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 *Empirical Software Engineering*, *Data Science*, *Data Mining*, and *Open Source Software*, 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
- Helped automate test procedures using **Python** that enhanced test coverage ~**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

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, LASSO), Random Forest (Decision Trees), Neural Network (Deep Learning
Unsupervised Learning	*Clustering* — K-means, C-means, Hierarchical, PAM; *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

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

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

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

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	*Knoxville, TN, USA*
2011	**Finalist**, Best B.Tech Project — 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: — See List of Publications in the next page or from the Google Scholar Page

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.

Journal Articles

[1] **Tapajit Dey** and Audris Mockus. "Deriving a Usage-Independent Software Quality Metric." International Journal of Empirical Software Engineering, 2019. (Accepted - In Press).

Conference Articles

[1] **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.

[2] **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.

[3] **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.

[4] **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.

Workshop Articles

[1] 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.

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).

[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.

Professional Activity

- Participated in multiple Hackathons and Webinars about the World of Code tool.

Referees

Audris Mockus
Ericsson-Harlan Mills Chair Professor
University of Tennessee, Knoxville
@audris@utk.edu

Russell Zaretzki
Associate Professor
University of Tennessee, Knoxville
@rzaretzk@utk.edu