Empirical Evaluation of Graph-Anonymized Metrics for JIT Defect Prediction

by

Akshat Malik

A thesis submitted to the

School of Computing

in conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

April 2024

Copyright © Akshat Malik, 2024

Abstract

SOFTWARE analytics data, housed in version control and issue report repositories, serves pivotal roles in organizational tasks like predicting code bugs and selecting code reviewers. Yet, individual organizations often lack the breadth of data needed for these analytics, making data sharing essential. However, privacy concerns impede such sharing, as it could expose sensitive organizational information or compromise product quality. Additionally, the risk of reverse-engineering training data from shared models further complicates data privacy.

To address these challenges, anonymization techniques like MORPH, LACE, and LACE2 provide privacy to defect prediction data. While effective, they often sacrifice metric performance due to a disregard for data relationships during anonymization.

To preserve these relationships, we propose graph anonymization techniques. These methods maintain data connections while ensuring privacy. Our research evaluates four specific graph anonymization methods—Random Add/Delete, Random Switch,

k-DA, and Generalization—across six large software projects. We assess privacy using the Increased Privacy Ratio (IPR) metric, finding that each method achieves privacy scores above 65%, with Random Add/Delete and Random Switch exceeding 80% for all projects.

Within-project analyses reveal minimal impacts on model performance, with median reductions of 1.45% in AUC, 5.35% in Recall, 2.29% in G-Mean, and 4.42% in FPR when privacy exceeds 65%. However, when privacy scores surpass 80%, performance declines further. Notably, graph anonymization outperforms tabular methods like MORPH and LACE, which significantly diminish model performance.

In cross-project analyses, graph anonymization maintains model performance with marginal reductions in metrics compared to non-graph methods. Combining graph with non-graph anonymization techniques mitigates performance declines observed with non-graph methods alone, highlighting the efficacy of graph anonymization in preserving privacy while sustaining model performance.

Our study underscores the effectiveness of graph anonymization in maintaining privacy without compromising model performance across within-project and cross-project settings. Even when employing diverse anonymization strategies, graph anonymization techniques consistently deliver robust performance, showcasing their viability in real-world applications.

Acknowledgments

I express my sincere appreciation to Professor Bram Adams and Professor Ahmed E. Hassan for their exceptional guidance and unwavering support throughout my thesis. Professor Adams' insightful feedback and innovative perspectives, alongside Professor Hassan's encouragement and constructive criticism, have been indispensable to the completion of this work. Their expertise and dedication not only shaped this thesis but also fostered my academic and personal growth. I am profoundly grateful for their mentorship and the invaluable lessons learned under their teaching.

I am profoundly grateful to the Natural Sciences and Engineering Research Council of Canada (NSERC) and Mitacs for their generous funding of my research. Their support has been crucial in advancing the work presented in this thesis. My heartfelt gratitude extends to Dominique and the team at BNC for their invaluable support throughout the journey.

I would like to express my sincere gratitude to Dr. Yuan Tian and Dr. Sean Kauffman

for their willingness to join the jury.

Thank you to my friends Harsh, Jay, Jovian, Shenyu, Yi Hao, Jaskirat, Divya, Rahul and all lab mates for the support through the highs and lows of this journey.

I would like to thank my wife, Gareema, for her unwavering support throughout this journey without which I would not have been able to complete this thesis.

I would further like to thank my family who supported me in this long and arduous journey, without their faith in me I would not have been able to complete this journey.

I would like to dedicate this thesis to my Dad.

Co-authorship

FOR all the thesis chapters and related publications, my contributions are designing and refining the research plan; conducting background literature reviews and related work elicitation; collecting datasets; conducting corresponding experiments and data analysis; and writing and polishing the manuscripts.

The works presented in this thesis are listed below:

- Akshat Malik, Bram Adams, and Ahmed E. Hassan. 2024. Towards Graph Anonymization of Software Analytics Data: Empirical Study on JIT Defect Prediction. Journal of Empirical Software Engineering (EMSE). Accepted.

- Akshat Malik, Bram Adams, and Ahmed E. Hassan. 2024. Pragmatic Evaluation of Graph Anonymized Metrics: Empirical Study on Cross-Project JIT Defect Prediction. Journal of Empirical Software Engineering (EMSE). Submitted.

Table of Contents

Abstract i

Acknowledgments iii

Co-authorship v

List of Tables viii

List of Figures xi

1 Introduction 1
- 1.1 Thesis Statement . . . 5
- 1.2 Thesis Overview . . . 6
- 1.3 Thesis Contributions . . . 12

2 Background and Related Work 14
- 2.1 Reverse Engineering Information from ML Models . . . 14
- 2.2 Existing Anonymisation Methods . . . 15
- 2.3 JIT Defect Prediction . . . 17
- 2.4 Data Privacy Concerns . . . 19
- 2.5 Graph Anonymization . . . 20

3 Proposed Graph Anonymization Techniques for Software Analytics Data 23
- 3.1 Step 1: Knowledge Graph Creation . . . 24
- 3.2 Step 2: Graph-Anonymization Algorithms . . . 27
- 3.3 Step 3: Generation of JIT metrics . . . 36
- 3.4 Evaluation of Graph Anonymized Metrics . . . 36

4 Empirical Evaluation of Graph Anonymized Metrics in Within-Project Setting 39
- 4.1 Introduction . . . 39
- 4.2 Empirical Study Design . . . 41

4.3 Empirical Study Results . . . 58
4.4 Discussion . . . 81
4.5 Threats to Validity . . . 83
4.6 Conclusion . . . 86

5 Empirical Evaluation of Graph Anonymized Metrics in Cross-Project Setting 87
5.1 Introduction . . . 87
5.2 Empirical Study Design . . . 88
5.3 Preliminary Study . . . 98
5.4 Empirical Study Results . . . 99
5.5 Conclusion . . . 122

6 Conclusions and Future Work 124
6.1 Thesis Statement . . . 125
6.2 Future Work . . . 126

Bibliography 127

List of Tables

3.1 Entities and their properties and links 26
3.2 Description and the method to derive the JIT defect prediction metrics (used by Kamei et al. [26]) calculated from the git-graph. 38

4.1 Summary of projects selected for the study. 42
4.2 Example for Query Generation with Equal Frequency Binning before and after anonymization. 45
4.3 “x/y" represents a configuration value x and its corresponding number of link changes y, for different IPR Privacy Levels. The cells in green represent the technique achieving that Privacy Level with the least number of links changed for a given project. 62
4.4 Median AUC score with the relative change in AUC from the non-anonymized baseline in parentheses at different Privacy Levels. 65
4.5 Median Recall score with the relative change in Recall from the non-anonymized baseline in parentheses at different Privacy Levels. 65
4.6 Median AUC score at Privacy Level II with the relative change compared to the non-anonymized baseline in parentheses. 77
4.7 Median Recall score at Privacy Level II with the relative change compared to the non-anonymized baseline in parentheses. 77
4.8 Overall statistical test results for each GA technique on 6 datasets compared to all three NGA techniques with at least Privacy Level I (i.e., 18 cases). They are presented as win/tie/loss. 79
4.9 Overall statistical test results for each GA technique compared to the Manual Down technique on the 6 datasets (i.e., 6 cases). They are presented as win/tie/loss. 80

5.1 Summary of projects shared in the replication package. 89
5.2 The JIT defect prediction metrics used in the study (as proposed by Kamei et al. [26]). 90
5.3 Average cross-project performance compared to the average within-project performance of the various datasets (preliminary study). 99

5.4 Performance metrics with relative change from the non-anonymized cross-project baselines, for different techniques, for all configurations where the training data's IPR score was greater than 65%. 101
5.5 Results of the statistical tests conducted between the row techniques (GA) and the other GA techniques for the AUC, Recall, FPR and G-Mean metrics. The results are reported as - "win", where the row technique is significantly higher, "tie", where there is no statistical significance and "loss" where the row technique's value is significantly less. 103
5.6 Results of the statistical tests conducted between the GA techniques and the other NGA techniques for the AUC, Recall, G-Mean and FPR metrics. The results are reported as - "win", where the row technique is significantly higher, "tie", where there is no statistical significance and "loss" where the row technique's value is significantly less. 105
5.7 Change in the AUC metric relative to the non-anonymized cross-project baseline with different anonymization techniques, where the training projects have IPR $\geq$ 80. 107
5.8 Results of the statistical tests conducted between the GA techniques and the MD model for the AUC, Recall, G-Mean and FPR metrics. The results are reported as - "win", where the row technique is significantly higher, "tie", where there is no statistical significance and "loss" where the row technique's value is significantly less. 107
5.9 Median performance and change in performance compared to the non-anonymized cross-project baseline for each technique pair (represented as Technique 1 & Technique 2) per project. 112
5.10 Results of the statistical tests for each technique pair compared to another across all datasets. The results are reported as - win, where the row technique provides a significantly higher value, tie, where there is no statistical significance, and loss where the row technique value is significantly less. 113
5.11 Results of the statistical tests for the AUC metric each ratio of a technique pair compared to all the other ratios across all datasets (6 datasets 5 comparisons with other ratios). The results are reported as - win, where the row technique provides a significantly higher value, tie, where there is no statistical significance, and loss where the row technique value is significantly less. The column headers signify the number of projects anonymized with that particular technique. 114
5.12 Relative change in the AUC metric relative to the non-anonymized cross-project baseline with different anonymization techniques. The last two columns are for the hybrid k-DA and Generalization techniques. 115

5.13 Median Performance with the change in performance from the non-anonymized cross-project baseline for each technique pair (represented as Technique 1 & Technique 2) across all datasets. 117
5.14 Results of the statistical tests for each technique pair compared to another across all datasets. The results are reported as - win, where the row technique provides a significantly higher value, tie, where there is no statistical significance, and loss where the row technique value is significantly lower. 120
5.15 Results of the statistical tests for each technique pair compared to another across all datasets. The results are reported as - win, where the row technique provides a significantly higher value, tie, where there is no statistical significance, and loss where the row technique value is significantly lower. 121

List of Figures

3.1 Example of a Knowledge Graph. The purple nodes are commits, the beige nodes are files and the green node is a person. 25
3.2 Random Add/Delete operation. The nodes in red (Commit) are the ones on which the operation is being carried out. 30
3.3 Random Switch operation. The nodes in red (Commit) are the ones on which the operation is being carried out. 31

4.1 Approach for answering the research questions 48
4.2 IPR scores of all the datasets for Random Add/Delete and Random Switch. The vertical-dotted lines are the two privacy levels. 59
4.3 IPR scores of all the datasets for k-DA and Generalization. The vertical-dotted lines are the two privacy levels. 61
4.4 The number of links changed in Apache Ignite to reach different privacy levels 63
4.5 AUC/Recall vs IPR with Random Add/Delete applied. The horizontal, coloured solid line is the performance of the non-anonymized (baseline) model. The vertical-dotted lines are the different privacy levels. . . 66
4.6 AUC/Recall vs IPR with k-DA applied. The horizontal, coloured solid line is the performance of the non-anonymized (baseline) model. The vertical-dotted lines are the different privacy levels. 69
4.7 AUC/Recall with IPR with Generalization. The horizontal, coloured solid line is the baseline (non-anonymized). The vertical-dotted lines are the different privacy levels. 71
4.8 AUC/Recall versus the number of links changed for Apache Groovy, for all the techniques. 74
4.9 AUC/Recall versus IPR score of Apache Ignite with graph and non-graph anonymization techniques. 78

5.1 General approach for answering the research questions 92

5.2 Violin plots of the performance metrics of different datasets at Privacy Level I. The red-dotted lines and green-dotted lines are the non-anonymized cross-project and within-project performance baseline, respectively. The remaining projects can be found in the replication package [2]. 102
5.3 The heatmaps show the median change in performance that comes with using data anonymized using two different anonymization techniques across all projects. The value of each cell is the median change in performance metric achieved by anonymizing the number of projects specified by the column value using the technique specified in the title, with the remaining anonymized using the row technique. The change percent is calculated with respect to each project's cross-project performance baselines as seen in Table 5.3. 110
5.4 The heatmaps show the median change in performance that comes with using data anonymized using two different anonymization techniques across all projects. The value of each cell is the median change in performance metric achieved by anonymizing the number of projects specified by the column value using the technique specified in the title, with the remaining anonymized using the row technique. The change percentage is calculated with respect to each project's cross-project performance baselines as seen in Table 5.3. 119

CHAPTER 1

Introduction

SOFTWARE analytics has become an essential part of any organization's software development process, where organizations use machine learning models to help teams to (amongst others) analyze the probability of bugs in code [5], the time required to fix them [58] and the causes of build failures [62]. Such models are trained on defect metrics, and development effort metrics and build artifacts using a variety of machine learning (ML) algorithms.

However, to function effectively, these tasks need a large amount of data. This sharing is crucial to have more diversity in data, provide bootstrapping data for newer projects and build better models. However, in an increasing number of cases, organizations are unwilling to share their data and models with each other, even within their

organizations and across different teams. This unwillingness is primarily due to privacy concerns organizations have about the possibility of extracting underlying data from a trained machine-learning model, as demonstrated in recent academic studies [49, 65]. These works demonstrate how, from a trained model, it is possible to reconstruct the properties of the data used to train the model. To achieve this, attackers train a “shadow model" to imitate the behaviour of a trained model on a dataset. Using this, they can determine whether a data point belongs to the training set. Leakage of such knowledge could be highly detrimental to an organization’s security and reputation. For example, revealing that a company fixes a hundred bugs a day while its competitors only fix ten could cause distress among its user base.

One of the most recently studied ways to deal with this threat is to anonymize ML models. There are multiple families of techniques to anonymize ML models, ranging from perturbing the models’ data input, where random noise is added into the training data so that the output is private, to perturbing the actual model training process (e.g., the intermediate values of an iterative ML algorithm using federated learning), or adding noise to the model outputs [3, 16]. Each of these families comprises a variety of specific techniques. This chapter focuses on the first family, i.e., anonymizing the training data, since anonymized data does not limit the ML algorithms that can be used on it, enabling data to be used with any training algorithm, making our approach flexible.

In terms of preserving privacy in data, several general-purpose anonymization techniques such as k-anonymity, l-diversity, t-closeness, personalized privacy and differential privacy have been proposed [32]. To measure how effectively an anonymization technique provides resistance to the data against attacks by nefarious users, privacy

scores can be used to estimate how different the anonymized data is from the non-anonymized data or how effectively information is hidden in the anonymized data. As pointed out by a recent survey [57], over time, several privacy measures have been proposed. For instance, a popular metric for privacy is the Increased Privacy Ratio (IPR) [39]. Based on some background information about the data, it measures what percentage of data has changed after anonymization, measuring how much information is being leaked about the sensitive attributes of the data. An IPR score of, for instance, 80% means that the anonymized dataset in 80% of the cases (or attacks) will result in different, i.e., safer, values for the sensitive attributes than the unanonymized data.

In the world of software analytics, Peters et al. propose the MORPH, LACE and LACE2 [39, 40, 41] algorithms, while Li et al. propose SRDO [28] for anonymization. They empirically validated that it is possible to provide privacy to defect prediction data without affecting its performance. In our replication using JIT defect prediction metrics, we indeed found that the MORPH, LACE and LACE2 algorithms provided high privacy, with scores in the range of 89% to 99%, which exceeds the threshold for adequate data sharing of 65% [41]. However, this high privacy came at the cost of decreased performance of the model, where, across all datasets, the AUC decreased by a median of 21.2%, 17.15% and 21.45%, and Recall by a median of 64.79%, 58.79% and 96.43%, respectively.

In essence, the existing tabular anonymization techniques just treat the software analytics data as a list of metrics, ignoring any logical connection between them. For example, in the case of JIT defect prediction, the number of lines modified by a commit should depend on the number of files modified by that commit. If the number of files modified by the commit changes, it will have a proportional change in the number of

lines changed by the commit. The tabular anonymization techniques mostly ignore such dependencies between the various data features, thereby losing context during anonymization.

A promising direction to address this lack of context is the use of graph anonymization techniques, which perform anonymization on a graph representation of the data. Assuming that the latter graph representation can be used to calculate the metrics of interest, graph anonymization promises to not only retain the structural knowledge of the graph better but also to yield proportional changes to all metrics derived from the graph, preserving the graph's overall utility [22].

While tabular data anonymization for software analytics has been evaluated in the context of traditional defect prediction, graph anonymization techniques have not been considered thus far. Hence, this thesis evaluates the use of graph anonymization techniques on Just-In-Time (JIT) Defect Prediction data, both in a within-project and cross-project setting. The latter setting also covers the case where different types of graph and tabular (or non-graph anonymization NGA) anonymization techniques might be used together. This popular software analytics task focuses on predicting defective commits using only commit-level information like the number of lines added, number of files changed and author experience. JIT defect prediction data is versatile and can be used for different software analytics tasks, like within-project and cross-project settings. We aim to evaluate to what extent graph anonymization can provide privacy to the JIT data while preserving the predictive power of the models trained on it. This would help us establish whether graph anonymization techniques retain the predictive capability of JIT data to be further used in different analytics tasks.

1.1 Thesis Statement

> Graph anonymization techniques are able to provide sufficient privacy to JIT defect prediction metrics both in a within-project and cross-project setting, without compromising their performance.

To evaluate the effect graph anonymization techniques have on the privacy and performance of the data, we adapt four graph anonymization techniques (Random Add/Delete, Random Switch, k-DA anonymity and Generalization) to the knowledge graph of the software analytics data collected for six large, long-lived projects, OpenStack, Qt, Apache Flink, Apache Ignite, Apache Groovy and Apache Cassandra, in a within-project (Chapter-4) and cross-project (Chapter-5) JIT defect prediction setting.

In the first thesis chapter, we explore how the JIT metrics can be modelled as a graph (i.e., git-graph). We then design algorithms for the four graph anonymization techniques that will be applied to the git-graph to provide privacy to it. We then, in a within-project setting, evaluate the change in performance that comes with using these graph anonymization techniques and compare the change in performance with the state-of-the-art techniques. In the second part of this thesis, we evaluate the cross-project performance of the graph anonymization techniques and compare them to the state-of-the-art techniques. In the latter chapter, we further evaluate their use in practical scenarios, where data is anonymized using different graph anonymization techniques.

1.2 Thesis Overview

In this section, we provide an outline of this thesis:

1.2.1 Chapter 2: Background and Related Work

Chapter 2 delves into the privacy concerns of practitioners when sharing models. We then discuss existing ways to provide privacy to software analytics data, as well as the specific use case of JIT defect prediction.

1.2.2 Chapter 3: Proposed Graph Anonymization Techniques for Software Analytics Data

We describe how to represent git metrics as a graph and how graph anonymization techniques work and can be used to provide privacy to JIT defect prediction metrics.

1.2.3 Chapter 4: Empirical Evaluation of Graph Anonymized Metrics in Within-Project Setting

We explore how graph anonymization techniques could be used to provide privacy to JIT Defect prediction metrics without compromising their within-project performance. We first propose a way to represent the commit metrics information, which is used for the JIT defect prediction models, as a graph of interconnected entities called git-graph. Furthermore, we show how from the graph representation we can regenerate the JIT defect prediction metrics.

We propose four graph anonymization techniques, Random Add/Delete, Random

Switch, k-Degree Anonymity and Generalization, to anonymize the entities in the graph so that the graph is provided with privacy without sacrificing its performance.

We perform our empirical study on six large, long-lived projects in a within-project JIT defect prediction setting: OpenStack, Qt [31], Apache Flink, Apache Ignite, Apache Groovy and Apache Cassandra [27]. We answer the following research questions:

1. **RQ1: How do different graph anonymization techniques affect the privacy of the JIT defect prediction data?**

   This question evaluates how effectively different graph anonymization techniques provide privacy to the JIT metrics generated from the anonymized knowledge graphs. Using the privacy metric Increased Privacy Ratio (IPR), it was found that all the techniques are able to provide privacy scores greater than 65% (adequate for sharing data) with minimal effort. Furthermore, Random Add/Delete and Random Switch are able to provide privacy scores greater than 80% to all the datasets, while k-DA and Generalization were able to do so for 3 and 4 datasets, respectively. Random Add/Delete is able to provide privacy scores close to 90% to all the datasets. All techniques change different amounts of links in the graph to reach the same privacy levels, with Generalization doing so with the least changes for 7 out of the 12 combinations of 2 privacy levels (Privacy Level I and II) and 6 datasets, followed by Random Add/Delete (3/12) and k-DA (2/12). Hence, graph anonymization techniques are effective in providing privacy to the JIT metrics.

2. **RQ2: How do different graph anonymization techniques affect the performance of JIT defect prediction models?**

   As the graph anonymization techniques change the graph representing software

analytics data to provide privacy, it is likely that these changes cause the performance of the metrics generated to degrade. With this second RQ, we want to quantify these techniques' impact on the model performance. Our case study shows that privacy does not come at the cost of performance. All techniques provide privacy scores greater than 65% with a median decrease of 1.45% and 5.35% in AUC (Area Under the Curve) and Recall, across all datasets. For 80% and greater privacy scores, the median loss in AUC is 6.44% and in Recall is 20.29%, across all datasets. Generalization performs significantly better than the other techniques in preserving the performance of the model. Graph anonymization techniques are able to preserve the model's performance while providing enough privacy to share the defect prediction data.

3. **RQ3: How do the graph anonymization techniques compare to the state-of-the-art anonymization techniques?**

   Since using graph anonymization techniques in software analytics is a novel approach, we want to compare how our attempt to graph-anonymize JIT defect prediction metrics compares to the state-of-the-art techniques that aim to provide privacy for JIT metrics. Across all projects, we find that MORPH, LACE and LACE2 were able to provide privacy scores greater than 89%; however, their AUC and Recall fell by a median of 21.15% and 80.35%, respectively, compared to the median loss in AUC of 6.44% and Recall of 20.29% for graph anonymization techniques. Furthermore, graph anonymization techniques provide greater control, where scores of 65% and greater (adequate for sharing) can be achieved with a 1.45% decrease in AUC, scores of 80% and greater (higher privacy threshold) can

be achieved with 6.4% decrease in AUC and scores of 88% (with Random Add/Delete) can be achieved with AUC reduced by 7.75%. Lastly, when compared to Manual Down (MD50), the graph anonymization (GA) techniques perform significantly better or similar in the majority of the cases for the AUC, G-Mean and FPR (False Positive Rate) metrics.

Our study shows that it is possible to use graph anonymization techniques to provide privacy while maintaining within-project performance. These techniques offer fine control over the trade-offs between privacy and performance. When compared to state-of-the-art techniques, MORPH, LACE and LACE2 provide more privacy than required for adequate sharing [41], with graph anonymization techniques having a lesser impact on the performance of the metrics.

1.2.4 Chapter 5: Empirical Evaluation of Graph Anonymized Metrics in Cross-Project Setting

While prior work showed the applicability of graph anonymization techniques to provide privacy without sacrificing performance in a within-project defect prediction setting, this chapter extends this work to evaluate how the graph anonymized metrics perform in a more practical setting of cross-project defect prediction, where data of different organizations might have been anonymized using a mixture of graph and non-graph anonymization techniques.

In an empirical evaluation on 77,469 commits of 6 large, long-lived projects, we answer the following research questions:

1. **RQ4.1: Do graph anonymized change metrics retain their performance in a**

cross-project defect prediction setting?

Evaluating the performance of graph anonymized metrics in a cross-project defect prediction setting will help us quantify the change in performance that comes with added privacy to the data. Although the graph anonymized metrics performed well in a within-project setting, performance might degrade in a cross-project setting as anonymization might make the metrics lose their predictive power. We find that graph anonymized metrics are able to provide privacy scores over 65% to all the projects used in the training of a cross-project defect prediction model, making them private enough to share, with a median decrease in AUC, Recall, FPR and G-Mean performance of 1.3%, 4.5%, 3.4% and 1.9%, across all projects. Interestingly, only Random Add/Delete and Random Switch are able to provide privacy scores greater than 80% to all the projects used in training, with a median decrease of 5.1% for AUC and 6% for G-Mean.

2. **RQ4.2: How do graph anonymized change metrics compare to state-of-the-art tabular anonymization techniques and unsupervised techniques when used in a cross-project setting?**

   When compared to tabular or non-graph anonymization (NGA) techniques, like MORPH, LACE and LACE2, it was found that the graph anonymization techniques were able to preserve significantly more model performance, with the NGA techniques decreasing AUC, Recall and G-Mean by a median of 15.5%, 31.5% and 22.1%, and increasing FPR by a median of 7% for privacy scores greater than 65%, across all datasets. The NGA techniques had inconsistent performance varying widely across the different configurations of these techniques, varying from 1.06x to 0.62x as opposed to graph techniques, which vary from 1x to 0.9x.

When compared to Manual Down, an unsupervised learning technique, the GA anonymized metrics perform better or similar for AUC, FPR and G-Mean, with the Gen technique performing better in the majority of the cases.

3. **RQ5: How does the performance of models change when trained using different types of graph anonymization techniques?**

   As graph anonymized data is shared across different organizations, it is not realistic to assume that the data would be anonymized using the same technique. For instance, it is possible that the data used in model training is anonymized using different GA techniques. Our study shows that projects anonymized using two different graph techniques incur an insignificant change in the performance of the trained model, with a median decrease of AUC by 1.9% and G-Mean by 2.5%. The performance of these combined predictors is similar to predictors made using data anonymized with one technique. Furthermore, we find that performance does not change when choosing different projects to be anonymized with different techniques, i.e., any project can be anonymized using any of the techniques.

4. **RQ6: How do predictors built using a combination of graph anonymization and non-graph anonymization techniques work together?**

   While graph anonymization techniques recently have been shown to be useful in privatizing software analytics data, it is highly likely that these GA techniques will have to be used in scenarios where some data is anonymized using GA techniques and some using NGA techniques. We want to evaluate whether the GA techniques can be used with the NGA techniques, and how that impacts the performance of these predictors. In our study, we find that these hybrid predictors

were able to reduce the NGA techniques' performance impact significantly, with the median AUC of these hybrid predictors dropping by 6% instead of 15.5%, and the median G-Mean by 8.8% instead of 22.1%. This shows that the combination of these two different classes of anonymization techniques can actually be beneficial, enabling highly private and performant datasets.

1.3 Thesis Contributions

This thesis explores the usage of graph anonymization techniques to provide privacy to software analytics data and evaluate the change in performance that comes with it. We make the following investigations in this thesis:

- We propose a novel way to represent the JIT defect prediction metrics as a knowledge graph consisting of *Commit*, *Person* and *File* nodes.
- We design algorithms for 4 graph anonymization techniques, *Random Add/Delete*, *Random Switch*, *k-Degree Anonymity* and *Generalization*, which can be applied to the graph representation of the JIT metrics to provide them with privacy.
- Empirically evaluate 56 configurations of 4 graph anonymization techniques to provide privacy to defect prediction data and measure how much privacy they provide to JIT defect prediction metrics.
- Further evaluating the change in performance that comes when using these metrics in a within-project and cross-project setting.
- Empirically comparing tabular anonymization techniques (MORPH, LACE and

LACE2) and unsupervised learning methods (Manual Down) to evaluate the effectiveness of graph anonymization techniques in a within-project and cross-project setting.

- We empirically evaluate how graph anonymization techniques can be used with other graph techniques with a similar decrease in performance to data anonymized using one graph technique.
- We empirically evaluate that graph anonymization techniques can be used with other non-graph anonymization techniques with better performance than a model trained using only non-graph anonymization techniques.

CHAPTER 2

Background and Related Work

In this chapter, we discuss existing works for anonymizing defect prediction data. We also discuss JIT defect prediction and how literature has shown that ML models can leak information. Finally, we give an overview of what is graph anonymization and the different techniques.

2.1 Reverse Engineering Information from ML Models

Fredrikson et al. [14] demonstrate a way of reconstructing the data on which a model is trained based on the confidence score that is assigned to every prediction. Using limited information about the training data and access to the ML model, the authors were able to extract sensitive information about the data used in training decisions.

The authors demonstrate the attacks on image and feature classification problems.

A similar possibility of attack was shown by Shokri et al. [50], who demonstrate how it is possible to extract information about the training data from a black box machine learning model. To do this, the authors use a machine learning model that aims to distinguish differences in model behaviour depending on whether the input was encountered by a model in its training or not. The paper treats the ML model as a black box, only interacting with its API. It was able to perform a membership inference attack on classification models provided by machine learning service providers like Google and Amazon, with an accuracy of 94% and 74%. This highlights the significant privacy implications that ML models can have.

These papers form the basis for our motivation to anonymize the data ML models are trained on to prevent attackers from learning about the sensitive attributes of the underlying software analytics data.

2.2 Existing Anonymisation Methods

Our work extends the work of Peters et al. with MORPH [39], LACE [40] and LACE2 [41] and the works by Li et al. on SRDO [28]. MORPH is an instance mutator algorithm that modifies all data points by a small random amount. The authors further extend this work in LACE, adding an instance pruner algorithm called CLIFF that removes instances based on how frequently an attribute occurs in a class (in defect prediction data, a class would be “defective" or “not defective"). To further enable the sharing of data among different projects, they propose LACE2, a multi-party privacy policy. It allows different data owners to work together to create a privatized cache where each

owner can add data based on what has already been added. Finally, SRDO is a privacy-preserving algorithm that uses a sparse representation based on double obfuscation to improve the techniques underlying MORPH's data mutation.

The fundamental difference between the above works and ours is that the former views data mostly as rows of attributes. However, we represent the data as a graph of interconnected entities, retaining the structural dependencies between the entities. When performing graph anonymization, the metrics derived from the graph should gain privacy without sacrificing their performance, because we believe that the graph anonymization techniques preserve context and information in the graph when altering it. Furthermore, we are evaluating the existing and proven graph anonymization techniques' ability to provide privacy to the graph of software analytics data. To our knowledge, this is the first instance of using graph anonymization on software analytics data to provide privacy.

