ОТЗЫВ

официального оппонента, кандидата технических наук
Павленко Евгения Юрьевича
на диссертационную работу Арутюнян Мариам Сероповны
«Статический анализ исходного и исполняемого кода на основе поиска клонов кода», представленную к защите на соискание ученой степени кандидата технических наук по специальности 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

Актуальность избранной темы диссертации

Статический анализ исходного и исполняемого кода на основе поиска клонов кода является одной из актуальных проблем современного программирования, связанной с анализом кода на предмет повышения качества, надежности и безопасности программного обеспечения (ПО). С увеличением сложности и объема программных систем растет необходимость в автоматизированных методах анализа кода, выявляющих те его фрагменты, которые дублируют друг друга, причем с каждым годом их влияние на сопровождение, оптимизацию и безопасность ПО становится все более заметным.

Поиск клонов кода является важной научно-технической задачей, поскольку наличие клонов затрудняет сопровождение программных систем, увеличивает вероятность появления ошибок и уязвимостей. В частности, клоны могут распространять устаревшие или уязвимые фрагменты кода, что представляет серьезную угрозу для кибербезопасности. Кроме того, выявление клонов играет значительную роль в задачах рефакторинга, в повышении производительности и сокращении объема программного кода.

Важность данной темы также обусловлена тем, что известные методы поиска клонов в исходном и исполняемом коде имеют значимые ограничения: низкую точность обнаружения сложных клонов (например, типа-3) и ограниченную масштабируемость. Разработка новых методов, направленных на достижение повышенной точности, полноты и эффективности анализа, представляет собой актуальную и востребованную научную задачу.

Таким образом, предложенная в диссертации тема обладает значительной практической и теоретической значимостью. В работе также рассматривается применение метода нахождения клонов кода для анализа изменений между версиями программ и для выявления статически связанных библиотек и уязвимостей, что делает исследование Арутюнян Мариам особенно ценным.

Обоснованность научных положений, выводов и рекомендаций, сформулированных в диссертации, их достоверность и новизна

В диссертационной работе представлены новые методы, направленные на решение задачи поиска клонов программного кода, включая:

– унифицированный метод поиска клонов в исходном и исполняемом коде, основанный на графах зависимостей программы;

– метод оптимизации программных реализаций циклической проверки избыточности (ЦПИ) путем поиска и замены неэффективных фрагментов кода;

– двухэтапный метод выявления изменений между версиями программ, интегрирующий метрический подход и разработанный метод поиска клонов кода на основе графов зависимостей управления и данных, обеспечивающий сопоставление функций в формате "многие ко многим" с высокоточной идентификацией совпадений на уровне программных инструкций;

– методы идентификации статически связанных библиотек и обнаружения клонов известных уязвимостей, с использованием разработанного метода поиска клонов фрагментов кода.

Предложенные подходы представляют собой значительный вклад в область анализа программного кода и обеспечения его безопасности.

Достоверность и обоснованность основных выводов и результатов диссертации

Достоверность и обоснованность результатов исследования подтверждается:

– выполненным сопоставлением научных положений и теоретических выводов с практическими результатами проведенных экспериментальных исследований;

– апробацией разработанных автором подходов на реальных программных системах, что подтверждает их практическую ценность и применимость;

– публикацией результатов в рецензируемых научных изданиях и их освещении в докладах на международных конференциях, что свидетельствует о признании научных результатов в профессиональном сообществе.

Теоретическая и практическая значимость

Теоретическая значимость работы заключается в разработке новых алгоритмов анализа кода, обладающих высокой точностью при идентификации клонов. Полученные результаты могут быть использованы при разработке новых

методов статического анализа программ и в целом расширяют область знаний в части программного обеспечения вычислительных систем, комплексов и компьютерных сетей.

Практическая значимость полученных результатов подтверждается внедрением созданных методов в программные системы **GenesISP, BinSide, GCC**, а также их применимостью в рамках российских и международных стандартов по безопасности программного обеспечения. Внедрение данных методов позволяет значительно повысить надежность программных систем и минимизировать риски, связанные с уязвимостями.

