Tapajit Dey

Empirical Software Engineering · Mining Software Repositories · Open Source Software · Data Analytics · Quantitative & Qualitative Methods

2 Rhebogue Road, Limerick, Ireland

(+353) 89 9528287 | tapajit.dey@lero.ie | tapjdey.github.io | tapjdey | tapajit-dey | Google Scholar - Tapajit Dey

Researcher with 2+ year of Post-Doctoral research experience and 3+ year of Industry work experience.

Work Experience

Lero, the Irish Software Research Institute - University of Limerick

Limerick, Ireland

Postdoctoral Fellow, with *Prof. Brian Fitzgerald* — *January 2022 - Present*

- **Research Project: Newcomers in Open Source as part of the *€6m TREES (Trustworthy Responsible Efficient Engineering of Software) Research Program*** - Facilitating the progress and long-term success of new open source software developers and projects — Worked with Prof. Fitzgerald to write the **Grant proposal** and develop the Statement of Work.
- As a steering committee member of *Lero Open Source Program Office (OSPO)* (since 2021), created and managing the GitHub organization for Lero and the OSPO website.
- Overseen Lero's efforts for ensuring Equality, Diversity, and Inclusion (EDI) as a member of the *Diversity Committee at Lero* (since 2022) and arranged EDI training sessions for the committee members.

Postdoctoral Researcher, with *Prof. Brian Fitzgerald* — *Sept. 2020 - Jan. 2022*

- **Research Project: InnerSource at Huawei** - Facilitating Innersource software development (adoption of Open Source culture within a company) for Huawei as part of the *TREES Research Program*.
- Became a member of the InnerSource Commons community, conducted a Global State of InnerSource survey with InnerSource Commons members in 2020, and presented the results in the InnerSource APAC summit in 2020.
- Developed several tools for helping Huawei's InnerSource adoption, organized InnerSource workshops at Huawei by distinguished professionals from Industry and Academia, and contributed a pattern for identifying candidate InnerSource projects to the InnerSource Commons community.

IBM India — India Software Lab : SRDC

Bangalore, India

Staff Research Engineer — *June 2012 - June 2015*

- **Primary Responsibility** — In charge of Release & Quality Assurance of IBM's Process Development Kits (PDK).
- Automated the test procedures using **Python** that enhanced test coverage by ∼**300x**, reduced turn-around-time by **75%**, improved efficiency by **50%**, and reduced customer issues drastically.
- Test methodology: "An Integrated Approach to Process Design Kit (PDK) Verification" accepted for **Patent Publication**.
- Mentored colleagues in conducting automated unit-testing of their models and conducted department-wide seminars on basic and advanced Python programming.

Python | Quality Assurance | Automation | Tool Design

Education

University of Tennessee, Knoxville (UTK)

Knoxville, TN, USA

Ph.D. in Computer Science, With *Dr. Audris Mockus* — *Aug. 2015 - Aug. 2020*

- *Thesis Title* — Modeling User-Affected Software Properties for Open Source Software Supply Chains.
- Conducted large-scale studies of open source software ecosystems and software supply chains with the goal of better understanding the effects of software dependencies and users on software popularity, quality, issues, and pull requests using data from World of Code, GitHub, and GHTorrent.
- *Research Focus* — Software Repository Mining, Data Analytics, Empirical Software Engineering.

Indian Institute of Technology (IIT) Kharagpur

Kharagpur, WB, India

Dual Degree (B.Tech.(Hons.) + M.Tech.) in Electronics and Electrical Communication Engineering — *July 2007 - May 2012*

- *Specialization* — Microelectronics and VLSI Design

Honors & Awards

2022	**Distinguished Reviewer Award**, The 19th Working Conference on Mining Software Repositories (MSR) 2022 - Technical Track	*Pittsburgh, PA, USA + Virtual*
2021	**Distinguished Reviewer Award**, The 18th Working Conference on Mining Software Repositories (MSR) 2021 - Technical Track	*Madrid, Spain - Virtual*
2019	**Best Paper Award**, Proceedings of the 15th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE - 2019)	*Recife, Brazil*
2017	**Best Graduate Research Assistant**, University of Tennessee, Knoxville — College of Engineering (Award Amount: $ 2500)	*Knoxville, TN, USA*
2011	**Finalist**, Best B.Tech Project, for designing an ultra low power watchdog timer — Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur	*Kharagpur, WB, India*

Teaching Experience

Introduction to Information Technology
University of Limerick

Co-Instructor — *Fall-2020 — Present*

- Co-Instructed the course *(with Prof. Brian Fitzgerald)* designed for **1st and 2nd-year undergraduates** intended to provide an introduction to modern information technology concepts and solutions.
- **Restructured course evaluation** due to COVID restrictions in AY 2020-2021, designed the end-of-semester evaluation projects, detailed grading rubrics, and took care of the grading.
- **Renovated the course** in Fall 2021 to incorporate latest developments in the field and to conform with the Universal Design for Learning (**UDL**) guidelines.

