2020 IEEE/ACM 17th International Conference on Mining Software Repositories (MSR 2020)

Seoul, South Korea 25 – 26 May 2020



IEEE Catalog Number: CFP2078C-POD **ISBN:**

978-1-7281-9840-8

Copyright © 2020, Association for Computing Machinery All Rights Reserved

*** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.

IEEE Catalog Number: ISBN (Print-On-Demand): ISBN (Online): ISSN: CFP2078C-POD 978-1-7281-9840-8 978-1-4503-7517-7 2574-3848

Additional Copies of This Publication Are Available From:

Curran Associates, Inc 57 Morehouse Lane Red Hook, NY 12571 USA Phone: (845) 758-0400 Fax: (845) 758-2633 E-mail: curran@proceedings.com Web: www.proceedings.com



2020 IEEE/ACM 17th International Conference on Mining Software Repositories (MSR) MSR 2020

Table of Contents

Message from the Chairs	xiv
Message from the Track Chairs	xviii
Organizing Committee	xxiii
Program Committee - Data Showcase	xxv
Program Committee - Education Track	xxvii
Program Committee - FOSS Awards	xxviii
Program Committee - Mining Challenge	xxx
Program Committee - Most Influential Paper (MIP) Award	xxxiv
Program Committee - MSR Awards	xxxv
Program Committee - Registered Reports	xxxvii
Program Committee - Technical Papers	xI
Additional Reviewers	xlix

Mining Challenge

The Software Heritage Graph Dataset: Large-Scale Analysis of Public Software Development History Antoine Pietri (Inria), Diomidis Spinellis (Athens University of Economics and Business), and Stefano Zacchiroli (Université de Paris and Inria)		
Cheating Death: A Statistical Survival Analysis of Publicly Available Python Projects Rao Hamza Ali (Chapman University, USA), Chelsea Parlett-Pelleriti (Chapman University, USA), and Erik Linstead (Chapman University, USA)		
An Exploratory Study to Find Motives behind Cross-Platform Forks from Software Heritage Dataset		
Avijit Bhattacharjee (University of Saskatchewan, Canada), Sristy Sumana Nath (University of Saskatchewan, Canada), Shurui Zhou (Carnegie Mellon University), Debasish Chakroborti (University of Saskatchewan, Canada), Banani Roy (University of Saskatchewan, Canada), Chanchal Roy (University of Saskatchewan, Canada), and Kevin Schneider (University of Saskatchewan, Canada)		
Exploring the Security Awareness of the Python and JavaScript Open Source Communities 10 Gábor Antal (University of Szeged), Márton Keleti (University of		

Szeged), and Péter Hegedűs (University of Szeged)

Technical Papers

A Large-Scale Comparative Evaluation of IR-Based Tools for Bug Localization Shayan Ali Akbar (Purdue University) and Avinash Kak (Purdue University)	21
A Machine Learning Approach for Vulnerability Curation Yang Cheng (Veracode), Andrew Edward Santosa (Veracode), Ming Yi Ang (Veracode), Abhishek Sharma (Veracode), Asankhaya Sharma (Veracode), and David Lo (Singapore Management University, Singapore)	32
A Soft Alignment Model for Bug Deduplication Irving Muller Rodrigues (Polytechnique Montreal, Canada), Daniel Aloise (Polytechnique Montreal, Canada), Eraldo Rezende Fernandes (UFMS, Brazil), and Michel Dagenais (Polytechnique Montreal, Canada)	43
A Study of Potential Code Borrowing and License Violations in Java Projects on GitHub Yaroslav Golubev (JetBrains Research, ITMO University, Russia), Maria Eliseeva (Higher School of Economics, Russia), Nikita Povarov (JetBrains, Russia), and Timofey Bryksin (JetBrains Research, Saint Petersburg State University, Russia)	54
A Study on the Accuracy of OCR Engines for Source Code Transcription from Programming Screencasts	65
An Empirical Study of Build Failures in the Docker Context Yiwen Wu (National University of Defense Technology, China), Yang Zhang (National University of Defense Technology, China), Tao Wang (National University of Defense Technology, China), and Huaimin Wang (National University of Defense Technology, China)	76
AIMMX: Artificial Intelligence Model Metadata Extractor Jason Tsay (IBM Research, USA), Alan Braz (IBM Research, Brazil), Martin Hirzel (IBM Research, USA), Avraham Shinnar (IBM Research, USA), and Todd Mummert (IBM Research, USA)	81
An Empirical Study of Method Chaining in Java Tomoki Nakamaru (The University of Tokyo), Tomomasa Matsunaga (The University of Tokyo), Tetsuro Yamazaki (The University of Tokyo), Soramichi Akiyama (The University of Tokyo), and Shigeru Chiba (The University of Tokyo)	93
An Empirical Study on Regular Expression Bugs Peipei Wang (North Carolina State University, USA), Chris Brown (North Carolina State University, USA), Jamie A. Jennings (North Carolina State University, USA), and Kathryn T. Stolee (North Carolina State University, USA)	103

