INFORMATION OF THE DOCTORAL THESIS

Research title: **APPLYING MODEL CHECKING AND FORMAL CONCEPT ANALYSIS TO CLASSIFY AND DETECT MALWARE**

Major: **COMPUTER SCIENCE**

Major code: **62.48.01.01**

PhD student: **NGUYEN THIEN BINH**

Scientific advisor: **Assoc. Prof. Dr. QUAN THANH THO**

University: **HCMC University of Technology – Vietnam National University Ho Chi Minh City**

The thesis summary:

To overcome the drawbacks of *signature matching* malware detection methods that widely used in industry, there is much research approaching the application of *model checking* to detect malware since this technique can logically represent malicious behaviors. However, model checking usually suffers from the infamous *state explosion* problem. Many studies have been conducted to address this, but none of them is dedicated for malware detection. By studying large amount of malware, we found that malicious behavior should not occupy in more than one code segment so-called ω-region. This provides a solid fundamental for the thesis to propose *incremental verification method*, which allows reducing *program model* complexity, thus helping to solve the state explosion problem.

In addition to the state explosion problem, model checking approach for malware detection encounters a major drawback that malware often employs *obfuscation techniques* to mask their harmful behavior. Despite some suggestions into the direction of improving *temporal logic* to solve this problem, each proposal following this direction can only handle one obfuscation technique with the requirement to update the *model checker*, resulting in enormous costs to handle one code obfuscation technique. Thus, the thesis studied the utilization of *abstract interpretation* in order to abstract the program into a minimal *intermediate representation*, eliminating most of the obfuscation techniques. Moreover, the thesis proposes HOPE framework, with the separation of the *deobfuscation* step and the model checking step.

The remaining problem of model checking for malicious code detection is that malicious behaviors are represented by logical formulae. Therefore, the typical data mining approaches based on feature extraction are not easily applied. The thesis solves this problem

with a framework called MarCHGen (Malware Conceptual Hierarchy Generation). In this framework, by extending *Formal Concept Analysis* (FCA), *Viral Logical Concept Analysis* (V-LCA) is proposed in the thesis to generate *viral concept lattice*. Then, the thesis proposes an *On-the-fly Conceptual Clustering* (OCC) technique to generate *malware concept hierarchy*. Finally, the malware concept hierarchy will be monitored by the *pre-large dataset management technique* to avoid re-clustering several times unnecessarily.

The main contributions:

1. Based on ω-region idea, the thesis solves the state explosion problem by proposing the incremental verification methods.
2. ω-region is constructed from a set of instructions called ω-instruction - a new concept proposed by the thesis - and the thesis uses a statistical approach to identify these ω-instructions.
3. The thesis solves the obfuscation problem by proposing an abstracted language, eliminating most common obfuscation techniques.
4. The thesis proposes HOPE framework with the separation of the deobfuscation step and the model checking step.
5. The thesis fulfills the demand of systematizing the viral logic formulae by proposing Viral Logical Concept Analysis (V-LCA) method.
6. The thesis proposes On-the-fly Conceptual Clustering algorithm to avoid calculating all formal concepts at the same time during cluster implementation. In addition, the thesis also proposes applying pre-large dataset management that reduces the cost of building the viral concept lattice and clustering the viral concept when the new malware is updated, or the old malware is removed.

Practical applications of the thesis's results:

Proposals in the thesis were experimented with practical datasets, which proved to be applicable in practice. Specifically, the proposed solution for state explosion problem when applying model checking for malware detection could be applied to detect malware in practice with reasonable running time and accurate results. Moreover, the proposed methods of V-LCA and OCC have proven the capability on practical by constructing the informative viral concept hierarchy from the real malware.

Further research of the thesis:

The research directions in the thesis are potential to further research, such as the followings.

1. The proposal of ω-instruction/ω-region can be expanded for checking other programs that are not malware.

2. Applying machine learning for constructing ω-instruction set.
3. Studying other clustering methods for improving the viral logical concept clustering method.
4. Increasing the number of viral abstraction techniques to generalize more viral logical formulae.
5. Applying machine learning for systematizing the viral logic formulae.

Scientific advisor	**PhD student**
Assoc. Prof. Dr. Quan Thanh Tho	Nguyen Thien Binh