LSE - Recrutement 2020-2021 - Phase I

Thomas Berlioz, thomas.berlioz, thomas.berlioz@epita.fr

January 2021

Lettre de motivation

Actuellement en première année du cycle ingénieur d'EPITA, je souhaite intégrer le LSE car je suis passionné par la sécurité informatique et que je pense pouvoir apporter à ce laboratoire un travail et une implication à l'image de ma motivation. Ce domaine me fascine car il me fait réaliser à quel point j'ai encore à apprendre chaque fois que je me pose une question en apparence simple et que je vais creuser le sujet.

J'ai commencé à m'y intéresser sérieusement avant le premier confinement avec des sites de challenges et des CTFs évènementiels. Étant curieux et avide de progresser, cela a développé une autonomie naturelle ainsi qu'une aisance à aller chercher l'information qui ont porté leurs fruits. En effet, en 9 mois, j'ai atteint le top 10 du CTF du LSE ($\sim$ 4500 points) avec récemment le challenge `Make Simple`, un exercice de *reverse* à 350 points qui utilise des techniques d'*anti-debug* à base de *threads*. Il était très intéressant d'apprendre leur manipulation dans `gdb` et d'apprendre à scripter en Python pour les exploiter plus facilement pendant le *debugging*. De plus, j'ai atteint le top 740 sur `RootMe` ($\sim$ 4600 points), j'ai pu faire mes premières boxes sur `HackTheBox` (25 machines *root*), et j'ai, entre autres, commencé mon ascension sur `Cryptohack`. J'ai aussi participé à quelques évènements comme le FCSC en Avril organisé par l'`ANSSI` (top 60 senior), le `TokyoWesternsCTF` en Septembre avec l'équipe du LSE ou encore le `HeroCTF` organisé par des étudiants de l'IUT de Vannes (top 5). À noter que j'utilise le pseudo **Ewaël** (ou **Ewael**) à chaque fois.

Loin de m'en lasser, j'ai compris que j'avais à peine effleuré la partie émergée de l'iceberg et j'ai plusieurs projets en tête pour approfondir les domaines qui m'intéressent. La première idée que j'ai eue était une *toolbox* pour les CTFs, afin d'automatiser la détection de vulnérabilités comme des *stack / heap buffer overflows* en système ou l'exploitation d'un *oracle* en cryptographie. Néanmoins j'ai réalisé avec des challenges plus difficiles ou des projets plus explorateurs comme celui de la phase I qu'il était plus intéressant de se concentrer sur un seul aspect plus technique que de rester en surface pour toucher à tout sur un projet de quelques mois. C'est pourquoi j'aimerais me concentrer sur une vulnérabilité connue comme une *CVE* ou une *0-day* et l'explorer. Il s'agirait non seulement de la comprendre, mais aussi de la reproduire, voire de construire mon propre *exploit*. L'objectif serait d'avoir un rapport complet qui couvre le contexte, la vulnérabilité en elle-même et ses conséquences, puis de pouvoir faire ma propre *PoC* (*Proof of Concept*), c'est-à-dire l'*exploit* en lui-même. En revanche, je n'ai pas encore choisi la vulnérabilité. J'ai commencé à regarder du côté de `VirtualBox` avec la *CVE-2018-2698* ou encore la *0-day e1000*

qui permettent de s'échapper de la machine virtuelle pour avoir un accès lecture / écriture dans l'hôte voire d'en prendre le contrôle. Néanmoins, je suis également intéressé par tout ce qui touche au *machine learning*. En effet j'avais codé tout le réseau de neurones d'un OCR fonctionnel lors du S3, et j'aimerais beaucoup me replonger dans ce domaine. J'ai d'ailleurs acheté *Practical Malware Analysis* car j'ai entendu que la classification de *malwares* utilisait des algorithme de *deep learning*, mais je n'ai pas encore pris le temps de m'y pencher sérieusement avec les CTFs. En résumé, si mes recherches concernent actuellement quelle vulnérabilité ou attaque j'aimerais reproduire et étudier en profondeur, je suis ouvert à beaucoup de choses. Je compte également me pencher sur du *browser exploit* dans un futur proche, et cela pourrait aussi être un sujet intéressant pour le LSE si je trouve une cible, car il mettrait en jeu des problématiques de *fuzzing* et aboutirait sur un papier intéressant sur les *internals* avec un *exploit* à la fin en bonus.

