Candidate Statement - Meiyappan Nagappan

I am an Assistant Professor in the David R. Cheriton School of Computer Science at the University of Waterloo (UW). My research is centered around the use of multi-faceted and multi-granular Software Engineering (SE) data to address the concerns of the various stakeholders such as developers, architects, and managers. I published more than 50 blind peer-reviewed journal and conference papers including in top SE venues such as TSE, ICSE, FSE, EMSE, JSS, MSR, ICSME and IEEE Software, with a H-Index of 27 and almost 2000 citations since 2014 according to Google Scholar. I received an ACM Distinguished paper award and a best paper award at the International Working Conference on Mining Software Repositories (MSR 15, 12). MSR is a CORE A conference and is one of the most attended venues after ICSE. In 2018, I was nominated by my peers for my body of work in software analytics and particularly app store analytics and was recognized by MSR with their Early Career Achievement Award [1], which is given to one researcher every year. I am supported by various funding agencies (NSERC, SSHRC, NSF) and industrial partners (SAP, BMO, CA) to the order of more than 800,000 CAD. I supervised/co-supervised 15 graduate students. I continue to collaborate with both industry and academic researchers from across the world in countries like Canada, Japan, Germany, Italy, US and Singapore.

Research: Since receiving my PhD 8 years ago, I have been working on analyzing data from Software Systems for solving SE problems. I do empirical software studies both at the macroscopic level using large data sets comprising thousands of case study subjects and microscopic level using data from one specific software project. I focus my research by looking at who are the stakeholders in SE and what problems they have. I use a combination of quantitative techniques (statistical and machine learning models) and qualitative techniques (interviews and surveys) to provide actionable evidence based intelligence. Below, I group my research contributions in terms of the various stakeholders whose problems I attempt to solve.

- **App Store Developers**: I was one of the first researchers to explore mobile app stores analytics. The biggest challenge in analyzing app store data was in curating 10's of thousands of apps and scaling the analysis to them. In one of my papers on this topic, along with the first graduate student that I mentored, we quantitatively looked at the ratings of 10K apps in Google Play [2]. We found that the way the average rating was calculated for apps was deeply flawed, since unlike products on Amazon.com, mobile apps are continuously evolving, with new versions rapidly replacing the old ones. Based on our findings we made several recommendations to the app developers and Google. The biggest impact that I see from this paper is that five years since the paper was published Google is now updating the ratings system to avoid the issue that we pointed out [3]. In another paper with the first undergraduate student that I mentored, who I later mentored for graduate school, we qualitatively analyzed over 4000 reviews to understand the common types of user complaints for mobile apps [4]. Data from Clarivate's (previously Thomson Reuters) InCites Journal Citation Report, shows that with over 250 citations this is the highest cited paper among the papers published in IEEE Software since 2014. This paper kick started an entire research area in SE on automatically analyzing user reviews from mobile apps. There are several companies that now provide this service to developers, including Google. Additionally, this paper resonated not only with SE researchers, but is cited by researchers from marketing journals to medical journals, who value the insights into what end-users care about in mobile apps. Given my pioneering work in mobile app store analytics, I was invited to share my perspectives on the future of app store analytics at the Leaders of Tomorrow Symposium co-located with the 23rd IEEE International Conference on Software Analysis, Evolution, and Reengineering [5]. Additionally, researchers from academic institutions like University College London, University of Zurich and College of William and Mary build on top of my work on app store analytics.

- **Software Managers and Architects**: In collaboration with a product team at Microsoft, I carried out an empirical investigation on the usage of 'Analytics Service' (AX) – a reporting service provided by Visual Studio Team Services (VSTS) to build dashboards and reports out of their work item tracking data [6]. In this industry track paper we wanted to understand why and how users interact with AX and what are the outcomes and business decisions taken by stakeholders from reports built using AX. The real world impact of this study is that, the results are used by the team in Microsoft to improve the visualizations, query support and customizability of AX, while also providing insights for other companies to build similar systems for their own developers. In another study of software architects, we tackle the enduring question [7] of whether they should write code or not [8] — the argument against architects writing code being that they have limited time. We mined and statistically modelled the data from five large-scale software systems and followed up with interviews of the architects, which we qualitatively analyzed. We found that implementing architectural features is more complex and error prone than software functionality, and the architects tend to introduce fewer bugs into the implementation of architectural features compared to the developers. Thus, contrary to conventional wisdom we recommended that architects should be coding architectural features. The impact of this paper is that the blog about it in the IEEE Software Blog platform was viewed over 33,000 times with almost 500 comments in an associated reddit post [9], with both developers and architects participating in the discussion.

