)^{-1} \right]^{-1}, & \text{otherwise} \end{cases} \quad (3-1)$$

式3-1中 $R(n, T_f)$ 为从函数 n 出发可达的目标函数集 T_f 中的所有目标函数 t_f 的集合。注意，AFLGo 是支持多个目标的，这也意味着目标函数可能不止一个。关于这里采用调和平均的计算方式，图3.2展示了一个例子。在图中 (a) 采用了算数平均，而 (b) 采取了调和平均。当有多个目标函数时，在调用图中利用算数平均计算函数间目标距离无法很好地区分当到达不同目标函数的最短路径总和恰好相等时，究竟哪个函数离其中某个目标函数更近一点。

基本块级目标函数距离则依托函数级目标距离来参与计算。基本块级目标距离确定了从基本块到调用函数的所有其他基本块的距离，并且要考虑被调用函数的函数级目标距离。基本块级距离决定了一个函数的 CFG 中任意两个基本块

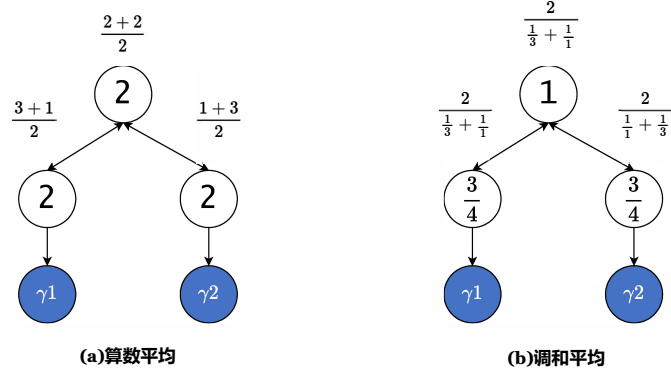


图 3.2 算数平均与调和平均的差异

之间的距离。因此，将基本块级距离 $d_b(m_1, m_2)$ 作为函数 i 的控制流图 G_i 中基本块 m_1 和 m_2 之间最短路径上的边数。设 $N(m)$ 为基本块 m 的一个调用函数集，使得 $\forall n \in N(m)$ 都有 $R(n, T_f) \neq \emptyset$ 。设 T 是 G_i 中一个基本块集，使得 $\forall m \in T$ 都有 $N(m) \neq \emptyset$ 。请注意，目标基本块 T_b 不一定需要存在于当前 CFG G_i 中，但 G_i 中可能存在可以调用可到达包含目标基本块 T_b 的函数的基本块集 T 。那么我们可以定义基本块 m 到目标基本块的 T_b 的基本块级目标距离 $d_b(m, T_b)$ 为：

$$d_b(m, T_b) = \begin{cases} 0, & \text{if } m \in T_b \\ c \cdot \min_{n \in N(m)} (d_f(n, T_f)), & \text{if } m \in T \\ [\sum_{t \in T} (d_b(m, t) + d_b(t, T_b))^{-1}]^{-1}, & \text{otherwise} \end{cases} \quad (3-2)$$

式3-2中的 c 为常数，在实际中取 $c = 10$ 。

在通过利用 CG 和每个函数的 CFG 计算得到程序中所有基本块的基本块级目标距离 $d_b(m, T_b)$ 后，AFLGo 就将该数值通过第二次对程序的编译插桩进所有基本块中。至此，AFLGo 的预处理阶段就结束了。

3.1.2 能量调度机制

在经过了预处理之后的二进制文件在被 AFLGo 测试时会收集执行测试的每个种子 s 的目标距离 $d(s, T_b)$ 。设种子 s 的执行轨迹为 $\xi(s)$ ，那么我们可以定义衡量种子执行与目标基本块 T_b 远近的种子目标距离 $d(s, T_b)$ 为：

$$d(s, T_b) = \frac{\sum_{m \in \xi(s)} d_b(m, T_b)}{|\xi(s)|} \quad (3-3)$$

有了种子的目标距离，也就很好相应的定义种子的归一化目标距离用以作为种子能量分配调度的评判指标。针对种子池 S 中所有种子，可以在种子 $s' \in S$ 执行时计算种子 s' 对应的归一化目标距离：

$$\tilde{d}(s, T_b) = \frac{d(s, T_b) - \min D}{\max D - \min D} \in [0, 1] \quad (3-4)$$

