ESEC/FSE 2019 Journal-First Presentation Proposal

1. **Title:** Toxic Code Snippets on Stack Overflow (Transactions on Software Engineering)
2. **Authors:** Chaiyong Ragkhitwetsagul, Jens Krinke, Matheus Paixao, Giuseppe Bianco, Rocco Oliveto
3. **Abstract:**

Online code clones are code fragments that are copied from software projects or online sources to Stack Overflow as examples. Due to an absence of a checking mechanism after the code has been copied to Stack Overflow, they can become toxic code snippets, e.g., they suffer from being outdated or violating the original software license. We present a study of online code clones on Stack Overflow and their toxicity by incorporating two developer surveys and a large-scale code clone detection. A survey of 201 high-reputation Stack Overflow answerers (33 percent response rate) showed that 131 participants (65 percent) have ever been notified of outdated code and 26 of them (20 percent) rarely or never fix the code. 138 answerers (69 percent) never check for licensing conflicts between their copied code snippets and Stack Overflow's CC BY-SA 3.0. A survey of 87 Stack Overflow visitors shows that they experienced several issues from Stack Overflow answers: mismatched solutions, outdated solutions, incorrect solutions, and buggy code. 85 percent of them are not aware of CC BY-SA 3.0 license enforced by Stack Overflow, and 66 percent never check for license conflicts when reusing code snippets. Our clone detection found online clone pairs between 72,365 Java code snippets on Stack Overflow and 111 open source projects in the curated Qualitas corpus. We analysed 2,289 non-trivial online clone candidates. Our investigation revealed strong evidence that 153 clones have been copied from a Qualitas project to Stack Overflow. We found 100 of them (66 percent) to be outdated, of which 10 were buggy and harmful for reuse. Furthermore, we found 214 code snippets that could potentially violate the license of their original software and appear 7,112 times in 2,427 GitHub projects.

4. **The original journal paper:**

IEEE Xplore: https://ieeexplore.ieee.org/document/8643998
Preprint: https://arxiv.org/abs/1806.07659

5. **Justification:**

The paper reports completely new results that are not reported before in any prior work. This study is among, if not the first, to address the important issues of toxic code snippets, including outdated and license-violating online code clones on Stack Overflow. The paper presents a study incorporating several state-of-the-art tools and techniques to automatically extract and filter online clone pairs between Stack Overflow and 111 open source projects in the curated Qualitas corpus. We analysed 2,289 non-trivial online clone candidates and reported 100 cases of outdated code snippets and 214 cases of potential software license violations. Our online surveys of 201 Stack Overflow high reputation users (33% response rate) and 87 Stack Overflow visitors are valuable since they support our findings of toxic code snippets, and also reveals the awareness of Stack Overflow answerers and visitors regarding the problems. Currently, there is no literature on finding the origins of code examples copied to Stack Overflow, their potential ramifications, and the awareness of Stack Overflow developers in doing so. Since Stack Overflow is actively used in research and software development, this study gives insights to the problem and lays a foundation to an automatic technique for finding toxic code snippets on Stack Overflow in the future.