The prior work [28, 39, 40, 41] also conduct their study on file-based bug prediction, i.e., using CK metrics (Chidamber and Kemerer object-oriented metrics) on the PROMISE dataset. The CK metrics are used to evaluate various aspects of the complexity and quality of object-oriented software systems [8], whereas we use JIT change metrics collected from the GitHub repositories of the projects. Lastly, these state-of-the-art techniques use the data elements' class labels, i.e., buggy or not, when performing anonymization operations to provide privacy, while the graph anonymization techniques do not use class labels, i.e., they are unsupervised.

Another approach is ManualDown [12], an unsupervised learning model that considers the top $x\%$ largest files that have been changed in a current time period as buggy

files. An interpretation of ManualDown in the context of JIT (inspired by other unsupervised JIT approaches [19, 64]) would consider the top $x\%$ commits with the most lines added as buggy. Basically, such a JIT variant of ManualDown would sort all commits in a studied time period by the “number of lines added" feature, then mark the top $x\%$ instances as buggy commits and the remaining ones as not buggy. ManualDown models do not require actual commit data other than the “number of lines added".

Yamamoto et al. [63] propose using federated learning to build a JIT defect prediction model. The study shows that building models using a federated learning approach is possible, not requiring any data to be shared among projects. However, recent works have shown that federated learning models are still open to ML model leaks [6, 37, 43] according to which researchers have been able to perform a membership inference attack. Basically, exploiting the cyclic nature of model learning and the exchange of weights between the different learners and central aggregation nodes. This would expose the un-anonymized training data to malicious users. Essentially, the repeated sharing of model weights via a central node that needs to be trusted by all parties is also the weak point of federated learning approaches from a privacy point of view. It does make attacks harder, but it does not remove the requirement for privatizing the models' input data. We believe that future work can use graph anonymized data on each learner node in a federated learning setting to provide even more privacy to the trained model.

2.3 JIT Defect Prediction

Software in its lifetime encounters a number of bugs, with developers then spending a large amount of time finding and fixing these bugs, which could be used in building

new features. The more time it takes to find a bug, the more associated costs with it. To remedy this problem, the software analytics research area of defect prediction began. Defect prediction aims to find defects early to reduce the time and resources spent on fixing the bugs. Defect prediction models find defective files using static code metrics [33] or commits using process/change metrics [26].

JIT defect prediction is a class of bug prediction techniques that predict bugs at the change or commit level using change or commit metrics. These change metrics are derived from the git history of a project and include metrics like the number of lines changed by a commit, the number of files modified by a commit, and prior developer experience. These metrics allow for fine-grained per-commit level bug prediction as opposed to more traditional bug prediction methods that predict bugs at the file or module level using cyclomatic complexity. JIT defect prediction allows for quicker resolution of bugs as it flags buggy commits, at patch submission time which a developer can respond to quicker [26].

However, to work effectively, these metrics need a large amount of historical data. Such data is unavailable for new or small-scale projects. This lack of data has led to using JIT prediction models for cross-project prediction, where data from one or more projects is used to predict bugs in another project [23, 69]. Cross-project defect prediction models leverage training data from one or more projects to predict buggy commits in a different set of projects. Cross-project models can be built in two ways, where one project is used as training data for finding bugs in another project or where multiple projects are used to find bugs in another project. In their study, Kamei et al. found that a model built using only one project as training data under-performs when compared to within-project prediction performance, while, on the other hand, models trained

using multiple projects' data are able to achieve performance similar to the within-project performance [23]. Privatized data that does not leak information and retains performance can be helpful in improving the applicability of cross-project defect prediction for the industry.

2.4 Data Privacy Concerns

Organizations have serious concerns about the sharing of sensitive data about their projects within teams and across different organizations. Any leak about the sensitive attributes of the data could be detrimental to an organization's reputation. For example, studies have shown that the number of lines changed is closely correlated to the number of bugs in code [26]. In case such information leaks, an organization might be considered as generating more defects, which could be detrimental to an organization's reputation.

Furthermore, organizations are unwilling to share trained ML models with one another, as research has shown that it is possible to extract information about the properties of the training data of a trained ML model. For instance, Yeom et al. [66], showed that overfitting on training data for an ML model can lead to both membership inference and attribute inference attacks. Studies have shown that even with black-box access to models, like classification models provided by Google and Amazon, it is possible to extract information about these models' training data with an accuracy of 94% and 74%, by distinguishing the model's behaviour on whether it encountered the particular element in its training [50]. Fredrikson et al[15] found that model inversion attacks, which train an attack model on the output of the target model, can be as effective as learning in ideal conditions. With some background knowledge about the target,

the model inversion algorithm is able to guess with 22% accuracy the target attribute value, which was almost as good as the regression model that was trained on the actual training data. This makes organizations unwilling to share data and models with one another.

2.5 Graph Anonymization

A graph [59] is a set of nodes (vertices) connected via edges (links). Graphs can be directed, where the relationship goes from one edge to another, or undirected, where the relationship is bidirectional. They provide a unique way to represent interconnected information between different entities. Information in a graph can be extracted by querying the nodes and their links. For example, a graph could link commit information (git information) to the pull request (code review) and, in turn, the report data (bug information) that led to the commit, providing the context for that commit.

Graph anonymization can be performed in two ways, i.e. online or offline. Online techniques anonymize data in real-time, where data is anonymized while querying it. Offline techniques anonymize all of the data together before releasing it for further use. This causes the offline techniques to be slower and computationally more intensive than the online techniques. However, offline techniques are better able to provide privacy and preserve utility as they consider the entire graph, allowing them to consider the overall properties of the underlying analytics data. For this reason, we focus on the offline method, anonymizing all the data at once.

Graph anonymization techniques are considered to be effective in providing privacy to the different entities of a graph while preserving the knowledge and context embedded in the graph [13, 35]. Many graph anonymization techniques have been

researched, and they can be divided into the following broad classification [22]:

1. **Naive ID Removal:** This technique removes the ID of all the graph entities. This is easy to apply; however, it does not provide much security to the graph entities as they can be easily re-identified [36, 51]. We do not consider these techniques.

2. **Edge Editing (EE) based techniques:** These techniques operate on a graph by changing its edges. The two ways these techniques can be implemented are Add/Delete and Switch. With Add/Delete, x random links are deleted from the graph and randomly added to other nodes. With Switch, the links of two randomly chosen nodes are swapped [67].

3. **k-Anonymity-based techniques:** k-Anonymity [10] techniques are most commonly used for anonymizing graphs. In this technique, a number k is selected, then the graph is modified until there are at least $k-1$ similar entities in the graph. k-anonymity techniques have various variants, which include: k-Neighborhood Anonymity [70], k-Degree Anonymity [29], k-Automorphic [71] and k-Isomorphic technique [7].

4. **Aggregation/Class/Cluster-based techniques:** These techniques group entities with similar structures into clusters, then modify the nodes in the clusters to achieve the same structure to provide them anonymity [55].

5. **Differential Privacy (DP) based techniques:** Differential privacy techniques guarantee the amount of privacy provided to a graph. It does so by capturing the graph structure in terms of certain statistics, then accurately regenerating the graph by adding a certain amount of noise in the statistics [47].

6. **Random Walk (RW):** This technique changes the edges between two nodes in the graph, which are a random i distance apart [34].

CHAPTER 3

Proposed Graph Anonymization Techniques for Software Analytics Data

Software analytics data focuses on extracting insights about the software development activities using data from (amongst others) code repositories, review repositories or CI build information. These data sources comprise different software artifacts that are inter-related. For example, in the case of CI build status prediction, the amount of code a commit change relates to that commit's build status. Such inter-relations provide important context that software analytics try to exploit, but which are ignored or even destroyed by traditional data anonymization techniques.

In this section, we provide a detailed overview of how we adapt existing graph anonymization techniques introduced previously to the context of software analytics models. In particular, we focus on the specific context of JIT defect prediction, as this is one of the

most canonical types of software analytics tasks. JIT defect prediction models take as input data a variety of metrics related to a submitted code change (commit) and try to predict the likelihood of this commit to introduce a defect.

JIT defect prediction metrics are linked to one another, wherein a change in the number of files changed by a commit will have a proportional change in the number of lines modified by the commit and the commit author's experience. Hence, instead of anonymizing individual metric values, we propose representing the underlying software analytics data as a graph and anonymizing the graph before calculating the JIT metrics. A change in any entity of the graph will have a proportional change in all the related entities. Therefore, we believe using graph anonymization techniques would preserve the structural knowledge between different graph entities while providing sufficient privacy.

To derive graph-anonymized change metrics for JIT defect prediction, we perform the following steps:

1. Creation of a knowledge graph of the involved repositories.
2. Anonymization of the graph using different algorithms.
3. Calculation of the change metrics from the anonymized graph.

3.1 Step 1: Knowledge Graph Creation

A knowledge graph refers to a graph of data that represents information about different nodes representing real-world entities, and links that define how those entities are related to one another [18]. A knowledge graph can be made for different domains like

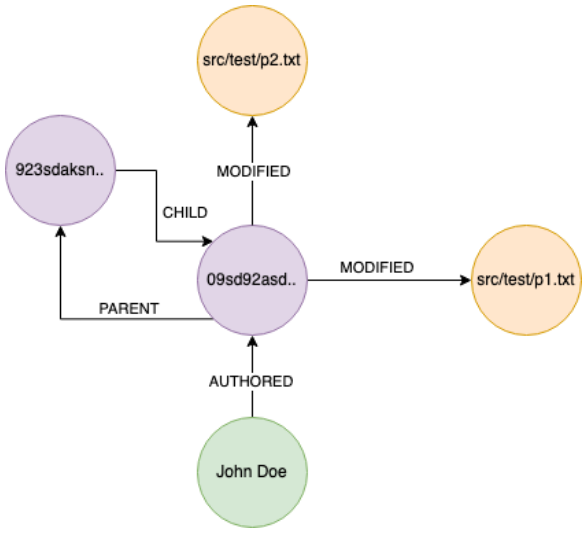


Figure 3.1: Example of a Knowledge Graph. The purple nodes are commits, the beige nodes are files and the green node is a person.

web search, social media, risk assessment and more. In the context of JIT defect prediction, we focus on a knowledge graph made from the information available in the git repository of the datasets to create a graph that we call “git-graph", as shown in Figure 3.1. The knowledge graph contains the minimum set of entities needed to calculate the change metrics needed for JIT defect prediction. The graph’s nodes and links help us understand the relationship between various entities and allow us to derive the JIT defect metrics (see Step-3 in Section 3.3). The various entities and their links are listed below:

1. **Commit:** This represents a commit in the git repository. This entity is linked to the *File* nodes, representing the files changed by the commit. The commit is also linked to the *Person* node that authored the specific commit.

2. **File:** This represents a file present in the code base. Each file is unique to the graph database, i.e., there cannot be duplicate nodes representing the same file.

Entity	Node Properties	Links to other entities (Linked Entity - Link Name - Link Properties)
Commit	ID - The hash of the commit Author-Date - The date the commit was created Bugcount - The number of bugs introduced by the commit Fixcount - The number of bugs fixed by the commit	Files - MODIFIED - Lines Added, Lines Deleted Commit - PARENT Commit - CHILD
File	ID - The complete path of the file File Name - The name of the the file Directory - The directory in which the file is stored Subsystem - The top level directory of a file	
Person	ID - The unique identifier for a person	Commit - AUTHORED

Table 3.1: Entities and their properties and links

Multiple commits in the graph can modify the same file. The changes a commit makes to a file are represented by the *Lines Added* and *Lines Deleted* properties of the link, representing the number of lines added and the number of lines deleted.

3. **Person:** A *Person* node represents a developer who worked on the repository. A person can be the one who authored a particular commit. In a graph, there is always one unique node for each person.

Table 3.1 lists all the node entities along with their properties and the links they have. The entities used to create a git-graph are sufficient to compute all the change metrics needed for our study, as highlighted in Table 3.2. It is possible to create more complex knowledge graphs that utilize bug reports, reviewer experience, review comments, etc. However, we limit our empirical study to the data needed to calculate the most commonly studied features for JIT models [26, 27, 52].

3.2 Step 2: Graph-Anonymization Algorithms

The graph anonymization techniques detailed below work on the *Commit* and/or *Person* nodes of a git-graph. We do not operate on *File* nodes, as they do not have any entity dependent on them, i.e., they are leaf nodes.

3.2.1 Selected Techniques

Out of the six categories for graph anonymization techniques mentioned in Section 2.5, we have selected the following four techniques (out of three categories) for our study:

1. **Edge Editing (EE) based techniques:** Ying et al. [67] propose two variants of Edge Editing techniques, i.e., Add/Delete and Switch. We are selecting the *Random Add/Delete* (RAD) and the *Random Switch* (RS) variant, because the latter is similar to the Random Walk technique and would also cover the Random Walk category.

2. **k-Anonymity [29]:** For k-anonymity, we have chosen *k-Degree Anonymity* (k-DA) as it represents a more modern variant that protects against the re-identification of nodes from attackers that have prior knowledge about the graph.

3. **Aggregation/Class/Cluster-based techniques [55]:** We selected *Generalization* (Gen) from this category as it was the only listed variant.

As the “Naive ID Removal” category offers no privacy to the nodes, therefore, we do not implement it [36]. “Differential privacy" offers mathematical guarantees of privacy, but it has only been recently adapted to graphs. The algorithm proposed by Sala et al. [47] converts the graph to a set of structural statistics, which are converted back

into a graph after adding a random amount of noise. However, when the graph is converted to a set of statistics, details like author-date, directory, and subsystem are lost. These details are essential in computing the JIT defect metrics, leading us to not select techniques from this category. We leave the adaptation of Differential Privacy for JIT prediction for future work.

In the next section, we describe in detail how each of the four selected anonymization techniques - *Random Add/Delete, Random Switch, k-Degree Anonymity* and *Generalization* operate and how they are applied to the software analytics data's knowledge graph.

3.2.2 Random Add/Delete

In Random Add/Delete, we anonymize x percent of the nodes of a particular type by modifying the node's links to another entity. The algorithm used is specified in Algorithm 1, and shows how the random adding/deleting of edges in the knowledge graph respects the types of the source and target nodes of said edges. We run the algorithm for each *Commit* node, where we modify the link it has to the modified *File* nodes (via the MODIFIED link), and for each *Person* node, for which we modify its *Commit* nodes (via the AUTHORED link).

From the x percent of nodes, the algorithm each time selects two nodes randomly to perform the anonymization. In Figure 3.2a, we see the unaltered graph to which the algorithm will be applied. Using the *randomlyDeleteAndGetNodes* function, it first randomly deletes nodes of the *dependNodeType* (e.g., File nodes of a Commit node) linked to the selected node and stores them in *nodeStore*. After randomly deleting the *File* nodes from the *Commit* nodes, we get Figure 3.2b.

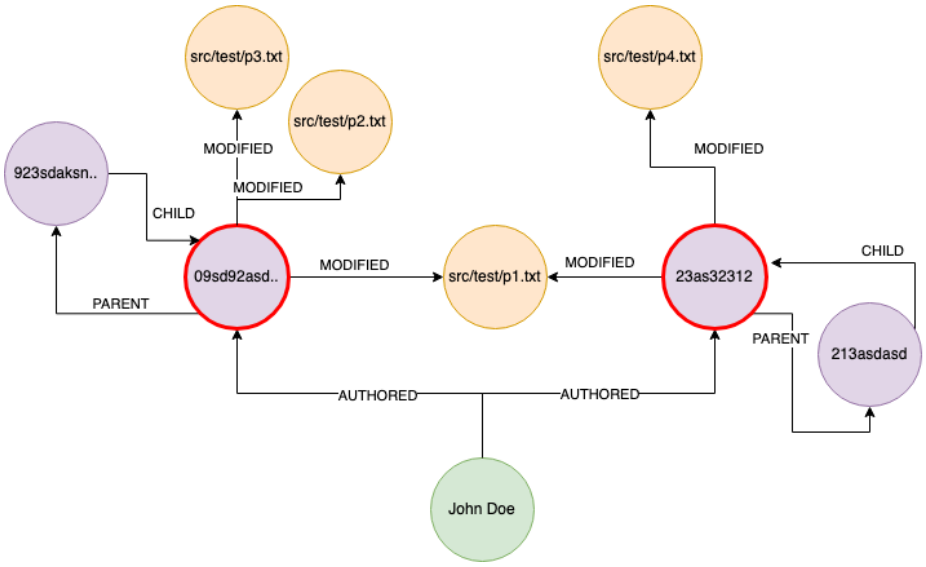
As seen in Figure 3.2c, all the deleted nodes from the *nodeStore* are then randomly linked back to the two nodes using the *randomlyAddNodes* function, leaving no node in the *nodeStore* . This operation is repeated until *x* percent of the nodes in the graph are anonymized. We select two nodes at a time to ensure that the structure of the nodes after the change is not dramatically different.

Algorithm 1 Algorithm for Random Add/Delete Technique

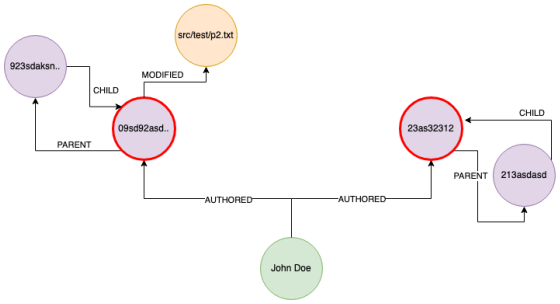
```
1:  procedure MAIN(percentToAnon)
2:      randomAddDelete(CommitNodeType, FileNodeType, percentToAnon)
3:      randomAddDelete(PersonNodeType, CommitNodeType, percentToAnon)
4:  end procedure
5:  procedure RANDOMADDDELETE(mainNodeType,dependNodeType, percentToAnon)
6:      modifyCount = count(mainNodeType) * percentToAnon
7:      for i in 0 -> modifyCount; i = i + 2
8:        // Select two random node of mainNodeType from the graph
9:        firstNode = getRandomNode(mainNodeType)
10:       secondNode = getRandomNode(mainNodeType)
11:       // Randomly delete linked nodes and return them
12:       nodeStore.add(randomlyDeleteAndGetNodes(firstNode, dependNodeType))
13:       nodeStore.add(randomlyDeleteAndGetNodes(secondNode, dependNodeType))
14:       // Randomly select nodes from the store and link them to the mainNodeType
15:       randomlyAddNodes(firstNode, nodeStore)
16:       randomlyAddNodes(secondNode, nodeStore)
17:     end
18: end procedure
```

3.2.3 Random Switch

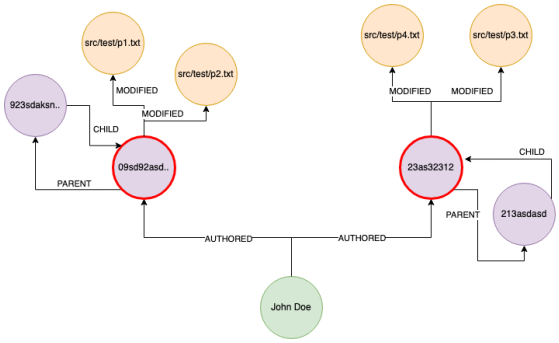
The Random Switch algorithm (Algorithm-2) works similarly to the Random Add/Delete algorithm. The algorithm selects two nodes randomly and switches all the links between them, unlike Random Add/Delete, which randomly changes some links. Figure 3.3 shows an example of the algorithm when applied to the *Commit* node and the *File* nodes they have modified. After randomly selecting two nodes, in Figure 3.3a, all



(a) Original graph

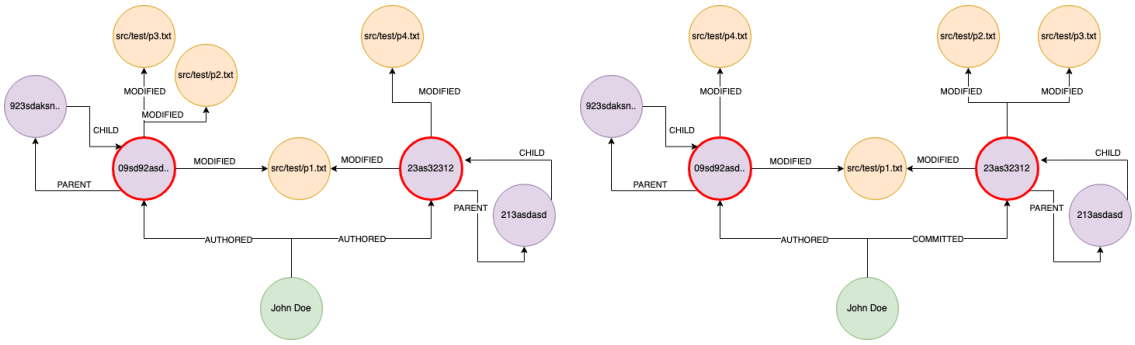


(b) After Delete Operation



(c) After Add Operation

Figure 3.2: Random Add/Delete operation. The nodes in red (Commit) are the ones on which the operation is being carried out.



(a) Original (b) After Switch Operation

Figure 3.3: Random Switch operation. The nodes in red (Commit) are the ones on which the operation is being carried out.

the links have been switched around, as seen in Figure 3.3b. This algorithm is also applied to the *Person* nodes and the links that they have authored.

3.2.4 k-Degree Anonymity

This technique requires that for every node, there are at least *k - 1* other nodes (excluding the current node) that have the same degree, yielding *k* nodes in total with that degree. Degree refers to the number of nodes a particular node is linked to. In our context, when operating on a *Commit* node, the degree is the number of *File* nodes it is linked to and for a *Person* node, it is the number of *Commit* nodes it is linked to. For example, when operating on the *Commit* nodes, in Figure 3.3a, the *Commit* in red on the left has a degree of 3, while the right one in red has a degree of 2.

In this case, if the value of *k* is chosen to be 3, then the graph should have at least 2 more commits that modify 3 files. If this requirement is met, then the graph satisfies the condition of k-Degree anonymity, making data 3-degree anonymous. k-DA anonymity yields a *1/k* probability of identifying which commit is modifying which file. It is important to note that it is not necessary that there should be *k* nodes for all

Algorithm 2 Algorithm for Random Switch technique

```
1:  procedure MAIN(percentToAnon)
2:      randomSwitch(CommitNodeType, FileNodeType, percentToAnon)
3:      randomSwitch(PersonNodeType, CommitNodeType, percentToAnon)
4:  end procedure
5:  procedure RANDOMSWITCH(mainNodeType,dependNodeType,percentToAnon)
6:      modifyCount = count(mainNodeType) * percentToAnon
7:      for i in 0 -> modifyCount; i = i + 2
8:        // Randomly select node from graph
9:        firstNode = getRandomNode(mainNodeType)
10:       secondNode = getRandomNode(mainNodeType)
11:       // Delete all dependentNode linked to the node
12:       firstNodeStore = deleteAndGetNodes(firstNode, dependNodeType)
13:       secondNodeStore = deleteAndGetNodes(secondNode, dependNodeType)
14:       // Switch the links, first to second node, and vice versa
15:       addNodes(firstNode, secondNodeStore)
16:       addNodes(secondNode, firstNodeStore)
17:     end
18: end procedure
```

the degrees from 1 to the max degree present in the graph, but rather that each degree that is present should have at least *k* nodes. For example, if a graph has degrees 2 and 4, k-DA ensures that degrees 2 and 4 have at least *k* nodes in them.

The algorithm for this technique is highlighted in Algorithm 3. It is applied on all *Commit* nodes and their linked *File* nodes, and on all *Person* nodes and their authored *Commit* nodes. The first step of the algorithm is to group the nodes based on the degree they have using the *GetDegree* function, which is stored in the *degreeStore* variable. The *degreeStore* variable has all the degrees that are present in the graph as keys (e.g., 2, 4, 10, etc.), and the value corresponding to a key is the list of nodes with that corresponding degree. The number of values that a corresponding key has helps us determine whether a key has *k* nodes in it or not.

Using the *GetDegreeBelowK*, we get all the degrees that do not have the required

k number of nodes in them. Let's call a degree from this set that does not have the required k nodes, as m , with a difference of x from k. Then for each such degree m , we find another degree l that has more than $k+x$ nodes in it. This ensures that the degree l should still have at least more than k nodes in it after the operation so that it still meets the condition of k-DA anonymity. From this selected degree l , we remove x nodes randomly and change their degree to match the degree m . This is carried out by the *changeDegree* function, which either randomly adds or deletes links to other nodes from the selected x nodes. By doing this operation, the nodes being modified will now belong to degree m, increasing the number of nodes in the degree m to k . Unlike Random Add/Delete and Random Switch, this technique works for all the nodes of a type in the graph until each node satisfies the k-degree constraint. It is important to ensure that correct k values are selected when anonymizing data. A good heuristic is to ensure that the k value should be greater than $\frac{\textit{Total number of commits}}{\textit{Number of unique degrees present in the graph}}$. This ensures that there are enough commits in the graph to ensure that each degree can have more than k commits.

3.2.5 Generalization

The generalization technique requires us to capture the structure of the graph and group nodes into clusters based on their structure. To capture the structure of the graph, we use the *Commit* node as a reference. The structure of a commit is represented by the number of *File* nodes it is linked to and the number of commits the author of the commit has created. Nodes with the same structure are grouped together into *cluster*. If a cluster has n nodes in it, it is called a full cluster; else, it is called an under-filled cluster.

Algorithm 3 Algorithm for K-Degree Anonymous

```
1:  procedure MAIN(k)
2:      KDegreeAnon(k, CommitNodeType, FileNodeType)
3:      KDegreeAnon(k, PersonNodeType, CommitNodeType)
4:  end procedure
5:
6:  procedure KDEGREEANON(k, nodeType, linkedNodeType)
7:      // degreeStore is a mapping of the degree to the list of nodes of that degree
8:      degreeStore = GetDegree(nodeType, linkedNodeType)
9:      degreeBelowK = GetDegreeBelowK(k,degreeStore)
10:     for i in degreeBelowK
11:       numNodesToAdd = k - count(degreeStore[i])
12:       surplusDegree = degreeWithCountGreaterThan(k+numNodesToAdd, degreeStore)
13:       for m in 0 -> numNodesToAdd
14:         randomNode = getRandomNode(degreeStore[surplusDegree])
15:         changeDegree(randomNode, i)
16:       end
17:     end
18: end procedure
19:
20: procedure GETDEGREE(nodeType, linkedNodeType)
21:     allNodes = getAllNodes(nodeType)
22:     for node in allNode
23:       linkedNodes = getLinks(node, linkedNodeType)
24:       degreeStore[count(linkedNodes)].add(node)
25:     end
26:     return degreeStore
27: end procedure
28:
29: procedure GETDEGREEBELOWK(k,degreeStore)
30:     for i in keys(degreeStore)
31:       If count(degreeStore[i]) < k
32:         lessThanKDegreeStore.add(i)
33:       end
34:     end
35:     return lessThanKDegreeStore
36: end procedure
```

The algorithm for capturing the structure of the graph and ensuring that there are enough elements in each cluster is specified in Algorithm 4. Using the *GetStructure* function, we capture the structure of each commit in the graph. We now attempt to create m clusters with at least n nodes in them, i.e., the graph should now have m new full clusters. The structure that is being targeted to have n nodes is found using the *GetStructWithLessThanKElem* function that finds the clusters that do not have n nodes in them, let's call this set z. *Using the countMoreThan* function, we find all the clusters that have more than n nodes in them. Then, for the nodes above n , we call the *modifyStructure*, which ensures that the node *nodeToModify* is changed to match the cluster structure z that is currently being targeted by randomly adding links or deleting links to nodes.

Algorithm 4 Algorithm for Generalization Technique

```
1:  procedure MAIN(targetCount)
2:      structStore = GetStructure()
3:      underFilledStruct = GetStructWithLessThanKElem(targetCount, structStore)
4:      for underFillStruct_i = keys(underFilledStruct)
5:       currCount = count(underFilledStruct[underFillStruct_i])
6:       overFillStructKey = countMoreThan(currCount+targetCount, structStore)
7:       for j in 0 to targetCount-currCount
8:        nodeToModify = structureStore[overFillStructKey]
9:        modifyStructure(nodeToModify, underFillStruct_i)
10:     end
11:     end
12: end procedure
13: procedure GETSTRUCTURE
14:     for not in getAllNode(CommitNodes)
15:      countFileNodes = count(getLinkedNodes(node, FileNodes))
16:      author = getAuthor(node)
17:      countCommitNodes = count(getLinkedNodes(author, CommitNodes))
18:      structureKey = countFileNodes+", "+countCommitNodes
19:      nodeStructure[structureKey].add(node)
20:     end
21: end procedure
```

3.3 Step 3: Generation of JIT metrics

The knowledge graph for the git repository has all the information needed to generate the typical change metrics needed for JIT defect predictions [26, 27, 52]. In Table 3.2, we show all JIT defect prediction metrics used for model training and how we calculate them from the git-graph generated for each project. The calculation of the metrics for each commit is done as defined in the last column of Table 3.2. For example, to calculate the *la* attribute for a commit, we take the sum of the *la_link* link property for each *File* node attached to the commit. Similarly, we generate all other metrics for a commit using this approach and repeat it for all commits in the graph.

3.4 Evaluation of Graph Anonymized Metrics

This chapter outlined the process of generating the graph anonymized JIT metrics. It details how from the git information, we can generate metrics from the knowledge graph consisting of the 'Commit', 'Author' and 'File' entities. The graph anonymization techniques have several parameters, i.e., configurations, that can be tweaked to configure the anonymization process to provide explicit control between the privacy provided and performance decrease to the metrics, allowing to obtain a trade-off.

This approach and its ability to tweak the privacy-performance trade-off form the core foundation of this thesis, and are evaluated in the following two chapters. In Chapter 4, we assess the privacy enhancements gained by the JIT metrics at different configuration levels. Furthermore, we assess how the performance of the models changes in a within-project setting and compare it to the state-of-the-art techniques. In Chapter 5, we evaluate the change in performance in a cross-project setting and

in settings where graph and non-graph anonymization techniques might be used together.