其中

$$\min D = \min_{s' \in S} [d(s', T_b)], \max D = \max_{s' \in S} [d(s', T_b)] \quad (3-5)$$

AFLGo 的整体能量调度思路就是通过给种子目标距离更近的种子分配更多的能量，从而实现对目标的导向覆盖。这一思路来源于对模糊测试的马尔科夫链建模，这部分已经在章节2.2.2做了详细的介绍。在 AFLGo 中引入了模拟退火机制来进一步增加对能量调度的导向性，以期处理探索-开发问题。

模拟退火算法是一种基于概率思想的全局优化算法，常用于解决组合优化问题。该算法基于统计力学中的退火过程，通过随机搜索来避免陷入局部最优解，从而寻找全局最优解。与经典的随机搜索方式相比，模拟退火的调度方式会以一定的概率接受较差的解，而这有使得模拟退火调度方式有跳出局部最优解达到全局最优解的可能性。接受较差解的概率由参数“温度”控制，随着时间的推延，温度逐渐下降能量调度也就越发降低接受较差解的可能性。随着温度最终的降低，模拟退火将找到一个全局的近似最优解。

在 AFLGo 中，初始温度 $T = T_0 = 1$ 。温度 $T = 1$ 时，AFLGo 对能量的调度更倾向于平均分配，不受种子目标距离的影响。此时是处于 AFLGo 的开发阶段；而当 $T = 0$ 时，AFLGo 对于能量的调度则更倾向于梯度下降，将最多的能量分配给种子目标距离更小的种子，而将较少的能量分配给种子目标距离较远的种子。此时是处于 AFLGo 的开发阶段。

AFLGo 采取指数型的冷却调度：

$$T = T_0 \times \alpha^k \quad (3-6)$$

其中 α 是一个小于 1 的常量，通常介于 0.8 到 0.99 之间； k 是当前温度循环次数。由于温度决定了能量分配的策略，于是将 $T_k = 0.05$ 设置为温度阈值。当 $T_k \leq 0.05$ 时，算法不再接受更差的解，进入开发阶段。

将模糊测试已经执行的时间 t 作为调整温度循环 T_k 的参数，公式如下：

$$\frac{k}{k_x} = \frac{t}{t_x} \quad (3-7)$$

其中 k_x 是温度到达阈值 T_k 时的温度循环次数； t_x 是此时模糊测试已经执行的时间。在 AFLGo 中，达到温度阈值时模糊测试已经执行的时间 t_x 由测试者预设，是已知量。

结合温度阈值 $T_k = 0.05 = \alpha^{k_x}$ ，那么要想计算某运行时刻 t 和温度 T 的关系。只需带入式3-7，有：

$$T = \alpha^k = \alpha^{\frac{t}{t_x} \times \frac{\log(0.05)}{\log(\alpha)}} = 20^{-\frac{t}{t_x}} \quad (3-8)$$