Характеристика опубликованности результатов и положений, выносимых на защиту

Основные результаты диссертационной работы Арутюнян М. С. докладывались на следующих конференциях: ежегодная научная сессия СНО ЕГУ 2016 (Ереван, Армения, 2016 г.), XIII Годичная научная конференция Российско-Армянского университета (Ереван, Армения, 2018 г.), международная конференция «Иванниковские чтения» (Великий Новгород, Россия, 2019 г.), XIV Годичная научная конференция Российско-Армянского университета (Ереван, Армения, 2019 г.), XXVIII Международная конференция студентов, аспирантов и молодых ученых «Ломоносов» (Москва, Россия, 2021 г.), международная конференция «Иванниковские чтения» (Нижний Новгород, Россия, 2021 г.), международная конференция «GNU Tools Cauldron 2023» (Кембридж, Великобритания, 2023 г.), международная конференция «VALID 2024» (Венеция, Италия, 2024 г.), международная конференция «FOSDEM 2025» (Брюссель, Бельгия, 2025 г.).

Основные результаты по теме диссертации изложены в 12 печатных работах, в том числе, в четырех (4) научных статьях в рецензируемых журналах, входящих в перечень рекомендованных ВАК РФ, в шести (6) работах, индексируемых в международной базе цитирования Scopus (две из которых опубликованы в журналах, входящих в первый квартиль SJR), также получено два свидетельства о государственной регистрации программ для ЭВМ.

Оценка структуры и содержания работы

Диссертационная работа состоит из введения, пяти глав, заключения, списка литературы из 202 наименований и приложения. Общий объем работы составляет 133 страницы, в том числе 27 рисунков и 24 таблицы.

Во введении автором обоснована актуальность темы диссертационного исследования, сформулирована его цель, определены основные понятия и научные задачи, сформулирована научная новизна и практическая значимость результатов, а также положения, выносимые на защиту.

Первая глава диссертации содержит обзор современных научно-технических работ по теме исследования, а также приведены необходимые определения терминов. Автором детально исследованы методы и инструменты поиска клонов, включая методы на основе текста, токенов, деревьев и графов, а также метрик. Выполнено обоснованное сравнение методов, отмечен недостаток, присущий большинству современных методов нахождения клонов, связанный со снижением значений точности и полноты для клонов типа-3. Автором также проанализированы работы по нахождению клонов и сравнению программ, а также инструменты идентификации статически связанных библиотек или их функций в исполняемых файлах.

Во второй главе автором описан унифицированный метод поиска клонов произвольных фрагментов кода в исходном и исполняемом коде, основанный на графах, содержащих зависимости управления и зависимости данных программы. Отмечается, что и фрагмент кода, и целевая программа могут быть представлены как в виде исходного кода, так и в виде исполняемого кода. Автором также уделено внимание аспектам практической реализации метода с использованием инструмента FCD и описана тестовая система для оценки качества инструментов нахождения клонов кода, которая автоматизирует генерацию тестов и вычисление нескольких метрик. Представлены результаты экспериментальных исследований инструмента FCD, полученные с использованием описанной тестовой системы и демонстрирующие высокую точность и полноту, а также масштабируемость инструмента. Автором выполнено сравнение инструмента FCD с близкими по тематике научными работами, а также выполнена оценка инструмента FCD на базе объемного набора данных BigCloneBench.

Третья глава содержит описание метода оптимизации программ, использующих вычисление циклической проверки избыточности (ЦПИ или Cyclic Redundancy Check, CRC), при помощи поиска клонов и подстановки эффективных реализаций ЦПИ с учетом аппаратной платформы. Использование символического выполнения в рамках метода позволяет выявить ЦПИ. Проведенные экспериментальные исследования реализованного инструмента продемонстрировали его эффективность в части ускорения работы программ.

В четвертой главе автором представлен двухэтапный метод выявления изменений между версиями программ, который сочетает метрический подход с разработанным методом поиска клонов кода на основе графов, содержащих зависимости управления и зависимости данных, для сопоставления функций в формате «многие ко многим» и отображения измененных инструкций. Метод обеспечивает эффективный поиск изменений, а также минимизацию числа ложных срабатываний.