Automatically Granted Permissions in Android Apps: An Empirical Study on their Prevalence and on the Potential Threats for Privacy
Behind the Intents: An In-Depth Empirical Study on Software Refactoring in Modern Code 125 Review 125 Matheus Paixao (University of Fortaleza), Anderson Uchôa (Pontifical 125 Catholic University of Rio de Janeiro), Ana Carla Bibiano (Pontifical 126 Catholic University of Rio de Janeiro), Daniel Oliveira (Pontifical 126 Catholic University of Rio de Janeiro), Jens Krinke (Ontifical 126 Catholic University of Rio de Janeiro), Jens Krinke (University 126 College London), and Emilio Arvonio (University of Salerno) 125
Beyond the Code: Mining Self-Admitted Technical Debt in Issue Tracker Systems
Boa Views: Easy Modularization and Sharing of MSR Analyses
Can We Use SE-Specific Sentiment Analysis Tools in a Cross-Platform Setting?
Capture the Feature Flag: Detecting Feature Flags in Open-Source
Challenges in Chatbot Development: A Study of Stack Overflow Posts
Characterizing and Identifying Composite Refactorings: Concepts, Heuristics and Patterns 186 Leonardo Sousa (Carnegie Mellon University), Diego Cedrim (Amazon), Alessandro Garcia (Pontifical Catholic University of Rio de Janeiro), Willian Oizumi (Pontifical Catholic University of Rio de Janeiro), Ana Bibiano (Pontifical Catholic University of Rio de Janeiro), Daniel Oliveira (Pontifical Catholic University of Rio de Janeiro), Miryung Kim (University of California, Los Angeles), and Anderson Oliveira (Pontifical Catholic University of Rio de Janeiro)
Detecting Video Game-Specific Bad Smells in Unity Projects

Detecting and Characterizing Bots that Commit Code Tapajit Dey (The University of Tennessee, USA), Sara Mousavi (The University of Tennessee, USA), Eduardo Ponce (The University of Tennessee, USA), Tanner Fry (The University of Tennessee, USA), Bogdan Vasilescu (Carnegie Mellon University, USA), Anna Filippova (GitHub, USA), and Audris Mockus (The University of Tennessee, USA)	209
Developer-Driven Code Smell Prioritization Fabiano Pecorelli (SeSa Lab - University of Salerno, Italy), Fabio Palomba (SeSa Lab - University of Salerno, Italy), Foutse Khomh (École Polytechnique de Montréal, Canada), and Andrea De Lucia (SeSa Lab - University of Salerno, Italy)	220
Did You Remember to Test Your Tokens? Danielle Gonzalez (Rochester Institute of Technology), Michael Rath (DLR Institute of Data Science, Technical University Ilmenau), and Mehdi Mirakhorli (Rochester Institute of Technology)	232
Embedding Java Classes with Code2vec: Improvements from Variable Obfuscation Rhys Compton (University of Waikato, New Zealand), Eibe Frank (University of Waikato, New Zealand), Panos Patros (University of Waikato, New Zealand), and Abigail Koay (University of Waikato, New Zealand)	243
Empirical Study of Restarted and Flaky Builds on Travis CI Thomas Durieux (INESC-ID and IST, University of Lisbon, Portugal), Claire Le Goues (Carnegie Mellon University), Michael Hilton (Carnegie Mellon University), and Rui Abreu (INESC-ID and IST, University of Lisbon, Portugal)	254
Ethical Mining - A Case Study on MSR Mining Challenges Nicolas E. Gold (University College London, UK) and Jens Krinke (University College London, UK)	265
Forking without Clicking: on How to Identify Software Repository Forks Antoine Pietri (Inria), Guillaume Rousseau (Université de Paris and Inria), and Stefano Zacchiroli (Université de Paris and Inria)	277
 From Innovations to Prospects: What is Hidden behind Cryptocurrencies? Ang Jia (Xi'an Jiaotong University, China), Ming Fan (Xi'an Jiaotong University, China), Xi Xu (Xi'an Jiaotong University, China), Di Cui (Xi'an Jiaotong University, China), Wenying Wei (Xi'an Jiaotong University, China), Zijiang Yang (Western Michigan University, America), Kai Ye (Xi'an Jiaotong University, China), and Ting Liu (Xi'an Jiaotong University, China) 	288
Improved Automatic Summarization of Subroutines via Attention to File Context Sakib Haque (University of Notre Dame, USA), Alexander LeClair (University of Notre Dame, USA), Lingfei Wu (IBM Research, USA), and Collin McMillan (University of Notre Dame)	300
Investigating Severity Thresholds for Test Smells Davide Spadini (Software Improvement Group & Delft University of Technology), Martin Schvarcbacher (Software Improvement Group), Ana-Maria Oprescu (University of Amsterdam), Magiel Bruntink (Software Improvement Group), and Alberto Bacchelli (University of Zurich)	311