- **Software Developers**: In work with my collaborator at University of Southern California and his student who I mentored throughout his Ph.D., we looked at 21 free-to-download mobile apps from the Google Play store and found that the advertisements (ads) in the apps had a significant amount of hidden costs [10]. We found

that response time increases, battery power decreases, network usage increases, and maintenance costs increase when ads were used. Hence, even though ads are the primary source of revenue from an app, the developers need to be careful about how they use ads in their apps, or they risk losing customers. This work was picked up by mainstream media like Forbes [11] and the Today Show [12]. In another study, I led a team across six universities where we analyzed over 10K C projects in GitHub to determine that there are certain use cases of the *goto* statement that are not as harmful as Edsger W. Dijkstra stated [13]. We found that developers routinely use the *goto* statement for error handling and memory cleanup operations since the C language does not have exception handling or finalize methods. Thus we recommend educators to teach future developers that they can use *goto* statements if they are careful and use it for simple purposes. The open access pre-print version of this paper was downloaded more than 9000 times and had over 600 comments on a Slashdot thread [14].

- **SE Researchers**: Millions of repositories from software forges, like GitHub, have been made available to researchers as potential study subjects through tools like GHTorrent. However, there are limited means of separating the signal (e.g. repositories containing engineered software projects) from the noise (e.g. repositories containing home work assignments). The proportion of noise in a random sample of repositories could skew the study and may lead researchers to reach unrealistic, potentially inaccurate, conclusions. Therefore, along with my students, I proposed a framework and presented a reference implementation of the framework as a tool called `reaper`, to enable researchers to select GitHub repositories that contain evidence of an engineered software project [15]. The results from this paper have been used and cited by almost 60 papers from institutions like North Carolina State University, Delft University, and University of Alberta in just the last two years. More recently, in collaboration with researchers at University of Illinois at Urbana-Champaign and the University of Michigan, we pointed out that there was no way for security researchers or white hat hackers to report vulnerabilities in open source projects to the developers of such projects when the projects are hosted in platforms like GitHub [16]. We found in our case study of hundreds of open source Java projects that 96.6% of the projects with a vulnerability did not have a security notification process in place to allow for private disclosure. The day after our paper was published GitHub coincidentally implemented the two recommendations we made in the paper [17]: the creation of a SECURITY.md file and the ability to privately discuss a security vulnerability with the developer.

Ongoing research: Currently in collaboration with researchers in affective computing and sociology, I started exploring the group dynamics on a massive scale using developers in an open source context (GitHub) where they seldom get a chance to meet face-to-face, as a case study example [18]. One of the projects that we are currently working on (submitted to TSE) is to examine if the personality of a developer plays any role in the acceptance of a pull request. The technical challenge that I am working on in such work is extracting the sentiment or personality information from just the online communication in platforms like GitHub. Additionally, I believe that by building tools that can help the various stakeholders in an industrial context, I can close the loop that starts with doing empirical studies, learning what the problems are, and implementing real world solutions for them. Hence, over the last three years at the UW, I started working closely with industry, some of whom support my work financially. We are currently building solutions that can help identify vulnerable commits (with SAP Labs), improve bug localization (for Bank of Montreal), identify commonly made mistakes that static analysis tools cannot find (for Apache Software Foundation projects with my Japanese collaborators), and identify vulnerable libraries included in mobile apps (in conversation with Snyk and Google) [19].

Teaching: I taught the 'Software Design and Architecture' (CS446) course over the last 3 years with an average student evaluation score of 4.0 – 4.1 out of 5.0. In this course the students learn about specific architecture styles, and design patterns and how and when to use them in a real world project — a mobile app. More than a dozen of the apps developed in the class have been published in the official Google Play store. I also taught the 'Empirical SE using Ultra Large Repositories' (CS846) course for graduate students with an average student evaluation score of 4.2 – 4.5. I taught a similar course in both Rochester Institute of Technology and Queen's University with some of the student projects being published in venues like FSE conference (2014) and EMSE journal (2017). I also taught on the topic of 'Trends and Challenges in SE Research for Mobile Apps' at the 2018 Summer School on SE hosted at the Free University of Bozen-Bolzano, Italy on July 9–12, 2018.

Service: With regards to external service I am currently an associate editor for the Empirical Software Engineering Journal (EMSE) and the Journal of Systems and Software (JSS). I served on the PC of several conferences like ICSE, MSR, and ICSME. I was recently invited to be the PC Co-Chair of the 16th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE) to be co-located with FSE 2020 in Sacramento, CA, and the 18th MSR conference, which will be co-located with ICSE 2021 in Madrid, Spain. I am also serving a three year term on the steering committee of the MSR conference. I am currently the Information Director of the TSE journal, and recently stepped down after two terms as the inaugural editor of the IEEE Software Blog, where almost a 100 posts were curated with almost 200,000 views in aggregate. At the UW School of CS, I served on the faculty hiring committee for the 2018 recruiting season and currently serve on the Grad Committee where we carry out admissions and scholarship review. At the UW Faculty of Math level, I served on the Strategic Plan Implementation working group for Research, Creativity, and Innovation in 2018–19.