Пятая глава содержит описание методов идентификации статически связанных библиотек и поиска копий известных уязвимостей с использованием разработанного метода поиска клонов фрагментов кода. Проведенные эксперименты относительно пакетов операционной системы Debian версии 12.6 и репозиториев GitHub продемонстрировали широкое использование устаревшего ПО, что увеличивает риски информационной безопасности.

В заключении содержатся выводы разработанных методов.

Работа носит законченный характер, написана научным стилем, достаточно подробно иллюстрирована. Результаты экспериментальных исследований позволяют сделать вывод о достижении цели, поставленной в диссертационном исследовании.

Исследования по теме диссертации проводились в рамках научных проектов, поддержанных следующими грантами: совместным грантом КН Армении и РФФИ 20RF-033 «Разработка и реализация масштабируемых методов анализа современных операционных систем» и грантом КН Армении 21SCG-1B003 «Разработать и реализовать систему анализа безопасности и сертификации программного обеспечения». Результаты диссертационной работы также были использованы в рамках гранта РФФИ 18-07-01153, «Исследование и разработка методов поиска ошибок на основе метода поиска клонов кода».

Тема диссертационного исследования Арутюнян М. С., направленность проведенных исследований и полученных результатов соответствуют паспорту специальности 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей.

Замечания и рекомендации

Несмотря на высокую научную ценность работы, имеются некоторые замечания:

1. Не вполне ясно, накладываются ли на графы, содержащие зависимости управления и зависимости данных, какие-либо ограничения, значимые для работы двухэтапного метода выявления изменений между версиями программ. Автору также следовало бы продемонстрировать временную оценку для различных по масштабу графов.

2. Следовало бы охарактеризовать полученные автором результаты относительно того, какие типы клонов обнаруживаются более эффективно, а также относительного какого типа клонов демонстрируется наибольший качественный прирост по сравнению с известными методами.

3. В работе недостаточно подробно обосновано, почему уравнения, полученные в рамках символического выполнения, являются простыми и не требуют использования SMT-решателей.

4. В методе выявления изменений между версиями программ считается семь различных хэш-значений. Чем обусловлена необходимость подсчета второго хэш-значения (на основе кодов операций ассемблера) при наличии подсчета первого хэш-значения (на основе инструкций ассемблера)?

5. В тексте диссертации присутствуют ошибки и опечатки, в частности, на стр. 31 написано «хранятся информация», на стр. 71 отмечено «добавленное условие всегда ложно, которую можно определить», на стр. 113 указано «два из находок».

Однако считаю, что отмеченные недостатки не снижают научной значимости работы, так как поставленная в работе цель достигнута. Диссертационная работа содержит детальный анализ предметной области и решаемой проблемы. Решение поставленной задачи отражается в результатах диссертации, выносимые положения аргументированы и обоснованы. Автореферат диссертации полностью отражает содержание диссертационного исследования и позволяет составить целостное представление о работе.

Заключение

Диссертационная работа **Арутюнян Мариам Сероповны «Статический анализ исходного и исполняемого кода на основе поиска клонов кода»** представляет собой является законченную научно-квалификационную работу, направленную на разработку методов статического анализа кода на основе поиска клонов. Работа выполнена на высоком научном уровне, содержит новые теоретические и практические результаты, имеет высокую значимость для развития методов анализа программного кода.

Учитывая актуальность, научную новизну, теоретическую и практическую значимость представленного исследования, а также степень проработки поставленных задач, считаю, что диссертация **соответствует требованиям**, предъявляемым к кандидатским диссертациям, а ее автор **Арутюнян Мариам Сероповна** заслуживает присуждения ученой степени **кандидата технических наук** по специальности 2.3.5 – Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей.

Официальный оппонент:
кандидат технических наук, доцент, доцент Высшей школы кибербезопасности федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого»

«21» марта 2025 г. Павленко Евгений Юрьевич