De manière générale, j'ai toujours aimé publier et rendre accessible ce que je faisais lorsque c'était possible, avec par exemple des *write-ups* après les CTFs ou des projets réalisés avec ou sans EPITA. J'aspire à partager mon travail et mes recherches avec cette communauté qui m'a aidé à poser les bases de mon avenir. Pouvoir faire partie d'une structure comme le LSE est une opportunité incroyable de parler en profondeur de ce que je fais seul depuis quelques mois, et je pense apporter une soif d'apprendre et une curiosité qui sont stimulantes dans un environnement de gens passionnés. J'ai toujours plein de questions à poser et j'aime m'impliquer dans celles des autres dès que je peux apprendre et progresser. Même en dehors des aspects techniques, je m'intègre bien dans tous les environnements dans lesquels je me retrouve car je m'adapte naturellement à mon entourage et au cadre dans lequel je suis. Mes projets de groupe se sont toujours très bien passés car je suis ouvert au dialogue et à la discussion pour peu qu'on ait cette même envie que moi d'être meilleur chaque jour. Enfin, et même si ce n'est pas l'activité principale du LSE, **j'adore** les CTFs. C'est mon activité principale dans mon temps libre et j'aimerais beaucoup m'impliquer encore plus avec l'équipe du LSE, pour attaquer de plus gros évènements à plusieurs. J'ai aussi quelques idées de challenges à implémenter à force d'en découvrir, et je pense pouvoir alimenter cette culture du CTF en intégrant le laboratoire.

De plus, le cadre du LSE est idéal pour progresser et se stimuler. J'ai toujours été tiré vers le haut en étant simplement entouré de gens passionnés et meilleurs que moi parce que j'ai envie d'être à la hauteur de ce qu'ils représentent pour moi, que ce soit en termes de connaissances, de motivation ou de curiosité et de culture. En plus de cela, ce sera une vraie nouveauté pour moi d'être enfin entouré de gens qui partagent la même capacité de s'investir vraiment dans ce qui nous plaît, et je pense que la stimulation intellectuelle d'être avec eux ne peut être que bénéfique. C'est aussi une occasion d'être entouré de tous les domaines de la cybersécurité qui me plaisent sans pour autant que je sache dans quoi m'orienter plus tard. Je pourrais m'investir dans mes projets et mes réflexions sans me soucier du fait que je connaisse déjà des gens du domaine ou non, ou qu'une voie soit plus facile d'accès qu'une autre. J'aime également l'idée de déjà poser un pied dans les attentes du monde professionnel. Le prisme du CTF et des projets personnels depuis lequel je vois le monde de la sécurité informatique est trop restreint pour m'orienter correctement et j'aimerais pouvoir me rendre compte du niveau et des exigences attendues, même au-delà du pur aspect technique et intellectuel, et des enjeux qu'on peut rencontrer.

D'ailleurs, à force de discuter et d'échanger je pense déjà avoir une certaine idée de mon projet professionnel. En effet, être dans la branche sécurité d'une grosse entreprise et ne m'occuper que de la protéger ne m'attire pas. J'aimerais mieux pouvoir faire partie d'un groupe plus pointu dans leur domaine, comme `Synacktiv` ou `Quarkslab`, voire l'`ANSSI` par exemple. Si j'avoue ne

pas avoir trouvé mon domaine, je commence à visualiser ce qui me plaît ou non dans l'expérience des gens que je rencontre pendant les évènements. Beaucoup de choses me plaisent, et beaucoup de gens m'inspirent, donc même s'il est encore compliqué pour moi de visualiser un seul aspect de la sécurité dans lequel me plonger, je sais que cela viendra à force d'en parler et de travailler sur des projets différents, et que le LSE peut aussi accélérer et aider ce processus.

Cordialement, avec tout mon respect et dans l'espoir de pouvoir en parler de vive voix un jour.

ACM Student Research Competition

A deep learning approach for OSINT : identifying a potential pedocriminal as such on social networks

Les réseaux sociaux sont très largement utilisés comme point de départ entre les pédocriminels et leurs potentielles victimes. En plus d'être un moyen facile de communiquer et d'établir un lien derrière un personnage virtuel fictif, ils permettent également d'afficher une adérence à des pratiques malsaines voire criminelles derrières des symboles qui sont propres au groupe en question. Par exemple, dans le cadre de la pédocriminalité, la spirale est une forme largement reconnue comme signe d'apologie de la pédophilie, selon un document interne du `FBI` publié par `WikiLeaks` en 2007.