Programming Languages and Systems
University of Tennessee

Teaching Assistant — *Spring-2017*

- Graded weekly assignments and the mid-term exam, conducted an introductory lecture on Python.

Fundamentals of Digital Archeology
University of Tennessee

Teaching Assistant — *Fall-2016, Fall-2015*

- Helped the professor by grading the assignments, overseeing class activities during his absence.

Operating Systems
University of Tennessee

Teaching Assistant — *Spring-2016*

- Helped the professor by grading the weekly assignments and grading the mid-term exam.

Supervising Experience

Doctoral Thesis Co-Supervisor
University of Limerick

- Supervising a Professional Doctorate Student together with Prof. Brian Fitzgerald and Prof. Kieran Conboy, and another Professional Doctorate Student together with Prof. Conor Ryan.
- Thesis titles: “Scaling Scrum with stable queuing systems - a mixed methodology evaluation” and “Design of AI Model for Fault Prediction and Debugging during Software Development Life Cycle”

Masters Thesis Supervisor
University of Limerick

Artificial Intelligence and Machine Learning — *Summer-2022*

- Supervised a total of **4 Masters students** for their Thesis.
- Thesis Topics - “Software product-user feedback sentimental analysis on Twitter at cross country level”, “Text Classification on GITHUB Projects”, “A Study of Amazon Products & Product Reviews: Product Popularity, Review Sentiment, and Market Basket Analysis”, “Movie Revenue Prediction using Machine Learning Algorithms”.

Masters Thesis Supervisor

University of Limerick

Artificial Intelligence and Machine Learning

Summer-2021

- Supervised a Masters students for their Thesis.
- Thesis Topics - "Detecting Toxic Comments on StackOverflow using BERT."

Masters Thesis Committee Member

University of Limerick

Artificial Intelligence and Machine Learning

Summer-2021 - Present

- Served at Thesis Committee Member of 5 Masters students and 3 Bachelors students.

Grants/Funding

Open Research Training Programme: Training for Open Research in an Irish Context (TROPIC)

University of Limerick

Funding Body - National Open Research Forum (NORF), Ireland

Amount - €190,000.

- The grant describes a plan for developing and rolling out an Open Science and Research Training curriculum in Ireland across all disciplines.
- Liaised with 10+ partners in Lero, University of Limerick Library, and University of Galway Library to draft the Expression of Interest for the grant.
- Liaised with 30+ partners across 5 institutes in Ireland to draft the final grant proposal.

Huawei Open Source Project

University of Limerick

Funding Body - Huawei

Amount - €293,032.

- Worked with Prof. Brian Fitzgerald and liaised with Huawei engineers to draft the work program and statement of work.

Skills

Programming Languages (According to level of Proficiency)

Python ●●●●●	**R** ●●●●●	**Shell Script(Bash)** ●●●●○
MySQL ●●●○○	**Perl** ●●○○○	**JavaScript/HTML/C/C++** ●●○○○

Software Tools and Machine Learning Techniques

Development Frameworks	R-Studio, Jupyter (IPython) Notebook, Visual Studio Code, Vim
Tools & Technologies	R-Shiny, MongoDB, Git, Docker, LaTeX
Supervised Learning	Regression — LM, GLM, GAM, Random Forest, Neural Network (Deep Learning), SVM
Unsupervised Learning	*Clustering* - K-means, Hierarchical etc.; *Dimension Reduction* - PCA, SVD; Association Rule
Bayesian Networks (BN)	BN structure search , BN inference (Causal and Non-Causal)

Research Projects

Facilitating Newcomers in Open Source

- Developing guidelines for new developers in Open Source to help them find and be accepted to suitable projects.
- Assisting developers in OSS projects to rise in rank to become core team members.
- Developing guidelines for projects to increase their attractiveness (magnetism) and developer retention (stickiness).

Open Source | Newcomers | Mixed Methods

Facilitating InnerSource at Huawei

- This project is focused on helping Huawei adopt InnerSource - an intra-organization open source development paradigm with a focus on inter-departmental sharing of code and ideas and the development of communities of interest.
- Conducted a cross-industry State of InnerSource survey, worked with Huawei developers to develop guidelines and a tool for selecting candidate projects (which was accepted as a Pattern in InnerSource Commons - LINK), InnerSource project success metrics, Incentive structure for motivating Huawei employees to work on InnerSource, conducted a series of tutorial events, workshops, and 4 training sessions by experienced professionals from industry and academia.