Attribute	Description	Calculation based on anonymized graph
la	Number of lines added by a commit	For all the *File* nodes modified by a *Commit* node, we add the 'la_link' property
ld	Number of lines deleted by a commit	For all the *File* nodes modified by a *Commit*, we add the 'ld_link' property
ent	Entropy or the spread of changes across file	This is calculated by taking the log of the proportion of change to a file as compared to all the files
nf	Number of files	This is the number of unique files changed by a commit. This is calculated by counting the number of *File* nodes a *commit* node has modified
nd	Number of directories	For all the files found in the 'nf', we make a unique set of the directory of each file
ns	Number of subsystem	For all the files found in the 'nf', we make a unique set of subsystems for each file. The subsystem is the root directory for a file
nuc	Number of unique changes	For each modified file for the current *Commit* node, we find the number of commits that have previously modified those *File* nodes
ndev	Number of unique developers that changed the file	For all the commits found in the 'nuc' attribute, we find all the unique authors for all the unique modifying authors
age	Average time interval between current changes and the last changes made to the file	For all the commits found in the 'nuc' attribute, we then find the latest commit for each file. After that, we take the average of the time that commit happened and the current commit
aexp	Experience of the author	The number of nodes the author of the commit has modified before this commit
asexp	Subsystem experience of the author	For all the commits found in 'aexp', we find the unique subsystems (which are same as the subsystems modified by the current commit) those commits have modified
arexp	Relative experience of the author	This is calculated by taking the 'aexp' commits, and then dividing them by their age [31]

Table 3.2: Description and the method to derive the JIT defect prediction metrics (used by Kamei et al. [26]) calculated from the git-graph.

CHAPTER 4

Empirical Evaluation of Graph Anonymized Metrics in Within-Project Setting

4.1 Introduction

As the usage of software analytics for understanding different organizational practices becomes prevalent, it is important that data for these practices is shared across different organizations to build a common understanding of software systems and processes. Yet, organizations are hesitant to share this data and trained models with one another due to concerns around privacy, e.g., because of the risk of reverse engineering the training data of the models.

To facilitate data sharing, tabular anonymization techniques like MORPH, LACE

and LACE2 have been proposed to provide privacy to defect prediction data. However, said techniques treat data points as individual elements, and lose the context between different features when performing anonymization. We study the effect of four anonymization techniques, i.e., Random Add/Delete, Random Switch, k-DA and Generalization, on the privacy score and performance in six large, long-lived projects. To measure privacy, we use the IPR metric, which is a measure of the inability of an attacker to extract information about sensitive attributes from the anonymized data. We find that all four graph anonymization techniques are able to provide privacy scores higher than 65% in all the datasets, while Random Add/ Delete and Random Switch are even able to achieve privacy scores of 80% and greater in all datasets. For techniques achieving privacy scores of 65%, the AUC and Recall decreased by a median of 1.45% and 5.35%, respectively. For techniques with privacy scores 80% or greater, the AUC and Recall of privatized models decreased by a median of 6.44% and 20.29%, respectively. The state-of-the-art tabular techniques like MORPH, LACE and LACE2 provide high privacy scores (89%-99%); however, they have a higher impact on performance with a median decrease of 21.15% in AUC and 80.34% in Recall.

Furthermore, since privacy scores 65% or greater are adequate for sharing, the graph anonymization techniques are able to provide more configurable results where one can make trade-offs between privacy and performance. When compared to unsupervised techniques like a JIT variant of ManualDown, the GA techniques perform comparable or significantly better for AUC, G-Mean and FPR metrics.

Our work shows that graph anonymization can be an effective way of providing privacy while preserving model performance.

4.2 Empirical Study Design

In this section, we give details about the design of the empirical study aiming to address the RQs for this chapter, i.e., the datasets used for our case study, the configuration of the various graph anonymization techniques, and our approach to answering the three research questions.

4.2.1 Subject System Selection

As our study focuses on graph anonymizing JIT metrics and on evaluating the impact on their privacy and performance, we use datasets that were made available by other JIT defect prediction research. One of the most common JIT defect prediction datasets was shared by McIntosh et al. [31], and has been used in later research [17, 42, 46]. It features JIT data for OpenStack and Qt. The other dataset we have used is ApacheJIT [27], which makes data available for popular 14 Apache projects.

In Table 4.1, we have listed all the projects selected for our study. As graph anonymization changes the structure of the code, we need projects whose trained models have high AUC and have a sizeable number of commits. To calculate the AUC of the models, we split the data into a chronological 80/20 train/test split. We perform out-of-sample bootstrapping training 100 models by randomly selecting 500 elements from the training split and testing on 500 randomly selected testing elements. We then take the average of these scores to find the average AUC. Because of this, we place two conditions on the datasets: a) have an average AUC greater than 60%, and b) the number of commits should be greater than 5,000. We want average AUC to be above 60% so that when anonymization is applied a model does not immediately become random (AUC 50%) but will retain some predictive power. Having more than 5,000 commits helps us

Dataset	Time Period	No. of Commits	Buggy Commits	Language
Qt	6/2011 - 3/2014	25,150	2002 (8%)	C++
OpenStack	11/2011 - 2/2014	12,374	1616 (13%)	Python
Cassandra	2003 - 2019	8,159	3,117 (38%)	Java
Flink	2003 - 2019	11,691	2,811 (24%)	Java
Groovy	2003 - 2019	8,059	1,614 (20%)	Java/Groovy
Ignite	2003 - 2019	12,036	2,439 (20%)	Java

Table 4.1: Summary of projects selected for the study.

ensure the graph has enough data points to be anonymized.

Based on this, we select the top 6 projects with the most commits. However, two selected Apache Hadoop projects recently went through a refactoring change that led the Hadoop repository to have 0 buggy commits, while the Hadoop-HDFS only had a total of only 603 commits (as opposed to 10,000). As such, we excluded these repositories. Although our criteria exclude the Qt dataset (the model has an AUC of less than 60%) we did include it in our study because of the massive size of this dataset and its important role as a benchmark in prior work [17, 42, 46].

4.2.2 Data Collection

The original datasets [27, 31] provide labels indicating which commits are buggy or not and the change metrics used for training the model. To determine the bug labels, both papers follow a similar methodology. The Apache dataset selects 14 popular Apache projects, based on stars on GitHub, while the OpenStack and Qt datasets were selected where at least 90% of the commits were part of a code review process. The respective authors then extract all the resolved or closed bugs filed in the projects' respective issue-tracking platforms, i.e., Gerrit for OpenStack/Qt and GitHub for Apache projects.

For these resolved bug reports, they obtain all commits referring to the corresponding bug report identifier, then consider the most recent such commit as the report's bug-fixing commit. On these bug-fixing commits, they apply filters to remove any erroneously detected commit, for example, commits changing trivial code (modifying only comments or removing white space) or changing over 100 files and/or 10,000 lines of code. Utilizing SZZ [56] on the bug-fixing commits, they were able to pinpoint the bug-inducing commits, which correspond to the labelled data points in our study.

However, the datasets did not provide the information needed by us to create the File and Person entities in the git-graph. To fetch those entities, we had to use the respective Gerrit (OpenStack [38] and Qt [45]) and PyDriller (Apache projects) APIs of the studied projects to extract the data needed to create the git-graph specified in Section 3.1. For each commit, we had to mine the number of files and the number of lines changed in them, as well as the identifier of the person who authored that commit.

Gerrit API: We call the Gerrit API for all the commits in the OpenStack and Qt datasets to obtain the details needed to form the git graph. After the knowledge graph has been created, we then label the *Commit* nodes as buggy or not according to the labels provided [31].

Git repository: Using the py-driller library [44], we mine each selected Apache repository and collect all the details needed for creating the git-graph. Using the labels provided by the ApacheJIT dataset, we mark the data as buggy or not.

4.2.3 Measuring privacy

Defining Privacy

Privacy refers to how much information can be disclosed about the members of the data. Ideally, we do not want information about sensitive attributes data to leak. However, it is hard to avoid all leaks, hence, we want to measure the amount of information that is leaking. We use the *la* feature as the sensitive feature because it has been observed that for defect prediction, the *number-of-lines-modified* attribute of a commit is closely related to bugs [26]. Furthermore, earlier work by Peters et al. on MORPH, LACE and LACE2 also used the number of source code lines feature as the main sensitive feature. We, therefore, use the lines added (*la*) and lines deleted (*ld*) attributes of a commit as the sensitive attributes in our study.

To measure privacy, we use the metric Increased Privacy Ratio (IPR) as used by Peters et al. [28, 41], similar to Adversarial Accuracy Gain [4]. This measure defines the ability of the attacker to extract information about sensitive attributes of the data, comparing the information gained by the attacker in the anonymized version to the non-anonymized version of the data. IPR can also be represented as a query made on the data before and after anonymization and whether this gives the same results or not.

A few other important definitions before defining IPR:

1. *Sensitive Attributes:* Attributes that we do not want attackers to identify in our dataset. These attributes can reveal sensitive information about the data. For our dataset, they are *la* and *ld*. When data is shared, it is assumed that the attacker does not have access to the sensitive attributes and is attempting to guess them.

2. *Quasi-identifier (QID's):* They, by themselves or along with other attributes from

commit_id	la	nf	nd	ns
commit-1	[1-10]	[0-3]	[0-2]	[0-1]
commit-2	[11-20]	[4-6]	[3-5]	[2-3]
commit-3	[1-10]	[0-3]	[0-2]	[2-3]
commit-4	[1-10]	[7-9]	[0-2]	[0-1]
commit-5	[21-30]	[4-6]	[6-8]	[0-1]
commit-6	[11-20]	[4-6]	[6-8]	[4-5]
commit-7	[11-20]	[7-9]	[6-8]	[2-3]

(a) Data with EFB applied before anonymization process

commit_id	la	nf	nd	ns
commit-1	[21-30]	[0-3]	[3-5]	[4-5]
commit-2	[11-20]	[7-9]	[6-8]	[2-3]
commit-3	[11-20]	[0-3]	[0-2]	[2-3]
commit-4	[11-20]	[0-3]	[6-8]	[0-1]
commit-5	[1-10]	[4-6]	[0-2]	[0-1]
commit-6	[11-20]	[4-6]	[6-8]	[4-5]
commit-7	[1-10]	[7-9]	[0-2]	[0-1]

(b) Data with EFB applied after anonymization process

Table 4.2: Example for Query Generation with Equal Frequency Binning before and after anonymization.

a dataset, can be used to identify entities in a dataset. For our dataset, it would be all the attributes except *la* and *ld*, therefore, *ns, nf, nd, etc.*

To calculate IPR in the dataset, we assume that the attacker has some knowledge about the value of the quasi-identifier of the entities they want to identify in the dataset. This knowledge becomes a query that we will perform on the data to calculate IPR.

Query generation

Similar to Peters et al. [41], before generating the queries to model an attacker's knowledge (created for IPR), we apply Equal Frequency Binning (EFB) to all attributes of the dataset under consideration. EFB ensures that data is divided into bins where the number of elements in each of the bins is equal, this process is automated. To explain the steps taken to create the query, we use Table 4.2 as an illustration. In this table, we take a small subset of all features and a few rows to explain how the queries used to calculate the IPR scores are constructed. Given this, we perform the following steps:

1. **Select the query size** to create queries. In this thesis, we construct queries of

sizes 1,2 and 4. For this example, we only show results for size 2. This means that this query will use 2 QIDs to query the data.

2. Randomly **select the quasi-identifiers** for query generation from all the available QIDs. In Table 3.2, we can see all the attributes available for selection. For this illustrative example, we use the sample shown in Table 4.2, from which we choose the attributes *nf* and *nd* as the two attributes.

3. For all selected attributes, **select one random sub-value range** from the possible ranges. In the sample Table 4.2, we randomly choose the range $[0-3]$ for the *nf* attribute and $[0-2]$ for the *nd* attribute.

In the example mentioned before, we hence generate a query with the conditions: $nf = [0-3]$ and $nd = [0-2]$. For each query, we ensure that the query at least returns one element and that the attribute value pairs are unique, i.e., there are no duplicate queries. If the query does not satisfy this condition, it is removed.

We generate queries with 1, 2 and 4 quasi-identifiers, and for each size, the number of queries is limited until all unique queries are exhausted or the number of queries generated reaches 1000 [41]. When generating queries, we only use the quasi-identifiers because the sensitive attributes are the ones the attacker is trying to identify.

Calculating IPR:

For each query generated previously, we calculate if the query resulted in a breach. A query causes a breach if the most common binned value of the sensitive attributes (S^*_{max}) for the rows returned from the privatized dataset (G^*_i) matches that of the non-privatized dataset (returning rows G_i with the most common binned value as S_{max}). If

this condition is met, the query is said to cause a breach.

$$\text{Breach}(G_i^*) = \begin{cases} 1, & \text{if } s_{\max(G_i)} = s^*_{\max(G_i^*)} \\ 0, & \text{otherwise.} \end{cases}$$

To calculate IPR, for all the queries, we find the percentage of queries that do not cause a breach. Using this, IPR is then calculated as:

$$\text{IPR} = (1 - (\text{Total number of breaches}/\text{Total number of queries run}) * 100)$$

Applying the query $nf = [0-3]$ and $nd = [0-2]$ on data before anonymization (Fig 4.2a) we get Commits-1 and Commits-3 with the most common value of the sensitive attribute (*la*) being [1—10]. However, when this is applied to the data after anonymization, as seen in Fig 4.2b, we only get Commit-3 with the most common value of the sensitive attribute (*la*) being $11-20$.

Given that this query results in a different value of the sensitive attribute after anonymization, this means that there was not a breach of information. This process is repeated for all the queries to calculate the IPR scores.

4.2.4 Empirical Evaluation

Overall Approach

Tan et al. [53] highlighted that training/testing splits should be done chronologically. They highlighted that in a realistic scenario, at a time t, information will be available

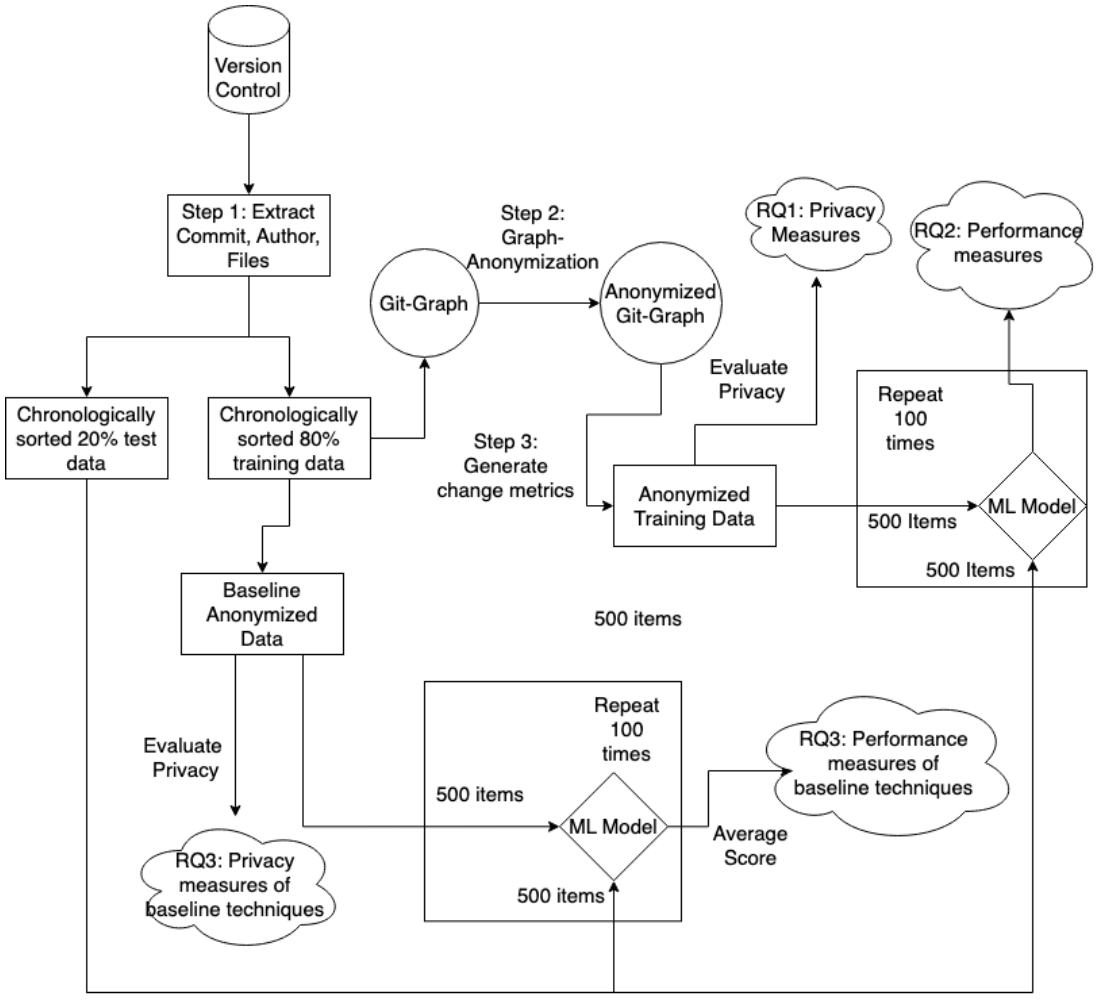


Figure 4.1: Approach for answering the research questions

only about the commits that were created before it. Therefore, only previous commits should be used to predict whether a new commit is defective. The authors further highlighted how techniques like cross-validation use information from future data to determine if the current commit is buggy, which yields abnormally high precision. To combat these problems, we sort each project's data into a chronological 80/20 train/test split.

Figure 4.1 shows the general approach taken to answer the different research questions. As highlighted before, the first 80% of the training data is used to construct the git-graph (Step 1 in Section 3.1). The graph generated from the training data is then anonymized using one of the graph anonymization techniques (Step 2 in Section 3.2). Using this graph, we generate the anonymized JIT defect metrics (Step 3 in Section 3.3), which are used to train the models and calculate the privacy gained. We train a large number of models on the subsets of the data using the out-of-sample bootstrapping technique [54], which produces more robust models with less bias and variance. Each bootstrap iteration randomly samples 500 items with replacement from the training and testing dataset to evaluate the model's accuracy. We repeat this 100 times to ensure robustness and take the average of all the performance scores. This would ensure that the technique obtains consistent results.

Baseline Techniques

Similar to Li et at., [28] and Yamamoto et al.,[63] to evaluate how our graph anonymization techniques compare with state-of-the-art privacy-preserving techniques, we use the different techniques provided by Peters et al., i.e., the MORPH, LACE and LACE2 algorithms [39, 40, 41]. We were unable to find the replication package for SRDO [28], therefore are not able to compare to it.

Similar to the approach for graph anonymization evaluation, we divide the baseline approach's data into chronologically sorted 80/20 train/test sets, then anonymize 80% of the training data. This anonymized data is used to calculate privacy and train models using out-of-sample bootstrapping. We train the model using 500 samples from the train data and use 500 samples from the test data to evaluate its performance. This

process is repeated 100 times to get models with fewer biases and variance. To ensure a fair comparison, the test bootstraps used are the same across the graph anonymization and baseline techniques. To tune the performance of the baseline techniques we perform parameter optimization by testing all value combinations between 0 to 1 with 0.2-point increments. For the LACE and LACE2 experiments, we keep the CLIFF prune rate parameter as 0.4, which is the default value in the replication package. We also add the default values provided by the package for all the baseline techniques. Based on this, we report the baseline techniques' privacy and performance.

Inspired by other unsupervised learning approaches for JIT [19, 64], in order to adopt the ManualDown technique of Fan et al. [12] for JIT, we mark the top 50% of commits in the test set as buggy, referring to this as "MD50". For a fair comparison, we use the same test split as used for the GA and NGA techniques to measure the performance of the MD50 model.

4.2.5 Graph Anonymization Configuration

Each graph anonymization technique can be applied in different amounts (levels) or configurations, allowing finer-grained control of the trade-off between privacy and utility. In this section, we will define how different graph anonymization techniques will be applied in terms of configuration levels.

Random Add/Delete: In our initial experiments, we observed that Random Add/Delete was able to provide a high amount of anonymization when applied in increments of 20% (20, 40, 60, 80, 100). However, this did not allow us to observe how small amounts of anonymization changed the privacy and AUC metrics. Therefore, to see fine-grained changes in privacy, we apply anonymizations in increments of 2% up to

20%, then in increments of 10%.

Therefore, the configuration levels are 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 30, 40, 50, 60, 70, 80, 90, and 100. When anonymization is applied, we select x% of the *Commit* nodes (x represents the level chosen) and perform the technique on the *File* nodes to which they are linked. We then select the *Person* nodes and change the *Commit* nodes for which they are the authors.

Random Switch: Similar to Random Add/Delete, we apply random switch at levels: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 30, 40, 50, 60, 70, 80, 90, and 100. This is also applied to the *Commit* nodes and the *Person* nodes in a configuration similar to Random Add/Delete.

k-DA Anonymization: In k-DA, we specify a number k, which becomes the minimum number of nodes each degree in the graph should have. For this, we specify k to have values of 2, 3, 5, 7, 10, 13, 15, 18, 21 and 24 [29]. The k-DA anonymity operations are applied to each *Commit* node and the *File* nodes to which it is connected, and for the *Person* nodes, the *Commit* nodes they have authored.

Generalization: Generalization creates clusters based on the graph structure. When analyzing the structure of a *Commit* node, we use the *File* node it is connected to and the author's experience, which is the number of nodes that the author has authored. Based on this structure, we ensure that each cluster has at least 10 commits. The number of clusters that need to be created are +10, +15, +20, +25, +30, +40, +50, +65, +80 and +100. This means that, for example, for a value of +20, the algorithm will ensure that it creates 20 new full clusters, each cluster containing at least 10 nodes. The clustering operation is performed on the *Commit* nodes alone. However, to make a cluster of a particular structure, we have to modify the *File* nodes that a *Commit* is connected to and the *Commit* nodes that its author has worked on.

4.2.6 Methodology for the Research Questions

In this section, we outline our approach to address the individual research questions we have posed.

RQ1: How do different graph anonymization techniques affect the privacy of the JIT defect prediction data?

Motivation: The main aim of our work is to evaluate the extent to which graph anonymization techniques can provide privacy to the JIT metrics generated from the git-graph. As the anonymization techniques change the graph data, we want to evaluate if they are effective in providing privacy to the data. What amount of graph anonymization needs to be applied to gain a sufficient amount of privacy? Does applying more anonymization always result in more privacy for the metrics?

Approach: To assess how different graph anonymization techniques provide anonymity to the git-graph, we first create the graph for each dataset in Neo4j. After the graph has been created, we make a backup of the graph. Then, for each dataset, we perform the following steps:

1. **Restore:** Using the backup created earlier, we recreate the database from the backup. This ensures that the graph is unaltered.

2. **Graph anonymize:** Depending on the anonymization technique currently being applied, we select the level (amount) of anonymization that needs to be applied to the graph. We then apply the anonymization techniques on the graph with the selected configuration value.

3. **Compute JIT metrics:** After anonymization, we compute the JIT metrics for each

commit that is present in the graph. We repeat this and the above two steps for all the configuration levels of the anonymization technique before moving to the next step.

4. **Non-anonymized baseline:** As the non-anonymized version acts as the baseline to calculate the performance and measure privacy, we compute the JIT metrics from a graph where no anonymization has been applied.

5. **Measure privacy:** Using the various files generated in the previous steps, we compute the privacy metrics for each file using the non-anonymized version as the baseline.

The above steps are repeated for each technique and dataset. After collecting all the files, we run scripts to analyze the privacy provided by each technique. All the privacy measures for different datasets are aggregated per technique to understand how different techniques provide privacy to different datasets.

For the IPR metric, we divide the privacy attained into different levels. We found that Peters et al. [41] called any privacy score above 65% adequate, while Li et al. [28] used an adequacy limit of 80%. Therefore, our privacy levels are as follows:

1. **Privacy Level I:** 65 - 80%

2. **Privacy Level II:** 80 - 100%

If any technique provides Privacy Level I to a dataset, it is deemed to be adequate for sharing and the technique is feasible to use. In case data owners may wish to anonymize data further, we discuss Privacy Level II.

Each anonymization technique alters the dataset differently for its configuration values. This makes it hard to compare the effect different values have on the dataset

since a configuration level of 20 for, say, k-DA does not reflect the same amount of graph anonymization as configuration level 20 of Random Switch. To allow for better comparison, we also calculate the number of graph links that were changed by each technique at each level when anonymizing the data. This helps to understand the number of links that were changed by each configuration value of a technique and how many link changes were needed to reach a particular privacy level. The number of links changed acts as a normalized scale to compare different techniques, instead of using their values.

RQ2: How do different graph anonymization techniques affect the performance of the JIT defect prediction model?

Motivation: Graph anonymization techniques augment the graph to add privacy. With this research question, we want to quantify how augmentation changes the performance of the metrics. Are the changes the graph anonymization techniques make to the graph always destructive, or do they, in any case, improve performance as well? This will also help us understand if the performance loss and privacy gains can be balanced. Do different performance measures degrade at the same rate with increasing anonymization levels?

Approach: Using the anonymized JIT metrics generated in the previous research question, we bootstrap the training of multiple Random Forest Classifiers [48]. The details of how the model is trained are highlighted in Section 4.2.4 above. Once the model has been trained, we calculate the AUC, Recall, FPR and G-Mean metrics of the model to evaluate the impact graph anonymization has on these metrics [26, 31].

The Area Under the Curve (AUC) has been used in many JIT studies to describe a

model's predictive power and is useful when there is an imbalanced training dataset. The value typically ranges from 0.5, equivalent to random guessing, to 1, which is perfect prediction power. The ROC curve, whose area is measured by AUC, plots a model's recall (Y axis) vs false positive rate (FPR), across different configurations of a model. Recall is the measure of the model's ability to correctly predict if a commit is buggy or not. It is important to understand how anonymization techniques change the ability of the data to detect bugs correctly. FPR is the measure of how many bugs are incorrectly identified as buggy. It is calculated as $\frac{FP}{FP+TN}$. The lower this measure, the better the model performance, as that would signify fewer commits are getting incorrectly classified as buggy, thereby wasting less time in fixing the wrong commits. Finally, we also use the G-Mean as it is useful in the evaluation of imbalanced datasets. It is the geometric mean of sensitivity (recall) and specificity ($\frac{TN}{FP+TN}$ (True Negative and False Positive)) of the model. The higher this metric, the better the model performance.

It is important to study the changes that anonymization brings to the models' performance as any anonymization applied without any utility function, i.e., model performance metric, can be harmful, as there would be no way to detect the effect of privacy. Privacy and utility cannot be considered in isolation [9].

As explained above in RQ1 design, the number of graph links changed acts as a normalized way to compare different configuration values for different techniques. Therefore, here too, we study the effect of graph link changes on the performance of the datasets. This is important as it will help us understand how destructive particular anonymization techniques are and establish a relation between the number of graph links changed and changes in the AUC and Recall.

We conduct a Kruskall-Wallis test ($\alpha = 0.05$) with Dunn's posthoc test to find the

best-performing technique in terms of preserving AUC, Recall, G-Mean and FPR. For each dataset, we take all the configuration values for each technique that are able to provide privacy scores greater than 65%. These filtered values are then compared to one another to find the best-performing technique in terms of AUC and Recall. We also perform the Spearman correlation between increasing configuration values and the performance metrics to quantify the possible relation between how the performance changes with increasing anonymization levels.

RQ3: How do the graph anonymization techniques compare to the state-of-the-art anonymization techniques?

Motivation: We want to compare how the graph anonymization techniques work in comparison to the state-of-the-art anonymization techniques for JIT models. Do the graph anonymization techniques achieve privacy scores that are comparable to the state-of-the-art techniques? Furthermore, we want to compare the difference in performance that comes with using these techniques when compared to graph anonymization techniques.

Approach: As highlighted in 4.2.4, to evaluate how the state-of-the-art techniques compare to graph anonymization techniques, we were able to find the replication packages for MORPH, LACE and LACE2 [60] (collectively referred to as non-graph anonymization (NGA) techniques) for evaluation on the JIT defect prediction metrics. As fine-tuning the NGA techniques results in different models, we compare the data points that are able to provide Privacy Level I and Privacy Level II and their median AUC, Recall, FPR and G-Mean scores for different datasets.

Furthermore, to understand the statistical difference between the AUC, Recall and

IPR scores of the NGA and GA (graph anonymization) techniques, we compare them in two different ways. All the comparisons are performed per dataset, and for each comparison, we consider the configuration values that have an IPR score greater than 65%. First, for the AUC, Recall, and IPR scores separately, we combine all of the different NGA techniques' values into one distribution (for one metric at a time, e.g., AUC) and compare that with all the GA technique values combined into one distribution (for one metric at a time, e.g., AUC) to understand, overall which group performs better using the Mann-Whitney U test. Second, we compare all of the techniques (NGA+GA) among themselves to evaluate how they compare to one another using the Kruskall-Wallis test with Dunn's posthoc test. This is to identify the technique that works best among all. For all tests, we use the $\alpha = 0.05$.

We compare the performance of the model trained using GA-anonymized data with the MD50 Manual Down model as well. For this MD50 model, as it requires no training, we only need the test data set [19, 64], then mark the top 50% of the test set commits with the highest "number of lines added" feature values as buggy, before comparing with the oracle for that test set. This process is repeated 100 times for each bootstrap iteration to obtain the performance distribution for the MD50 model.

The graph techniques have multiple configurations, each with its own bootstrapped model performance. Hence, for each dataset anonymized with a GA technique, we find the best-performing configuration value (at least Privacy Level I) out of all the configurations. We do this by statistically comparing the bootstrap performance distributions across each configuration, giving us the best-performing configuration for each GA technique and dataset combination. The performance of each such combination is then compared to the AUC, G-Mean, Recall and FPR performance metrics of MD50

using the Wilcoxon signed-rank test ($\alpha = 0.05$) with Bonferroni correction.

