Tapajit Dey

PhD Candidate · Computer Science — Data Analytics

1402 Highland Ave, Apt 6, Knoxville, TN - 37916, USA

☎ (+1) 865-361-3643 | @ tapjdey@gmail.com, tdey2@vols.utk.edu | tapjdey.github.io | tapjdey | tapajit-dey | Google Scholar - Tapajit Dey

Summary

- Computer Science PhD candidate passionate about *Data Mining* and *Data Science*, graduating in *Summer 2020.*
- Experienced in Software Repository Mining, Quantitative Research methods including various Supervised and Unsupervised Machine Learning techniques, Software Supply Chain, and Empirical Software Engineering.

Education

University of Tennessee, Knoxville (UTK)

Knoxville, TN, USA

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

- *Thesis Topic* — Modeling User-Affected Software Properties for Open Source Software Supply Chains
- *Research Focus* — Software Repository Mining, Data Analytics, Empirical Software Engineering
- Overall *GPA* — 3.91/4.00
- *Relevant Courses* — Data Mining Methods, Evidence Engineering, Customer Analytics (*Haslam College of Business, UTK*)

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
- Overall *GPA* — 8.67/10.00
- *Relevant Courses* — Data Structure and Object Representation, Advanced Operating System Design, Artificial Intelligence

Work Experience

University of Tennessee, Knoxville

Knoxville, TN, USA

Graduate Research and Teaching Assistant — *August 2015 - Present*

- TA Courses: Fundamentals of Digital Archeology, Programming Languages and Systems, Operating Systems

IBM India — India Software Lab : SRDC

Bangalore, India

Staff Research Engineer — *June 2012 - June 2015*

- **Primary Responsibility** — Key member of the team in charge of Quality Assurance of IBM's 200mm Process Development Kits (PDK), containing device models (primarily MOSFETs) and support components for chip design and layout
- Automated the test procedures using **Python** and created an end-to-end tool for running simulations, parsing the result, and creating a PDF report; 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** (Ref. No. IN820131127)

Python | Quality Assurance | Automation | Tool Design

IBM India — India Software Lab : SRDC

Bangalore, India

Internship — *May 2011 - July 2011*

- Designed a Python based GUI for simulating and plotting the results of simulation for SOI (Silicon On Insulator) devices

Python | GUI

Skills

Programming Languages (According to level of Proficiency)

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

Software Tools and Machine Learning Techniques

Development Frameworks	R-Studio, Jupyter (IPython) Notebook, Vim, Sublime
Other Tools	MongoDB, Git, Docker, LaTeX
Supervised Learning	Linear Regression (LM, GLM), 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 Specific Tools	World of Code tool, GHTorrent, Google Big Query
Languages Spoken	English, Bengali, Hindi

Relevant Research Projects

Effects of Software Usage on Perceived Quality of a Software

- Proposed an usage-independent *actionable* and *easy-to-use measure* of software quality
- Used **Bayesian Networks** and **Random Forest** to model the interrelationship between software faults, complexity, and usage
- Developed the model by studying a proprietary mobile software and verified the external validity of the concept for popular NPM packages

Software Usage | Software Quality | Bayesian Network | Random Forest | R

User Classification based on Participation Patterns in NPM Ecosystem

- Explored User participation to NPM packages inside and outside of the users' individual supply chains, i.e. 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)
- Classified the Users based on their participation patterns using C-means clustering algorithm

Software Supply Chain | Software Dependency | NPM Ecosystem | Issues | Pull Requests | Clustering | R | Python

Identifying Bots that Commit Code

- Proposed 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.
- 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

Representing Software Developers' and Projects' specific expertise using Vector Embedding

- Used Libraries imported in code files edited/authored by developers/projects to measure their expertise.
- Used LSI and Doc2Vec for representing their expertise for over a skill space.