接下来，使用温度结合种子目标距离来进行能量调度。对给定的种子 s ，可以定义其被分配的能量分配因子 $p(s)$ 为：

$$p(s) = (1 - \tilde{d}(s, T_b)) \cdot (1 - T) + 0.5T \quad (3-9)$$

式中的 $\tilde{d}(s, T_b)$ 即由式3-4计算得到; 当前温度 T 由式3-8计算得到。

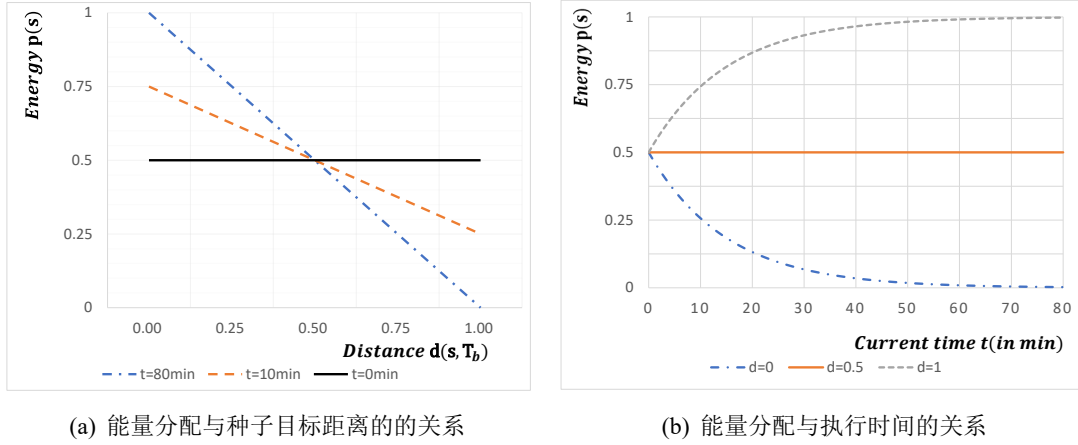


图 3.3 模拟退火能量分配方式

则最终的种子被分配能量 $\hat{p}(s)$ 为:

$$\hat{p}(s) = p_{aff}(s) \cdot 2^{10 \cdot p(s) - 5} \quad (3-10)$$

式中的 $p_{aff}(s)$ 是原有框架 AFL 为种子赋予的原始能量。

图3.3展示了在设置时间阈值 $t_x = 45$ 时模拟退火调度下种子能量与归一化种子目标距离 $d(s, T_b)$ 和模糊测试已经执行时间 t 之间的关系。其中 (a) 图展现了能量分配机制在不同的已执行时间条件下针对不同的归一化种子目标距离的种子的能量分配方式; (b) 图展示了不同的归一化种子目标距离的种子随着执行时间的增加分配能量的变化。

3.2 定向适应度指标设计

AFLGo 的能量调度策略以及设计方式奠定了 DGF 的基本流程框架。但是, 在 DGF 关于距离的定义机制应该是公平且不带有偏见性的。也即, DGF 应该有一个很好的基于距离的机制, 可以通过考虑到目标的所有轨迹并避免对某些轨迹的偏见来指导定向模糊测试。此外, 作为灰盒模糊测试, DGF 应该在静态分析的开销和实用程序之间取得平衡。

出于以上原则, 实际上上述 AFLGo 的距离设计和能量调度的机制会具有一定的偏见性。这也是本文针对 AFLGo 设计改进的基础, 将在以下小节分别阐述。

3.2.1 距离定义

让我们回到 AFLGo 对于函数级距离的定义并重新审视它: 函数级距离 $d_f(n, n')$ 定义为调用图 CG 中函数 n 和 n' 之间最短路径上的边数。事实上这一定义中包含了一项默认信息, 即从一个函数 n 到另一个函数 n' 的单位距离是调用图中相邻函数之间的一条边, 即单位“1”。

这样的默认条件只考虑了函数间的调用关系。我们再回到 AFLGo 对模糊测试过程的马尔科夫链建模过程: 种子变异导致的程序执行路径转移概率, 为马尔可夫链上的状态转移概率。那么函数间的函数级距离事实上也应该反映函数间互相调

