2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR 2021)

Virtual Conference
17 – 19 May 2021

IEEE

IEEE Catalog Number: CFP2178C-POD
ISBN: 978-1-6654-2985-6

Copyright © 2021 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. 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:	CFP2178C-POD
ISBN (Print-On-Demand):	978-1-6654-2985-6
ISBN (Online):	978-1-7281-8710-5
ISSN:	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

2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)

MSR 2021

Table of Contents

Message from the MSR 2021 Chairs xiv
Message from the MSR 2021 Data Showcase Track Co-Chairs xvii
Message from the MSR 2021 Hackathon Co-Chairs xviii
Message from the MSR 2021 Mining Challenge Track Co-Chairs xix
Message from the MSR 2021 Registered Reports Track Co-Chairs xx
Message from the MSR 2021 Shadow PC and Advisors Co-Chairs xxi
Organizing Committee xxii
Program Committee xxiv

Technical Papers

Mining Workflows for Anomalous Data Transfers 1
Huy Tu (North Carolina State University, USA), George Papadimitriou (University of Southern California, Information Sciences Institute, USA), Mariam Kiran (Energy Sciences Network, Lawrence Berkeley National Labs, USA), Cong Wang (RENCI, University of North Carolina Chapel Hill, USA), Anirban Mandal (RENCI, University of North Carolina Chapel Hill, USA), Ewa Deelman (University of Southern California, Information Sciences Institute, USA), and Tim Menzies (North Carolina State University, USA)

PSIMiner: A Tool for Mining Rich Abstract Syntax Trees from Code 13
Egor Spirin (JetBrains Research, Higher School of Economics, Russia), Egor Bogomolov (JetBrains Research, Higher School of Economics, Russia), Vladimir Kovalenko (JetBrains Research, JetBrains N.V., The Netherlands), and Timofey Bryksin (JetBrains Research, Saint Petersburg State University, Russia)

Comparative Study of Feature Reduction Techniques in Software Change Prediction 18
Ruchika Malhotra (Delhi Technological University, India), Ritvik Kapoor (Delhi Technological University, India), Deepti Aggarwal (Delhi Technological University, India), and Priya Garg (Delhi Technological University, India)

Automatic Part-of-Speech Tagging for Security Vulnerability Descriptions 29
Sofonias Yitagesu (Tianjin University, China), Xiaowang Zhang (Tianjin University, China), Zhiyong Feng (Tianjin University, China), Xiaohong Li (Tianjin University, China), and Zhenchang Xing (Australian National University, Australia)

Identifying Critical Projects via PageRank and Truck Factor ..41
Rolf-Helge Pfeiffer (IT University of Copenhagen, Denmark)

An Empirical Study of Developer Discussions on Low-Code Software Development Challenges ..46
Md Abdullah Al Alamin (University of Calgary), Sanjay Malakar (Bangladesh University of Engineering and Technology, Bangladesh), Gias Uddin (University of Calgary), Sadia Afroz (Bangladesh University of Engineering and Technology, Bangladesh), Tameem Bin Haider (Bangladesh University of Engineering and Technology, Bangladesh), and Anindya Iqbal (Bangladesh University of Engineering and Technology, Bangladesh)

Learning Off-By-One Mistakes: An Empirical Study ..58
Hendrig Sellik (Delft University of Technology, The Netherlands), Onno van Paridon (Adyen N.V., The Netherlands), Georgios Gousios (Delft University of Technology, The Netherlands), and Maurício Aniche (Delft University of Technology, The Netherlands)

The Secret Life of Hackathon Code: Where Does it Come from and Where Does it Go? ..68
Ahmed Imam (University of Tartu, Estonia), Tapajit Dey (Lero—the Irish Software Research Centre, University of Limerick, Ireland), Alexander Nolte (University of Tartu; Carnegie Mellon University, USA), Audris Mockus (University of Tennessee, USA), and James Herbsleb (Carnegie Mellon University, USA)

gambit – An Open Source Name Disambiguation Tool for Version Control Systems ..80
Christoph Gote (ETH Zurich, Switzerland) and Christian Zingg (ETH Zurich, Switzerland)