4.3 Empirical Study Results

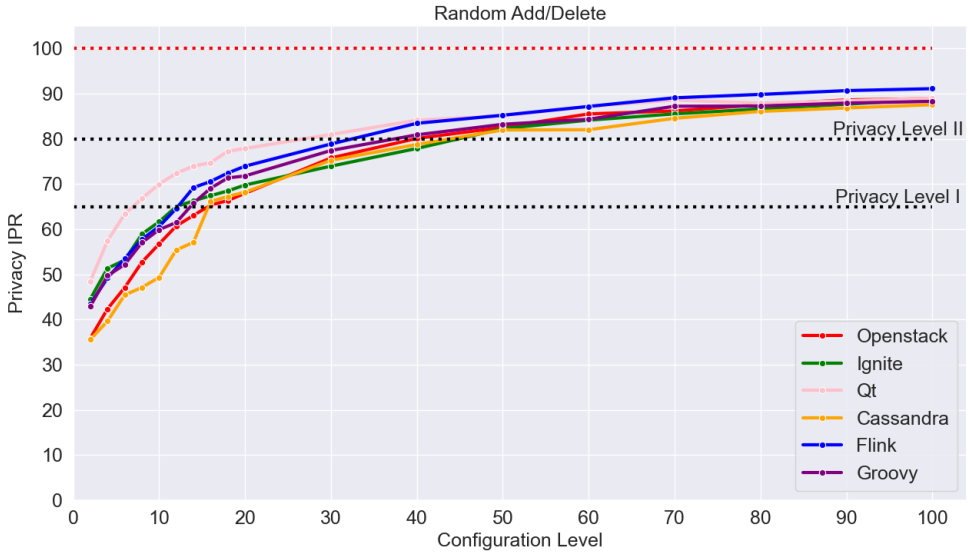
In this section, we discuss the results of the research questions we set out to answer.

4.3.1 RQ1: How do different graph anonymization techniques affect the privacy of the JIT defect prediction data?

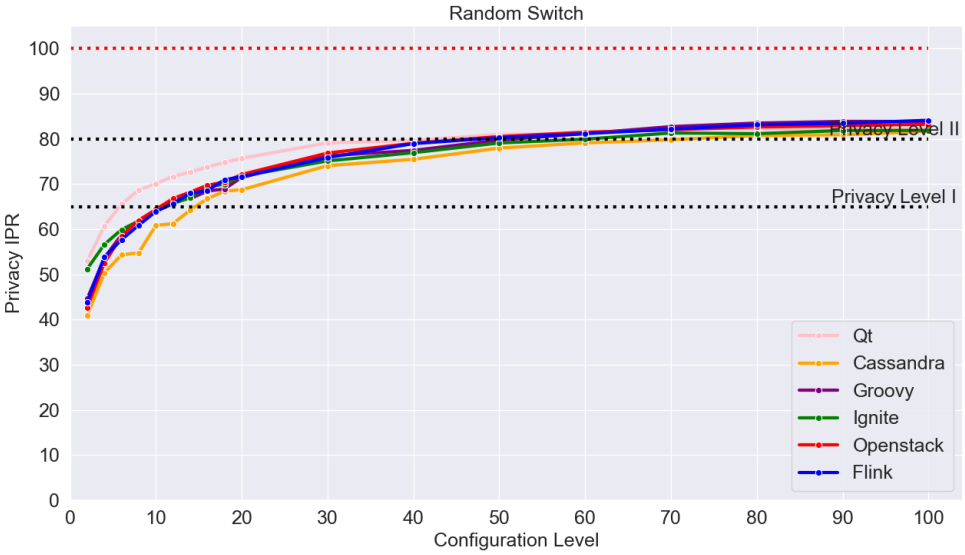
Overall, our results show that graph anonymization techniques are effective in providing privacy to the JIT metrics. Each technique is able to provide Privacy Level I (IPR scores of 65% or greater) to all the datasets. Furthermore, for every dataset, there is more than one technique that is able to provide Privacy Level II (IPR scores of 80% or greater).

Figure 5.2 shows that the Random Switch and Random Add/Delete techniques are able to provide both Privacy Levels I and II to all the datasets. We see that both techniques have a similar effect, where with increasing levels of anonymization, the privacy scores keep increasing. At 100% anonymized nodes, Random Add/Delete is able to provide privacy scores of almost 90%.

However, for both techniques, we see an elbow-shaped curve in the amount of privacy gained with increasing levels, indicating diminishing returns in privacy gains with increasing anonymization levels. This is because even when increasingly more nodes are changing, the anonymization may affect nodes that have already been altered and therefore do not dramatically increase the privacy scores anymore. We see this similar trend for all techniques, where with increasing anonymization levels, they gain lesser



(a) Random Add/Delete



(b) Random Switch

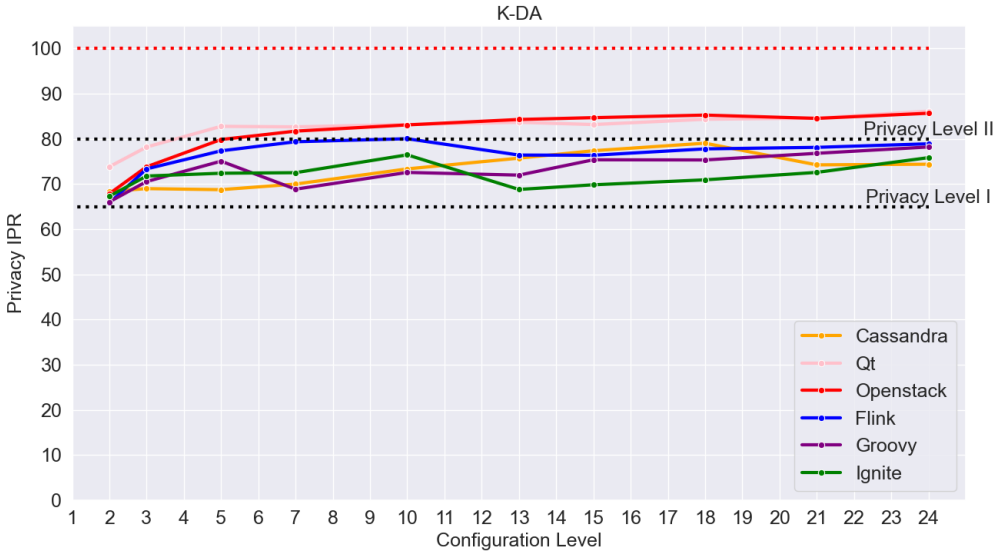
Figure 4.2: IPR scores of all the datasets for Random Add/Delete and Random Switch. The vertical-dotted lines are the two privacy levels.

amounts of privacy.

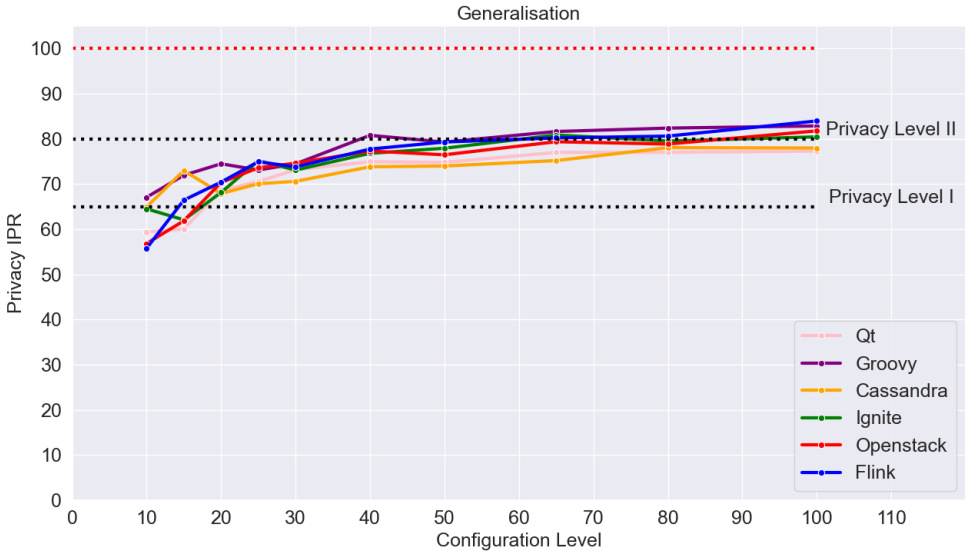
With little effort spent on anonymization, both Random Switch and Random Add/Delete techniques are able to achieve high privacy scores. To gain Privacy Level I, less than 20% of the nodes are to be anonymized for both. However, there is a difference in the number of nodes to be anonymized to gain Privacy Level II, where Random Switch requires anonymization levels of 60/70% as compared to Random Add/Delete, which does for 40/50%, i.e., with lesser nodes anonymized. This difference is because Random Add/Delete causes more change in the graph as it destroys and redistributes the links for the various nodes. This is in contrast to Random Switch, which simply switches the nodes around, keeping the same distribution of links.

While k-DA only reaches Privacy Level II for half of the datasets, it is able to provide Privacy Level I to all the datasets at the lowest configuration level ($k = 2$). Figure 4.3a shows how all projects quickly increase their IPR above 65%, yet taper off relatively quickly. This is because, unlike Random Add/Delete and Random Switch, k-DA operates on all the nodes in the graph that do not satisfy the k-DA anonymity condition. As all the nodes in the graph are modified, it leads the graph to gain privacy quicker than Random Add/Delete or Random Switch, which only operates on x% of the nodes at a time.

With increasing values of k, the gains in privacy of k-DA are limited. One of the possible reasons is that at higher values of k, the same characteristics of the graph are further amplified, causing a limited change in the metrics. For example, at $k = 5$, there are 5 commits that modify 10 files, making the nf attribute of those commits 10. At $k = 21$, there are now 21 commits that modify 10 files. While according to the IPR metrics, some query results still do change as there is an increase in privacy measures. However,



(a) k-DA



(b) Generalisation

Figure 4.3: IPR scores of all the datasets for k-DA and Generalization. The vertical-dotted lines are the two privacy levels.

overall the benefits from higher values of k are limited as the same characteristics of the graph are retained. This is the reason why only 3/6 datasets gain Privacy Level II.

Similar to k-DA, in Figure 4.3b, we see that Generalization provides higher privacy scores than Random Add/Delete and Random Switch at the lowest configuration level. At configuration level 20, it is able to provide Privacy Level I to all the datasets. Furthermore, at configuration level 100, it provides privacy scores greater than 80% to 4 out of the 6 datasets.

	Random Add Delete		Random Switch		k-DA		Generalization	
	Level I	**Level II**	**Level I**	**Level II**	**Level I**	**Level II**	**Level I**	**Level II**
Cassandra	16/10492	50/24136	16 /18048	80 /83594	2 /6081	NA	15 /9424	NA
Flink	14 /11904	40 /36966	12 /27840	50 /113154	2 /8060	10 /40255	15 /6813	65 /32996
Groovy	14 /5140	40 /11610	12 /7120	60 /41296	2 /5757	NA	10 /3218	40 /13082
Ignite	12 /7996	50 /36306	12 /22762	70 /129756	2 /9449	NA	20 /5668	65 /41083
OpenStack	16 /14712	50 /48366	12 /36540	60 /171220	2 /9725	7 /74104	20 /12155	100 /47938
Qt	8 /46936	30 /144314	6 /82972	40 /491580	2 /119411	5 /514410	20 /12912	NA

Table 4.3: "x/y" represents a configuration value x and its corresponding number of link changes y, for different IPR Privacy Levels. The cells in green represent the technique achieving that Privacy Level with the least number of links changed for a given project.

Across all techniques, Generalization requires changing the least amount of links to achieve Privacy Level I (4/6)**, followed by k-DA** (2/6)**.** In Table 4.3, we can see the number of links changed by each technique for each dataset to achieve a particular Privacy Level. For Privacy Level II, it is able to provide privacy with the least link changes for 3 out of the 6 datasets, with Random Add/Delete doing so for the remaining three datasets. For Privacy Level II, we see that Generalization and Random Add/Delete perform equally well, providing privacy with the least links changed to three datasets each. Every technique changes a different number of links to reach a particular privacy level. Across all the datasets, Random Switch changes the most links to achieve Privacy Level I (Table 4.3). For instance, in Figure 4.4, we see the number of links changed to achieve

different privacy levels for the Apache Ignite dataset. Generalization is able to provide almost Privacy Level I for 4,285 links changed in the Apache Ignite datasets. At that same number of links changed, privacy provided by Random Add/Delete (59%), and Random Switch (51%) is lower.

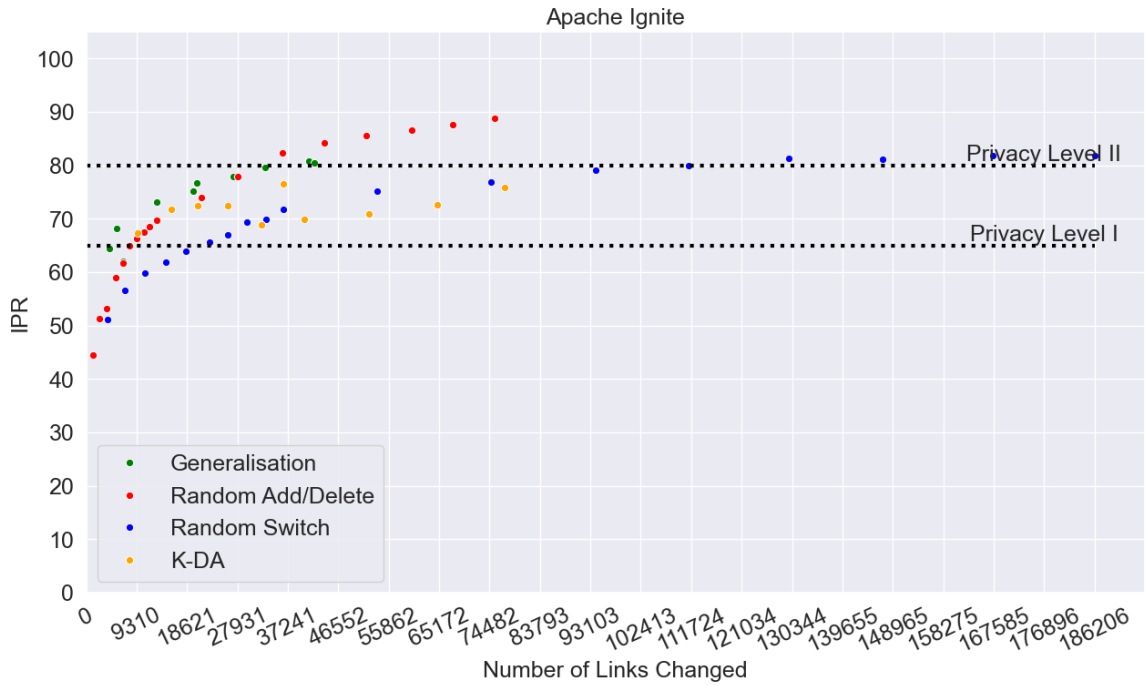


Figure 4.4: The number of links changed in Apache Ignite to reach different privacy levels

To understand how invasively graph anonymization changes the distribution of values of a given feature between the non-anonymized data and anonymization configuration level 1, we compare, for each feature, its non-anonymized distribution of feature values to the anonymized distribution of feature values. For example, we compare OpenStack's k-DA anonymized *la* feature's distribution ($k = 2$) with the non-anonymized distribution. Wilcoxon tests (with $\alpha = 0.01$) on each dataset and for each feature show that each technique at its lowest configuration level changes at least 3

features' distribution in a significant manner. This shows that anonymization causes either the quasi-identifiers or the sensitive feature distribution to change, leading the query to give different results and an increase in IPR scores.

Summary of Research Question 1

Graph anonymization techniques are effective in providing privacy to the JIT metrics. All techniques are able to provide privacy scores greater than 65% to all the datasets. For each dataset, there is one technique that achieves a privacy score greater than 80%. Random Switch and Random Add/Delete are able to provide Privacy Level II to all the datasets, with only Random Add/Delete being able to provide privacy scores close to 90% to all datasets. Generalization is able to provide both privacy levels with the least number of links changed in (7/12) cases, followed by Random Add/Delete (3/12) and k-DA (2/12).

4.3.2 RQ2: How do different graph anonymization techniques affect the performance of the JIT defect prediction model?

Across all projects, graph anonymization techniques that reach Privacy Level I, see a median decrease in AUC of 1.45% and in Recall of 5.35% (average of 1.98% AUC and 7.71% Recall). Furthermore, reaching Privacy Level II involves a median drop in AUC of 6.44% and in Recall of 20.29% (average of 7.28% AUC and of 22.42% Recall). In Table 4.4 and Table 4.5, we can see the effect of the median AUC and Recall to achieve a particular privacy level for all the datasets. The best-performing techniques at their best-performing configuration are able to reach Privacy Level I, on average,

with a 1.3% increase in AUC and a 4.45% increase in Recall across all datasets. The best-performing techniques at their best-performing configuration reach Privacy Level II across all datasets with an average decrease of 1.37% in AUC and 5.51% in Recall.

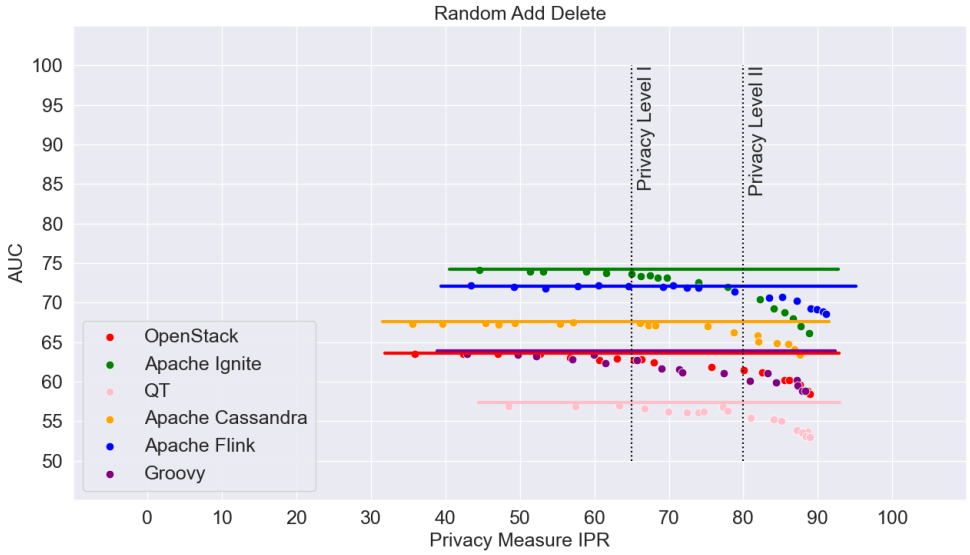
	Random Add Delete		Random Switch		k-DA		Generalization	
	Level I	**Level II**	**Level I**	**Level II**	**Level I**	**Level II**	**Level I**	**Level II**
Cassandra	65.85 (-2.53)	64.78 (-4.11)	63.1 (-6.67)	57.98 (-14.24)	67.6 (0.13)	NA	67.52 (0.01)	NA
Flink	70.68 (-1.92)	69.24 (-3.93)	68.68 (-4.7)	64.54 (-10.46)	71.04 (-1.45)	70.26 (-2.52)	72.19 (0.36)	71.81 (-0.17)
Groovy	60.63 (-5.06)	59.89 (-6.22)	60.34 (-5.63)	57.65 (-9.84)	61.54 (-3.36)	NA	62.8 (-1.73)	62.26 (-2.56)
Ignite	72.0 (-2.94)	68.38 (-7.82)	67.5 (-8.88)	59.76 (-19.32)	73.46 (-0.61)	NA	73.73 (-0.51)	71.5 (-3.52)
OpenStack	61.19 (-3.85)	60.23 (-5.36)	59.72 (-6.28)	56.18 (-11.84)	59.6 (-6.41)	58.62 (-7.95)	63.95 (0.39)	62.76 (-1.48)
Qt	55.42 (-3.37)	53.77 (-6.24)	54.62 (-5.0)	51.8 (-9.89)	54.69 (-4.84)	54.58 (-5.02)	56.42 (-1.46)	NA

Table 4.4: Median AUC score with the relative change in AUC from the non-anonymized baseline in parentheses at different Privacy Levels.

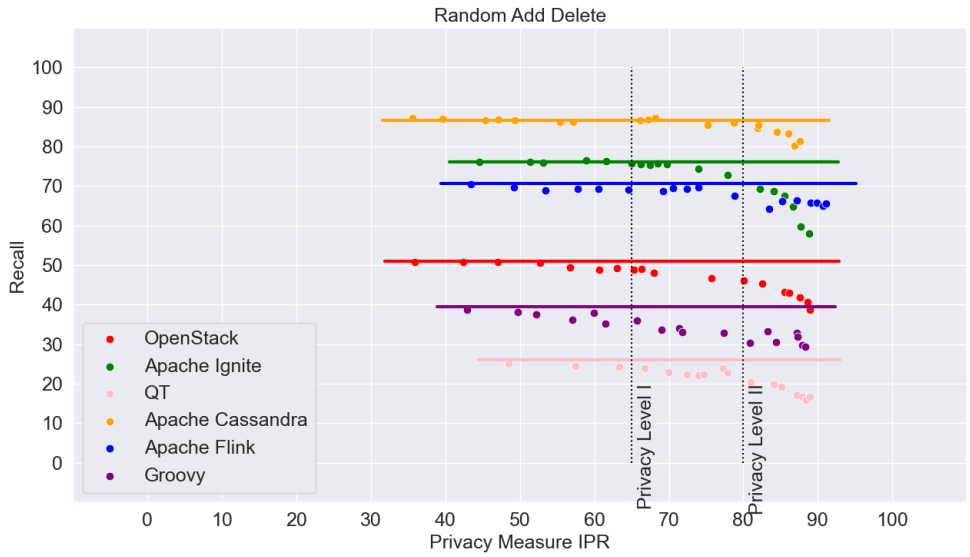
	Random Add Delete		Random Switch		k-DA		Generalization	
	Level I	**Level II**	**Level I**	**Level II**	**Level I**	**Level II**	**Level I**	**Level II**
Cassandra	85.39 (-1.43)	83.49 (-3.62)	85.88 (-1.17)	78.06 (-10.17)	85.18 (-2.02)	NA	84.45 (-2.65)	NA
Flink	66.26 (-6.31)	65.72 (-7.07)	67.2 (-5.41)	66.08 (-6.98)	61.84 (-11.88)	59.78 (-14.82)	66.43 (-6.2)	66.4 (-6.24)
Groovy	32.82 (-16.9)	30.49 (-22.81)	35.56 (-9.7)	32.77 (-16.79)	33.72 (-13.04)	NA	37.27 (-6.12)	36.94 (-6.96)
Ignite	72.8 (-4.24)	66.14 (-13.0)	71.65 (-5.38)	59.61 (-21.28)	68.06 (-9.76)	NA	64.95 (-14.38)	57.79 (-23.82)
OpenStack	45.42 (-10.78)	42.98 (-15.58)	44.93 (-11.83)	39.1 (-23.28)	32.72 (-35.73)	30.3 (-40.48)	52.06 (1.96)	46.36 (-9.2)
Qt	20.32 (-21.85)	16.96 (-34.75)	18.97 (-27.82)	13.75 (-47.68)	13.21 (-49.64)	12.72 (-51.49)	21.4 (-17.29)	NA

Table 4.5: Median Recall score with the relative change in Recall from the non-anonymized baseline in parentheses at different Privacy Levels.

Random Add/Delete is able to provide Privacy Level I with a median decrease of 1.35% and 3.93% in AUC and Recall (on average 1.63% AUC and 6.05% Recall), and Privacy Level II with a median decrease of 5.36% and 15.58% in AUC and Recall (average 5.54% AUC and 16.65% Recall). As seen in Figure 4.5, Random Add/Delete provides an IPR score of almost 90% for all the datasets at 100% anonymized nodes. However, across projects, at this privacy score (near 90%), on average, the AUC and Recall decrease by 7.59% and 20.45%, respectively. Since any dataset with privacy scores of 65% or greater can be further shared, the graph anonymization techniques provide users with the ability to tweak the amount of anonymization performed to find a good



(a) AUC Score and Privacy



(b) Recall and Privacy

Figure 4.5: AUC/Recall vs IPR with Random Add/Delete applied. The horizontal, coloured solid line is the performance of the non-anonymized (baseline) model. The vertical-dotted lines are the different privacy levels.

trade-off between performance and privacy.

Random Switch has a higher impact on the performance than Random Add/Delete for all the datasets, where AUC and Recall decrease by a median value of 3.25% and 5.06% for Privacy Level I and 11.25% and 20.12% for Privacy Level II across all datasets[1]. One reason for this difference could be that Random Add/Delete is able to provide privacy with fewer link changes (as seen in Table 4.3) than Random Switch. This implies that Random Add/Delete causes less change to the graph structure and therefore does not degrade performance as severely as Random Switch.

For Random Add/Delete and Random Switch, we find that there is always a strong negative linear correlation (Spearman correlation $\geq$ 0.8) between the anonymization level and the AUC and Recall. As seen in Figure 4.5 for Random Add/Delete, we see that with increasing privacy scores, the AUC and Recall keep on falling. At 100% anonymized nodes, on average, the AUC is reduced by 7.59%, and Recall by 20.45% across all datasets. We observe similar trends for Random Switch as well, where, at 100% anonymization level, on average, the AUC decreases by 15.18% and Recall decreases by 27.54%. Random Switch has a more dramatic drop in AUC and Recall with an increase in privacy. The k-DA and Generalization techniques show either a weak or moderate negative correlation (with the exception of k-DA for OpenStack and Flink, which show a strong correlation).

We believe that the AUC and Recall keep falling with increasing anonymization because, as Random Add/Delete and Random Switch randomly changes the structure of the data, they break the context that is present in the graph. Changes to this context progressively cause the model to be confused and lead to poor performance. At lower

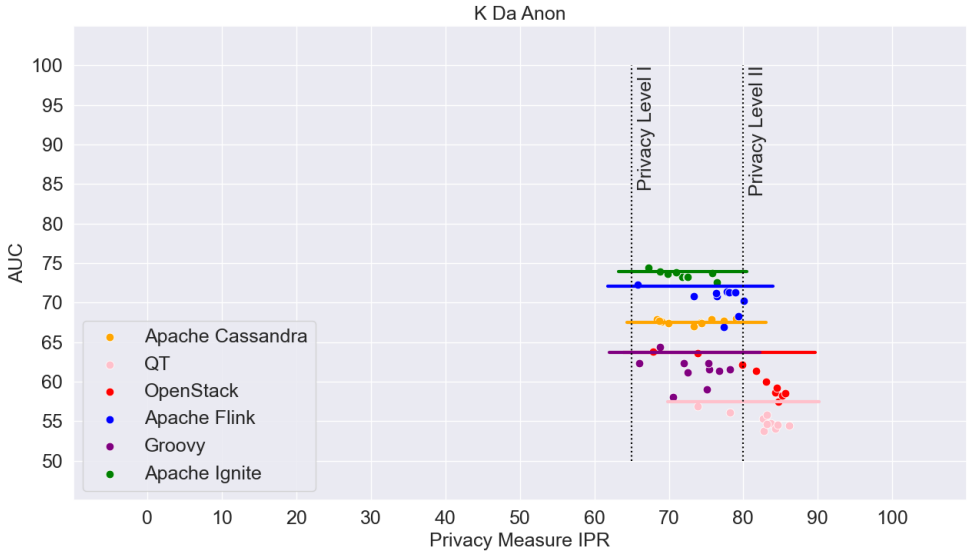
[1]To save space, we do not show the corresponding plots for Random Switch. They can be found in the online replication package [1].

configuration levels, as the changes are less invasive, the performance of the model is not significantly impacted. As the amount of anonymization increases, more changes are made to the graph, which comes at the cost of performance.

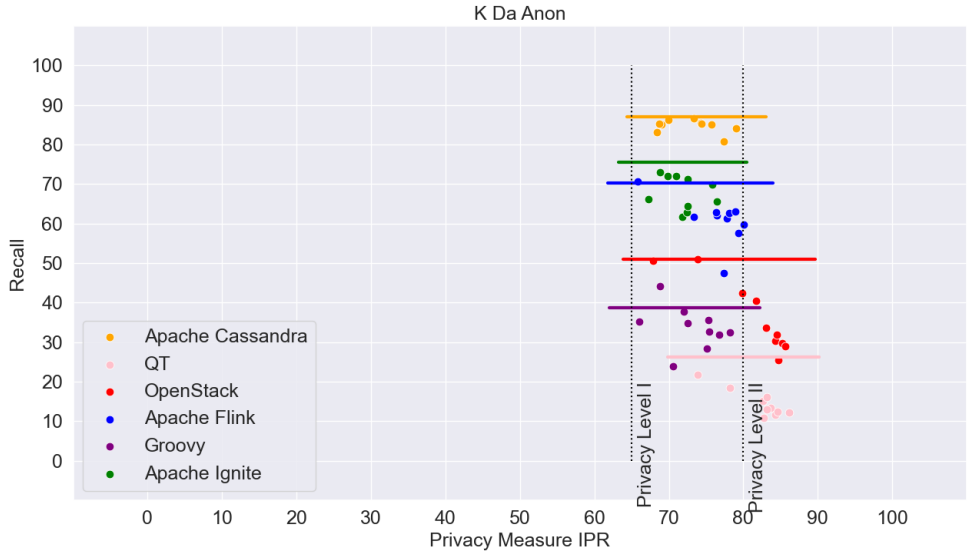
The k-DA technique's impact on Recall (Figure 4.6b) is the highest of all the techniques, with a median decrease in Recall of 9.78% and 42.83% for the two privacy levels across all datasets. For Privacy Level II, even the best-performing configuration for k-DA decreases Recall by up to 25% on average for OpenStack, QT and Flink. At the same privacy level, other techniques incur less loss in Recall; for example, Random Add/Delete drops recall by 12% on average for the same three datasets. However, k-DA's impact on AUC (Figure 4.6a) is not the highest, with AUC decreasing by a median of 0.9% and 5.49% across the two privacy levels. k-DA is only able to provide Privacy Level II to 3 datasets (Flink, OpenStack and QT), with a median AUC loss in the range of 2.52 - 7.95%.