InnerSource | Survey | Tool design | Consulting

Vector Representation of developers, projects, and APIs - Skill Space

- Aim - Define a feasible representation of developers' and projects' expertise in specific focus areas of software development by gauging their fluency with different APIs.
- Method - Defined the concept of *Skill Space*, operationalized based on the World of Code data containing information on the APIs extracted from changes to source code files in 17 programming languages.
- Result - Established a proof-of-concept for *Skill Space*, which provides a vector representation (derived using the Doc2vec approach) for individual developers, projects, programming languages, or APIs, with the topology of the resulting representations reflecting the conceptual and practical (API-related) relationships among these four entities.

Doc2Vec | Developer Expertise | Python | Open Source Software | World of Code

Study of Hackathon Code Creation and Reuse

- Conducted archival analysis of over 22,000 hackathons and a survey of 500+ developers for identifying the origin of the code used in the projects, the subsequent reuse of code developed during hackathons (by tracking blob reuse), factors that affect code reuse.
- Worked closely with researchers from CMU, University of Tartu (Estonia), and University of Tennessee in deciding the research goals and mentored a Masters student.

Hackathon | Code Reuse | Mining Software Repositories | Empirical Study | Survey | World of Code

Study of Code-Commit Bots

- Developed a systematic approach to detect bots that commit code in Git repositories using committer name, commit messages, files modified by the commit, and projects associated with the commits.
- Characterized bots based on the commits, identified frequently modified file types.
- Collaborated with other Graduate and Undergraduate Students, an industry professional from GitHub, and a professor from Carnegie Mellon University.

Bot Detection | Code Commits | Text Analysis | Random Forest | Predictive Modeling | Python | R

Various studies of the NPM Ecosystem using *World of Code*

- Explored User participation to NPM packages inside and outside of the users' individual supply chains (packages the user's projects depend on directly or indirectly - dataset consisted of **1,376,946** issues and pull-requests(PR) created for **4433** NPM packages and **272,142** issue creators)
- Identified various social and technical factors affecting Pull Request Quality (chance of the Pull Request being accepted within a month) using a dataset containing **483,988** Pull Requests.
- Explored the relationship between package dependencies and package popularity, and proposed an usage-independent *actionable* and *easy-to-use measure* of software quality using data from NPM packages and a proprietary mobile software.

Software Supply Chain | Software Dependency | Issues | Pull Requests | Clustering | R | Python | Bayesian Network

Professional Services and Memberships

- **PC Member** - 46th International Conference on Software Engineering (ICSE) 2024 (Technical Track).
- **PC Member** - 45th International Conference on Software Engineering (ICSE) 2023 (Technical Track, Poster Track, Artifact Evaluation Track).
- **Publicity and Social Media Co-Chair** - 20th Working Conference on Mining Software Repositories (MSR) 2023.
- **PC Member** - 20th Working Conference on Mining Software Repositories (MSR) 2022 (Mining Challenge Track).
- **Publicity and Social Media Co-Chair** - 19th Working Conference on Mining Software Repositories (MSR) 2022.
- **PC Member** - 19th Working Conference on Mining Software Repositories (MSR) 2022 (Technical Track).
- **Shadow PC Mentor** - 19th Working Conference on Mining Software Repositories (MSR) 2022.
- **Proceedings Chair** - 18th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE - 2022).
- **PC Member** - 18th Working Conference on Mining Software Repositories (MSR) 2021 (Technical Track).
- **PC Member** - 17th International Conference on Open Source Systems (OSS) 2021.
- Reviewed a number of papers for a number of journals, including: *Empirical Software Engineering*, ACM Transactions on Software Engineering and Methodology (*TOSEM*), IEEE Transactions on Engineering Management, IEEE's Transactions on Software Engineering (*TSE*), *IEEE Software*, ACM Transactions on Internet of Things (TIOT), Journal of Systems and Software, etc.
- Member of **ACM**.
- Member of **InnerSource Commons**, a global forum of InnerSource practitioners.
- Member of **Lero Open Source Programme Office (OSPO)** and **Lero Diversity Committee**.
- Organized the World of Code Hackathon in Limerick, Ireland in 2022.

Publications: Tapajit Dey

Book Chapters

[1] Sadika Amreen, Bogdan Bichescu, Randy Bradley, **Tapajit Dey**, Yuxing Ma, Audris Mockus, Sara Mousavi, and Russell Zaretzki. "A Methodology for Measuring FLOSS Ecosystems." In "Towards Engineering Free/Libre Open Source Software (FLOSS) Ecosystems for Impact and Sustainability", pp. 1-29. Springer, Singapore, 2019. LINK

Journal Articles

[1] **Tapajit Dey**, Willem Jiang, Brian Fitzgerald "Knights and Gold Stars: A Tale of InnerSource Incentivization", Accepted in *IEEE Software*. LINK. Impact Factor: 3 (June 2022)