Ce projet a pour ambition de construire un réseau de neurones capable d'identifier un profil comme potentiellement pédocriminel à partir d'informations publiques sur ses réseaux sociaux. On ne prétend **pas** présenter un outil fiable juridiquement mais simplement être capable d'identifier un profil comme **potentiellement** pédocriminel, c'est-à-dire un profil enclin à la pédophilie mais qui s'affiche suffisamment pour qu'on suppose qu'il soit en contact avec d'autres, et qu'il ait des activités illicites comme la pédopornographie, voire criminelle s'il passe à l'acte. En pratique, cela permettrait d'identifier directement si un individu d'une enquête est lié ou non au groupe traqué. En plus de pouvoir identifier des liens entre les individus positifs (i.e. reconnu comme potentiel pédocriminel) par le fait de suivre ou d'être suivi sur les réseaux sociaux, chacun représente une donnée supplémentaire dans l'entrainement de l'intelligence artificielle. Ainsi, plus on l'utilise, meilleur le réseau est, puisqu'il suffit d'ajouter le profil aux données d'entrainement une fois que la réponse a été confirmée ou infirmée.

Le *deep learning* est ici bien plus intéressant qu'un algorithme basique puisqu'il requiert seulement une base de données de profils qu'il est très facile de trouver sur les réseaux sociaux. Sans *deep learning*, il aurait fallu constituer d'énormes banques de mots et de symboles importants ainsi que de comptes types avec lesquels un profil positif pourrait intéragir, différencier manuellement chacune de ces interactions (suivre, être suivi, aimer, partager...), et les analyser une par une, là où l'intelligence artificielle ne fera ce processus que pendant l'apprentissage et de plus en plus finement. Par conséquent, c'est un outil capable de gérer plusieurs réseaux sociaux à condition que les données ait été préalablement traitées pour avoir le même format en entrée. Cela permet donc de traquer un suspect sur plusieurs réseaux sociaux à la fois, et si l'on souhaite entrainer cet outil pour qu'il soit plus performant sur un réseau social en particulier, il suffit d'adapter le format en entrée pour qu'il prenne en compte les spécificités de la plateforme

en question, comme le fait de pouvoir *retweet* sur `Twitter` ou le fait de pouvoir enregistrer en favori sur `Instagram`. Le profil type créé par l'intelligence artificielle est donc dynamique. C'est intéressant car cela pourrait permettre de faire tourner l'outil sur une base de profils dont on ne connait pas les résultats à chaque fois qu'un profil différent est ajouté à la base d'entrainement afin d'affiner des résultats vagues.

C'est un projet intéressant techniquement car on peut rapidement constituer des paliers et des axes d'amélioration. On peut commencer par ne gérer que les comptes suivis par exemple, avec une sortie binaire. Ensuite, on peut ajouter d'autres données en entrée petit à petit, comme les interactions avec un compte (publications aimées, commentées...), puis affiner la sortie en passant d'une sortie binaire à un indice entre 1 et 4. L'objectif final étant de gérer toutes les informations publiques d'un compte pour avoir un pourcentage en sortie, 100% étant un profil potentiellement pédocriminel ou affilié à la pédophilie de manière sûre. Le modèle utilisé serait un *ANN* (*Artificial Neural Network*) car j'en ai déjà manipulés dans le passé et que cela semble suffisant dans un sens où l'entrée est déjà formattée : il n'y pas de bruit comme dans un fichier audio ou une photo avec des éléments parasites.

C'est une problématique qui me tient à cœur car elle est encore trop peu mise en avant sur les réseaux sociaux. En plus de traiter de la cybercriminalité, il met en avant les profils auxquels il faut faire attention en tant qu'utilisateur. C'est un point important : on utilise seulement des informations accessibles en tant qu'utilisateur de la plateforme. Il rentre ainsi dans le cadre de l'*OSINT* (*Open Source Intelligence*) en constituant un outil utile, par exemple dans le contexte d'une enquête où il serait important de savoir facilement si un ou plusieurs individus sont impliqués ou non à partir de leurs profils sur les réseaux sociaux. Cela ne constitue évidemment pas une preuve et ne permet pas d'éviter une vérification humaine, mais c'est toujours du temps gagné, donc des traumatismes évités voire des vies sauvées. On peut également souligné que ce projet se prête bien au travail en équipe puisque si le réseau de neurones constitue l'organe principal, il faudrait également développer un outil pour extraire facilement et formatter les données d'un profil à partir d'un identifiant unique comme le nom d'utilisateur et le réseau social concerné.