Need for Tweet: How Open Source Developers Talk about Their GitHub Work on Twitter 322 Hongbo Fang (Carnegie Mellon University, USA), Daniel Klug (Carnegie Mellon University, USA), Hemank Lamba (Carnegie Mellon University, USA), James Herbsleb (Carnegie Mellon University, USA), and Bogdan Vasilescu (Carnegie Mellon University, USA)
On the Prevalence, Impact, and Evolution of SQL Code Smells in Data-Intensive Systems 327 Biruk Asmare Muse (Polytechnique Montréal, Canada), Mohammad Masudur Rahman (Polytechnique Montréal, Canada), Csaba Nagy (Università della Svizzera italiana, Switzerland), Anthony Cleve (University of Namur, Belgium), Foutse Khomh (Polytechnique Montréal, Canada), and Giuliano Antoniol (Polytechnique Montréal, Canada)
On the Relationship between User Churn and Software Issues
 PUMiner: Mining Security Posts from Developer Question and Answer Websites with PU Learning
 Painting Flowers: Reasons for Using Single-State State Machines in Model-Driven Engineering
 Polyglot and Distributed Software Repository Mining with CROSSFLOW
RTPTorrent: An Open-Source Dataset for Evaluating Regression Test Prioritization
SoftMon: A Tool to Compare Similar Open-Source Software from a Performance Perspective . 397 Shubhankar Suman Singh (Indian Institute of Technology Delhi, India) and Smruti R. Sarangi (Indian Institute of Technology Delhi, India)
The Impact of a Major Security Event on an Open Source Project: The Case of OpenSSL 409 James Walden (Northern Kentucky University)
The Scent of Deep Learning Code: An Empirical Study

The State of the ML-Universe: 10 Years of Artificial Intelligence & Machine Learning Software Development on GitHub Danielle Gonzalez (Rochester Institute of Technology, USA), Thomas Zimmermann (Microsoft Research, Redmond), and Nachiappan Nagappan (Microsoft Research, Redmond)	431
Traceability Support for Multi-Lingual Software Projects Yalin Liu (University of Notre Dame, USA), Jinfeng Lin (University of Notre Dame, USA), and Jane Cleland-Huang (University of Notre Dame, USA)	443
Using Large-Scale Anomaly Detection on Code to Improve Kotlin Compiler <i>Timofey Bryksin (JetBrains Research, Saint Petersburg State</i> <i>University), Victor Petukhov (JetBrains), Ilya Alexin (Saint</i> <i>Petersburg State University), Stanislav Prikhodko (Saint Petersburg</i> <i>State University), Alexey Shpilman (JetBrains Research, Higher School</i> <i>of Economics), Vladimir Kovalenko (JetBrains Research), and Nikita</i> <i>Povarov (JetBrains)</i>	455
Using Others' Tests to Identify Breaking Updates Suhaib Mujahid (Concordia University), Rabe Abdalkareem (Queen's University), Emad Shihab (Concordia University), and Shane McIntosh (McGill University)	466
Visualization of Methods Changeability Based on VCS Data Sergey Svitkov (Saint-Petersburg State University) and Timofey Bryksin (JetBrains Research, Saint-Petersburg State University)	477
What Constitutes Software? An Empirical, Descriptive Study of Artifacts Rolf-Helge Pfeiffer (IT University of Copenhagen)	481
 What is the Vocabulary of Flaky Tests? Gustavo Pinto (Federal University of Pará), Breno Miranda (Federal University of Pernambuco), Supun Dissanayake (University of Adelaide), Marcelo d'Amorim (Federal University of Pernambuco), Christoph Treude (University of Adelaide), and Antonia Bertolino (ISTI — CNR) 	492
Data Showcase	
20-MAD - 20 Years of Issues and Commits of Mozilla and Apache Development Maëlick Claes (University of Oulu) and Mika Mäntylä (University of Oulu)	503
A C/C++ Code Vulnerability Dataset with Code Changes and CVE Summaries Jiahao Fan (New Jersey Institute of Technology, USA), Yi Li (New Jersey Institute of Technology, USA), Shaohua Wang (New Jersey Institute of Technology, USA), and Tien N. Nguyen (The University of Texas at Dallas, USA)	508
A Complete Set of Related Git Repositories Identified via Community Detection Approaches Based on Shared Commits <i>Audris Mockus (The University of Tennessee, USA), Diomidis Spinellis</i>	513