Similar to k-DA, Generalization is able to provide Privacy Level I with a median increase of 0.1% AUC and a median decrease of 5.15% Recall. For Privacy Level II, it sees a median decrease of 1.89% in AUC and 8.57% in Recall. As seen in Figure 4.7, Generalization is unable to provide Privacy Level II to Cassandra and Qt projects even at configuration level 100. However, the privacy scores reach about 77%, with the AUC being impacted by 0.19% and -2.03%, respectively.

Generalization preserves significantly more AUC and Recall than other techniques. For a particular project, when Generalization is used for anonymization, there is a statistical difference in the AUC when compared to the other anonymization techniques,



(a) AUC Score and Privacy



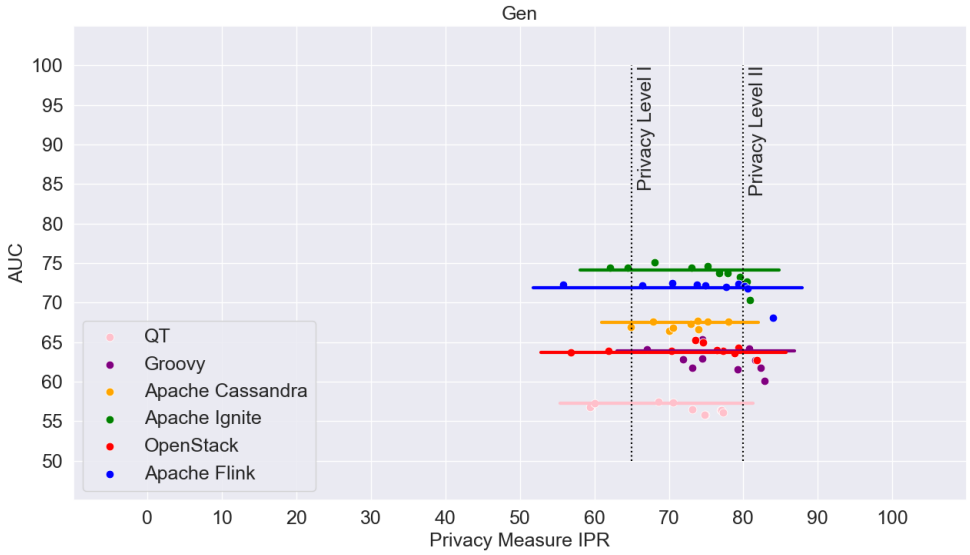
(b) Recall and Privacy

Figure 4.6: AUC/Recall vs IPR with k-DA applied. The horizontal, coloured solid line is the performance of the non-anonymized (baseline) model. The vertical-dotted lines are the different privacy levels.

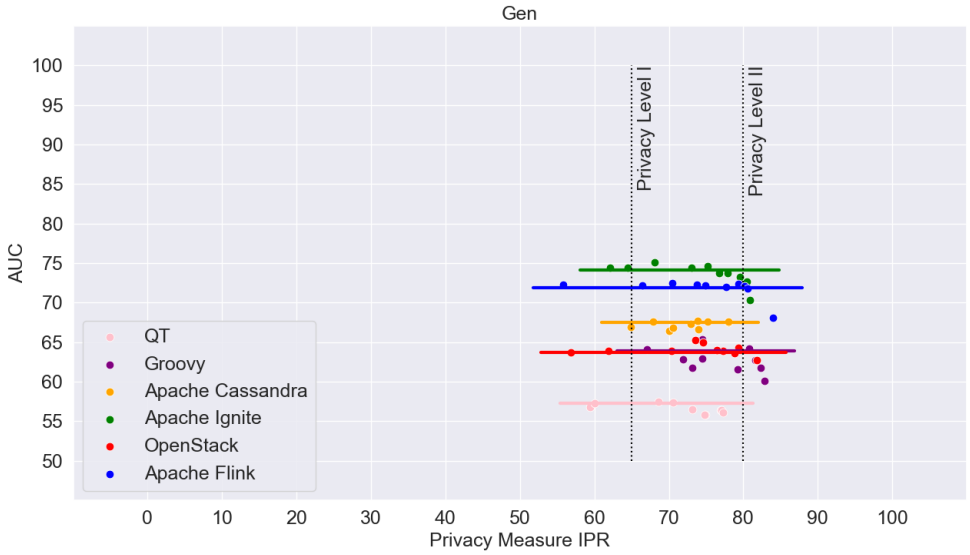
with AUC being higher for Generalization in 4/6 projects, followed by k-DA, which performs better in 2/6 projects. For Recall, Generalization's scores are statistically significantly higher when compared to other techniques in 3/6 projects, with Random Add/Delete being in 1/6 projects. In the remaining two projects, we observe insignificant differences. One reason why Generalization is able to provide privacy with the least change in performance is that it provides the different Privacy Levels with the least number of changes (Table 4.3). Therefore, the data derived from the anonymized graph undergoes fewer changes compared to others and loses its predictive power.

The graph anonymization techniques across all projects and configuration points decrease the G-Mean by a median of 2.29% and 11.76% for Privacy Level I and Level II. G-Mean shows similar trends as AUC and Recall, where with increasing anonymization levels, the model's performance G-Mean score decreases. Generalization and k-DA perform the best in terms of preserving G-Mean, being significantly better in 4 and 2 out of 6 cases, respectively. The FPR decreases (i.e., improves) by a median of 4.42% and 3.13% for Privacy Level I and Level II. However, we should keep in mind that at the same time, the Recall score decreases by 5.31% and 20.29% for Privacy Level I and Privacy Level II, reducing the performance of the model as a whole. For example, we see that k-DA statistically performs the best for FPR, reducing it (i.e., improving it) by a median of 11% and 52% for Privacy Level I and Privacy Level II. Yet, simultaneously, the Recall decreases by a median of 9.78% and 42.83%, negating the benefits. We see similar behaviour of the FPR metric for other techniques. The graphs and tables for these two metrics can be found in our online appendix [1].

In Figure 4.8, we can see how many links are changed by the techniques for the Apache Groovy dataset and the effect on the AUC and Recall of the model. For a value



(a) AUC



(b) Recall

Figure 4.7: AUC/Recall with IPR with Generalization. The horizontal, coloured solid line is the baseline (non-anonymized). The vertical-dotted lines are the different privacy levels.

of 12k changed links in Figure 4.8, Generalization and Random Switch have less impact on the AUC as compared to Random Add/Delete and k-DA. Both Generalization (0.52%) and Random Switch (1.06%) change the AUC negligibly. However, the IPR score provided by Generalization at the same level is 80% compared to 71% for Random Switch. Furthermore, at 12k changed links, Random Add/Delete is able to provide IPR scores of 80%, but the loss in AUC is almost 6%. We see similar trends for the other datasets, where for a similar number of changed links, Generalization is able to provide high privacy scores with the least impact on performance, followed by Random Add/Delete. The graphs for the remaining datasets can be found in our online replication package [1].

We note that the recall metric observes a higher impact in certain cases, especially Qt. To calculate how much the recall metric changes with changing anonymization level, for each technique, we calculate the effect size of each configuration point when compared to the non-anonymized baseline. To calculate this for each dataset, we compare the distribution of the 100 bootstrapped models for each configuration point (for example, RAD 30) to the non-anonymized performance and calculate the effect size. We find that the change in Qt's recall is not significantly different than for the other projects. For example, in Table 7 we see that Random Add/Delete decreases the Qt project's Recall more when compared to the other projects (-21% for Privacy Level II). Yet, for Qt, when statistically comparing the 18 configuration points with Random Add/Delete to the non-anonymized baseline, the effect size for the recall is insignificant, negligible or small for 13 configurations. This is similar to OpenStack's 12, Flink's 13, Groovy's 14 and Ignite's 12 configuration points. We see similar behaviour for other techniques, where the relative change in Qt's recall score is similar to other projects.

For example, for k-DA it is small in 1/10 cases same as Flink and OpenStack. Only for Random Switch the change in recall is higher compared to others, being small in 9/18 cases compared to the others being small in 12 or more. To conclude, even though the effect on Qt's recall may seem higher, the change is not substantially different from what other projects observe and depends on the technique being used.

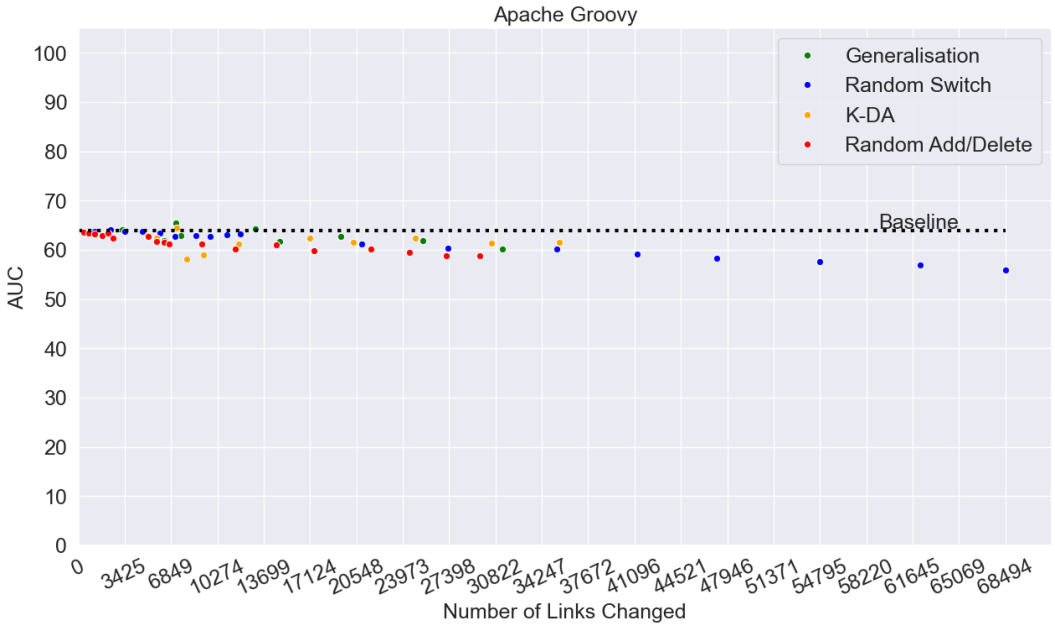
Summary of Research Question 2

The graph anonymization techniques, across all datasets, are able to provide Privacy Level I with a median drop in AUC of less than 2% and Recall of less than 6%, and Privacy Level II median drop in AUC of less than 7% and Recall of 21%. Random Add/Delete and Random Switch see a linear decrease in performance with increasing anonymization. k-DA sees the largest drop in Recall out of all the techniques. Generalization is able to provide all privacy levels with the least significant drop in AUC and Recall.

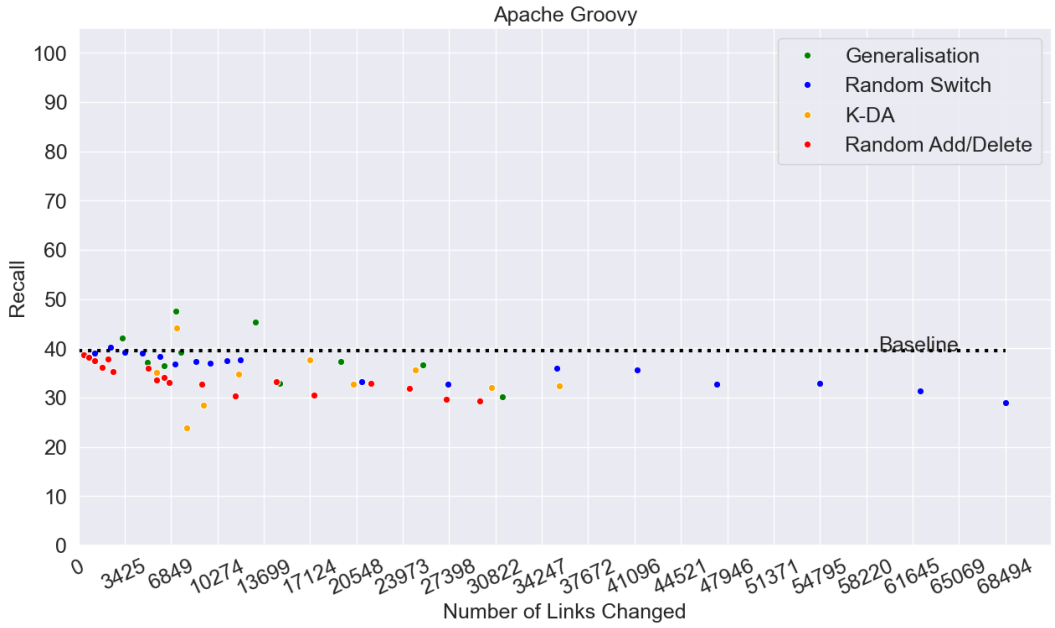
4.3.3 RQ3: How do the graph anonymization techniques compare to the state-of-the-art anonymization techniques?

For each project, with statistical significance, the MORPH, LACE and LACE2 (NGA) techniques provide higher privacy scores when compared to the graph anonymization techniques, with the NGA privacy scores in the range of 89% - 99% (Privacy Level II). As seen in Figure 4.9 for Apache Ignite, the NGA techniques are able to provide privacy scores in the range of 91% to 94% for all configurations. The graphs for other projects are available in the online replication package [1].

In Table 4.8, we see the aggregated results for the different GA techniques when



(a) AUC Score



(b) Recall

Figure 4.8: AUC/Recall versus the number of links changed for Apache Groovy, for all the techniques.

compared to the three NGA techniques across the six datasets. The results are reported as win/tie/loss with respect to the row header technique, where win indicates that the distribution was statistically significantly higher for the row header technique (e.g., Random Add Delete), tie indicates that there was no statistical difference, and loss means that there was a statistically significant drop for the row header technique (e.g., Random Add Delete).

GA techniques provide significantly less privacy than MORPH, LACE and LACE2, with the latter techniques having significantly higher privacy scores in more than 16/18 cases. Only Random Add/Delete has comparable privacy results to MORPH, where the privacy results are statistically insignificant to MORPH in 5/6 cases.

On the other hand, the impact of NGA techniques on the AUC (Table 4.6), and Recall (Table 4.7), is significantly higher than that of the GA techniques. As seen in Table 4.8, for AUC, the results are either insignificant or favour GA techniques, with Generalization performing better than NGA techniques in 17/18 cases. For Recall, we see similar results where each GA technique performs better in at least 14/18 cases, except k-DA, which does so in 10/18 cases. Interestingly, only for the Qt dataset, MORPH (NGA) performs better for Recall than k-DA, Random Add/Delete and Random Switch (GA techniques), however, it still does not perform better than Generalization in that case.

Across all datasets, the NGA techniques provide Privacy Level II with AUC and Recall dropping by a median of 21.15% and 80.34% (on average 19.81% and 66.63%). In contrast, for the same privacy level, the GA techniques suffer only a median loss of 6.44% and 20.29% for AUC and Recall, respectively. As NGA techniques always

provide Privacy Level II, they have the same impact on performance for the lower privacy levels. The GA techniques provide Privacy Level I with a 1.45% and 5.35% median decrease in AUC and Recall. For example, as seen in Figure 4.9, the best-performing GA technique for Apache Ignite provides Privacy Level II with a 5% decrease in AUC and a 9% decrease in Recall. In contrast, the best-performing NGA techniques do so with a 24% and 73% decrease in AUC and Recall. Furthermore, for Privacy Level I, the impact of the best-performing GA technique on both performance metrics is less than 1%, for instance, Random Add/Delete.

The NGA techniques decrease the G-Mean by a median of 70.07% across all projects and configuration points. For the FPR metric, the NGA techniques decrease it by 82.43% (i.e., improving it). However, as highlighted in RQ2, FPR needs to be considered along with Recall, where the gains in FPR are negated by a corresponding decrease in Recall, making the model's performance degrade as a whole. As seen in Table 4.8, statistical tests show that for both G-Mean and FPR metrics, the GA techniques perform significantly better than the NGA techniques in the majority of cases (80% and higher). The plots for each dataset can be found in our online replication package [1].

The NGA techniques do not offer any choice over the amount of anonymization and privacy. They provide high privacy with severe performance degradation and do not offer any control over either. In contrast, GA techniques offer choices where the end users can choose Privacy Level I with almost no change in performance or Privacy Level II with slight changes in performance, simply by varying the configuration level values. Graph anonymization techniques are able to provide privacy scores close to 90%, with a lower impact on performance. For example, the Random Add/Delete technique is able to provide the Apache Groovy dataset with an IPR score of 88% and an AUC

	RAD	RS	K-DA	Gen	MORPH	LACE	LACE2
Cassandra	64.785 (-4.11)	57.98 (-14.24)	NA	NA	53.87 (-20.32)	64.89 (-4.02)	54.5 (-19.39)
Flink	69.24 (-3.93)	64.535 (-10.46)	70.26 (-2.52)	71.81 (-0.17)	51.46 (-28.46)	55.69 (-22.58)	52.26 (-27.33)
Groovy	59.89 (-6.22)	57.65 (-9.84)	NA	62.265 (-2.56)	49.66 (-22.28)	51.155 (-19.95)	50.00 (-21.74)
Ignite	68.38 (-7.82)	59.765 (-19.32)	NA	71.505 (-3.52)	51.4 (-30.64)	51.465 (-30.56)	52.76 (-28.81)
OpenStack	60.23 (-5.36)	56.175 (-11.84)	58.62 (-7.95)	62.76 (-1.48)	52.73 (-17.14)	58.06 (-8.77)	50.0 (-21.43)
Qt	53.78 (-6.24)	51.80 (-9.89)	54.58(-5.02)	NA	50.32 (-12.47)	52.11 (-9.35)	50.00 (-13.02)

Table 4.6: Median AUC score at Privacy Level II with the relative change compared to the non-anonymized baseline in parentheses.

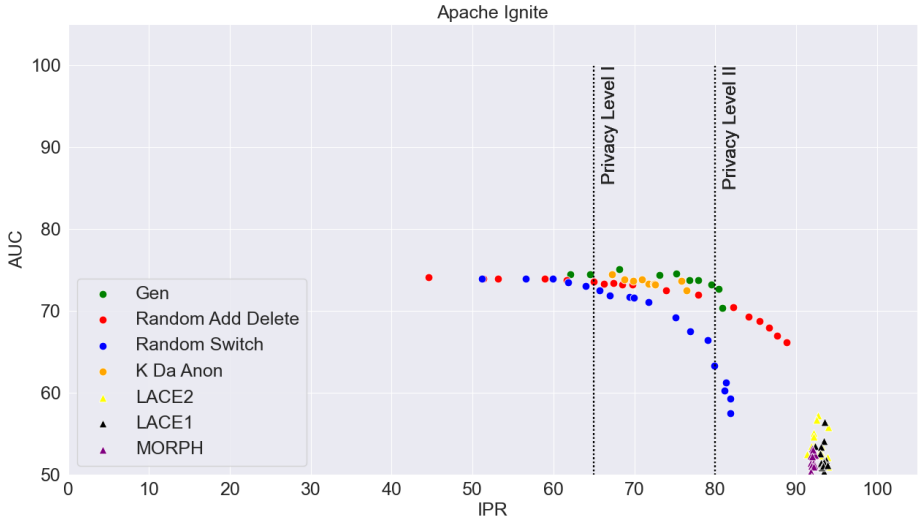
	RAD	RS	K-DA	Gen	MORPH	LACE	LACE2
Cassandra	83.49 (-3.62)	78.06 (-10.17)	NA	NA	27.57 (-68.27)	57.75 (-33.54)	14.665 (-83.12)
Flink	65.72 (-7.07)	66.08 (-6.98)	59.78 (-14.82)	66.4 (-6.24)	6.66 (-90.6)	17.14 (-75.8)	7.735 (-89.08)
Groovy	30.49 (-22.81)	32.77 (-16.79)	NA	36.935 (-6.96)	18.32 (-53.85)	9.87 (-75.14)	0.255 (-99.36)
Ignite	66.14 (-13.0)	59.61 (-21.28)	NA	57.79 (-23.82)	10.11 (-86.67)	3.755 (-95.05)	7.92 (-89.56)
OpenStack	42.98 (-15.58)	39.095 (-23.28)	30.3 (-40.48)	46.36 (-9.2)	37.69 (-25.97)	36.32 (-28.66)	0.025 (-99.95)
Qt	16.965 (-34.75)	13.75 (-47.68)	12.725 (-51.49)	NA	31.66 (20.47)	19.16 (-27.09)	0.035 (-99.87)

Table 4.7: Median Recall score at Privacy Level II with the relative change compared to the non-anonymized baseline in parentheses.

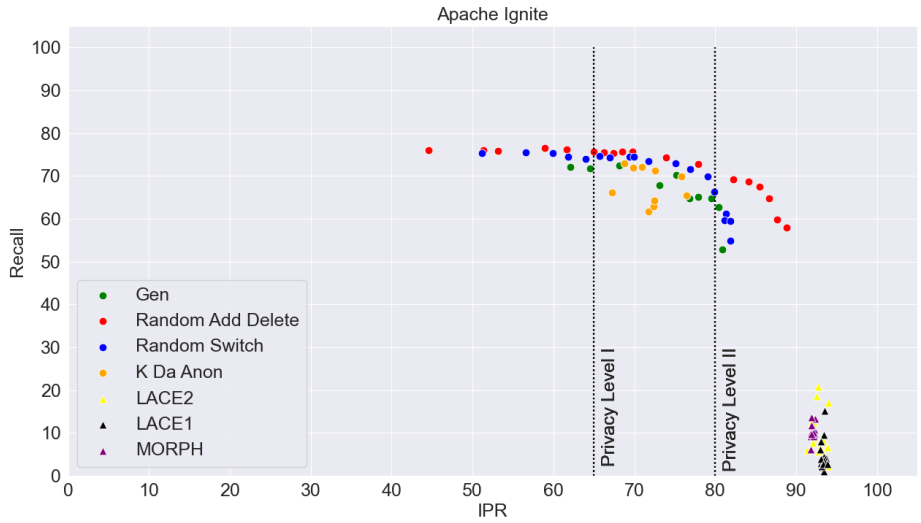
score of 60% (-5.7% from the baseline). This is in contrast to non-graph anonymization techniques where the best performing NGA technique provides a privacy score of 97.5% but the AUC is reduced to 52% (-17.7% decrease from baseline).

It is important to highlight that even though GA techniques provide statistically lower privacy, they are still able to provide Privacy Level I and Privacy Level II to all the datasets, which is private enough to be shared with others [41]. GA techniques achieve the same privacy levels as NGA techniques, with significantly higher AUC, Recall and G-Mean scores.

In Table 4.9, we show the comparison of the graph techniques to MD50. The results are reported as the aggregation of #win/tie/loss for the six studied projects, where a win means that a best-performing GA configuration performed significantly better than MD50 for a project, a tie means that a best-performing GA configuration and



(a) AUC Score



(b) Recall

Figure 4.9: AUC/Recall versus IPR score of Apache Ignite with graph and non-graph anonymization techniques.

	Overall AUC	Overall Recall	Overall G-Mean	Overall FPR	Overall Privacy
Random Add Delete	15/3/0	14/3/1	17/1/0	3/3/12	0/5/13
Random Switch	12/6/0	14/3/1	17/1/0	2/4/12	0/1/17
k-DA	16/2/0	10/7/1	15/3/0	3/7/8	0/2/16
Gen	17/1/0	14/4/0	18/0/0	2/6/10	0/1/17

Table 4.8: Overall statistical test results for each GA technique on 6 datasets compared to all three NGA techniques with at least Privacy Level I (i.e., 18 cases). They are presented as win/tie/loss.

MD50 technique performed equally well for a project, and a loss means that a best-performing GA configuration performed significantly worse than the MD50 technique for a project. The last row in Table 4.9 shows the results of the MD50 model when compared to a JIT model trained on non-anonymized data.

For the AUC metric, we find that all the GA-anonymized metrics perform equally well when compared to the MD50 model, being significantly better in 2/6 cases and comparable in 3/6 cases. For the G-Mean metric, all the GA techniques perform better in at least 4/6 cases, being insignificant in the other two cases (except Random Switch). The performance for AUC and G-Mean is similar to the performance achieved for models trained on non-anonymized data. For recall, we observe that the graph techniques perform better in just 1 or 2 cases, being worse in others. However, the FPR metric shows that even though MD50 has a better recall, its FPR rate is very high, with all the GA techniques performing at least as good in terms of FPR as MD50 for 5 cases, with k-DA and Gen even outperforming MD50 in 4 or more cases. This difference is due to the MD50 model marking 50% of the commits as buggy, which ends up marking a substantial number of commits as buggy while they are not actually buggy. For MD, the benefits of recall are counteracted by the increase in FPR.

Table 4.9 also shows that the performance of a model trained on graph anonymized

metrics relative to MD50 is similar to that of models trained using non-anonymized data, i.e., the best performing GA anonymized model for a dataset and technique performs similar to the non-anonymized JIT defect prediction models, incurring no additional performance penalty (except FPR), while providing an IPR score of more than 65%. Overall, the models trained on graph anonymized data outperform the MD50 models in terms of FPR and G-Mean, while in more than 50% of the cases performing better or similar to MD50 models in terms of AUC.

	Overall AUC	**Overall Recall**	**Overall G-Mean**	**Overall FPR**
Random Add Delete	2 / 3 / 1	1 / 1 / 4	4 / 2 / 0	2 / 3 / 1
Random Switch	2 / 3 / 1	2 / 0 / 4	4 / 1 / 1	2 / 3 / 1
k-DA	2 / 3 / 1	0 / 2 / 4	4 / 2 / 0	4 / 1 / 1
Gen	2 / 3 / 1	1 / 1 / 4	4 / 2 / 0	5 / 0 / 1
Non-anonymized data	2 / 3 / 1	1 / 1 / 4	4 / 2 / 0	4 / 1 / 1

Table 4.9: Overall statistical test results for each GA technique compared to the Manual Down technique on the 6 datasets (i.e., 6 cases). They are presented as win/tie/loss.

Summary of Research Question 3

NGA techniques are able to provide significantly higher IPR scores when compared to GA techniques, with scores ranging between 89% to 98% for all datasets. However, this privacy (Privacy Level II) comes at the cost of a significant performance decrease, where NGA techniques reduce AUC and Recall by a median of 19.82% and 66.64%, compared to 7.28% and 22.43% for GA techniques, respectively. Furthermore, NGA offers no ability to control the trade-offs between privacy and performance. When compared to MD50, GA techniques perform better or similar for at least 5/6 datasets for AUC, G-Mean and FPR.

4.4 Discussion

Implications for Researchers: Our work shows that graph anonymization techniques are able to provide increased privacy to JIT defect prediction data without impacting the trained models' performance. Unlike tabular techniques, the graph anonymization techniques are able to maintain performance, because, upon a change to the graph, they do not randomly alter the values of other features, but retain their logical context. Future research should explore the use of graph anonymization techniques to provide privacy for other types of software analytics data and models, for example, build status prediction [62], effort estimation [58], and more.

We would also recommend future work to combine multiple anonymization techniques together, for example, combining Random Add/Delete and Generalization techniques to observe how they change the performance of the models when applied one

after the other on the data. Further exploration of whether techniques from the same graph anonymization category exhibit similar results would be useful in understanding how differently the techniques of the same category help data gain privacy.

Implications for Developers: Developers can use graph anonymization to provide privacy without sacrificing performance. For IPR scores greater than 65%, any of the graph anonymization techniques can be used without significant change in performance. In our case studies, we observed that applying the Random Add/Delete and Random Switch techniques to 20% of the nodes and for k-DA and Gen the configuration values 3 and 20, respectively, can be chosen to provide IPR scores 65% and greater.

For higher privacy requirements, we would recommend applying Random Add/Delete and Random Switch to 50% and 80% of the nodes, with Random Add/Delete having a lesser impact on performance. There is no consistent configuration value for k-DA and Gen that provides IPR scores of 80% or greater. We would recommend evaluating with higher values, k>10 and Gen>50. If the task is performance-sensitive, we suggest using Generalization, as our study shows that it provides significantly more performance as compared to the other techniques. Future work can explore if there are values where the behaviour of the Gen and k-DA techniques becomes consistent.

We also observe that privacy and performance have an almost inverse relation, since when privacy scores increase the performance stays constant for some time before decreasing, and vice versa. Higher privacy levels provided by RAD and RS are always accompanied by higher performance degradation. Trade-offs need to be made between both, where projects that have higher performance can be anonymized to a higher degree without the models becoming unusable. However, the same should not be done for projects with lower performance, where anonymization can make the

models unusable.

Challenges: The main challenge of GA techniques is the additional setup required to create the underlying data graph and generate the metrics. Since drift might require retraining models on future versions of the underlying git data, future work should look into a way to incrementally update the graph, allowing more swift re-training or fine-tuning of models. This would ease the use and adoption of graph anonymization techniques for providing privacy. Finally, we believe that future research should explore the integration of the three families of anonymizing ML models, i.e., (1) graph anonymizing input data (our focus), (2) adopting specialized learning algorithms like federated learning [63], and (3) anonymizing the output of model predictions [16].

4.5 Threats to Validity

In this section, we highlight the threats to the validity of our chapter are:

4.5.1 Construct Validity

Git-Graph specific results: It is possible that if the graph anonymization techniques are used on a different knowledge graph, the result of our study would be different. However, as graph anonymization techniques have been able to provide privacy to knowledge graphs of different domains, we believe that they would be able to provide privacy to knowledge graphs for different software analytics tasks made up of different entities (Reviewers, Comments, etc.). Verification of such applicability has been left for future work.

4.5.2 Internal Validity

Input threat: The process of graph anonymization and generating the JIT defect prediction metrics is time-consuming. Therefore, we do not repeat the graph anonymization at the same configuration level multiple times. However, we mitigate this threat by repeating the anonymization for multiple different configuration levels of the anonymization techniques. In doing so, we are able to establish that the results of anonymization are not an anomaly, as we see similar behaviour trends across different datasets.