[2] Ahmed Imam, **Tapajit Dey**, Alexander Nolte, Audris Mockus, James D Herbsleb "One-off Events? An Empirical Study of Hackathon Code Creation and Reuse", Empirical Software Engineering 27, no. 7 (2022): 1-49. LINK. Impact Factor: 3.762 (2021)

[3] Yuxing Ma, **Tapajit Dey**, Chris Bogart, Sadika Amreen, Marat Valiev, Adam Tutko,d David Kennard, Russell Zaretzki, Audris Mockus "World of Code: Enabling a Research Workflow for Mining and Analyzing the Universe of Open Source VCS data", Empirical Software Engineering 26, no. 2 (2021): 1-42. LINK. Impact Factor: 3.762 (2021)

[4] Andrey Krutauz, **Tapajit Dey**, Peter Rigby, and Audris Mockus. "Do Code Review Measures Explain the Incidence of Post-Release Defects?" Empirical Software Engineering 25, no.5 (2020): 3323–3356. LINK. Impact Factor: 3.762 (2021).
Also accepted in Journal-First Track in ESEC/FSE 2020 conference.

[5] **Tapajit Dey** and Audris Mockus. "Deriving a Usage-Independent Software Quality Metric." Empirical Software Engineering (2020): 1-46. LINK. Impact Factor: 3.762 (2021)

Conference Articles

[1] **Tapajit Dey**, Andrey Karnauch, and Audris Mockus "Representation of Developer Expertise in Open Source Software." In 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE) (pp. 995-1007). IEEE. LINK

[2] Ahmed Imam, **Tapajit Dey**, Alexander Nolte, Audris Mockus, James D Herbsleb "The Secret Life of Hackathon Code - Where does it come from and where does it go?" In 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR) (pp. 68-79). IEEE. LINK

[3] Ahmed Imam, **Tapajit Dey** "Tracking Hackathon Code Creation and Reuse." In 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR) (pp. 615-617). IEEE. LINK

[4] **Tapajit Dey** and Audris Mockus "Effect of Technical and Social Factors on Pull Request Quality for the NPM Ecosystem." In Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), pp. 1-11. 2020. LINK

[5] **Tapajit Dey**, Sara Mousavi, Eduardo Ponce, Tanner Fry, Bogdan Vasilescu, Anna Filippova, and Audris Mockus "Detecting and Characterizing Bots that Commit Code." In Proceedings of the 17th International Conference on Mining Software Repositories (pp. 209–219). Association for Computing Machinery, 2020. LINK

[6] Tanner Fry, **Tapajit Dey**, Andrey Karnauch, and Audris Mockus "A Dataset and an Approach for Identity Resolution of 38 Million Author IDs extracted from 2B Git Commits." In Proceedings of the 17th International Conference on Mining Software Repositories (pp. 518–522). Association for Computing Machinery, 2020. LINK

[7] **Tapajit Dey**, Yuxing Ma, and Audris Mockus. "Patterns of effort contribution and demand and user classification based on participation patterns in npm ecosystem." In Proceedings of the Fifteenth

International Conference on Predictive Models and Data Analytics in Software Engineering, pp. 36-45. ACM, 2019. LINK

[8] **Tapajit Dey** and Audris Mockus. "Modeling Relationship between Post-Release Faults and Usage in Mobile Software." In Proceedings of the 14th International Conference on Predictive Models and Data Analytics in Software Engineering, pp. 56-65. ACM, 2018. LINK

[9] **Tapajit Dey** and Audris Mockus. "Are software dependency supply chain metrics useful in predicting change of popularity of npm packages?." In Proceedings of the 14th International Conference on Predictive Models and Data Analytics in Software Engineering, pp. 66-69. ACM, 2018. LINK

[10] **Tapajit Dey**, Jacob Logan Massengill, and Audris Mockus. "Analysis of popularity of game mods: A case study." In Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts, pp. 133-139. ACM, 2016. LINK

Workshop Articles

[1] **Tapajit Dey**,Bogdan Vasilescu, and Audris Mockus "An Exploratory Study of Bot Commits." In Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops (pp. 61–65). Association for Computing Machinery, 2020. LINK

[2] Yuxing Ma, **Tapajit Dey**, and Audris Mockus. "Modularizing global variable in climate simulation software: position paper." In Proceedings of the International Workshop on Software Engineering for Science, pp. 8-11. ACM, 2016. LINK

Works Under Review/In Progress

[1] A systematic literature review of 338 papers on OSS developers' onboarding and offboarding process.

[2] Designing a tool for helping OSS newcomers find suitable project based on their expertise.

[3] Managing developers' and software firms' impression for mitigating bias due to false stereotypes.

Citations: — See List of Publications from the Google Scholar Page for citation information.

References: — Available on request.