用（状态转移）的概率。

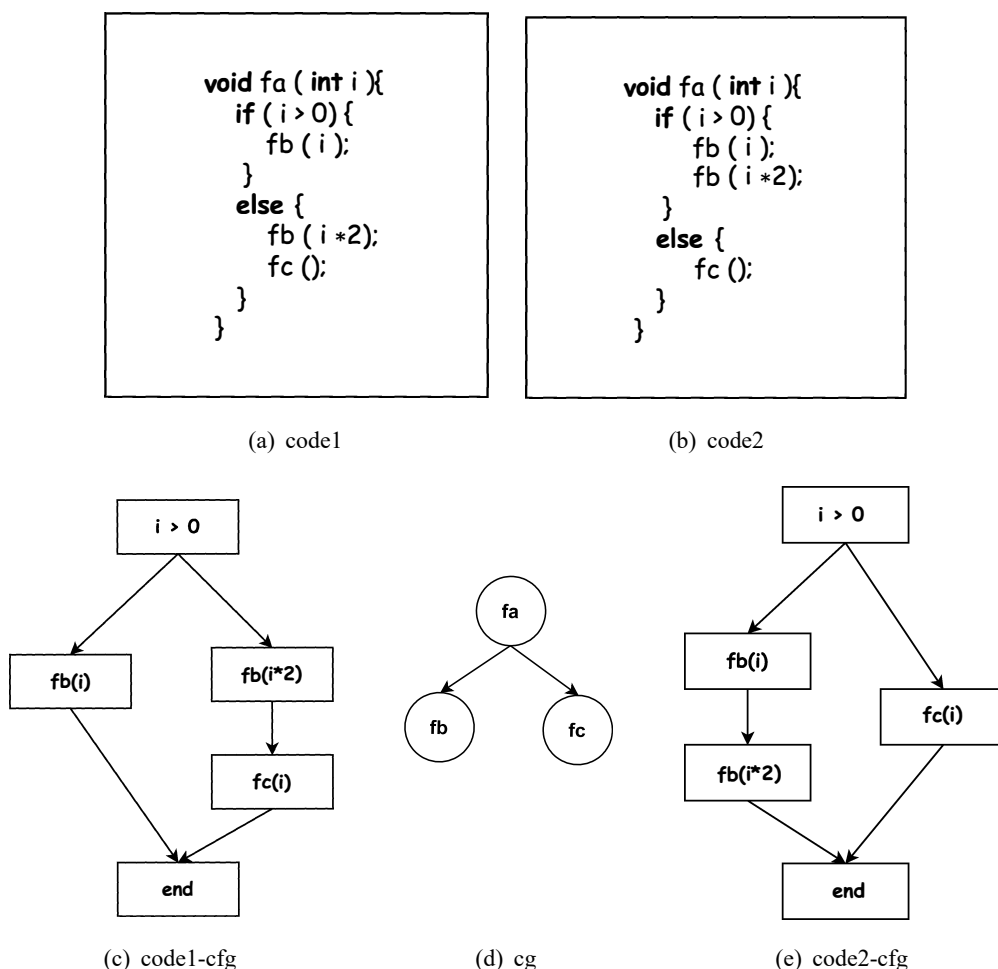


图 3.4 代码示例

图3.4展现了这样的一个例子：(a) 和 (b) 是两个不同的函数调用代码，但是它们的调用图 CG 都如 (d) 图所示一样。但是如果仔细追究其所对应的控制流图 CFG，也即 (c) 和 (e) 时，可以知道在这个两个代码示例中，函数 fa 对函数 fb, fc 的调用概率（状态转移概率）是不同的。在 (a) 的情况下 fa 必然会调用 fb ，而不一定调用 fc ；在 (b) 的情况下 fa 并不一定会调用 fb 。但是仅凭 (d) 中的调用图 CG 是无法反映这样的区别的，也即这两种状况下被计算的得到的函数级距离是相同的。我们期望的理想状态应该是在 (a) 的情况下， $fa \rightarrow fb$ 的距离应当小于 (b) 情况下的 $fa \rightarrow fb$ 的距离。

这就使得我们得重新定义一个更精确的函数级距离。因此，我们采取使用两个额外的指标来增加由调用函数（称作“caller”）和被调用函数（称作“callee”）之间的直接调用关系定义的距离。

- (1) 给定调用函数 caller 中针对某个被调用函数 callee 的调用点的出现次数 C_N 。被调用函数 callee 的出现次数越多，就越有可能在执行时被调用函数 caller 调用，从而使调用函数 caller 到被调用函数 callee 之间的距离变小。为了显

示这一效果，定义一个因子：

$$\Phi(C_N) = \frac{\varphi \cdot C_N + 1}{\varphi \cdot C_N} \quad (3-11)$$

其中 φ 为常量，一般取 $\varphi = 2$ 。

- (2) 调用函数 **caller** 中至少包含一个被调用函数 **callee** 调用点的基本块的个数 C_B 。随着调用函数中更多分支 (或者说，基本块) 具有调用点，也就意味着更多不同的执行路径将包括被调用函数。为了显示这一效果，定义一个因子：

