# 关于二进制文件比对技术优化的研究

石永生
江苏信息职业技术学院　江苏无锡　214153

**摘　要：**二进制文件比技术用于评估两个版本不同或编译环境不同的二进制文件之间的差异，所选择的比较粒度一般是基本块粒度或函数粒度。这种方法对于软件工程领域和安全领域十分重要，可以用于分析软件以识别是否有恶意代码插入或发现软件版本之间的差异。本文简要介绍了当前业界流行的二进制文件比技术以及常见的分析流程。二进制文件比对技术一般分为两步，第一步使用二进制文件相似性检测工具以得到基本块或函数之间的相似度，第二步则基于基本块相似度或函数相似度这样的节点信息以及控制流图或函数调用图的图结构信息来进行图匹配的操作，本文重点在于第二部分，选取函数作为粒度进行分析。本文分析了三种图匹配算法，并基于findutils二进制文件对该三种算法进行了探究与实验，并且提出二进制文本比对技术的准确度主要受到二进制文件的符号表和外部函数引用的影响。最后，本文提出了一种结合节点相似性和图结构的新方法，并引入了函数数量更多、图结构更复杂的objdump二进制文件进行实验，最后得出新的算法在一定程度上获得了比之前的技术要好的结果。

**关键词：**二进制文件比对；二进制文件相似性检测；图匹配算法

## 引言

二进文件比对技术目的在于精确衡量两个给定二进制代码之间的相似度，并能在基本块或函数级别上提供详尽的匹配结果。该技术通过精确描述两个输入二进制文件之间在整个程序层面上的差异，从而生成基本块或函数之间最佳的匹配。这种分析方法不仅能够提供关于二进制规模的精确、详细和量化的数据，而且能够清晰地发现代码在不同版本或不同优化级别之间的变化情况。

二进制文件比对技术已在多种重要的安全场景中得到广泛应用，其中包括识别已更改的基本块或函数[1]、对恶意软件进行深入分析[2]、检测文件之间的克隆行为[3]等。在当前的二进制文件比对实践中，尚存在许多需要改进的地方。这些地方包括处理由于代码混淆技术而引起的复杂性、适应不同架构下编译的二进制文件、嵌入模型的泛化性问题，以及整体可扩展性的不足。尤其是在面对大规模二进制文件时，由于图匹配本质上是一个NP-hard问题，快速而准确地找到最佳匹配解决方案变得尤为重要。

在传统的二进制文件比对技术（例如BinDiff[4]）中，分析通常在从静态分析中得到的函数调用图和控制流图上进行。首先提取这些图，然后应用图同构算法和特定的启发式方法来建立图间的映射，从而得到分析结果。这些传统方法主要关注函数或基本块的语法特征而忽视了语义层面，因此在代码被混淆或语法被改变的情况下（如使用不同编译器的优化级别时），其效果可能会受到严重影响。

在机器学习广泛应用的背景下，当前二进制文件比对技术通常使用基于学习的方式来解决二进制相似性检测的问题，并且这也是该文章的重点。典型的基于学习的二进制文件比对技术的流程包括两个阶段：二进制相似性检测和图匹配。二进制相似性检测部分又包含三步，第一步是反汇编，将二进制代码反汇编成人类可解释的汇编代码。第二步是嵌入向量生成，可以使用自然语言处理和图表示学习技术将代码信息嵌入到向量表示中，这些嵌入主要为了包含汇编代码的语义信息以及结构信息。第三步是基于嵌入向量求余弦相似度量化二进制函数或基本块之间的相似性。这里的相似性指的是语义相似性，意味着两个在语法上不同的二进制文件如果具有完全相同的功能，则被视为相似。由于程序的调用关系为一个图结构，程序内的基本块或函数可以被视为图上的节点，因此在下游我们可以将该文件比对问题转化为

**基金项目：**“江苏省高职院校工程技术研究开发中心（苏教科函〔2023〕11号）”和“江苏省示范性虚拟仿真实训基地培育项目（苏教职函〔2023〕30号）”

**作者简介：**石永生（1970—），男，汉族，江苏省无锡市，副教授，教师，硕士研究生，主要研究领域为计算机网络技术和物联网应用技术。

图匹配问题，结合前文的基本块或函数相似度与图匹配算法来解决问题。图匹配算法包括最大权重匹配算法[5]、最大公共边子图算法[4, 6]以及网络对齐算法[7-8]。总的来说，目标是在两个输入的二进制文件之间找到最佳匹配。