Escaping the Time Pit: Pitfalls and Guidelines for Using Time-Based Git Data ..85
Samuel W Flint (University of Nebraska--Lincoln, USA), Jigyasa Chauhan (University of Nebraska--Lincoln, USA), and Robert Dyer (University of Nebraska--Lincoln, USA)

Attention-Based Model for Predicting Question Relatedness on Stack Overflow ..97
Jiayan Pei (South China University of Technology, China), Yimin Wu (South China University of Technology, China; Research Institute of SCUT in Yangjiang, China), Zishan Qin (South China University of Technology, China), Yao Cong (South China University of Technology, China), and Jingtao Guan (Research Institute of SCUT in Yangjiang, China)

An Empirical Study on the Usage of BERT Models for Code Completion ..108
Matteo Ciniselli (USI Università della Svizzera italiana, Switzerland), Nathan Cooper (William and Mary, USA), Luca Pascarella (USI Università della Svizzera italiana, Switzerland), Denys Poshyvanyk (William and Mary, USA), Massimiliano Di Penta (University of Sannio, Italy), and Gabriele Bavota (USI Università della Svizzera italiana, Switzerland)

On Improving Deep Learning Trace Analysis with System Call Arguments ..120
Quentin Fournier (Polytechnique Montréal, Canada), Daniel Aloise (Polytechnique Montréal, Canada), Seyed Vahid Azhari (Ciena, Canada), and François Tetreault (Ciena, Canada)

An Empirical Study of OSS-Fuzz Bugs .131
Zhen Yu Ding (Motional (work done at Carnegie Mellon University), USA) and Claire Le Goues (Carnegie Mellon University, USA)

An Exploratory Study of Log Placement Recommendation in an Enterprise System .143
Jeanderson Cândido (Delft University of Technology, The Netherlands), Jan Haesen (Adyen N.V., The Netherlands), Maurício Aniche (Delft University of Technology, The Netherlands), and Arie van Deursen (Delft University of Technology, The Netherlands)

On the Naturalness and Localness of Software Logs .155
Sina Gholamian (University of Waterloo, Canada) and Paul A. S. Ward (University of Waterloo, Canada)

Automatically Selecting Follow-up Questions for Deficient Bug Reports .167
Mia Mohammad Imran (Virginia Commonwealth University, USA), Agnieszka Ciborowska (Virginia Commonwealth University, USA), and Kostadin Damevski (Virginia Commonwealth University, USA)

Self-Admitted Technical Debt in R Packages: An Exploratory Study .179
Melina Vidoni (RMIT University, School of Computing Technologies)

Architecture Smells and Pareto Principle: A Preliminary Empirical Exploration .190
Alexandra-Maria Chaniotaki (Athens University of Economics & Business, Greece) and Tushar Sharma (Siemens Technology, USA)

Technical Debt in the Peer-Review Documentation of R Packages: A rOpenSci Case Study .195
Zadia Codabux (University of Saskatchewan, Canada), Melina Vidoni (RMIT University, Australia), and Fatemeh H. Fard (University of British Columbia, Canada)

How Java Programmers Test Exceptional Behavior .207
Diego Marcilio (Università della Svizzera italiana, Switzerland) and Carlo A. Furia (Università della Svizzera italiana, Switzerland)

A Replication Study on the Usability of Code Vocabulary in Predicting Flaky Tests .219
Guillaume Haben (University of Luxembourg), Sarra Habchi (University of Luxembourg), Mike Papadakis (University of Luxembourg), Maxime Cordy (University of Luxembourg), and Yves Le Traon (University of Luxembourg)

Leveraging Models to Reduce Test Cases in Software Repositories .230
Golnaz Gharachorlu (Simon Fraser University, Canada) and Nick Sumner (Simon Fraser University, Canada)