Dependent metric selection: Similar to previous work for anonymization of defect prediction metrics by Peters et al. [39, 40, 41], we only used bug-proneness as the target-dependent variable for the model. The results might change if a different dependent variable is used, for example, effort-aware bug-proneness [20]. However, we hypothesize that the results will not dramatically change by graph anonymization of the metrics, based on findings like the fact that anonymization improves the FPR of the JIT model by about 3%. This indicates that graph anonymized metrics do not dramatically change the models' ability to avoid classifying non-buggy commits as buggy, thereby avoiding wasting developers' effort. We propose future work to evaluate the performance of the models with effort-aware (and other) dependent variables.

Hyperparameter tuning: It is possible that tuning the hyperparameters of the model would result in different performance metrics. Falessi et al., [11], in their work for understanding the effect of dormant defects, do not perform hyperparameter tuning. This is because their aim is not to produce the most optimal models, but instead to highlight the fact that dormant defects decrease the recall of defect prediction. Similarly, the aim of our study is not to produce the most optimal models but rather to assess the change that comes with privatizing the datasets. We measure the change

in model performance anonymization brings compared to the non-anonymized baseline. A model that has been tuned can be expected to show a better performance, albeit with a similar trend of change in performance metrics as our current results.

Testing/training split: We use an 80/20 chronologically sorted training/testing split to train the models.[53]. We then use out-of-sample bootstrapping to ensure that the models have less bias and variance [54]. It would also ensure that the model's performance is good or bad based on the random draw of elements but rather consistent overall.

4.5.3 External Validity

Privacy metrics: We only use IPR for measuring privacy results, which has been used in multiple previous works [28, 39, 40, 41]. However, it is possible that a different privacy metric may produce different results. We leave the evaluation of privacy with different metrics to future work.

Project-specific: It is possible that our results might change if different projects were studied or projects in languages other than Java, C++ and Python. However, the graph anonymization techniques and the git-graph generated are language-independent. They only operate on the knowledge graph generated from the repositories, regardless of the underlying programming language used. We used projects of different sizes and use cases to ensure wider applicability. Therefore, we believe that the findings should be valid even for different projects.

Random Forest Classifier: It is possible that the results might change if a different classifier is used. However, Random Forest is commonly used for JIT defect prediction models and has been found to perform at least as good as deep learning models [68].

4.6 Conclusion

Our case study on 6 large, long-lived projects with 4 graph anonymization techniques shows that they are an effective way to provide privacy to the data while maintaining performance. All 6 datasets under study were able to gain privacy scores greater than 65% with no significant loss in performance and greater than 80% with slight changes in performance. Out of the four graph-anonymization techniques, Generalization was the best performing in terms of preserving AUC and Recall, followed by Random Add/Delete. All techniques were able to provide Privacy Level II at lower configuration levels without significant anonymization effort.

When compared to state-of-the-art anonymization techniques like MORPH, LACE and LACE2, the graph-anonymization techniques were able to provide privacy scores greater than 80% with significantly lesser change in performance metrics. Compared to MD50, all the GA techniques perform better in at least 4/6 cases for G-Mean and FPR, (except RS and RAD, which perform better in 2/6 cases for FPR) while performing similar or better for AUC in 5/6 cases. This shows that graph-anonymization techniques are an effective way to provide privacy to data.

As of now, we have been able to demonstrate that graph-anonymization techniques are effective in preserving the data's within-project predictive capabilities; however, in the next chapter, we evaluate this in a cross-project defect prediction setting. Where we validate whether even after graph-anonymization, the data is able to retain its predictive power in a cross-project setting.

CHAPTER 5

Empirical Evaluation of Graph Anonymized Metrics in Cross-Project Setting

5.1 Introduction

In the previous chapter, graph anonymized process metrics have been shown to be useful in providing privacy without sacrificing the within-project defect prediction capabilities of the just-in-time (JIT) defect prediction models. In this chapter, we evaluate the change in performance that comes with using graph anonymized metrics in a cross-project setting, and in settings where different organizations could possibly use different anonymization techniques (either graph or non-graph/tabular techniques) on the data they are contributing towards the training process. We want to evaluate whether gains in privacy, measured in terms of IPR (i.e., a measure of the amount of

change in data after anonymization), adversely affect the cross-project performance of the anonymized models.

Our empirical study on 6 large, long-lived open-source projects shows that in a cross-project prediction setting when all the training datasets have privacy scores of 65% or greater the graph anonymization techniques reduce the AUC, Recall, FPR and G-Mean of the models only by a median of 1.3%, 4.5%, 3.4% and 1.9%. This impact is significantly lower when compared to non-graph anonymization techniques, which reduce the AUC, Recall, and G-Mean by a median of 15.5%, 31.5% and 22.1%, and increase FPR by 7%. When the graph anonymization techniques are compared to an unsupervised learning technique like Manual Down, they perform similar or better in all cases for AUC, FPR and G-Mean.

Furthermore, we find that when projects are anonymized using two different graph anonymization techniques, their performance is similar to models trained using only one anonymization technique. When graph and non-graph anonymization techniques are used together, we found that they perform significantly better than only using non-graph anonymization techniques, with a median decrease in AUC and G-Mean of 6% and 15.8% (compared to 15.3% and 31.5%). Hence our empirical study is able to show that graph anonymization techniques are an effective choice for anonymizing data for cross-project prediction, even in environments where mixed types of anonymization techniques have been used.

5.2 Empirical Study Design

In this section, we describe the approach taken to answer the RQs for this chapter, this involves the datasets used, and the way the different cross-project models are built and

evaluated.

5.2.1 Dataset

Using the data created in Chapter-4, we were able to get JIT defect prediction projects for 6 large long-lived projects [30]. The selection criteria for the projects involve having an AUC of greater than 60% in within-project defect prediction, to ensure that when anonymizing the models they do not lose their predictive power immediately, and having more than 5000 commits, to ensure there are enough commits to generate a graph and anonymize. These conditions excluded the Qt project, yet we decided to still include this project due to its importance in other studies. This dataset further provides data that was anonymized using 4 graph anonymization techniques: Random Add/Delete, Random Switch, k-degree Anonymization and Generalization. This dataset includes projects shared by McIntosh et al. [31] (OpenStack and Qt) and Keshavarz et al. (Apache Projects) [27] for JIT defect prediction. The details of the projects used can be seen in Table 5.1, and correspond to the data set used in Chapter 4. For each dataset, the JIT features used can be found in Table 5.2. These features are commonly used in other JIT defect prediction studies [23, 26, 27, 52].

Dataset	Time Period	No. of Commits	Buggy Commits	Language
Qt	2011 - 2014	25,150	2002 (8%)	C++
OpenStack	2011 - 2014	12,374	1616 (13%)	Python
Cassandra	2003 - 2019	8,159	3,117 (38%)	Java
Flink	2003 - 2019	11,691	2,811 (24%)	Java
Groovy	2003 - 2019	8,059	1,614 (20%)	Java/Groovy
Ignite	2003 - 2019	12,036	2,439 (20%)	Java

Table 5.1: Summary of projects shared in the replication package.

Attribute	Description
la	Number of lines added by a commit
ld	Number of lines deleted by a commit
churn	Total number of changed lines by a commit
ent	Entropy or the spread of changes across file
nf	Number of files changed by the commit
nd	Number of directories changed by the commit
ns	Number of subsystems changed by the commit
nuc	Total number of previous changes on the files modified by the commit
ndev	Number of unique developers that changed the file
age	Average time interval between current changes and the last changes made to the file
aexp	Experience of the author, i.e., the number of commits created by the author
asexp	Subsystem experience of the author, i.e., the number of commits created by the author for the current subsystems under change
arexp	Relative experience of the author, i.e., the number of commits created by the author relative to the time passed

Table 5.2: The JIT defect prediction metrics used in the study (as proposed by Kamei et al. [26]).

5.2.2 Overall Approach

Figure 5.1 shows the general approach taken to answer the different research questions. Similar to Peters et al. [39, 40, 41] and as highlighted by Kamei et al. [23], we train cross-project models by choosing one project as the test project, while the remaining projects are used to train the models. We train Random Forest Classifiers as opposed to using deep learning techniques because studies by Pornprasit et al. [42] and Zeng et al. [68] show that traditional classifiers perform better than deep learning techniques at within-project and cross-project JIT defect prediction and can be trained in less time.

For the training of models, we use the out-of-sample bootstrapping technique, which produces more robust models that have less variance and bias. For each bootstrap iteration, from each project's training split, we select 500 random elements. These elements from different projects are combined together to train a classifier. To test the

performance, we take 500 random samples from the testing project and test the classifier's performance. This process is repeated 100 times and the final performance metrics are reported as the average of these 100 runs, which ensures that the performance results are consistent. The test set data chronologically comes after the training data.

In earlier work, Peters et al. [41] anonymize data using MORPH, LACE and LACE2 to train models by first combining all of the non-anonymized training data together, then anonymizing it to build the training model. Collecting all non-anonymized data together implicitly implies that private data has been shared across different organizations without any anonymization, violating the requirement for private sharing of data. Therefore, to remedy this, we anonymize each project's data separately, before combining it together to train models. This ensures that non-anonymized data is not shared across different organizations and that only privatized data is shared and used for training the models.

The defective and non-defective commits for classification are unbalanced with there being a lesser number of defect-introducing commits. Imbalance classes can lead to a decrease in performance [25]. To address this problem, we balance the classes using the SMOTEENN variant of SMOTE [21]. We over-sample the minority class, the defective class, using the SMOTEENN technique. This is performed on the combined anonymized training elements for each iteration of the bootstrap.

5.2.3 Baseline techniques

To evaluate how effective graph anonymization techniques are when compared to tabular techniques, we compare their performance to the state-of-the-art tabular (non-graph) anonymization techniques, MORPH, LACE and LACE2 [39, 40, 41]. These techniques have been used in other studies to compare the performance of anonymization

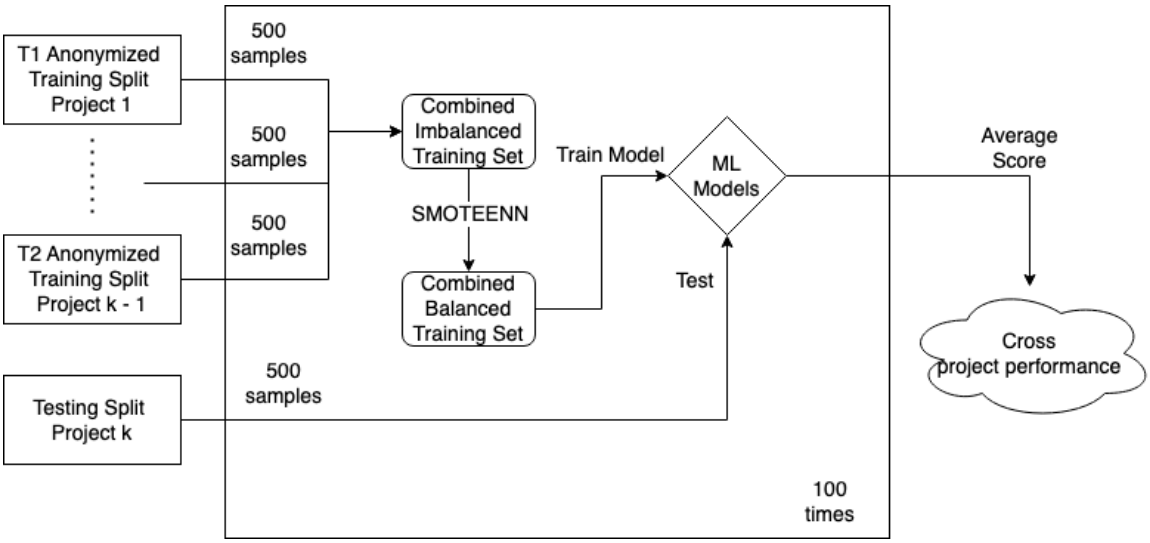


Figure 5.1: General approach for answering the research questions

methods [28, 30, 63]. Using the data from Chapter-4, we run the NGA anonymization techniques [60] on each project's data once for each configuration value. We fine-tune the techniques by testing all their configuration values between 0 to 1 in steps of 0.1. The data is split into the same 80/20 train/test split as the graph anonymized (GA) data, where only the training split is anonymized. When building models using these NGA techniques, we follow the same approach of bootstrapping the models and using SMOTTEENN for balancing classes as highlighted in Figure 5.1.

We also evaluate the performance of the GA metrics against the MD model in a cross-project setting as well. As the MD model does not require any training data, it only needs the testing data. The MD model marks the top $x\%$ of commits as buggy, for our case study we chose the value of x as 50 [19, 30, 64]. To ensure a fair comparison, we use the test splits generated above, using 500 elements to evaluate the performance of the model. The elements are sorted based on the *number of lines added* feature and the top *50%*, i.e., 250 are marked as buggy and the remaining as not buggy. This process

is repeated 100 times to get a better estimate of model performance.

5.2.4 Measuring Cross-Project Prediction Performance

To evaluate the performance of the models, we use the AUC, Recall, FPR and G-Mean performance metrics used by other studies [26, 30, 31]. We do not use precision as the dataset is imbalanced. The AUC or Area Under the Curve is used to define a model's overall predictive power and is useful in cases where there is an imbalanced dataset. The value ranges from 0.5, indicating random guessing, to 1, indicating perfect discerning capability. Recall is the ratio of true positives to true positives and false negatives, and it measures the number of actual defects that are found. FPR is the measure of how many bugs are incorrectly identified as buggy. It is calculated as $\frac{FP}{FP+TN}$. The lower this measure, the better the model performance, as that would signify fewer commits are getting incorrectly classified as buggy, thereby wasting less time in fixing the wrong commits. Finally, we also use the G-Mean as it is useful in the evaluation of imbalanced datasets. It is the geometric mean of sensitivity (recall) and specificity ($\frac{TN}{FP+TN}$) of the model. The higher this metric, the better the model performance. Observing change in Recall evaluates how anonymization changes the model's ability to correctly predict whether a commit is a bug or not.

5.2.5 Methodology for Research Questions

RQ4.1: Do graph anonymized change metrics retain their performance in a cross-project defect prediction setting?

Motivation: In the previous Chapter-4, we evaluate the performance of the JIT defect prediction metrics in a within-project setting. We want to evaluate how the performance of the graph anonymized metrics changes in a cross-project setting. Does graph anonymization make defect prediction harder in a cross-project defect prediction?

RQ4.2: How do graph anonymized change metrics compare to the state-of-the-art anonymization techniques and unsupervised learning techniques when used in a cross-project setting?

Motivation: As graph anonymization of software analytics data is a novel approach, we want to compare it to state-of-the-art techniques in a cross-project setting, since prior work did this comparison in a within-project setting. This will help us evaluate the effectiveness of the GA techniques when compared to NGA techniques in a cross-project setting. Comparing to MD will also help us determine whether the GA techniques perform better than MD model.

Approach for RQ4.1 and RQ4.2: To evaluate the impact of GA and NGA techniques on the performance of the models, we use the general evaluation approach highlighted in Figure 5.1. The projects are anonymized using different GA and NGA techniques using all their different configuration values. The models are built where one project is used as a test set and the remaining are used as anonymized training projects. For example, if the Qt project is selected for testing, all the remaining projects become the

training set. This process is repeated for all the anonymization techniques and all their configuration values, treating each project as the test set and the remaining projects as the anonymized training set.

Since an IPR score greater than 65% has been defined as adequate for sharing data between anonymization [30, 41], for the two sub-RQs, we only discuss configurations where all the training projects achieve a privacy score greater than 65%. This ensures that all the trained models are private enough to be shared, regardless of their performance change.

To test for statistical significance between the performance numbers of the models, we use the Mann-Whitney U ($\alpha = 0.05$) test [61] with Bonferroni correction to determine if there is a statistically significant difference between the two groups being compared. For the AUC, Recall, FPR and G-Mean performance metrics, we compare a technique (for example, Random Add/Delete) with all other techniques to find if there is a statistically significant difference in performance.

RQ5: How does the performance of models change when trained using different types of graph anonymization techniques?

Motivation: As different organizations have different requirements for privacy, they might use different graph anonymization techniques on their data. In this RQ, we want to evaluate how the performance of a model changes when the defect predictors are generated using different graph anonymization techniques. Is it feasible to use different techniques together to build a classifier? Do these predictors offer a better balance in performance and privacy as compared to predictors built from only one technique?

Approach: We want to evaluate how using a model trained on data anonymized

using two different techniques impacts performance. Therefore, in this RQ, we pair up the 4 GA techniques together, yielding 6 combinations (for example, RAD and RS). We explain the approach taken to train the models below for the 6 projects, using the example of RAD and RS techniques:

1. Train the project where all the 5 training projects are anonymized using the RS technique, and none using the RAD. Test the performance of the model using the 1 test project selected.

2. Increase the number of RAD anonymized projects by 1, making the RS anonymized projects to be 4 (RAD is 1). Using this, train a model by combining the data anonymized using the two techniques.

3. Repeat the above step, by incrementing the RAD anonymized projects by 1, till all the projects are now anonymized using RAD (i.e., RAD 0, 1, 2, 3, 4, 5 anonymized projects), and none using RS.

The above approach is repeated for all 6 projects where one project is treated as a test project and the others as train projects, and using all of the technique pairs. This approach allows us to evaluate how the performance changes with different numbers of projects chosen to be anonymized using either one of the two techniques.

Furthermore, for each technique pair and ratio of techniques, we also change the projects chosen for the two techniques keeping the test project the same. For example, for the case where 1 project is anonymized using RAD, we train one model where Cassandra is anonymized using RAD, another where Ignite is anonymized using RAD and so on. This process is then repeated by selecting another test project and changing the projects chosen to anonymized by either of the techniques. This helps us establish

whether the change in performance is due to a more influential project or whether it is independent of the project chosen. To analyze statistical differences in the performance of different technique pairs, we use the Mann-Whitney U test ($\alpha = 0.05$).

All the technique pairs (6) with all the datasets (6), ratios (6) and the choosing different projects for the same ratio (30) yield a large combination. (6,480). To limit this, we perform the tests with one configuration value of each technique. For GA techniques, this configuration point is chosen where the privacy gains stabilize i.e., where increasing levels only yield diminishing returns (from chapter 4's RQ1). The configuration value chosen for RAD is 50, RS is 50, k-DA is 18 and Gen is 50.

RQ6: How do predictors build using a combination of graph anonymization and non-graph anonymization techniques work together?

Motivation: Graph anonymization techniques' ability to provide privacy to defect prediction data has only recently been proposed [30]. It is possible that they might be used in a scenario where certain datasets are anonymized using GA techniques and others using NGA techniques. With this RQ, we want to identify the change in performance that comes when certain projects are anonymized using GA techniques, while others use NGA techniques. How does the performance of the model change when data has been anonymized using two different classes of anonymization techniques, i.e., tabular and graph techniques? Are these two techniques compatible with one another or not?

Approach: Similar to the previous RQ, a model is built using two anonymization techniques. However, in this case, one technique is one of the four GA techniques,

and the other is one of the three NGA techniques, giving us 12 combinations. We select one GA and NGA technique, suppose RAD and LACE2, and change the number of projects anonymized using each of the techniques and observe how their performance changes. We first anonymize all projects using the NGA technique (LACE2) and none using the GA technique. At each step, we increase the number of GA anonymized metrics, going from 0 to 5 in increments of 1.

For the GA techniques, we choose the configuration values selected in the previous RQ, i.e., for RAD we use 50, for RS we use 50, for k-DA we use 18, and for Gen we use 50. For the NGA techniques, we chose the default configuration values provided by their replication package [60].

5.3 Preliminary Study

Before performing anonymized cross-project defect prediction, it is necessary to ensure that the cross-project prediction is viable with non-anonymized projects on our study's data to evaluate whether the cross-project performance is comparable to the within-project performance of the models. This will also act as the baseline to which the anonymized cross-project model performance will be compared, the training data in this case is not anonymized. We use a similar approach as highlighted in Section 5.2, where we select one project for testing and all remaining projects are used to bootstrap 100 training models whose average score is then reported. We compare the AUC and G-Mean performance metrics of the cross-project and within-project models.

For the non-anonymized projects, their cross-project AUC and G-Mean improved by a median of 1.1% and 2.1% compared to their within-project performance. In Table 5.3, we see the performance for the cross-project defect prediction. For both the

Project	AUC - Cross	AUC - Within	G-Mean - Cross	G-Mean - Within
Cassandra	67.72	67.52	65.31	64.04
Flink	71.21	72.08	71.26	71.85
Groovy	63.81	63.68	59.25	57.98
Ingnite	75.18	73.91	75.2	73.65
OpenStack	66.11	63.68	65.67	62.12
Qt	63.22	57.47	58.91	46.13

Table 5.3: Average cross-project performance compared to the average within-project performance of the various datasets (preliminary study).

OpenStack and Qt projects, in a cross-project setting, their AUC increases by 3.8% and 10% and their G-Mean increases by 5% and 27% compared to the within-project performance, respectively. The Qt and OpenStack projects have 8% and 13% of commits as buggy commits, which is lower than other projects, which have between 20-38%. When cross-project models are built for this project, the classifiers are better able to understand the difference between buggy and non-buggy commits as they see more examples of buggy commits. We see a slight decrease in G-Mean and AUC performance for the Apache Flink project, of about 0.8% in both cases.

On conducting the Wilcoxon signed-rank test ($\alpha = 0.05$) between cross-project and within-project AUC and G-Mean metrics, we find that there is no significant difference. This shows that cross-project defect prediction is as effective as the within-project defect prediction capability for the six projects under study.

5.4 Empirical Study Results

In this section, we discuss the results of the research questions we set out to answer.

5.4.1 RQ4.1: Do graph anonymized change metrics retain their performance in a cross-project defect prediction setting?

For all test projects, the graph anonymization techniques achieved privacy scores of greater than 65% for all the training projects. When compared to the non-anonymized cross-project performance baseline, the AUC, Recall, FPR and G-Mean decreased by a median of 1.3%, 4.5%, 3.4% and 1.9% across all configuration values. Graph anonymization techniques (as seen in Table 5.4) are able to provide privacy to JIT defect prediction data used in a cross-project setting without compromising the performance of the model.

The GA techniques have mostly consistent performance across different configuration values, staying around the non-anonymized baselines, as seen in Figure 5.2. The figure shows the distribution of performance for the GA and NGA techniques across all their configurations. In this research question, we focus on the GA techniques i.e., the first 4 plots. We do observe that the RAD and RS techniques have longer tails than the k-DA and Gen techniques, particularly for Apache Flink, indicating a wider range of performance. Although that is true, the median decrease for AUC across all projects for RAD, RS, k-DA, and Gen is similar being 1.6%, 2.1%, 1.0% and 1.1%. The G-Mean observed a similar median decrease, 1.8%, 2%, 2.3% and 1.9%.

In Table 5.5, we see that the performance of the anonymized cross-project models for AUC, Recall and G-Mean is indistinguishable from one another in the majority (60% and more) of the cases. A technique is individually compared with the other three GA techniques for the six datasets using the Mann-Whitney U test ($\alpha = 0.05$ and Bonferroni Correction). The k-DA technique performs better in most cases, performing better in 5/18 cases for AUC and G-Mean.

Dataset	RAD	RS	k-DA	Gen
Cassandra	64.77 (-4.36%)	66.67 (-1.55%)	69.48 (2.61%)	67.0 (-1.06%)
Flink	69.57 (-2.3%)	67.72 (-4.89%)	69.53 (-2.37%)	70.12 (-1.52%)
Groovy	63.32 (-0.77%)	64.06 (0.39%)	62.08 (-2.72%)	62.34 (-2.3%)
Ignite	74.2 (-1.3%)	73.92 (-1.68%)	75.3 (0.16%)	75.44 (0.35%)
OpenStack	65.38 (-1.1%)	64.48 (-2.47%)	65.21 (-1.37%)	65.57 (-0.82%)
Qt	62.58 (-1.01%)	61.8 (-2.25%)	62.68 (-0.86%)	62.68 (-0.85%)

(a) Median AUC value

Dataset	RAD	RS	k-DA	Gen
Cassandra	42.83 (-18.0%)	50.3 (-3.7%)	55.5 (6.26%)	48.56 (-7.02%)
Flink	64.74 (-9.51%)	60.9 (-14.87%)	62.69 (-12.37%)	65.56 (-8.36%)
Groovy	39.47 (-2.33%)	41.89 (3.65%)	35.64 (-11.8%)	35.48 (-12.2%)
Ignite	73.06 (-1.77%)	72.03 (-3.16%)	73.34 (-1.4%)	73.1 (-1.72%)
OpenStack	57.61 (-6.0%)	59.18 (-3.45%)	52.9 (-13.69%)	57.84 (-5.62%)
Qt	41.47 (0.92%)	40.62 (-1.14%)	39.3 (-4.34%)	38.66 (-5.9%)

(b) Median Recall value

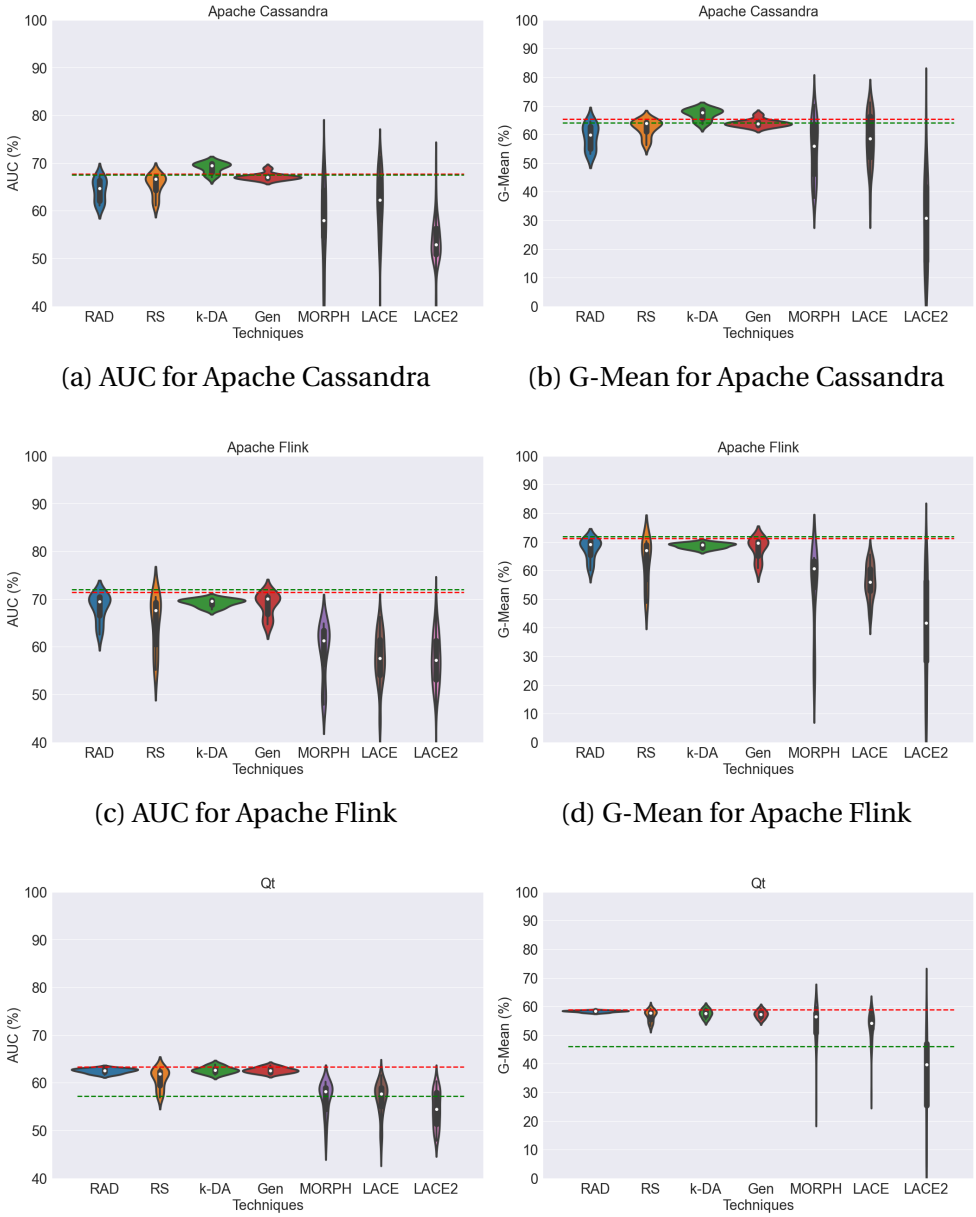
Dataset	RAD	RS	k-DA	Gen
Cassandra	13.43 (-19.73%)	17.3 (3.44%)	16.3 (-2.57%)	14.38 (-14.08%)
Flink	25.62 (-11.93%)	25.42 (-12.63%)	25.16 (-13.51%)	24.9 (-14.42%)
Groovy	12.65 (-0.39%)	14.44 (13.66%)	11.42 (-10.12%)	10.76 (-15.24%)
Ignite	25.11 (4.62%)	24.22 (0.94%)	22.88 (-4.69%)	22.46 (-6.44%)
OpenStack	27.24 (-6.04%)	29.74 (2.6%)	22.29 (-23.11%)	25.85 (-10.83%)
Qt	15.73 (6.97%)	18.27 (24.29%)	14.01 (-4.69%)	13.47 (-8.37%)

(c) Median FPR value

Dataset	RAD	RS	k-DA	Gen
Cassandra	59.87 (-8.33%)	63.9 (-2.17%)	67.72 (3.7%)	63.76 (-2.38%)
Flink	69.06 (-3.09%)	66.96 (-6.03%)	68.92 (-3.28%)	69.67 (-2.23%)
Groovy	58.18 (-1.81%)	59.6 (0.59%)	55.78 (-5.86%)	55.96 (-5.55%)
Ignite	74.08 (-1.49%)	73.82 (-1.83%)	75.15 (-0.07%)	75.4 (0.26%)
OpenStack	64.53 (-1.74%)	63.79 (-2.86%)	63.72 (-2.98%)	64.81 (-1.32%)
Qt	58.36 (-0.93%)	57.79 (-1.9%)	57.62 (-2.19%)	57.24 (-2.84%)

(d) Median G-Mean value

Table 5.4: Performance metrics with relative change from the non-anonymized cross-project baselines, for different techniques, for all configurations where the training data's IPR score was greater than 65%.



