Cybersecurity Internship (M2)

EP CIDRE Inria / CentraleSupelec / CNRS / Univ. de Rennes 1 – IRISA

Caractérisation d'applications malveillantes par apprentissage

Pierre-François Gimenez, Valérie Viet Triem Tong
EPC INRIA CIDRE, Rennes, France
valerie.viettriemtong@centralesupelec.fr

Keywords: Security - Malware - Machine Learning

Contexte et travaux antérieurs

Depuis plusieurs années, l'équipe Inria Cidre s'intéresse à l'étude des codes malveillants, que ceux-ci visent les système d'exploitation Windows ou Android. L'étude de ces codes vise à être capable de reconnaître un code connu comme étant malveillant, puis de détecter un code suspect même si celui-ci ne correspond pas à un code déjà étudié.

L'étude précise d'un malware particulier repose sur une l'analyse statique décrivant le code malveillant en profondeur: *le graphe de flot de contrôle, les API suspectes utilisées, le nombre de chemins d'exécution menant au code suspect, la présence et la qualification des conditions de déclenchement, les protections mises en place par le malware (chiffrement, packing)...* L'analyse dynamique enrichit cette connaissance en décrivant le comportement du malware (le graphe de flot de contrôle du code qui est exécuté, l'impact sur le système, le déroulement de l'attaque, etc.). La richesse de ces données dépend du degré de protection mis en place pour prévenir l'analyse du code malveillant.

Toutes ces données sont utiles pour le calcul de *signatures* caractéristiques des codes malveillants étudiés. Ces signatures permettent d'identifier rapidement des codes malveillants similaires aux codes précédemment étudiés. En revanche, elles ne permettent pas de détecter de *nouveaux* codes malveillants.

Dancs ce contexte, des chercheurs de l'équipe CIDRE et du Laboratoire Haute Sécurité, ont développé MoM, une plateforme d'analyse de malware visant les OS Windows [1] qui permet d'étudier l'impact d'un malware et en particulier d'un ransomware sur une machine Windows. Cette plateforme permet d'obtenir des informations statiques et dynamiques sur le malware, sur son comportement et sur l'impact de l'exécution malveillante sur l'OS (en particulier sur le système de fichier).

Objectif du stage

L'objet du stage est de proposer une approche de détection de codes malveillants utilisant des algorithmes d'apprentissage automatique [2, 3]. Dans l'objectif de réduire à la fois la complexité du stage et le temps de calcul nécessaire à l'obtention de nouvelles données, la détection de malware se basera sur un sous-ensemble des analyses de MoM [4]. Le stagiaire aura alors l'occasion de s'intéresser au fonctionnement technique de quelques analyses statiques et/ou dynamiques effectuées par MoM, ce qui lui permettra ensuite de choisir un modèle d'apprentissage automatique avec un certain recul.

Si l'étudiant souhaite poursuivre dans la recherche après le stage, des opportunités de poursuite en thèse sur un sujet similaire sont envisageables.

Profil recherché: De formation Bac+5 en informatique ou mathématiques, maîtrisant la programmation C, Java et Python, ayant de bonnes connaissances en statistiques et en apprentissage automatique.

Lieu: Le stage sera gratifié. Le stage durera de 5 à 6 mois à partir de février 2021 suivant les contraintes de la formation de l'étudiant. Il se déroulera au sein de l'équipe Inria Cidre à Rennes, sur le campus de Beaulieu.

Contacts
N'hésitez *surtout* pas à nous contacter pour tout renseignement supplémentaire.

pierre-francois.gimenez@centralesupelec.fr
valerie.viettriemtong@centralesupelec.fr

References

[1] Yassine Lemmou, Hélène Le Bouder, and Jean-Louis Lanet. Discriminating Unknown Software Using Distance Model. In *ICACSIS 2019 : 11th International Conference on Advanced Computer Science and Information Systems*, Bali, Indonesia, October 2019. IEEE.

[2] Daniele Ucci, Leonardo Aniello, and Roberto Baldoni. Survey of machine learning techniques for malware analysis. *Computers & Security*, 81:123 – 147, 2019.

[3] Daniel Gibert, Carles Mateu, and Jordi Planes. The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. *Journal of Network and Computer Applications*, 153:102526, 2020.

[4] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.