## 一、二进制文件相似性检测技术

近年来，二进制文本嵌入和二进制相似性检测的研究领域已经取得了显著的进展。在这一领域内，二进制相似性检测主要是通过计算嵌入的二进制反汇编后的汇编代码的向量的余弦距离来确定的。这些方法在嵌入生成的技术手段上各有侧重，但大体上可以分为三个主要类别：基于自然语言处理的方法、基于图表示学习的方法，以及同时利用这两种技术的综合方法。

基于自然语言处理（NLP）的方法将反汇编的二进制代码视为标记序列，并应用典型的NLP方法如标记嵌入来捕捉序列的语义。这些方法包括InnerEye[9]、PalmTree[10]和Asm2Vec[11]等。Transformer架构的流行也激发了研究者创建更强大的二进制嵌入，这些嵌入能够学习控制流图信息并提高模型的可转移性。例如，jTrans[12]和BinShot[13]等研究通过利用Transformer架构，不仅增强了嵌入模型对控制流图的理解能力，还改进了模型在不同二进制环境中的适应性和准确性。这些进展显著提升了二进制分析工具的性能，使其在面对复杂的代码变化和架构差异时，仍能有效地进行相似性检测和差异分析。

基于图表示学习的方法通常首先从反汇编代码中提取函数调用图和控制流图信息，然后使用图神经网络（Graph Neural Networks）和孪生神经网络（Siamese Neural Networks）来学习嵌入并确定函数或基本块之间的相似性。这些方法包括XBA[14]、Asteria[15]和Gemini[16]等。通过这种方式，研究者可以利用图神经网络捕捉代码的结构特征，而孪生神经网络则有助于比较和量化不同二进制代码间的相似性。这种方法特别适用于分析复杂的程序结构和函数调用关系，提供了一种有效的手段来理解和比较具有不同结构的二进制文件。通过这些技术的应用，二进制文件比对技术在处理大规模和复杂数据时的性能和准确性得到了显著提升。

## 二、图匹配算法

二进制文件比对工具一般包含两个部分，第一个部分是二进制代码相似性检测，在上文已经提及，而第二部分则是图匹配算法，由于每一个二进制文件，即每一个程序具有很明显的函数调用图和控制流图，因此我们可以将二进制文件比对问题转变为图匹配问题。图匹配算法一般分为以下三类：最大权重匹配算法[5]、最大公共边子图算法[4, 6]，以及网络对齐算法[7-8]，如图1可以明显的看出来这些算法的区别。

最大权重匹配算法（MWM），也被成为二分图匹配算法，改算法为匹配两个不同函数调用图提供了一种简单直接的解决方案。该方法忽略了图本身的结构，仅通过寻找能够最大化匹配函数间相似度总和的一对一映射来进行工作。这种算法可以通过应用匈牙利算法（Hungarian Algorithm）在多项式时间内找到解决方案。另外，最大公共边子图算法（MCS）解决了最大权重匹配可能忽视的问题，即边的匹配。这是一个NP-hard问题，其目标是寻找两个图间对应的节点，使得当这些节点被对齐时，能够最大化重叠边的数量。这种方法被BinDiff[4]和DeepBinDiff[6]等工具采用。这些算法通常是通过迭代扩展已有的部分解，来探索当前已经匹配的节点周围（已匹配函数的调用者或被调用者）的潜在匹配来实现的。然而，这些解决方案通常难以达到全局最优，因为它们容易在局部最优解中停滞不前。最后，网络对齐算法（NAP）是另一种解决图匹配问题的方法，被QBinDiff[7]等工具使用。网络对齐问题的核心是寻找两个给定图的节点间的对应关系，目标是全局最大化它们结构的相似性。这通常涉及到引入额外信息，比如节点的属性或边的权重，以增强对齐效果。在二进制差异分析的场景中，这涉及到利用函数或基本块的语义信息以及图的结构信息。此类算法的大多数研究都集中生物化学面，如用于解决蛋白质匹配问题的IsoRank[8]算法。这些方法提供了一种复杂但有效的方式来处理二进制文件的分析和对比，尽管在处理时可能面临计算上的挑战。

## 三、算法研究与实验

### （一）数据集构建

为了评估不同图匹配算法在跨版本差异方面的表现，本研究使用了一个广泛使用的二进制集findutils[17]，共包含4个二进制文件。收集了从4.2.33版本到最新的4.9.0版本的多个版本，涵盖了17年的时间跨度。这个二进制集在多个相关研究中被广泛用于评估，因此被认为是进行深入研究的合适起点。这些二进制文件采用GCC v5.4编译，以及为了模拟更真实的即工业界会提供的二进制文件，我们为这些二进制文件去除了符号表（Symbol Table）和外部函数引用（External Function References）。