(a) AUC for Apache Cassandra

(b) G-Mean for Apache Cassandra

(c) AUC for Apache Flink

(d) G-Mean for Apache Flink

(e) AUC for Qt

(f) G-Mean for Qt

Figure 5.2: Violin plots of the performance metrics of different datasets at Privacy Level I. The red-dotted lines and green-dotted lines are the non-anonymized cross-project and within-project performance baseline, respectively. The remaining projects can be found in the replication package [2].

	AUC	Recall	FPR	G-Mean
RAD	1/14/3	4/12/2	4/6/8	2/12/4
RS	2/12/4	5/12/1	0/5/13	3/12/3
k-DA	5/12/1	3/11/4	8/10/0	5/11/2
Gen	2/14/2	0/13/5	9/9/0	2/13/3

Table 5.5: Results of the statistical tests conducted between the row techniques (GA) and the other GA techniques for the AUC, Recall, FPR and G-Mean metrics. The results are reported as - “win", where the row technique is significantly higher, “tie", where there is no statistical significance and “loss" where the row technique’s value is significantly less.

Only Random Add/Delete and Random Switch are able to provide IPR scores of over 80% to all the training datasets, with the AUC decreasing by a median of 3.6% for RAD and 9.7% for RS. The recall decreases by a median of 7.9% and 12.2%, and the G-Mean decreases by a median of 4.3% and 11.5% for RAD and RS, compared to the non-anonymized cross-project baseline. On the other hand, the FPR metric decreases (i.e., improves) by a median of 6.3% and 19.5% for RAD and RS. The RAD and RS techniques are able to provide IPR scores of 80% and greater to all the training projects at the same configuration level across all projects (RAD at 50 and higher and RS at 80 and higher). The performance decreases with increased privacy because as the data becomes more private, i.e., an increase in IPR scores, the anonymization destroys the patterns in the data that identify buggy and non-buggy commits, and decreases the performance of the models.

The k-DA and Gen techniques are able to provide scores greater than 80% to 3 and 4 projects, however, it does not happen at the same configuration value and thereby not all training projects are able to achieve IPR scores of 80% or greater at the same time. In RQ5, we explore ways that a model can be constructed for k-DA and Gen, where some

projects are anonymized using different techniques so that all training projects have IPR scores of 80% or greater.

Both the RAD and RS techniques see a progressive decrease in their cross-project performance metrics with increasing anonymization levels. For example for Apache Flink with RAD, when 40% of the nodes are anonymized, the AUC decreases by a median of 3.5%, while for 100% of anonymized nodes, the AUC decreases by 12.5%. The impact of RS is even higher on performance, with it decreasing by 8.8% at 40% anonymized nodes and 22.5% at 100% anonymized nodes. This linear decrease is because with increasing anonymization levels, the techniques destroy the context and information present in the software analytics knowledge graph to provide privacy. As such, this increase in privacy decreases the predictive capability of the metrics generated from the anonymized knowledge graph.

5.4.2 RQ4.2: How do graph anonymized metrics compare to state-of-the-art tabular anonymization techniques when used in a cross-project setting?

In Figure 5.2, we see that all NGA techniques, i.e., MORPH, LACE and LACE2 are able to provide IPR $\geq$ 65 to all projects, while the AUC decreases by a median of 12.1%, 13.1% and 20.2%, and G-Mean decreases by a median of 13.5%, 16% and 44.4%, respectively, across all datasets. The Recall decreases by a median of 5.6%, 35.7% and 54.6%, while FPR changes by a median of 81.1%, 27.7% and -67%. We see that the NGA techniques have different effects on Recall and FPR, where LACE2 improves FPR but decreases Recall. However, all the NGA techniques decrease G-Mean by a median of

22%. For all projects, when we perform the Mann Whitney-U tests ($\alpha = 0.05$ with Bonferroni correction) between all the GA and NGA techniques combined for the performance metrics, the impact of the NGA techniques on AUC and G-Mean is significantly higher than the GA techniques, with the effect size always being large for both AUC and G-mean.

	AUC	**Recall**	**FPR**	**G-Mean**
RAD	17/1/0	9/7/2	9/4/5	15/3/0
RS	17/1/0	9/7/2	9/4/5	16/2/0
k-DA	18/0/0	7/9/2	9/6/3	16/2/0
Gen	18/0/0	6/9/3	9/5/4	15/3/0

Table 5.6: Results of the statistical tests conducted between the GA techniques and the other NGA techniques for the AUC, Recall, G-Mean and FPR metrics. The results are reported as - “win", where the row technique is significantly higher, “tie", where there is no statistical significance and “loss" where the row technique’s value is significantly less.

Furthermore, performing statistical tests for the individual techniques shows that the GA techniques retain significantly more performance when compared to the three NGA techniques across the 6 projects. In Table 5.6, with each of the four GA techniques having a higher AUC in at least 17/18 cases, higher Recall in at least 6/18 of the cases, lower FPR in at least 9/18 cases and higher G-Mean in 15/18 cases. The GA techniques in more than 80% of cases preserve more AUC and G-Mean. However, for FPR and Recall the results between GA and NGA are mixed, where GA performs better in at least 50% of the cases (except k-DA for Recall) and loses in at most 27% of cases.

The NGA techniques, when compared to GA techniques, end up with a high variance in performance, i.e., inconsistent performance, as seen by the wide shape of the violin plot for the NGA techniques. In 25/318 data points across all datasets and

techniques (with different configurations), the performance of the NGA anonymized model is better than the non-anonymized baseline (more than 1%), and in 272 data points, it is worse (more than -1%). For example, for Apache Cassandra, the non-anonymized cross-project AUC and Recall scores are 67.7% and 52.2%. However, the AUC for MORPH ranges from 1x to 0.69x, for LACE from 1.03x to 0.77x, and for LACE2 from 0.89x and 0.62x. This difference is even wider for Recall, with MORPH from 1.62x to 0.34x, LACE being from 1.55x to 0.39x, and LACE2 being from 0.76x to 0.07x. The graph anonymization techniques have a smaller range, for example, the RAD technique has an AUC in the range of 1x to 0.9x and Recall in the range of 0.96x to 0.63x. However, for all datasets, it was observed that the median AUC and recall of the NGA techniques were lower than graph anonymizing techniques, being 0.92x - 0.83x times the AUC and 0.92x - 0.57x times the recall.

The NGA techniques consistently have privacy scores greater than 90% for all the training projects. However, this increased privacy comes with poor performance, and without the ability to tweak privacy scores to obtain better performance. As scores above 65% and above are deemed private enough to share [41], the users should have the ability to choose higher model performance in exchange for lower privacy scores, yet still above 65%. On the other hand, the GA techniques offer better control, with the ability to achieve adequate privacy scores with lower performance drops than NGA techniques. As seen in Table 5.7, the Random Add/Delete and Random Switch techniques can provide privacy scores greater than 80% to all training projects with more AUC than the NGA techniques.

The GA anonymized cross-project prediction models perform better or are similar to the MD models for AUC, FPR and G-Mean metric in all cases. Table 5.8 shows

	RAD	RS	MORPH	LACE	LACE2
Cassandra	61.41 (-9.32%)	61.84 (-8.69%)	54.78 (-19.11%)	61.96 (-8.5%)	53.3 (-21.3%)
Flink	65.15 (-8.51%)	55.6 (-21.92%)	59.15 (-16.94%)	56.5 (-20.66%)	64.18 (-9.88%)
Groovy	63.06 (-1.18%)	62.08 (-2.71%)	57.41 (-10.03%)	53.63 (-15.95%)	52.36 (-17.94%)
Ignite	72.1 (-4.1%)	67.08 (-10.77%)	61.51 (-18.18%)	66.06 (-12.12%)	61.53 (-18.16%)
OpenStack	64.22 (-2.85%)	58.76 (-11.12%)	57.46 (-13.08%)	56.45 (-14.61%)	53.61 (-18.91%)
Qt	62.25 (-1.53%)	57.49 (-9.06%)	57.32 (-9.33%)	58.24 (-7.87%)	56.74 (-10.24%)

Table 5.7: Change in the AUC metric relative to the non-anonymized cross-project baseline with different anonymization techniques, where the training projects have IPR ≥ 80.

the results of conducting the Wilcoxon signed rank test ($\alpha = 0.05$) between the performance metrics of the GA model and MD model for the 6 projects. When the performance of CP models is compared to the WP models [30], the CP models in no case perform worse than the MD model for AUC, FPR and G-Mean, surpassing their WP performance. The Gen anonymized models perform best when compared to MD models.

	AUC	Recall	FPR	G-Mean
RAD	2 / 4 / 0	1 / 0 / 5	5 / 1 / 0	4 / 2 / 0
RS	2 / 4 / 0	1 / 0 / 5	5 / 1 / 0	4 / 2 / 0
k-DA	3 / 3 / 0	1 / 0 / 5	5 / 1 / 0	4 / 2 / 0
Gen	4 / 2 / 0	1 / 0 / 5	5 / 1 / 0	4 / 2 / 0

Table 5.8: Results of the statistical tests conducted between the GA techniques and the MD model for the AUC, Recall, G-Mean and FPR metrics. The results are reported as - “win", where the row technique is significantly higher, “tie", where there is no statistical significance and “loss" where the row technique’s value is significantly less.

Implications: We observe that even in a cross-project setting, the graph anonymization techniques are able to mostly preserve performance while providing enough privacy to allow the sharing of the data. The NGA techniques achieve high privacy (IPR $\geq$ 89%), however, they perform significantly worse than the graph anonymization techniques in terms of preserving model prediction performance. This shows that GA anonymized

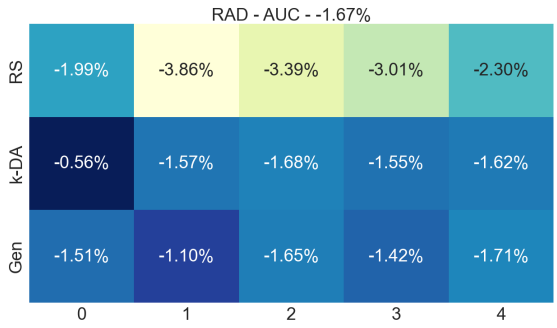
metrics are a good choice for providing privacy to software defect prediction data without compromising performance. The GA models perform better or are similar to the MD models too. Kamei et al. [24] suggest that using a large number of projects for cross-project defect prediction can be useful in building a universal predictor. As we show graph anonymization techniques are able to provide privacy to the defect prediction data without sacrificing its performance, organizations should be able to share private data while preserving their cross-project prediction capability.

Summary of Research Question 4

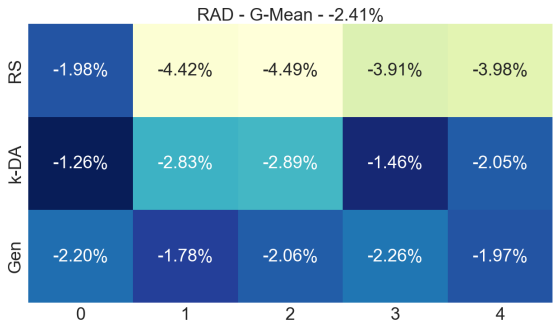
Graph anonymization techniques mostly retain the performance of the JIT defect prediction metrics in a cross-project setting, for providing IPR scores greater than 65% they decrease AUC and G-Mean by a median value of 1.3% and 1.9% and for providing IPR scores greater than 80% they decrease AUC by a median of 6.6% and G-Mean by 7.9%. The k-DA technique preserves significantly more AUC & G-Mean, and the RS technique preserves significantly more Recall compared to other GA techniques. When compared to the NGA techniques, the GA techniques obtain significantly higher AUC and Recall scores, while offering more control over privacy as the GA techniques offer IPR scores from 65% to 85% as opposed to NGA techniques, which offer only IPR scores greater than 89%, but with higher performance impact.

5.4.3 RQ5: How does the performance of models change when built using different types of graph anonymized techniques?

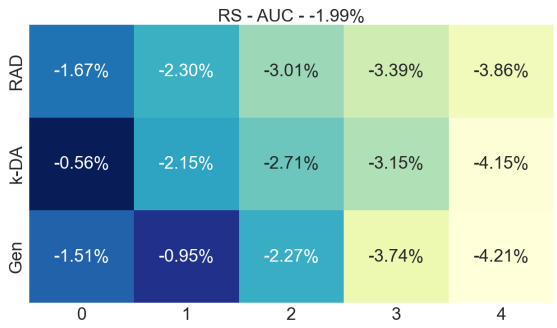
Our study shows that it is possible to build models that are trained on data anonymized using different graph anonymization techniques and still have performance similar to a model built using only one of the anonymization techniques. The pairs of techniques decrease the AUC, Recall, FPR and G-Mean by 1.9%, 4%, 2.2% and 2.5%. This decrease is similar compared to the median decrease that individual graph anonymization techniques incur. For example, k-DA and Gen techniques individually decrease the AUC by a median of 0.56% and 1.51%, which is similar to when paired together,



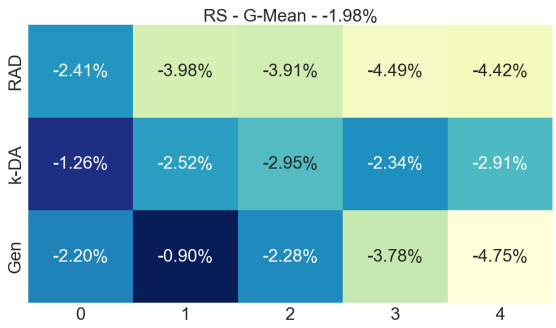
(a) Median AUC change with Random Add/Delete



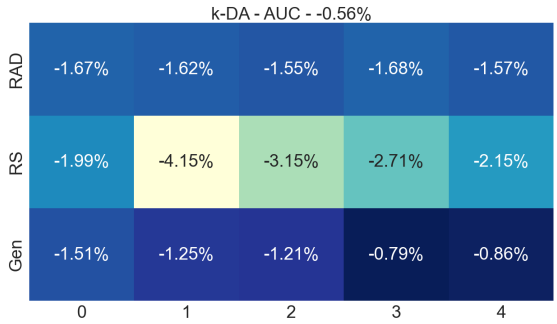
(b) Median G-Mean change with Random Add/Delete



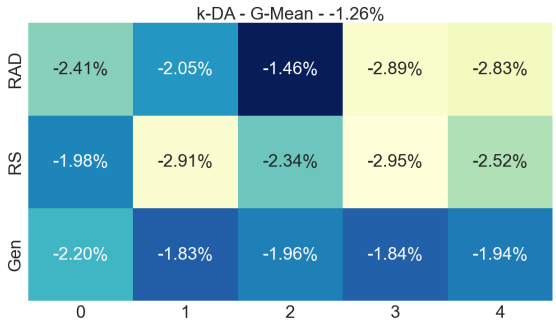
(c) AUC score for Cassandra with Random Switch



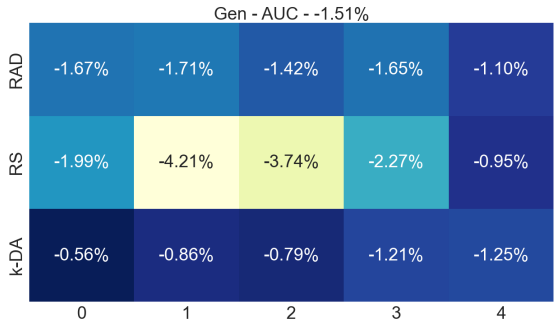
(d) G-Mean score for Cassandra with Random Switch



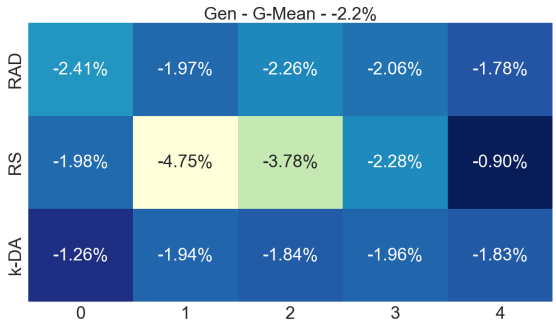
(e) AUC score for Cassandra with k-DA



(f) G-Mean score for Cassandra with k-DA



(g) AUC score for Cassandra with Gen



(h) G-Mean score for Cassandra with Gen

Figure 5.3: The heatmaps show the median change in performance that comes with using data anonymized using two different anonymization techniques across all projects. The value of each cell is the median change in performance metric achieved by anonymizing the number of projects specified by the column value using the technique specified in the title, with the remaining anonymized using the row technique. The change percent is calculated with respect to each project's cross-project performance baselines as seen in Table 5.3.

which yields a median decrease of 1.07%. In Table 5.9, we see the median change in performance for each dataset with the different ratios of the different technique pairs.

In Figure 5.3, we see the median change in performance across all datasets when

using data that is anonymized using two different anonymization techniques. For example, in subfigure 5.3a, the title of the heatmap of the graph tells the main anonymization technique (Random Add/Delete) that will be used along with the technique defined as the row header technique (for example, the row 3 technique is Gen). Each column value represents the number of datasets out of the 5 training datasets that have been anonymized using the title's anonymization technique, starting with 0. The value in *row Gen* and *column 0* has no data anonymized using Random Add/Delete, i.e., all

Technique Pair	Median AUC	Median Recall	Median FPR	Median G-Mean
RAD & RS	65.12 (-4.27%)	47.34 (-10.81%)	16.81 (-1.35%)	61.53 (-6.48%)
RAD & k-DA	67.48 (-0.79%)	51.48 (-3.02%)	16.52 (-3.05%)	64.96 (-1.26%)
RAD & Gen	66.2 (-2.68%)	47.44 (-10.63%)	15.03 (-11.8%)	62.71 (-4.68%)
RS & k-DA	67.18 (-1.24%)	53.83 (1.41%)	19.26 (13.06%)	65.24 (-0.83%)
RS & Gen	65.4 (-3.85%)	47.16 (-11.16%)	16.16 (-5.16%)	61.75 (-6.14%)
k-DA & Gen	68.74 (1.07%)	53.88 (1.52%)	16.39 (-3.81%)	66.8 (1.53%)

(a) Cassandra project

Technique Pair	Median AUC	Median Recall	Median FPR	Median G-Mean
RAD & RS	66.22 (-6.7%)	55.46 (-22.27%)	23.08 (-21.47%)	64.99 (-8.23%)
RAD & k-DA	68.76 (-3.12%)	62.06 (-13.01%)	25.11 (-14.56%)	68.16 (-3.76%)
RAD & Gen	69.25 (-2.44%)	63.22 (-11.39%)	24.94 (-15.14%)	68.71 (-2.98%)
RS & k-DA	66.24 (-6.68%)	58.28 (-18.31%)	25.58 (-12.98%)	65.45 (-7.58%)
RS & Gen	66.89 (-5.76%)	59.6 (-16.46%)	25.34 (-13.78%)	65.97 (-6.85%)
k-DA & Gen	69.94 (-1.46%)	64.57 (-9.48%)	23.84 (-18.88%)	69.43 (-1.96%)

(b) Flink project

Technique Pair	Median AUC	Median Recall	Median FPR	Median G-Mean
RAD & RS	62.79 (-2.04%)	37.44 (-9.13%)	11.98 (-7.78%)	57.0 (-4.26%)
RAD & k-DA	62.94 (-1.81%)	39.33 (-4.54%)	13.4 (3.19%)	57.99 (-2.6%)
RAD & Gen	63.41 (-1.08%)	38.72 (-6.02%)	11.86 (-8.66%)	58.05 (-2.5%)
RS & k-DA	63.46 (-1.0%)	45.64 (10.76%)	18.02 (38.76%)	60.81 (2.13%)
RS & Gen	63.88 (-0.35%)	41.62 (1.01%)	14.2 (9.28%)	59.56 (0.03%)
k-DA & Gen	61.24 (-4.47%)	31.91 (-22.55%)	9.4 (-27.64%)	53.3 (-10.48%)

(c) Groovy project

Technique Pair	Median AUC	Median Recall	Median FPR	Median G-Mean
RAD & RS	73.2 (-2.64%)	72.3 (-2.95%)	26.08 (8.24%)	73.06 (-2.76%)
RAD & k-DA	74.15 (-1.39%)	75.88 (1.87%)	28.06 (16.43%)	73.96 (-1.55%)
RAD & Gen	74.44 (-0.99%)	75.44 (1.27%)	26.24 (8.88%)	74.35 (-1.04%)
RS & k-DA	73.0 (-2.92%)	73.94 (-0.75%)	28.74 (19.27%)	72.76 (-3.16%)
RS & Gen	73.76 (-1.9%)	73.75 (-0.99%)	25.82 (7.14%)	73.65 (-1.98%)
k-DA & Gen	75.6 (0.55%)	72.5 (-2.67%)	21.76 (-9.73%)	75.44 (0.41%)

(d) Ignite project

Technique Pair	Median AUC	Median Recall	Median FPR	Median G-Mean
RAD & RS	64.21 (-3.0%)	56.99 (-7.29%)	27.97 (-3.82%)	63.26 (-3.79%)
RAD & k-DA	65.38 (-1.22%)	55.86 (-9.12%)	25.02 (-13.96%)	64.31 (-2.21%)
RAD & Gen	65.47 (-1.09%)	61.98 (0.83%)	31.04 (6.76%)	65.11 (-0.98%)
RS & k-DA	63.55 (-3.99%)	58.24 (-5.25%)	31.14 (7.1%)	62.79 (-4.52%)
RS & Gen	63.86 (-3.53%)	65.81 (7.06%)	37.08 (27.49%)	63.36 (-3.66%)
k-DA & Gen	65.16 (-1.55%)	55.89 (-9.08%)	25.68 (-11.69%)	64.22 (-2.33%)

(e) Openstack project

Technique Pair	Median AUC	Median Recall	Median FPR	Median G-Mean
RAD & RS	61.57 (-2.79%)	40.32 (-3.08%)	17.02 (14.12%)	57.26 (-2.93%)
RAD & k-DA	62.18 (-1.82%)	40.25 (-3.23%)	15.88 (6.51%)	57.62 (-2.31%)
RAD & Gen	62.74 (-0.95%)	41.33 (-0.65%)	15.93 (6.87%)	58.41 (-0.98%)
RS & k-DA	61.58 (-2.77%)	41.84 (0.58%)	18.67 (25.22%)	57.82 (-1.98%)
RS & Gen	62.1 (-1.97%)	42.89 (3.1%)	18.18 (21.9%)	58.48 (-0.86%)
k-DA & Gen	62.66 (-1.07%)	38.23 (-8.1%)	12.9 (-13.45%)	57.18 (-3.06%)

(f) Qt project

Table 5.9: Median performance and change in performance compared to the non-anonymized cross-project baseline for each technique pair (represented as Technique 1 & Technique 2) per project.

are anonymized using Generalization. *Column 2* of *row Gen* specifies that two projects have been anonymized using Random Add/Delete. The values found in subfigure 5.3a (RAD) for row RS can be found in subfigure 5.3c (RS) in the row for RAD. This is because when 1 dataset is anonymized using RAD and 4 using RS, this state is present in both

the RAD and RS heatmaps.

We find that the change in performance of the combined models is statistically insignificant to their individual technique graph anonymization counterparts. Using the Wilcoxon signed rank test ($\alpha = 0.05$ with Bonferroni Correction), we observe that there is no significant change in the AUC and G-Mean of the models. However, whenever Random Switch is used in one of the pairs, there is a slight decrease in performance where the pair performs worse in one or two cases with a small effect size. When these pairs are compared to non-graph anonymization techniques, we observe that for AUC and G-Mean they perform better than the NGA techniques in all cases (except one, where the G-Mean is indifferent in one case for k-DA and Gen).

When compared to each other, for AUC the RAD & Gen and k-DA & Gen pairs perform better in $4/5$ **cases, and for G-Mean the RAD & Gen and RAD & k-DA pair perform better in** $2/5$ **cases.** In Table 5.10, we see the results for all the technique pairs. Again, we see that whenever RS is used in a pair, the AUC performs similar or better. On the other hand, whenever Gen is added, that pair performs better.

Technique Pair	AUC	Recall	FPR	G-Mean
RAD & RS	0/1/4	0/1/4	1/2/2	0/0/5
RAD & k-DA	3/0/2	2/2/1	1/2/2	2/3/0
RAD & Gen	4/1/0	2/3/0	1/2/2	2/3/0
RS & k-DA	0/1/4	3/2/0	5/0/0	1/4/0
RS & Gen	2/0/3	2/3/0	4/0/1	1/2/2
k-DA & Gen	4/1/0	0/1/4	0/0/5	1/4/0

Table 5.10: Results of the statistical tests for each technique pair compared to another across all datasets. The results are reported as - win, where the row technique provides a significantly higher value, tie, where there is no statistical significance, and loss where the row technique value is significantly less.

For two technique pairs, we find that changing the mixture ratio (i.e., the number of projects anonymized by a technique) has a statistical difference in performance, where increasing the ratio of the more performant technique results in a better model. In Table 5.11, we see the results for AUC when comparing are made with different technique ratios of the same technique pair. Taking the example of the technique pair RAD and RS, when we compare all the different ratios (i.e., RAD 0 and RS 5 to RAD 5 and RS 0, a total of 6 ratios) for all the projects (i.e., 6 projects * 5 comparisons = 30 total comparisons), we find that increasing the count of RAD anonymized projects increases the performance of the model. However, for the RAD and k-DA pair we observe that decreasing k-DA reduces the performance. This effect is even more pronounced for the RS and k-DA pairs. The only pair that has a moderate change in performance with changing ratios is the k-DA and Gen pair as both techniques are performant. For all the comparisons, the effect size is negligible (30%), small (46%) or medium (13%) in 89% of the cases.

T1, T2	T1=0, T2=5	T1=1, T2=4	T1=2, T2=3	T1=3, T2=2	T1=4, T2=1	T1=5, T2=0
RAD & RS	2/9/19	5/13/12	10/12/8	8/14/8	14/14/2	15/10/5
RAD & k-DA	13/12/5	9/19/2	7/19/4	3/20/7	4/17/9	1/19/10
RAD & Gen	12/15/3	14/15/1	9/16/5	6/15/9	4/17/9	0/12/18
RS & k-DA	23/2/5	20/4/6	14/8/8	6/10/14	5/9/16	1/9/20
RS & Gen	21/4/5	20/7/3	14/7/9	6/9/15	7/6/17	1/9/20
k-DA & Gen	8/15/7	10/17/3	7/20/3	8/14/8	5/14/11	5/14/11

Table 5.11: Results of the statistical tests for the AUC metric each ratio of a technique pair compared to all the other ratios across all datasets (6 datasets 5 comparisons with other ratios). The results are reported as - win, where the row technique provides a significantly higher value, tie, where there is no statistical significance, and loss where the row technique value is significantly less. The column headers signify the number of projects anonymized with that particular technique.

As combining techniques is feasible, it makes it possible to build predictors with

k-DA and Gen techniques as base techniques supplemented with RAD and RS, where all projects are able to achieve privacy scores greater than 80%. The k-DA and Gen techniques individually were unable to provide IPR scores greater than 80% to all the datasets at the same configuration level. The projects that were unable to reach an IPR score greater than 80% can be simply substituted with RAD and RS anonymized datasets. The results can be seen in Figure 5.12, where we see that the AUC of the Generalization hybrid classifier for Cassandra is 68.6% as opposed to using just RAD, which was 61.41%. The k-DA and Generalization hybrid techniques, mixed with both RAD and RS, decrease the AUC by a median of 1% and 1.55% respectively, across all projects while all the datasets have an IPR score ≥ 80.

	k-DA (hybrid)	**Gen (hybrid)**
Cassandra	68.62 (1.3%)	66.52 (-1.8%)
Flink	69.52 (-2.4%)	66.66 (-6.4%)
Groovy	63.12 (-1.1%)	62.95 (-1.3%)
Ignite	74.51 (-0.9%)	75.02 (-0.2%)
OpenStack	64.31 (-2.7%)	64.66 (-2.2%)
Qt	63.45 (0.4%)	63.25 (0%)

Table 5.12: Relative change in the AUC metric relative to the non-anonymized cross-project baseline with different anonymization techniques. The last two columns are for the hybrid k-DA and Generalization techniques.

Summary of Research Question 5

Datasets anonymized using two different graph anonymization techniques can be combined together to build models that have similar performance as a classifier built using either one technique. This enables the k-DA and Generalization techniques to be supplemented with projects anonymized with RAD/RS to allow all training datasets to have privacy scores of greater than 80% with similar performance scores. Furthermore, which datasets were anonymized did not play a significant role in the gain in performance. The best-performing technique pair for AUC is k-DA and Gen, and for Recall is RAD and Gen.

5.4.4 RQ6: How do predictors built using a combination of graph anonymization and non-graph anonymization techniques work together?

The model trained using data where projects are anonymized using GA techniques or NGA techniques performs better or is similar in terms of AUC than only NGA techniques models in all cases. For G-Mean, they perform similar in most cases (better in a few cases). However, when compared to only graph anonymized models, these pair's performance is similar or worse for AUC and G-Mean. All the GA and NGA pairs decrease AUC, Recall, FPR and G-Mean by 6.05%, 16.6 %, 3.7% and 8.8%. In Table 5.13, we see the change in performance from the non-anonymized baseline. The table of remaining projects can be found in our online replication package [2].

