L’intelligence Artificielle Pour le Traitement des Données de Sécurité

Ce document décrit un sujet de thèse de doctorat portant sur les défis et les objectifs liés à l’utilisation des systèmes et algorithmes d’intelligence artificiel pour l’étude, la conception et le développement d’une plateforme de traitement, d’analyse et de structuration des données de sécurité. L’encadrement s’effectuera entre SAP (industriel) et EURECOM (académique). Le candidat étudiera en particulier les algorithmes d’analyse, de classification et de regroupement des méga-données (big data), leur applicabilité aux données de sécurité (alertes de menaces, vulnérabilités, virus, traces d’attaques, etc.) qui peut engendrer plusieurs défis notamment en termes d’hétérogénéité de ces données, de la diversité des sources d’informations et du degré de confidentialité concernant ces données de sécurité. L’objectif de cette thèse est en relation avec les besoins internes de SAP, ainsi que de leurs clients concernant la simplification du traitement des données massives de sécurité.

1. Introduction

Gérer la sécurité dans les grandes entreprises et les systèmes complexes ne se résume plus aux méthodes de défense classiques du type anti-virus, pare feu, DMZ, systèmes d’intrusions, etc. L’origine de la menace n’est pas uniquement liée aux failles et vulnérabilités des systèmes, mais elle concerne le comportement humain des employés. Avec les attaques du type ingénierie sociale, l’employé est devenu la passerelle idéale pour les attaques sophistiquées. La sécurité informatique, se transforme alors en cyber sécurité, qui prend en compte les menaces venant du cyber espace. Le cyber espace regroupe globalement trois types d’activités, le cyber crime, le cyber activisme et le cyber espionnage. Ces trois piliers des cyber menaces visent aveuglément tous les systèmes et toutes les organisations privées et publiques sans distinction. Afin de contrer ces nouvelles menaces extrêmement sophistiquées qui pourraient être assimilées à une cyber guerre moderne, le concept d’intelligence contre la cyber menace ou « Cyber Threat Intelligence » (CTI) a été défini. Le concept du CTI est de

collecter toutes les informations liées aux menaces éventuelles le plus en amont possible. Les méthodes utilisées se rapprochent beaucoup de celles utilisées dans les services de renseignement gouvernementaux. Afin que le CTI soit réellement efficace, il faut couvrir un maximum de sources d'informations possibles (internes et externes). Les informations peuvent être diverses et variées avec des degrés de pertinence très variables. La plupart d'entre elles sont non structurées, redondantes et très volumineuses. Les données proviennent de sources multiples comprenant celles officielles comme des institutions gouvernementales (par exemple le NIST[1] aux états Unis, ou le CERT-FR[2] en France), des institutions privées spécialisée dans la sécurité et la Cyber Sécurité (par exemple Symantec, Fire Eye, Mc Affee, etc.), des communautés organisées à but non lucratif (par exemple la plateforme open source MISP[3]), et d'autres sources diverses comme les réseaux sociaux [1], les échanges inter organisationnels ou inter entreprise, les rapports des chercheurs indépendants (consultants en sécurité), ainsi que des lieux d'activités des cyber criminels (tel que le Dark Web). L'information peut aussi provenir de l'intérieur de l'organisation via des alertes internes diffusées par les employés ou les systèmes de sécurité. La liste est encore longue, mais ce qui rend la tâche de cette collecte et d'analyse d'information encore plus ardue, c'est l'hétérogénéité du type la donnée récoltées ainsi que leur format. En effet, les informations diffusées peuvent concerner des nouvelles menaces de type virus, vulnérabilité, bug, campagne d'hameçonnage (Phishing), information confidentielle volée, liste noire d'adresses IP et de domaines mal intentionnés, traces d'attaques, activités suspectes, etc. et peuvent également avoir différent niveaux de confidentialité.