Which Contributions Count? Analysis of Attribution in Open Source .242
Jean-Gabriel Young (University of Vermont, USA), Amanda Casari (Open Source Programs Office, Google, USA), Katie McLaughlin (Open Source Programs Office, Google, Australia), Milo Z. Trujillo (University of Vermont, USA), Laurent Hébert-Dufresne (University of Vermont, USA), and James P. Bagrow (University of Vermont, USA)

On the Use of Dependabot Security Pull Requests .254
Mahmoud Alfadel (Concordia University, Canada), Diego Elias Costa (Concordia University, Canada), Emad Shihab (Concordia University, Canada), and Mouafak Mkhallalati (Concordia University, Canada)

S3M: Siamese Stack (Trace) Similarity Measure 266
Aleksandr Khvorov (JetBrains, ITMO University, Russia), Roman Vasiliev (JetBrains, Russia), George Chernishev (Saint-Petersburg State University, Russia), Irving Muller Rodrigues (Polytechnique Montreal, Canada), Dmitrij Koznov (Saint-Petersburg State University, Russia), and Nikita Povarov (JetBrains, Russia)

Challenges in Developing Desktop Web Apps: A Study of Stack Overflow and GitHub 271
Gian Luca Scoccia (University of L'Aquila, Italy), Patrizio Migliarini (University of L'Aquila, Italy), and Marco Autili (University of L'Aquila, Italy)

Studying the Change Histories of Stack Overflow and GitHub Snippets 283
Saraj Singh Manes (Carleton University, Canada) and Olga Baysal (Carleton University, Canada)

TNM: A Tool for Mining of Socio-Technical Data from Git Repositories 295
Nikolai Sviridov (ITMO University, Russia), Mikhail Evtikhiev (JetBrains Research, Russia), and Vladimir Kovalenko (JetBrains Research, The Netherlands)

Mining the ROS Ecosystem for Green Architectural Tactics in Robotics and an Empirical Evaluation 300
Ivano Malavolta (Vrije Universiteit Amsterdam, The Netherlands), Katerina Chinnappan (Vrije Universiteit Amsterdam, The Netherlands), Stan Swanborn (Vrije Universiteit Amsterdam, The Netherlands), Grace Lewis (Software Engineering Institute, Carnegie Mellon University, USA), and Patricia Lago (Vrije Universiteit Amsterdam, The Netherlands)

Mining API Interactions to Analyze Software Revisions for the Evolution of Energy Consumption 312
Andreas Schuler (University of Applied Sciences Upper Austria, Austria; Johannes Kepler University, Austria) and Gabriele Kotsis (Johannes Kepler University, Austria)

Googling for Software Development: What Developers Search For and What They Find 317
Andre Hora (Universidade Federal de Minas Gerais (UFMG), Brazil)

Fast and Memory-Efficient Neural Code Completion 329
Alexey Svyatkovskiy (Microsoft), Sebastian Lee (University of Oxford), Anna Hadjitofi (University of Edinburgh), Maik Riechert (Microsoft), Juliana Vicente Franco (Microsoft), and Miltiadis Allamanis (Microsoft)

Identifying Versions of Libraries used in Stack Overflow Code Snippets 341
Ahmed Zerouali (Vrije Universiteit Brussel, Belgium), Camilo Velázquez-Rodríguez (Vrije Universiteit Brussel, Belgium), and Coen De Roover (Vrije Universiteit Brussel, Belgium)

Can I Solve it? Identifying APIs Required to Complete OSS Tasks 346
Fabio Santos (Northern Arizona University, USA), Igor Wiese (Universidade Tecnológica Federal do Paraná, Brazil), Bianca Trinkenreich (Northern Arizona University, USA), Igor Steinmacher (Northern Arizona University, USA; Universidade Tecnológica Federal do Paraná, Brazil), Anita Sarma (Oregon State University, USA), and Marco Gerosa (Northern Arizona University, USA)

Data Balancing Improves Self-Admitted Technical Debt Detection .358
Murali Sridharan (University of Oulu, Finland), Mika Mantyla (University of Oulu, Finland), Leevi Rantala (University of Oulu, Finland), and Maelick Claes (University of Oulu, Finland)