$$\Psi(C_B) = \frac{\psi \cdot C_B + 1}{\psi \cdot C_B} \quad (3-12)$$

其中 ψ 为常量，一般取 $\psi = 2$ 。

现在，有了这两个影响因子，我们可以重新定义函数级距离 $d(f_1, f_2)$ ：

$$d(f_1, f_2) = \Psi(f_1, f_2) \cdot \Phi(f_1, f_2) \quad (3-13)$$

注意这里的 f_1, f_2 是调用函数与被调用函数的关系。我们再以图3.4中的例子来做验证。对于 (a) 中的情况， $d(fa, fb) = \frac{2 \cdot 2 + 1}{2 \cdot 2} \cdot \frac{2 \cdot 2 + 1}{2 \cdot 2} = 1.56$ ， $d(fa, fc) = \frac{2 \cdot 1 + 1}{2 \cdot 1} \cdot \frac{2 \cdot 1 + 1}{2 \cdot 1} = 2.25$ 。对于 (b) 中的情况， $d(fa, fb) = \frac{2 \cdot 2 + 1}{2 \cdot 2} \cdot \frac{2 \cdot 1 + 1}{2 \cdot 1} = 1.87$ ， $d(fa, fc) = \frac{2 \cdot 1 + 1}{2 \cdot 1} \cdot \frac{2 \cdot 1 + 1}{2 \cdot 1} = 2.25$ 。

可以看到，这一机制成功的区分了这样的概率偏差导致的函数级调用距离的不同。为此，在 AFLGo 中的函数级目标距离（也即式3-1），基本块级目标距离（也即式3-2）中函数间的距离都要以此距离为基础做修改。

3.2.2 可达目标函数集覆盖率

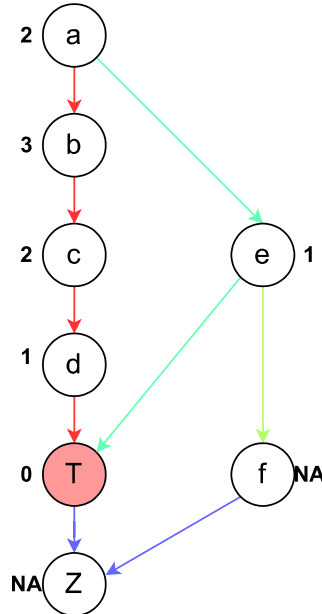


图 3.5 种子路径示例

让我们再重新审视种子目标距离的定义，即式3-3。其采取的计算方式是种子路径经过的基本块的基本块级目标距离总和除以经过的基本块数。注意，由于轻量化插桩实现的机制，在实际计算过程中，种子目标距离计算时除以的是种子路

径上经过的基本块级目标距离大于等于 0 的基本块总数。

在图3.5中提供了这样的一个例子。这是程序的部分调用图 CG。为了简单起见，将每个函数中的基本块假设为 1，那么我们就用函数级的目标距离来代替基本块级目标距离来计算种子的目标距离。如图所示，假设 T 为目标函数，其余每个函数的目标距离已计算得出并在图中标出，其中“NA”为不可达目标函数。假设三个种子的执行路径分别如下：

- 种子 S1: $a \rightarrow b \rightarrow c \rightarrow d \rightarrow T \rightarrow Z$
- 种子 S2: $a \rightarrow e \rightarrow T \rightarrow Z$
- 种子 S3: $a \rightarrow e \rightarrow f \rightarrow Z$

那么依据式3-3，可以分别计算出种子目标距离 $d(S1, T) = \frac{2+3+2+1+0}{5} = 1.6$, $d(S2, T) = \frac{2+1+0}{3} = 1$, $d(S3, T) = \frac{2+1}{2} = 1.5$ 。

按照 DGF 关于距离的定义机制是公平且不带有偏见性的原则，我们显然可以看出 AFLGo 设计的种子目标距离是具有一定的偏见性的：即 AFLGo 更倾向于选择路径更短的种子。我们可以看到，能够到达目标函数的长路径种子 S1 的距离甚至比不能到达目标函数的短路径种子 S3 要远！这也意味着，如果实际漏洞存在于较长的执行路径的话，AFLGo 的偏向性会使得其花费更多的时间资源来探查到此漏洞。这显然与 DGF 的目标不符，也与设计原则冲突。究其原因，是 AFLGo 采取了最短路径算法的距离定义和能量调度策略使得 AFLGo 倾向于更快地到达目标，而不是尽可能的覆盖测试所有可能到达目标的路径。

于是我们可以补充定义一个可达目标函数集覆盖率的指标来引导种子尽可能覆盖更多可能到达目标的路径。首先我们引入可达目标函数集 $\zeta(T_f)$ 的概念， $\zeta(T_f)$ 是被测试程序中所有可以到达目标函数集 T_f 中任一函数 t_f 的函数集。