LSI | Doc2Vec | Expertise | Python | Open Source Software

Predicting Pull Request Quality

- Investigated the factors affecting Pull Request Quality (chance of the Pull Request being accepted within a month) using Random Forest models
- Replicated a study that explored a similar question by investigating a larger dataset (**483,988** Pull Requests from **4218** popular NPM packages) and improved the performance by adding additional factors, achieving an *AUC-ROC value of* ***0.95***, **6%** higher than the previous study

Pull Requests | Pull Request Quality | Predictive Modeling | NPM Ecosystem | Random Forest | R | Python

Software Supply Chain: Exploring the Effects of Interdependencies in a Software Ecosystem on Popularity

- Created Software Supply Chain for NPM Ecosystem by analyzing dependencies of packages
- Explored the effects of dependencies on Package popularity, measured by downloads

Software Supply Chain | Software Dependency | NPM Ecosystem | Software Popularity | R

Honors & Awards

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*
2005	**Ranked 4th**, Regional Mathematical Olympiad, (Selection test for International Mathematical Olympiad)	*India*

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** and Audris Mockus. "Deriving a Usage-Independent Software Quality Metric." Empirical Software Engineering (2020): 1-46. LINK

Conference Articles

[1] **Tapajit Dey**, Sara Mousavi, Eduardo Ponce, Tanner Fry. Bogdan Vasilescu, Anna Filippova, and Audris Mockus "Detecting and Characterizing Bots that Commit Code", Accepted in Mining Software Repositories Conference, 2020. — A proposed systematic approach to detect bots using author names, commit messages, files modified by the commit, and projects associated with the commits; and characterization of the bots found based on the time patterns of their code commits and the types of files modified. LINK

[2] 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", Accepted in Mining Software Repositories Conference (Data Showcase Track), 2020. LINK

[3] **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

[4] **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

[5] **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

[6] **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" — Invited article in 2nd International Workshop on Bots in Software Engineering, 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

Pre-print Articles (Not Peer-reviewed)

[1] **Tapajit Dey** and Audris Mockus. "A Matching Based Theoretical Framework for Estimating Probability of Causation." arXiv preprint arXiv:1808.04139 (2018). — Proposing a novel way to estimate probability of causation using statistical matching for experimental and observational scenarios. LINK

[2] Yuxing Ma, **Tapajit Dey**, Jarred M. Smith, Nathan Wilder, and Audris Mockus. "Crowdsourcing the discovery of software repositories in an educational environment." PeerJ Preprints 4 (2016): e2551v1. LINK

[3] **Tapajit Dey** and Audris Mockus "Which Pull Requests Get Accepted and Why? A study of popular NPM Packages", arXiv preprint arXiv:2003.01153 — A study of 483,988 PRs for 4218 NPM projects that identified 14 different factors that have an impact on the acceptance of Pull Requests. LINK

Works Under Review/In Progress

[1] Andrey Krutauz, **Tapajit Dey**, Peter Rigby, and Audris Mockus. ”The Impact of Code Review Measures on Post-Release Defects: Replications and Bayesian Networks", Under Review in the International Journal of Empirical Software Engineering journal. — A critical review of the effect of code review on post-release software defects, a replication of an earlier study, and addressing the limitations of that study using Bayesian Network modeling.

[2] 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", Under Review in the International Journal of Empirical Software Engineering journal. — A broad description of the World of Code dataset and the associated tool.

[3] **Tapajit Dey**, Andrey Karnauch, and Audris Mockus "Skill Spaces: Representation of Expertise in Open Source Software" Under review — Representing Software Developers' and projects' specific expertise using LSI and Doc2Vec embedding over a 200 dimensional skill space

[4] **Tapajit Dey** and Audris Mockus "Effect of Software Dependencies on Software Popularity", In progress — A study highlighting the effect of the number of dependents and their popularity on the popularity of a software package.

Other Activity

- Participated in multiple Hackathons and Webinars about the World of Code tool.
- Shared Multiple Datasets related to published work for ease of community access and enabling replication:
  - `https://zenodo.org/record/3701819`
  - `https://zenodo.org/record/3648702`
  - `https://zenodo.org/record/3653283`
  - `https://zenodo.org/record/3700859`
  - `https://zenodo.org/record/3699665`
  - `https://zenodo.org/record/3694401`