JITLine: A Simpler, Better, Faster, Finer-Grained Just-in-Time Defect Prediction .369
Chanathip Pornprasit (Monash University, Australia) and Chakkrit Tantithamthavorn (Monash University, Australia)

Rollback Edit Inconsistencies in Developer Forum .380
Saikat Mondal (University of Saskatchewan, Canada), Gias Uddin (University of Calgary, Canada), and Chanchal K. Roy (University of Saskatchewan, Canada)

What Code Is Deliberately Excluded from Test Coverage and Why? .392
Andre Hora (Universidade Federal de Minas Gerais (UFMG), Brazil)

Waiting Around or job Half-Done? Sentiment in Self-Admitted Technical Debt .403
Gianmarco Fucci (University of Sannio, Italy), Nathan Cassee (Eindhoven University of Technology, The Netherlands), Fiorella Zampetti (University of Sannio, Italy), Nicole Novielli (University of Bari, Italy), Alexander Serebrenik (Eindhoven University of Technology, The Netherlands), and Massimiliano Di Penta (University of Sannio, Italy)

Mining DEV for Social and Technical Insights about Software Development .415
Maria Papoutsoglou (University of Cyprus & Aristotle University of Thessaloniki), Johannes Wachs (Vienna University of Economics and Business & Complexity Science Hub), and Georgia M. Kapitsaki (University of Cyprus)

How Do Software Developers Use GitHub Actions to Automate Their Workflows? .420
Timothy Kinsman (University of Adelaide, Australia), Mairieli Wessel (University of São Paulo, Brazil), Marco A. Gerosa (Northern Arizona University, USA), and Christoph Treude (University of Adelaide, Australia)

Practitioners' Perceptions of the Goals and Visual Explanations of Defect Prediction Models .432
Jirayus Jiarpakdee (Monash University, Australia), Chakkrit Tantithamthavorn (Monash University, Australia), and John Grundy (Monash University, Australia)

Does Code Review Promote Conformance? A Study of OpenStack Patches .444
Panyawut Sri-iesaranusorn (Nara Institute of Science and Technology, Japan), Raula Gaikovina Kula (Nara Institute of Science and Technology, Japan), and Takashi Ishio (Nara Institute of Science and Technology, Japan)

Revisiting Dockerfiles in Open Source Software Over Time .449
Kalvin Eng (University of Alberta, Canada) and Abram Hindle (University of Alberta, Canada)

Characterising the Knowledge about Primitive Variables in Java Code Comments .460
Mahfouth Alghamdi (The University of Adelaide, Australia), Shinpei Hayashi (Tokyo Institute of Technology, Japan), Takashi Kobayashi (Tokyo Institute of Technology, Japan), and Christoph Treude (The University of Adelaide, Australia)

Predicting Design Impactful Changes in Modern Code Review: A Large-Scale Empirical Study .471
Anderson Uchôa (Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil), Caio Barbosa (Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil), Daniel Coutinho (Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil), Willian Oizumi (Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil), Wesley K. G. Assunção (Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil), Silvia Regina Vergilio (Federal University of Paraná (UFPR), Brazil), Juliana Alves Pereira (Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil), Anderson Oliveira (Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil), and Alessandro Garcia (Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil)

Mining Energy-Related Practices in Robotics Software .483
Michel Albonico (Technological Federal University of Paraná, Brazil), Ivano Malavolta (Vrije Universiteit Amsterdam, The Netherlands), Gustavo Pinto (Federal University of Pará, Brazil), Emitza Guzman (Vrije Universiteit Amsterdam, The Netherlands), Katerina Chinnappan (Vrije Universiteit Amsterdam, The Netherlands), and Patricia Lago (Vrije Universiteit Amsterdam, The Netherlands)