算法 3-2: 计算可达目标函数集的不动点算法

```

Input:  $\mathbb{T}$  // a finite set of targets
Output:  $\mathbb{R}$  // a finite set of target-reachable functions

1  $\mathbb{R} \leftarrow \emptyset$ ;
2  $Queue \leftarrow \mathbb{T}$ ;
3 while !IsEmpty( $Queue$ ) do
4    $Function \leftarrow Dequeue(Queue)$ ;
5   for  $func \in CallGraph$  do
6     if IsCall( $func, Function$ ) then
7       // to be unique
8       if  $func \in Queue$  then continue;
9       Enqueue( $Queue, func$ );
10       $\mathbb{R} \leftarrow \mathbb{R} \cup func$ ;
10 return  $\mathbb{R}$ ;

```

算法3-2展示了一个简单的依据调用图 CG 计算 $\zeta(T_f)$ 的不动点算法。利用此算

法我们可以计算得到可达目标函数集，由此可以定义种子的可达目标函数集覆盖率：

$$C_s(s) = \frac{|\xi(s) \cap \zeta(T_f)|}{|\zeta(T_f)|} \quad (3-14)$$

由此我们就得到了可达目标函数集覆盖率。不过由于在实际测试中， $\zeta(T_f)$ 一般为固定大小，故我们实际采用可达目标函数集覆盖量 $r(s) = |\xi(s) \cap \zeta(T_f)|$ 来做衡量即可。

在引入可达目标函数集覆盖量这一指标后，我们通过对其进行归一化处理得到 $\tilde{r}(s)$ ，将其引入能量调度公式（即式3-9），得到更新后的能量调度公式：

$$p(s) = \tilde{r}(s) \cdot (1 - \tilde{d}(s, T_b)) \cdot (1 - T) + 0.5T \quad (3-15)$$

3.3 本章小结

本章详细介绍了本文重点研究的灰盒模糊测试工具 AFLGo 及其框架结构。通过距离计算机制和能量调度机制详细分析了 AFLGo 实现对被测试程序的定向策略。而后，本章结合示例指出其设计机制上的不合理之处，并针对性的进行了定向指标的补充设计和修改，包括距离的重新定义和可达函数集覆盖率的设计。

第四章 基于 AFLGo 的定向模糊测试系统的实现

4.1 需求分析

4.2 架构设计

4.3 静态分析器的改进

4.4 定向模糊测试工具

4.5 本章小结

第五章 系统测试

5.1 系统测试概述

5.1.1 系统测试目标

5.1.2 系统测试环境

5.2 功能测试

5.3 实验评估

5.4 本章小结



第六章 总结与展望

6.1 总结

6.2 展望

结束语

致 谢

本论文采用 \LaTeX 模版编写的，是基于南京邮电大学 2021 年理工艺教类的 Word 模板进行严格迁移编写的。本模板地址<https://github.com/dhiyu/NJUPT-Bachelor>感谢imguozr(<https://github.com/imguozr/NJUPTThesis-Bachelor>)和lemoxiao(<https://github.com/lemoxiao/NJUPTThesis-Scholar>)的工作，为本模板的形成奠定了大量的基础。