为了收集基准真值数据集（Ground Truth Dataset），该处将依赖源代码级别的匹配信息以及符号表信息（Symbols Info）和调试信息（Debug Info），因为本研究选

择函数级别粒度，因此最后得到的是两个二进制文件中的函数的匹配。具体来说，针对两个输入的二进制文件，首先基于反汇编工具以及文件所包含的信息，可以从二进制中提取源文件名称以及函数名称，之后匹配函数名一致的函数，这些匹配作为初步的基准真值集合。接下来，将人工检查每一对匹配的函数对的源代码，排除那些虽然函数名相同但语义含义有所变化的函数对。为确保所提取基准真值数据集的准确性（Accuracy）和完备性（Soundness），我们具体采取了以下策略：（1）仅收集函数名完全一致的函数作为匹配对，忽视那些函数名经过修改但是代码内容相似的函数；（2）通过人工检查严格去除那些函数名完全一致但在执行结果上有所差异的匹配对。值得注意的是，人工检查过程具有一定的主观性，因此为了最大限度地保证基准真值数据集的准确性，我们认为即使执行结果一致，但是代码不完全一致，则不认为是匹配的。此外，本工作在技术上使用IDA Pro v7.6[18]作为反汇编工具，并且在函数相似性测量上统一采用了jTrans[12]的函数嵌入技术和余弦距离算法。这种方法的选择是因为本工作的焦点主要集中在分析不同图匹配算法的有效性和准确性上。

**（二）评估指标**

本研究采用精确度（Precision）和召回率（Recall）这两个指标来衡量图匹配算法生成的结果的有效性。图匹配算法产生的匹配结果M可以表述为一组函数匹配对，这些匹配对的总数为x。同理，两个二进制文件的基准真值数据集G也可以表述为一组函数匹配对，其总数为y。对于匹配结果中与基准真值相交的部分，我们将其定义为$Mc=M \cap G$。由于基准真值集可能不完整，存在一些匹配对Mu，这些对中的函数的任意一个在基准真值中并未出现。这种情况可能由于基准真值收集过程的保守性所导致。因此，从M中剔除Mu后，剩余的部分M-Mu被确认为阳性匹配对的总和。在评估过程中，精确度是指正确识别的匹配对占图匹配算法提出的所有匹配对的比例，而召回率则是指正确识别的匹配对占基准真值数据集中所有匹配对的比例。这两个指标共同评价了图匹配算法在实际应用中的准确性和完整性。通过这种方式，我们能够更全面地理解不同算法在处理实际二进制文件比对时的表现。最后，我们在以下方程中定义精确度（Precision）、召回率（Recall）和F1分数（F1-Score）分别对应公式1，2和3.

**结束语**

在本文中，我们深入研究了二进制文件比对技术的核心问题，并提出了一种新的方法，通过结合节点相似性和图结构分析来优化比对过程。本研究的主要目标是探索其他因素对该技术的影响以及影响程度，并且提出一种新的方法用于提高二进制文件比对技术的准确度和效率。通过一系列详尽的实验，我们观察到不同图匹配算法对二进制文件比对技术的影响各异。我们还发现，移除符号表（Symbol Table）和外部函数引用（External Function References）对传统技术的影响明显大于对基于学习的算法的影响。

基于上述研究我们提出了结合不同算法优点的Seed-and-extend算法，使用多个不同的数据集进行测试，结果显示，相比于现有技术，我们的方法在准确度和效率上都有显著提升，这些实验验证了方法的实际应用潜力。此外还提出了优化二进制文件比对技术的重点应该在于提升二进制相似性检测技术上，并且搭配更简单的匹配算法，以实现其在大规模的文件上依旧保持较高分析效率。本研究的成果不仅推动了二进制文件比对技术的理论和实践发展，也为相关领域的研究人员和实践者提供了新的工具和方法。未来，我们将继续探索更多创新的技术，以解决二进制比对中尚未充分解决的问题，特别是在新兴的编程语言和平台上的应用。

**参考文献**

[1]Diffing-with-kam1n0[Z]. https://www.whitehatters.academy/diffing-with-kam1n0/.

[2]FARHADI M R, FUNG B C M, CHARLAND P, et al. BinClone: Detecting Code Clones in Mal- ware[J]. 2014 Eighth International Conference on Software Security and Reliability, 2014: 78-87.

[3]LUO L, MING J, WU D, et al. Semantics-based obfuscation-resilient binary code similarity com- parison with applications to software plagiarism detection[J]. Proceedings of the 22nd ACM SIG- SOFT International Symposium on Foundations of Software Engineering, 2014.

[4]BinDiff[Z]. https://www.zynamics.com/bindiff.html.

[5]KUHN H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics (NRL), 1955, 52.