MSR Challenge

On the Rise and Fall of Simple Stupid Bugs: A Life-Cycle Analysis of SStuBs .495
Balázs Mosolygó (University of Szeged, Hungary), Norbert Vándor (University of Szeged, Hungary), Gábor Antal (University of Szeged, Hungary), and Péter Hegedűs (University of Szeged, Hungary; MTA-SZTE Research Group on Artificial Intelligence, Hungary)

How Effective is Continuous Integration in Indicating Single-Statement Bugs? .500
Jasmine Latendresse (Concordia University, Canada), Rabe Abdalkareem (Queen's University, Canada), Diego Elias Costa (Concordia University, Canada), and Emad Shihab (Concordia University, Canada)

Applying CodeBERT for Automated Program Repair of Java Simple Bugs .505
Ehsan Mashhadi (University of Calgary, Canada) and Hadi Hemmati (University of Calgary, Canada)

A Large-Scale Study on Human-Cloned Changes for Automated Program Repair .510
Fernanda Madeiral (KTH Royal Institute of Technology, Sweden) and Thomas Durieux (KTH Royal Institute of Technology, Sweden)

Mea Culpa: How Developers Fix Their own Simple Bugs Differently from Other Developers .515
Wenhan Zhu (University of Waterloo, Canada) and Michael W. Godfrey (University of Waterloo, Canada)

PySStuBs: Characterizing Single-Statement Bugs in Popular Open-Source Python Projects .520
Arthur V. Kamienski (University of Alberta, Canada), Luisa Palechor (University of Alberta, Canada), Cor-Paul Bezemer (University of Alberta, Canada), and Abram Hindle (University of Alberta, Canada)

On the Distribution of "Simple Stupid Bugs" in Unit Test Files: An Exploratory Study .525
Anthony Peruma (Rochester Institute of Technology, USA) and Christian D. Newman (Rochester Institute of Technology, USA)

On the Effectiveness of Deep Vulnerability Detectors to Simple Stupid Bug Detection 530
Jiayi Hua (Beijing University of Posts and Telecommunications, China) and Haoyu Wang (Beijing University of Posts and Telecommunications, China)

MSR Data

AndroidCompass: A Dataset of Android Compatibility Checks in Code Repositories 535
Sebastian Nielebock (Otto-von-Guericke University Magdeburg, Germany), Paul Blockhaus (Otto-von-Guericke University Magdeburg, Germany), Jacob Krüger (Otto-von-Guericke University Magdeburg, Germany), and Frank Ortmeier (Otto-von-Guericke University Magdeburg, Germany)

Denchmark: A Bug Benchmark of Deep Learning-Related Software 540
Misoo Kim (Sungkyunkwan University, Republic of Korea), Youngkyoung Kim (Sungkyunkwan University, Republic of Korea), and Eunseok Lee (Sungkyunkwan University, Republic of Korea)

DUETS: A Dataset of Reproducible Pairs of Java Library-Clients 545
Thomas Durieux (KTH Royal Institute of Technology, Stockholm, Sweden), César Soto-Valer (KTH Royal Institute of Technology, Stockholm, Sweden), and Benoit Baudry (KTH Royal Institute of Technology, Stockholm, Sweden)

KGTorrent: A Dataset of Python Jupyter Notebooks from Kaggle 550
Luigi Quaranta (University of Bari, Italy), Fabio Calefato (University of Bari, Italy), and Filippo Lanubile (University of Bari, Italy)

A Traceability Dataset for Open Source Systems 555
Mouna Hammoudi (Johannes Kepler University, Austria), Christoph Mayr-Dorn (Johannes Kepler University, Austria), Atif Mashkoor (Johannes Kepler University, Austria), and Alexander Egyed (Johannes Kepler University, Austria)

Sampling Projects in GitHub for MSR Studies 560
Ozren Dabic (Software Institute, USI Università della Svizzera italiana, Switzerland), Emad Aghajani (Software Institute, USI Università della Svizzera italiana, Switzerland), and Gabriele Bavota (Software Institute, USI Università della Svizzera italiana, Switzerland)