参考文献

- [1] Miller B P, Fredriksen L, So B. An empirical study of the reliability of unix utilities[J]. Communications of the ACM, 1990, 33(12): 32-44.
- [2] Manès V J, Han H, Han C, et al. The art, science, and engineering of fuzzing: A survey[J]. IEEE Transactions on Software Engineering, 2019, 47(11): 2312-2331.
- [3] Wang P, Zhou X, Lu K, et al. The progress, challenges, and perspectives of directed greybox fuzzing: arXiv:2005.11907[EB/OL]. arXiv, 2022.
- [4] Cadar C, Dunbar D, Engler D R. Klee: unassisted and automatic generation of high-coverage tests for complex systems programs[C]//Proceedings of the 8th USENIX conference on Operating systems design and implementation. 2008: 209-224.
- [5] Ganesh V, Leek T, Rinard M. Taint-based directed whitebox fuzzing[C]//2009 IEEE 31st International Conference on Software Engineering. IEEE, 2009: 474-484.
- [6] Ma K K, Yit Phang K, Foster J S, et al. Directed symbolic execution[C]//Static Analysis: 18th International Symposium, SAS 2011, Venice, Italy, September 14-16, 2011. Proceedings 18. Springer, 2011: 95-111.
- [7] Marinescu P D, Cadar C. Katch: High-coverage testing of software patches[C]//Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering. 2013: 235-245.
- [8] Jin W, Orso A. Bugredux: Reproducing field failures for in-house debugging[C]//International Conference on Software Engineering. 2012.
- [9] Böhme M, Pham V T, Nguyen M D, et al. Directed greybox fuzzing[C/OL]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas Texas USA: ACM, 2017: 2329-2344. DOI: 10.1145/3133956.3134020.
- [10] Heartbleed - a vulnerability in openssl[EB/OL]. [2023-05-13]. <https://heartbleed.com/>.
- [11] Chen H, Xue Y, Li Y, et al. Hawkeye: Towards a desired directed grey-box fuzzer[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018: 2095-2108.
- [12] Wang H, Xie X, Li Y, et al. Tpestate-guided fuzzer for discovering use-after-free vulnerabilities [C]//Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. 2020: 999-1010.
- [13] Lee G, Shim W, Lee B. Constraint-guided directed greybox fuzzing.[C]//USENIX Security Symposium. 2021: 3559-3576.
- [14] Huang H, Guo Y, Shi Q, et al. Beacon: Directed grey-box fuzzing with provable path pruning [C]//2022 IEEE Symposium on Security and Privacy (SP). IEEE, 2022: 36-50.
- [15] Zhu X, Wen S, Camtepe S, et al. Fuzzing: a survey for roadmap[J]. ACM Computing Surveys (CSUR), 2022, 54(11s): 1-36.
- [16] Writing an llvm pass[EB/OL]. [2023-05-13]. <https://llvm.org/docs/WritingAnLLVMPass.html/>.
- [17] Libxml2 is the xml c parser and toolkit developed for the gnome project[EB/OL]. [2023-05-13]. <https://gitlab.gnome.org/GNOME/libxml2/-/wikis/home>.
- [18] Householder A D, Foote J M. Probability-based parameter selection for black-box fuzz testing [R]. CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 2012.
- [19] Rebert A, Cha S K, Avgerinos T, et al. Optimizing seed selection for fuzzing[C]//23rd {USENIX} Security Symposium ({USENIX} Security 14). 2014: 861-875.
- [20] Böhme M, Cadar C, Roychoudhury A. Fuzzing: Challenges and reflections[J]. IEEE Software, 2020, 38(3): 79-86.
- [21] Tsankov P, Dashti M T, Basin D. Secfuzz: Fuzz-testing security protocols[C]//2012 7th International Workshop on Automation of Software Test (AST). IEEE, 2012: 1-7.
- [22] Myers G J, Sandler C, Badgett T. The art of software testing[M]. John Wiley & Sons, 2011.
- [23] Beizer B. Black-box testing: techniques for functional testing of software and systems[M]. John Wiley & Sons, Inc., 1995.
- [24] Peach[EB/OL]. [2023-05-13]. <https://github.com/MozillaSecurity/peach>.
- [25] Spike[EB/OL]. [2023-05-13]. <https://github.com/guilhermeferreira/spikepp>.

- [26] Godefroid P, Levin M Y, Molnar D A, et al. Automated whitebox fuzz testing.[C]//NDSS: volume 8. 2008: 151-166.
- [27] Jha S, Jin Z, Lerner S, et al. Sage: A framework for succinct analysis of syntactic greedy equivalence[J]. arXiv preprint arXiv:1611.06279, 2016.
- [28] american fuzzy lop[EB/OL]. [2023-05-13]. <https://lcamtuf.coredump.cx/afl/>.
- [29] King J C. Select—a formal system for testing and debugging programs by symbolic execution [C]//Proceedings of the 5th international conference on Software engineering. IEEE Press, 1976: 238-249.
- [30] Bhme M, Pham V T, Roychoudhury A. Coverage-based greybox fuzzing as markov chain[J]. IEEE Transactions on Software Engineering, 2016.
- [31] Chen H, Li Y, Chen B, et al. Fot: A versatile, configurable, extensible fuzzing framework[C]// Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2018: 867-870.