(Athens University of Economics and Business, Greece), Zoe Kotti (Athens University of Economics and Business, Greece), and Gabriel John Dusing (The University of Tennessee, USA)

A Dataset and an Approach for Identity Resolution of 38 Million Author IDs Extracted from 2B Git Commits
A Dataset for GitHub Repository Deduplication
A Dataset of Dockerfiles
A Dataset of Enterprise-Driven Open Source Software
A Mixed Graph-Relational Dataset of Socio-Technical Interactions in Open Source Systems 538 Usman Ashraf (Johannes Kepler University, Austria), Christoph Mayr-Dorn (Johannes Kepler University, Austria), Alexander Egyed (Johannes Kepler University, Austria), and Sebastiano Panichella (Zurich University of Applied Sciences, Switzerland)
On the Shoulders of Giants: A New Dataset for Pull-Based Development Research
AndroZooOpen: Collecting Large-Scale Open Source Android Apps for the Research Community 548
Pei Liu (Monash University, Australia), Li Li (Monash University, Australia), Yanjie Zhao (Monash University, Australia), Xiaoyu Sun (Monash University, Australia), and John Grundy (Monash University, Australia)
Dataset of Video Game Development Problems
Employing Contribution and Quality Metrics for Quantifying the Software Development Process
Themistoklis Diamantopoulos (Aristotle University of Thessaloniki), Michail D. Papamichail (Aristotle University of Thessaloniki), Thomas Karanikiotis (Aristotle University of Thessaloniki), Kyriakos C. Chatzidimitriou (Aristotle University of Thessaloniki), and Andreas L. Symeonidis (Aristotle University of Thessaloniki)

GitterCom - A Dataset of Open Source Developer Communications in Gitter	3
 Hall-of-Apps: The Top Android Apps Metadata Archive	8
How Often Do Single-Statement Bugs Occur? The ManySStuBs4J Dataset	3
JTeC: A Large Collection of Java Test Classes for Test Code Analysis and Processing	8
LogChunks: A Data Set for Build Log Analysis	3
Software-Related Slack Chats with Disentangled Conversations	8
TestRoutes: A Manually Curated Method Level Dataset for Test-to-Code Traceability	3

Registered Reports

An Empirical Study on the Impact of Deimplicitization on Comprehension in Programs Using Application Frameworks	. 598
Jürgen Cito (TU Wien and MIT), Jiasi Shen (MIT), and Martin Rinard (MIT)	
Determining the Intrinsic Structure of Public Software Development History Antoine Pietri (Inria), Guillaume Rousseau (Université de Paris and	602

Inria), and Stefano Zacchiroli (Université de Paris and Inria)

Do Explicit Review Strategies Improve Code Review Performance?	606
Large-Scale Manual Validation of Bugfixing Changes Steffen Herbold (University of Goettingen), Alexander Trautsch (University of Goettingen), and Benjamin Ledel (University of Goettingen)	611
Multi-Language Design Smells: A Backstage Perspective Mouna Abidi (Ecole Polytechnique de Montréal), Moses Openja (Ecole Polytechnique de Montréal), and Foutse Khomh (Ecole Polytechnique de Montréal)	615
The Impact of Dynamics of Collaborative Software Engineering on Introverts: A Study Protocol	619

Author Index		623
	 	•••••••••••••••••••••••••••••••••••••••