The Wonderless Dataset for Serverless Computing 565
Nafise Eskandani (Technische Universität Darmstadt, Germany) and Guido Salvaneschi (University of St.Gallen, Switzerland)

AndroCT: Ten Years of App Call Traces in Android 570
Wen Li (Washington State University, USA), Xiaoqin Fu (Washington State University, USA), and Haipeng Cai (Washington State University, USA)

Search4Code: Code Search Intent Classification Using Weak Supervision 575
Nikitha Rao (Microsoft Research), Chetan Bansal (Microsoft Research), and Joe Guan (Microsoft)

Andromeda: A Dataset of Ansible Galaxy Roles and Their Evolution .580
Ruben Opdebeeck (Vrije Universiteit Brussel, Belgium), Ahmed Zerouali (Vrije Universiteit Brussel, Belgium), and Coen De Roover (Vrije Universiteit Brussel, Belgium)

ManyTypes4Py: A Benchmark Python Dataset for Machine Learning-Based Type Inference .585
Amir M. Mir (Delft University of Technology, The Netherlands), Evaldas Latoškinas (Delft University of Technology, The Netherlands), and Georgios Gousios (Delft University of Technology, The Netherlands)

QScored: A Large Dataset of Code Smells and Quality Metrics .590
Tushar Sharma (Siemens Technology, USA) and Marouane Kessentini (University of Michigan, USA)

Apache Software Foundation Incubator Project Sustainability Dataset .595
Likang Yin (University of California, Davis), Zhiyuan Zhang (University of California, Davis), Qi Xuan (Zhejiang University of Technology), and Vladimir Filkov (University of California, Davis)

AndroR2: A Dataset of Manually-Reproduced Bug Reports for Android Apps .600
Tyler Wendland (University of Minnesota, USA), Jingyang Sun (University of Bristish Columbia, Canada), Junayed Mahmud (George Mason University, USA), S. M. Hasan Mansur (George Mason University, USA), Steven Huang (University of Bristish Columbia, Canada), Kevin Moran (George Mason University, USA), Julia Rubin (University of Bristish Columbia, Canada), and Mattia Fazzini (University of Minnesota, USA)

GE526: A Dataset of Open-Source Game Engines .605
Dheeraj Vagavolu (Indian Institute of Technology Tirupati, India), Vartika Agrahari (Indian Institute of Technology Tirupati, India), Sridhar Chimalakonda (Indian Institute of Technology Tirupati, India), and Akhila Sri Manasa Venigalla (Indian Institute of Technology Tirupati, India)

EQBENCH: A Dataset of Equivalent and Non-Equivalent Program Pairs .610
Sahar Badihi (University of British Columbia, Canada), Yi Li (Nanyang Technological University, Singapore), and Julia Rubin (University of British Columbia, Canada)

MSR Data Hackathon

Tracking Hackathon Code Creation and Reuse .615
Ahmed Imam (University of Tartu, Estonia) and Tapajit Dey (Lero—the Irish Software Research Centre, University of Limerick, Ireland)

Building the Collaboration Graph of Open-Source Software Ecosystem .618
Elena Lyulina (JetBrains Research, Russia) and Mahmoud Jahanshahi (Babol, Iran)

Tracing Vulnerable Code Lineage .621
David Reid (University of Tennessee, USA), Kalvin Eng (University of Alberta, Canada), Chris Bogart (Carnegie Mellon University, USA), and Adam Tutko (University of Tennessee, USA)

An Exploratory Study of Project Activity Changepoints in Open Source Software Evolution . . . 624
James Walden (Northern Kentucky University, USA), Noah Burgin (University of Tennessee, USA), and Kuljit Kaur (Guru Nanak Dev University, India)

The Diversity-Innovation Paradox in Open-Source Software . . . 627
Mengchen Sam Yong (Carnegie Mellon University, USA), Lavínia Paganini (Federal University of Pernambuco, Brazil), Huilian Sophie Qiu (Carnegie Mellon University, USA), and José Bayoán Santiago Calderón (University of Virginia, USA)

Author Index . . . 631