In Figure 5.4, we see the change in performance that comes with using a predictor built using a combination of graph anonymized projects with non-graph anonymized datasets for cross-project defect prediction. The format of these graphs is similar to

	Median AUC	Median Recall	Median FPR	Median G-Mean
RAD & MORPH	61.9 (-6.49%)	53.23 (-13.4%)	29.04 (-0.15%)	60.44 (-8.08%)
RAD & LACE	60.88 (-8.01%)	47.39 (-22.91%)	26.2 (-9.92%)	58.78 (-10.61%)
RAD & LACE2	63.16 (-4.58%)	58.58 (-4.7%)	32.42 (11.5%)	61.78 (-6.04%)
RS & MORPH	58.2 (-12.07%)	62.3 (1.34%)	45.9 (57.84%)	56.94 (-13.42%)
RS & LACE	57.43 (-13.23%)	57.94 (-5.73%)	43.09 (48.18%)	56.37 (-14.28%)
RS & LACE2	59.2 (-10.56%)	70.8 (15.18%)	52.7 (81.22%)	56.7 (-13.78%)
k-DA & MORPH	62.38 (-5.76%)	45.71 (-25.64%)	21.93 (-24.59%)	59.52 (-9.48%)
k-DA & LACE	60.76 (-8.2%)	41.86 (-31.9%)	20.54 (-29.38%)	57.3 (-12.87%)
k-DA & LACE2	64.04 (-3.26%)	49.47 (-19.52%)	23.42 (-19.46%)	61.94 (-5.8%)
Gen & MORPH	62.25 (-5.95%)	49.39 (-19.65%)	23.96 (-17.61%)	60.24 (-8.39%)
Gen & LACE	60.32 (-8.87%)	44.88 (-26.99%)	22.3 (-23.31%)	57.38 (-12.74%)
Gen & LACE2	63.85 (-3.54%)	50.76 (-17.41%)	24.2 (-16.76%)	62.16 (-5.47%)

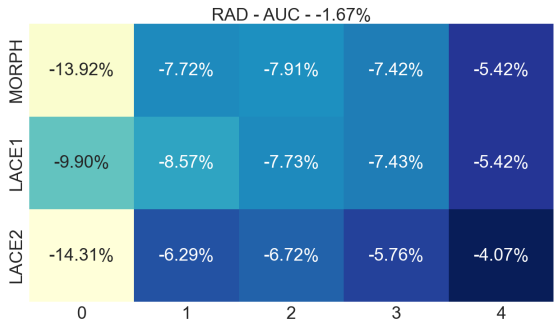
(a) OpenStack project

	Median AUC	Median Recall	Median FPR	Median G-Mean
RAD & MORPH	65.81 (-7.28%)	57.84 (-18.93%)	26.42 (-10.12%)	64.72 (-8.62%)
RAD & LACE	63.48 (-10.57%)	51.28 (-28.12%)	24.32 (-17.25%)	61.64 (-12.97%)
RAD & LACE2	67.01 (-5.59%)	58.72 (-17.7%)	22.1 (-24.8%)	65.58 (-7.39%)
RS & MORPH	61.92 (-12.76%)	52.96 (-25.76%)	30.49 (3.76%)	59.94 (-15.36%)
RS & LACE	60.76 (-14.4%)	49.7 (-30.33%)	27.14 (-7.64%)	58.72 (-17.08%)
RS & LACE2	62.31 (-12.21%)	51.86 (-27.31%)	24.95 (-15.11%)	59.96 (-15.34%)
k-DA & MORPH	66.44 (-6.4%)	58.94 (-17.39%)	25.51 (-13.2%)	65.65 (-7.3%)
k-DA & LACE	62.95 (-11.31%)	48.75 (-31.67%)	23.16 (-21.18%)	61.04 (-13.81%)
k-DA & LACE2	68.0 (-4.19%)	56.92 (-20.22%)	22.14 (-24.67%)	66.59 (-5.98%)
Gen & MORPH	65.1 (-8.29%)	56.9 (-20.25%)	27.91 (-5.04%)	64.38 (-9.1%)
Gen & LACE	61.99 (-12.67%)	49.08 (-31.2%)	25.32 (-13.83%)	60.3 (-14.85%)
Gen & LACE2	65.85 (-7.23%)	54.86 (-23.1%)	23.54 (-19.9%)	64.35 (-9.13%)

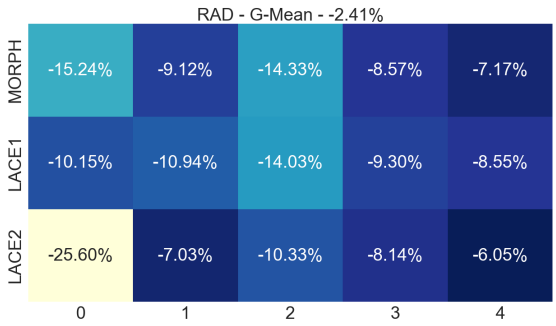
(b) Flink project

Table 5.13: Median Performance with the change in performance from the non-anonymized cross-project baseline for each technique pair (represented as Technique 1 & Technique 2) across all datasets.

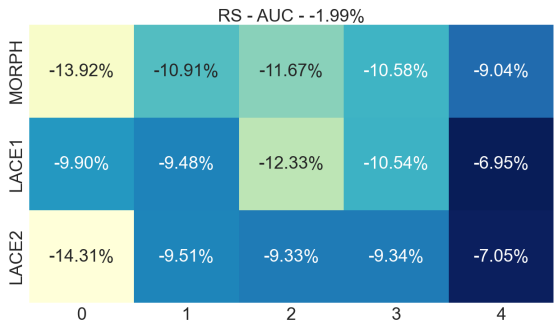
the graphs used in the previous RQ, where, for example, in subfigure 5.4a, the title of the heatmap of the graph tells the main/title anonymization technique (graph techniques, for example, Random Add/Delete) that will be used along with the technique defined as the row header technique (tabular techniques, for example, the row 3 technique is LACE2). Each column value represents the number of datasets that have been anonymized using the title anonymization technique, starting with 0 i.e., the value in *row LACE2* and *column 0* means no data has been anonymized using Random Add/Delete. *Column 1* of *row LACE2* specifies that one data set has been anonymized using Random Add/Delete.



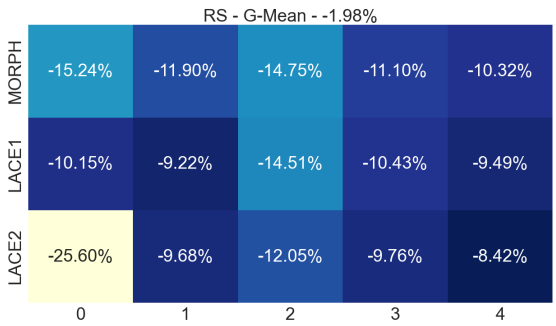
(a) Median AUC change with Random Add/Delete



(b) G-Mean score for Cassandra with Random Add/Delete

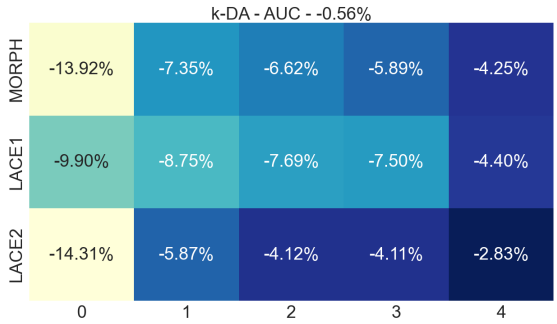


(c) AUC score for Cassandra with Random Switch

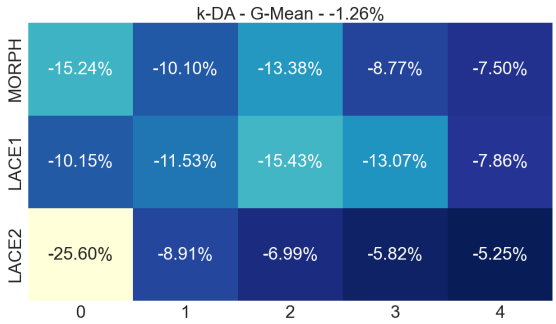


(d) G-Mean score for Cassandra with Random Switch

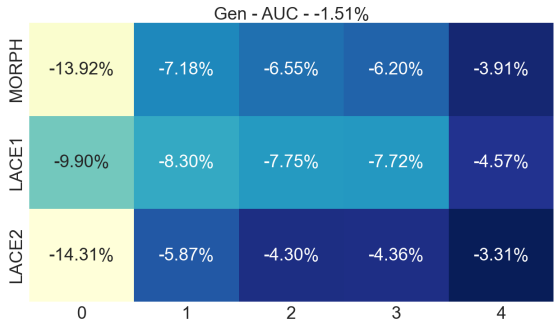
Even adding one graph-anonymized project to the other non-graph anonymized



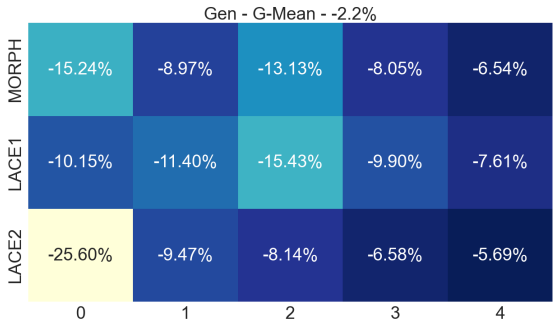
(e) AUC score for Cassandra with k-DA



(f) G-Mean score for Cassandra with k-DA



(g) AUC score for Cassandra with Gen



(h) G-Mean score for Cassandra with Gen

Figure 5.4: The heatmaps show the median change in performance that comes with using data anonymized using two different anonymization techniques across all projects. The value of each cell is the median change in performance metric achieved by anonymizing the number of projects specified by the column value using the technique specified in the title, with the remaining anonymized using the row technique. The change percentage is calculated with respect to each project's cross-project performance baselines as seen in Table 5.3.

projects it increases the performance and reduces the performance decrease compared to using only tabular techniques by half its initial value. For example, in subfigure 5.4e, we see that when all data is anonymized using MORPH (row MORPH and column 0), the AUC is decreased by a median of 13.92%. In the next column, column 1 of the MORPH row, the AUC median decrease changes to 7.35%. The performance impact was reduced by 6.5%. We see similar behaviour across different datasets

and different technique combinations as well, where with an increasing number of graph anonymized projects, the performance metrics improve from their value when all projects are anonymized using tabular techniques.

In Table 5.14 we see this in further detail with statistical evidence, showing how changing the mixture ratio for a technique pair correlates with a change in the performance of the model. We see that adding one graph anonymized project results in the pair performing better in 6 more cases. With increasing amounts of GA technique, the model performance keeps improving, with all GA anonymized performing best in the majority (76.6% and more) of the cases.

T1, T2	T1=0, T2=5	T1=1, T2=4	T1=2, T2=3	T1=3, T2=2	T1=4, T2=1	T1=5, T2=0
RAD & MORPH	1/7/22	8/9/13	10/8/12	10/4/16	18/6/6	25/2/3
RAD & LACE	4/6/20	10/4/16	6/8/16	9/8/13	21/4/5	25/0/5
RAD & LACE2	1/2/27	10/10/10	12/6/12	10/7/13	18/5/7	23/2/5
RS & MORPH	2/13/15	10/10/10	5/8/17	5/11/14	17/7/6	25/3/2
RS & LACE	8/11/11	6/13/11	4/6/20	5/9/16	20/6/4	23/3/4
RS & LACE2	5/5/20	9/10/11	5/10/15	9/7/14	18/7/5	23/3/4
k-DA & MORPH	1/4/25	8/5/17	10/6/14	12/4/14	20/4/6	27/1/2
k-DA & LACE	2/5/23	8/8/14	9/7/14	10/6/14	20/4/6	25/2/3
k-DA & LACE2	0/1/29	10/6/14	16/5/9	9/8/13	16/6/8	25/2/3
Gen & MORPH	2/3/25	9/7/14	8/5/17	12/3/15	21/4/5	27/0/3
Gen & LACE	3/3/24	11/5/14	8/4/18	9/7/14	24/1/5	25/0/5
Gen & LACE2	0/1/29	9/6/15	11/9/10	10/8/12	20/5/5	25/1/4

Table 5.14: Results of the statistical tests for each technique pair compared to another across all datasets. The results are reported as - win, where the row technique provides a significantly higher value, tie, where there is no statistical significance, and loss where the row technique value is significantly lower.

We hypothesize that the combined models have better performance compared to only tabular techniques anonymized models because when they are used with the more performant graph techniques, the resulting model performs better than only using the tabular techniques. Even though the tabular techniques anonymized projects reduce the overall performance, the performant graph anonymized projects allow the

model to retain its predictive capability.

	AUC	Recall	FPR	G-Mean
RAD & MORPH	3/3/5	1/5/5	3/4/4	1/6/4
RAD & LACE	3/1/7	1/4/6	3/5/3	1/6/4
RAD & LACE2	5/4/2	5/5/1	3/4/4	7/3/1
RS & MORPH	0/1/10	7/3/1	1/0/10	0/3/8
RS & LACE	0/2/9	8/2/1	2/0/9	0/6/5
RS & LACE2	1/1/9	11/0/0	0/0/11	1/8/2
k-DA & MORPH	5/4/2	0/4/7	9/1/1	1/6/4
k-DA & LACE	4/5/2	0/1/10	9/2/0	0/6/5
k-DA & LACE2	10/1/0	2/3/6	10/1/0	7/4/0
Gen & MORPH	5/4/2	4/4/3	3/4/4	6/4/1
Gen & LACE	4/5/2	1/7/3	3/5/3	2/7/2
Gen & LACE2	10/1/0	5/4/2	6/2/3	10/1/0

Table 5.15: Results of the statistical tests for each technique pair compared to another across all datasets. The results are reported as - win, where the row technique provides a significantly higher value, tie, where there is no statistical significance, and loss where the row technique value is significantly lower.

When comparing the GA and NGA pairs to one another, we find that for AUC, LACE2 paired with Gen and k-DA performs better than others in 10/11 **cases.** For G-Mean, the Gen and LACE2 pair perform the best in 10/11 cases, with the LACE2 pairs with RAD and k-DA performing better in 7/11 cases. LACE2 performs better than other NGA techniques when used with GA anonymized projects.

Implications: The results of our study help show two things:

1. **Graph and tabular anonymization can work together:** Our case study shows that two different classes of anonymization techniques can be combined together to make a predictor that can perform better than using only tabular techniques which have a higher performance impact. In certain cases, the hybrid models perform better than either of the two techniques in terms of performance.

2. **Highly private projects (90% or higher score), with less performance, can be made performant by adding graph anonymized projects:** The MORPH, LACE and LACE2 techniques provide high privacy scores to datasets, with scores ranging from 89% to 99%. However, as seen in RQ4.2, their performance impact is severe. Although there might be cases where certain projects require very high privacy levels. Our study shows that they can be made performant in a cross-project setting by using graph anonymized metrics which provide the same Privacy Level II and have better performance. This would allow more private projects to be included in a model, without affecting the models' cross-project performance.

Summary of Research Question 6

Combining graph anonymized and tabular technique anonymized datasets produces predictors that are significantly more performant than only using tabular anonymization techniques, however, they do not perform better than only graph anonymized models. In certain cases, these models perform better than using either of the two techniques.

5.5 Conclusion

Our empirical study shows that graph anonymized metrics perform well in a cross-project prediction setting only, decreasing the AUC, Recall, FPR and G-Mean by a median of 1.3%, 4.5%, 3.4% and 1.9% across all datasets and configuration values They perform significantly better than NGA techniques in terms of AUC, G-Mean (80% or more) and FPR (50% or more) in the majority of the cases. This shows that they are an effective choice for anonymizing data, and can be used successfully both in a within

and cross-project setting. When compared to MD, it performs better for FPR and G-Mean in the majority (66% or more) of the cases. However, for AUC the Gen technique performs better in the majority of the cases (66% or more) while the other techniques alternatively perform better or similarly.

GA techniques can also be used in practical scenarios where projects might be anonymized using different anonymization techniques. When GA metrics are combined with projects anonymized using another GA technique, the performance of all the models is similar to the model performance when data is anonymized using a single GA technique. The RAD & Gen pair is performing best in terms of AUC and G-Mean. Furthermore, when these metrics are used along with NGA techniques, the combined models perform better than models trained using data anonymized only using NGA techniques. The Gen & LACE2 pair performs best for both AUC and G-Mean.

Our work shows that GA anonymization techniques are performant and versatile, being compatible with other anonymization techniques. This should encourage academia and industry to make use of and experiment with GA techniques.

CHAPTER 6

Conclusions and Future Work

THIS thesis addresses the challenge of preserving privacy in JIT defect prediction metrics, contrasting the use of traditional tabular or non-graph anonymization methods with graph anonymization techniques. While former methods offer a high degree of privacy, they often compromise model performance. In comparison, graph anonymization techniques maintain similar levels of privacy but with significantly better performance across various settings, including within-project, cross-project, and cross-technique analyses.

6.1 Thesis Statement

> Graph anonymization techniques are able to provide sufficient privacy to JIT defect prediction metrics both in a within-project and cross-project setting, without compromising their performance.

This thesis empirically shows that graph anonymization not only effectively provides privacy to all datasets, achieving privacy scores above 65% without notable performance loss and scores above 80% with minimal performance impact, but also outperforms the state-of-the-art techniques in terms of AUC, Recall, FPR and G-Mean metrics, especially with techniques like Generalization and Random Add/Delete.

When compared to anonymization techniques like MORPH, LACE, and LACE2, graph anonymization shows a better ability to balance high privacy scores with less performance degradation. Specifically, compared to MD50, graph anonymization excels in the majority of scenarios for G-Mean and FPR metrics, and shows comparable or better AUC performance.

In cross-project settings, graph anonymization reduces AUC, Recall, FPR, and G-Mean by median values of 1.3%, 4.5%, 3.4%, and 1.9%, respectively, showcasing better performance over non-graph anonymization (NGA) techniques, particularly for AUC and G-Mean in 80% or more cases, and FPR by 50% or more cases.

Graph anonymization techniques also demonstrate versatility in practical applications, showing consistent model performance when combining metrics from projects anonymized with different techniques. The Random Add/Delete and Generalization

combination is notably effective for AUC and G-Mean. Additionally, models that integrate graph anonymization metrics with those from NGA techniques surpass the performance of models using only NGA techniques, with the Generalization and LACE2 pair excelling in AUC and G-Mean.

Our findings advocate for the performance and adaptability of graph anonymization techniques, suggesting their potential for broader adoption in both academic and industrial contexts.

6.2 Future Work

Potential directions for future research include:

- Differential privacy techniques have great applicability to provide privacy to tabular data. We propose future research to use them on the graph generated for JIT defect prediction metrics as they can offer mathematical guarantees for privacy.
- The knowledge graph suggested in the thesis can be extended by adding additional software analytics data sources, for example, review/pull request information, build CI data results and more.
- We would urge future research to use graph anonymization techniques to provide privacy to other software analytics data and models, for example, build status prediction and task effort estimation.

Bibliography

[1] Akshat. Replication package for chapter 4: Towards graph-anonymization of software analytics data. https://github.com/SAILResearch/replication-24-akshat-towards-graph-anonymization.

[2] Ahmed E Hassan Akshat Malik, Bram Adams. Replication package for chapter 5: Pragmatic evaluation of graph anonymized metrics. https://github.com/SAILResearch/replication-24-akshat-pragmatic-graph-anonymization.

[3] Mohammad Al-Rubaie and J. Morris Chang. Privacy-preserving machine learning: Threats and solutions. *IEEE Security and Privacy*, 17(2):49–58, 2019.

[4] Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 70–78, 2008.

[5] Saikat Chakraborty, Rahul Krishna, Yangruibo Ding, and Baishakhi Ray. Deep learning based vulnerability detection: Are we there yet? *IEEE Transactions on Software Engineering*, 48(9):3280–3296, 2022.

[6] Jiale Chen, Jiale Zhang, Yanchao Zhao, Hao Han, Kun Zhu, and Bing Chen. Beyond model-level membership privacy leakage: an adversarial approach in federated learning. In *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, pages 1–9, 2020.

[7] James Cheng, Ada Wai-chee Fu, and Jia Liu. K-isomorphism: privacy preserving network publication against structural attacks. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 459–470, 2010.

[8] S.R. Chidamber and C.F. Kemerer. A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, 20(6):476–493, 1994.

[9] Aritra Dasgupta, Min Chen, and Robert Kosara. Measuring privacy and utility in privacy-preserving visualization. In *Computer Graphics Forum*, volume 32, pages 35–47. Wiley Online Library, 2013.

[10] Khaled El Emam and Fida Kamal Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637, 2008.

[11] Davide Falessi, Aalok Ahluwalia, and Massimiliano DI Penta. The impact of dormant defects on defect prediction: A study of 19 apache projects. *ACM Trans. Softw. Eng. Methodol.*, 31(1), sep 2021.

[12] Yi Fan, Chenxi Lv, Xu Zhang, Guoqiang Zhou, and Yuming Zhou. The utility challenge of privacy-preserving data-sharing in cross-company defect prediction: An empirical study of the cliff and morph algorithm. In *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 80–90, 2017.

[13] Tomas Feder, Shubha U. Nabar, and Evimaria Terzi. *Anonymizing Graphs*. Oct 2008. ADS Bibcode: 2008arXiv0810.5578F type: article.

[14] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. page 1322–1333, 2015.

[15] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, page 17–32, USA, 2014. USENIX Association.

[16] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Advances in neural information processing systems*, 27, 2014.

[17] Thong Hoang, Hong Jin Kang, David Lo, and Julia Lawall. Cc2vec: Distributed representations of code changes. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pages 518–529, 2020.

[18] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid,

Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4), jul 2021.

[19] Qiao Huang, Xin Xia, and David Lo. Revisiting supervised and unsupervised models for effort-aware just-in-time defect prediction. *Empirical Software Engineering*, 24:2823–2862, 2019.

[20] Qiao Huang, Xin Xia, and David Lo. Revisiting supervised and unsupervised models for effort-aware just-in-time defect prediction. *Empirical Software Engineering*, 24:2823–2862, 2019.

[21] Imbalanced-Learn. Smoteenn. https://imbalanced-learn.org/stable/references/generated/imblearn.combine.SMOTEENN.html.

[22] Shouling Ji, Prateek Mittal, and Raheem Beyah. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys and Tutorials*, 19(2):1305–1326, 2017.

[23] Yasutaka Kamei, Takafumi Fukushima, Shane Mcintosh, Kazuhiro Yamashita, Naoyasu Ubayashi, and Ahmed E. Hassan. Studying just-in-time defect prediction using cross-project models. 21(5):2072–2106, oct 2016.

[24] Yasutaka Kamei, Takafumi Fukushima, Shane Mcintosh, Kazuhiro Yamashita, Naoyasu Ubayashi, and Ahmed E. Hassan. Studying just-in-time defect prediction using cross-project models. *Empirical Softw. Engg.*, 21(5):2072–2106, oct 2016.

[25] Yasutaka Kamei, Akito Monden, Shinsuke Matsumoto, Takeshi Kakimoto, and

Ken-ichi Matsumoto. The effects of over and under sampling on fault-prone module detection. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, pages 196–204, 2007.

[26] Yasutaka Kamei, Emad Shihab, Bram Adams, Ahmed E Hassan, Audris Mockus, Anand Sinha, and Naoyasu Ubayashi. A large-scale empirical study of just-in-time quality assurance. *IEEE Transactions on Software Engineering*, 39(6):757–773, 2012.

[27] Hossein Keshavarz and Meiyappan Nagappan. Apachejit: A large dataset for just-in-time defect prediction. In *Proceedings of the 19th International Conference on Mining Software Repositories*, MSR '22, page 191–195, New York, NY, USA, 2022. Association for Computing Machinery.

[28] Zhiqiang Li, Xiao-Yuan Jing, Xiaoke Zhu, Hongyu Zhang, Baowen Xu, and Shi Ying. On the multiple sources and privacy preservation issues for heterogeneous defect prediction. *IEEE Transactions on Software Engineering*, 45(4):391–411, 2019.

[29] Kun Liu and Evimaria Terzi. Towards identity anonymization on graphs. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 93–106, 2008.

[30] Akshat Malik, Bram Adams, and Ahmed E. Hassan. Towards graph-anonymization of software analytics data: Empirical study on jit defect prediction. *Empirical Software Engineering*, 2024.

[31] Shane McIntosh and Yasutaka Kamei. Are fix-inducing changes a moving target?

a longitudinal case study of just-in-time defect prediction. *IEEE Transactions on Software Engineering*, 44(5):412–428, 2018.

[32] Ricardo Mendes and João P. Vilela. Privacy-preserving data mining: Methods, metrics, and applications. *IEEE Access*, 5:10562–10582, 2017.

[33] Tim Menzies, Jeremy Greenwald, and Art Frank. Data mining static code attributes to learn defect predictors. *IEEE Transactions on Software Engineering*, 33(1):2–13, 2007.

[34] Prateek Mittal, Charalampos Papamanthou, and Dawn Song. Preserving link privacy in social network based systems. *arXiv preprint arXiv:1208.6189*, 2012.

[35] R. Mortazavi and S. H. Erfani. Gram: An efficient (k, l) graph anonymization method. *Expert Systems with Applications*, 153:113454, Sep 2020.

[36] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy*, pages 173–187. IEEE, 2009.

[37] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019.

[38] OpenStack. Openstack gerrit. https://review.opendev.org/q/status:open+-is:wip.

[39] Fayola Peters and Tim Menzies. Privacy and utility for defect prediction: Experiments with morph. In *2012 34th International Conference on Software Engineering (ICSE)*, pages 189–199, 2012.

[40] Fayola Peters, Tim Menzies, Liang Gong, and Hongyu Zhang. Balancing privacy and utility in cross-company defect prediction. *IEEE Transactions on Software Engineering*, 39(8):1054–1068, 2013.

[41] Fayola Peters, Tim Menzies, and Lucas Layman. Lace2: Better privacy-preserving data sharing for cross project defect prediction. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 1, pages 801–811, 2015.

[42] Chanathip Pornprasit and Chakkrit Kla Tantithamthavorn. Jitline: A simpler, better, faster, finer-grained just-in-time defect prediction. page 369–379. IEEE Computer Society, May 2021.

[43] Anastasia Pustozerova and Rudolf Mayer. Information leaks in federated learning. In *Proceedings of the Network and Distributed System Security Symposium*, volume 10, 2020.

[44] PyDriller. Pydriller. https://pydriller.readthedocs.io/en/latest/.

[45] Qt. Qt gerrit. https://codereview.qt-project.org/.

[46] Gema Rodríguez-Pérez, Meiyappan Nagappan, and Gregorio Robles. Watch out for extrinsic bugs! a case study of their impact in just-in-time bug prediction models on the openstack project. *IEEE Transactions on Software Engineering*, 48(4):1400–1416, 2022.

[47] Alessandra Sala, Xiaohan Zhao, Christo Wilson, Haitao Zheng, and Ben Y Zhao. Sharing graphs using differentially private graph models. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 81–98, 2011.

[48] scikit learn. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[49] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[50] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[51] Mudhakar Srivatsa and Mike Hicks. Deanonymizing mobility traces: Using social network as a side-channel. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 628–637, 2012.

[52] Sadia Tabassum, Leandro L. Minku, and Danyi Feng. Cross-project online just-in-time software defect prediction. *IEEE Transactions on Software Engineering*, 49(1):268–287, 2023.

[53] Ming Tan, Lin Tan, Sashank Dara, and Caleb Mayeux. Online defect prediction for imbalanced data. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 2, pages 99–108, 2015.

[54] Chakkrit Tantithamthavorn, Shane McIntosh, Ahmed E Hassan, and Kenichi Matsumoto. An empirical comparison of model validation techniques for defect prediction models. *IEEE Transactions on Software Engineering*, 43(1):1–18, 2016.

[55] Brian Thompson and Danfeng Yao. The union-split algorithm and cluster-based

anonymization of social networks. In *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security*, pages 218–227, 2009.

[56] Jacek undefinedliwerski, Thomas Zimmermann, and Andreas Zeller. When do changes induce fixes? In *Proceedings of the 2005 International Workshop on Mining Software Repositories*, MSR '05, page 1–5, New York, NY, USA, 2005. Association for Computing Machinery.

[57] Isabel Wagner and David Eckhoff. Technical privacy metrics: A systematic survey. *ACM Comput. Surv.*, 51(3), jun 2018.

[58] Peter A Whigham, Caitlin A Owen, and Stephen G Macdonell. A baseline model for software effort estimation. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 24(3):1–11, 2015.

[59] Wikipidea. Graph definition. https://en.wikipedia.org/w/index.php?title=Graph.

[60] Wikipidea. Lace. http://lace.readthedocs.io/en/latest/readme.html.

[61] Wikipidea. Mann whitney u. https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test.

[62] Timo Wolf, Adrian Schroter, Daniela Damian, and Thanh Nguyen. Predicting build failures using social network analysis on developer communication. In *2009 IEEE 31st International Conference on Software Engineering*, pages 1–11, 2009.

[63] Hiroki Yamamoto, Dong Wang, Gopi Krishnan Rajbahadur, Masanari Kondo, Yasutaka Kamei, and Naoyasu Ubayashi. Towards privacy preserving cross project

defect prediction with federated learning. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 485–496, 2023.

[64] Yibiao Yang, Yuming Zhou, Jinping Liu, Yangyang Zhao, Hongmin Lu, Lei Xu, Baowen Xu, and Hareton Leung. Effort-aware just-in-time defect prediction: Simple unsupervised models could be better than supervised models. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE 2016, page 157–168, New York, NY, USA, 2016. Association for Computing Machinery.

[65] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018.

[66] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018.

[67] Xiaowei Ying and Xintao Wu. Randomizing social networks: a spectrum preserving approach. In *proceedings of the 2008 SIAM International Conference on Data Mining*, pages 739–750. SIAM, 2008.

[68] Zhengran Zeng, Yuqun Zhang, Haotian Zhang, and Lingming Zhang. Deep just-in-time defect prediction: How far are we? In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2021, page 427–438, New York, NY, USA, 2021. Association for Computing Machinery.

[69] Yunhua Zhao, Kostadin Damevski, and Hui Chen. A systematic survey of just-in-time software defect prediction. *ACM Comput. Surv.*, 55(10), feb 2023.

[70] Bin Zhou and Jian Pei. Preserving privacy in social networks against neighborhood attacks. In *2008 IEEE 24th International Conference on Data Engineering*, pages 506–515. IEEE, 2008.

[71] Lei Zou, Lei Chen, and M Tamer Özsu. K-automorphism: A general framework for privacy preserving network publication. *Proceedings of the VLDB Endowment*, 2(1):946–957, 2009.