References

[1] https://cs.uwaterloo.ca/news/professor-mei-nagappan-receives-early-career-achievement

[2] Israel J. Mojica Ruiz, Meiyappan Nagappan, Bram Adams, Thorsten Berger, Steffen Dienst, Ahmed E. Hassan, "Examining the Rating System Used in Mobile-App Stores", in IEEE Software, vol. 33, no. 6, pp. 86-92, Nov.-Dec. 2016.

[3] https://android-developers.googleblog.com/2019/05/whats-new-in-play.html

[4] Hammad Khalid, Emad Shihab, Meiyappan Nagappan, Ahmed E. Hassan, "What Do Mobile App Users Complain About?", in IEEE Software, vol. 32, no. 3, pp. 70-77, May-June 2015.

[5] Meiyappan Nagappan, Emad Shihab, "Future Trends in Software Engineering Research for Mobile Apps", 23rd International Conference on Software Analysis, Evolution, and Re-engineering (SANER), Osaka, Japan pp. 21 - 32, Mar 13 - 18, 2018.

[6] Pavneet Singh Kochhar, Stanislaw Swierc, Trevor Carnahan, Hitesh Sajnani, Meiyappan Nagappan, "Understanding the Role of Reporting in Work Item Tracking Systems for Software Development: An Industrial Case Study", In Proceedings of the *34th IEEE International Conference on Software Maintenance and Evolution (ICSME 2018)*, Madrid, Spain, Sept 23 - 29, 2018.

[7] https://blogs.msdn.microsoft.com/architectsrule/2010/05/07/do-architects-need-to-code/

[8] Inayat Rehman, Mehdi Mirakhorli, Meiyappan Nagappan, Azad Aralbay, Matthew Thornton, "Roles and Impacts of Hands-on Software Architects in Five Industrial Case Studies", In Proceedings of the *40th International Conference on Software Engineering (ICSE 2018)*, Gothenburg, Sweden, May 27 - Jun 3, 2018.

[9] https://www.reddit.com/r/programming/comments/47c174/why_should_software_architects_write_code_an/

[10] Jiaping Gui, Stuart Mcilroy, **Meiyappan Nagappan**, William G.J. Halfond, "Truth in Advertising: The Hidden Cost of Mobile Ads for Software Developers", In Proceedings of the *37th International Conference on Software Engineering (ICSE 2015)*, Florence, Italy, May 16 - 24, 2015.

[11] https://www.forbes.com/sites/alexkonrad/2015/04/01/free-app-ads-kill-phone-battery-and-data/

[12] https://www.today.com/money/your-free-apps-could-be-costing-you-more-your-paid-t12926

[13] **Meiyappan Nagappan**, Romain Robbes, Yasutaka Kamei, Éric Tanter, Shane McIntosh, Audris Mockus, Ahmed E. Hassan, "An Empirical Study of Goto in C Code from GitHub Repositories", In Proceedings of the *9th joint meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE 2015)*, Bergamo, Italy, August 30 - September 04, 2015.

[14] https://developers.slashdot.org/story/15/02/12/1744207/

[15] Nuthan Munaiah, Steven Kroh, Craig Cabrey, Meiyappan Nagappan, "Curating GitHub for engineered software projects", Empirical Software Engineering. 22(6): 3219-3253.

[16] Brandon Carlson, Kevin Leach, Darko Marinov, Meiyappan Nagappan, Atul Prakash, "Open Source Vulnerability Notification", In: Bordeleau F., Sillitti A., Meirelles P., Lenarduzzi V. (eds) Open Source Systems. OSS 2019. IFIP Advances in Information and Communication Technology, vol 556. Springer.

[17] https://www.youtube.com/watch?v=sGC2rwOiaWc&feature=youtu.be&t=1676

[18] Jesse Hoey, Tobias Schröder, Jonathan Morgan, Kimberly B Rogers, Deepak Rishi, Meiyappan Nagappan, "Artificial Intelligence and Social Simulation: Studying Group Dynamics on a Massive Scale", Small Group Research, 49(6), 647683.

[19] Nikita Volodin, "Detecting Vulnerable JavaScript Libraries in Hybrid Android Applications", Master's Thesis at the University of Waterloo (2019).