Post-Binary Code Similarity Analysis: Scalable Explanation and Verification

Overview:

The PI's career goal is to advance and promote binary code analysis research that can address the grand challenges of our time, such as, protecting critical cyber infrastructures, safeguarding financial assets, and diagnosing cyber attacks, all of which depend upon binary code analysis. As one of the key techniques, binary code similarity analysis provides the ability to identify similarities and differences between two or more pieces of binary code. It enables many real-world applications in scenarios where source code may not be available such as patch analysis, bug search, and malware analysis. Though state-of-the-art binary code similarity analysis methods achieve high accuracy, they face a critical post-analysis challenge, i.e., *tedious human efforts of understanding and verifying the results*. When being applied to real applications, e.g., bug search for Internet-of-Things (IoT) firmware, it can easily generate thousands of highly similar candidates. With that, security analysts can not verify them in a timely manner, which is known as the threat alert fatigue problem. Fortunately, *automatic explanation and verification* methods can revolutionize this process. On one hand, the explanation methods can explain why the models classify two pieces of code as similar. On the other hand, the verification methods can confirm if they are equivalent or not, beyond similarity.

As a crucial step towards the PI's career goal, this research proposes research and education coherent objectives. Particularly, the research objective is to pioneer the effort of designing efficient and scalable explanation and verification methods especially for binary code similarity analysis. The educational objective is to use this proposed research to attract and train a diverse group of participants, including K-12, undergraduate, female, and underrepresented minority (URM) populations in the Science, Technology, Engineering, and Mathematics (STEM) field.

Intellectual Merit:

This project will lead to an automatic and scalable explanation and verification framework that advances the post-binary code similarity analysis with the following innovations:

- **Automatic Explanation for Binary Code Similarity Analysis**, which formulates the state-of-the-art binary code similarity analysis methods into a problem of explaining the combination of graph neural network and natural language processing (NLP) methods.
- **Scalable Code Equivalence Verification**, which takes the scalable concolic execution as the foundation and builds upon it with novel techniques aiming to scalably verify the equivalence between two pieces of binary code.
- **High-Performance Computing for Explanation and Verification**, which aims to make the computation of post-binary code similarity analysis techniques close to real-time with world-class Graphics Processing Unit (GPU) computing infrastructure.

Broader Impact:

This project will result in a scalable explanation and verification library that serves as a foundational tool for fellow science and engineering practitioners from academia, national laboratories and industry. With a commitment to helping K-12, undergraduate, female, and URM populations in the STEM field through the proposed investment and rewarding education plan, this project provides a comprehensive road map to prepare the next-generation of software security professional workers and researchers. This project will contribute to the US national goal to increase participation in science and engineering, which is of paramount value to America's success in addressing global challenges, building a stronger and more diversified workforce and meeting the needs of the global innovation economy. This project will also lead to new courses and redesigned core courses for the PI's home department which is the process of improving the curriculum. To benefit the society at large, the PI will share project data, open source software, and publications with the broader research community.