Le but ultime des systèmes de CTI est de digérer ce flot de données, de l'analyser et de générer des alertes voir des prédictions liées à des dangers imminents voir futur en optimisant le taux de faux positifs. Il doit aussi offrir une vue globale (et graphique) de tous les risques en temps réel et prioritaires les alertes afin de fournir une réponse rapide aux problèmes encourus. Idéalement, cette analyse des menaces doit être prise en charge par des analystes de sécurité qui ont la capacité intellectuelle de trier, analyser et évaluer la pertinence de l'information qu'il reçoit. Sauf que les informations sont tellement importantes que l'automatisation des traitements devient nécessaire afin de couvrir toutes les sources et réduire le risque au minimum. L'automatisation du traitement des données basé sur des règles programmées est possible si celles-ci sont assez homogènes, ce qui n'est pas le cas ici, d'où la nécessité d'une intervention humaine accrue. Une étude récente [2], sur les différentes solutions de CTI disponibles actuellement, montre que d'une part, la plupart des offres ne fournissent qu'une vue partielle des menaces (elle se focalisent sur un type de menace) et que d'autre part elles se contentent de collecter de l'information mais pas de l'analyser. De plus, cette étude révèle que les tâches manuelles nécessitant une intervention humaine sont assez importantes et qu'elles peuvent générer une saturation sur le traitement des données. Plus généralement les solutions CTI sont majoritairement issues d'offres commerciales non collaboratives dont le code source est souvent confidentiel et les fonctionnalités assez rudimentaires. Très peu de solutions issues de la recherche scientifique ont été menées autour des technologies CTI

[1] https://csrc.nist.gov/
[2] https://www.cert.ssi.gouv.fr/
[3] http://www.misp-project.org/

négligeant ainsi l'aspect algorithmique avancé sur le traitement, l'analyse et la structuration des données. L'aspect lié à l'accès à ces données de sécurité et leur niveau de confidentialité vis-à-vis des différent acteurs qui les collectent, les analysent ou les échangent avec des tiers reste aussi négligé par les solutions existantes. Par exemple, lors d'un accord d'échange de traces d'attaques entre organismes, il est très important de garder certaines données internes confidentielles (comme les adresses IP, noms d'utilisateurs, version des serveurs, topologie du réseau, rôles dans l'entreprise, mots de passes, certificats, etc.) vis-à-vis de certaines entités qui pourraient les analyser. Par ailleurs, ces données sont souvent difficiles à être détectées automatiquement en utilisant des filtres traditionnels.

Le département recherche à SAP a développé des solutions basées sur la technologie de l'apprentissage automatique « Machine Learning » (ML) permettant de mieux structurer les données de sécurité hétérogènes et d'identifier et de prédire certaines attaques [3]. D'autres nouveaux projets internes sont en cours et nécessitent plus de travaux de recherche afin de proposer des solutions innovantes et efficaces pour rendre les solutions CTI plus performantes et mieux automatisées. D'autre part, EURECOM dispose déjà d'une grande expertise dans la protection des données provenant de multiples sources, des risques de fuites d'informations confidentielles et de la conception de nouveaux mécanismes d'analyse et de traitement de ces données confidentielles [14].

La thèse proposée s'inscrit dans cette démarche d'innovation et d'enrichissement des travaux de recherche dans le domaine de la CTI. Les solutions qui seront développées vont se baser sur l'exploitation des capacités avancées des nouveaux algorithmes des systèmes d'Intelligence Artificielle qui seront capables de traiter des données massives, hétérogènes et à degré d'accès variable.

2. Etat de l'Art

La thématique du CTI a été appréhendée par le monde de la recherche principalement sur l'aspect de l'échange de la donnée cyber sécurité. Plusieurs études comme [4,5,6] ont d'abord insisté sur la nécessité d'échanger les données de sécurité entre les différents acteurs et sources (définis dans la section précédente). Afin que cet échange soit fait de façon optimale plusieurs standards et structures d'échanges de données de sécurité ont été proposés. On peut citer par exemple Structured Threat Information eXpression (STIX) [7], or Trusted Automated eXchange of Indicator Information (TAXII) [8], Open Incident of Compromise (OpenIOC)[4] , Incident Object Description Exchange Format (IODEF) [9], Cyber Observable eXpression (CybOX) [10], ainsi que d'autres taxonomies [11] proposent même d'intégrer plusieurs standards à la fois. Grace à ces standards, les données de sécurité sont mieux structurées, ainsi plusieurs solutions commerciales les ont intégrées dans leur plateforme. Cependant ces standards n'ont pas proposé de solutions pour transformer la donnée brute en données structurées, ni d'algorithmes d'analyse de ces données complexe.

Paradoxalement, les solutions commerciales de plateformes CTI sont très nombreuses et variées, la plupart d'entre elles n'offrent pas de données variées et complètes, mais plutôt une mise à disposition de l'information collectée par l'outil de sécurité dont il dépend. Par exemple la plateforme CTI

[4] OpenIOC https://www.fireeye.com/blog/threat-research/2013/10/openioc-basics.html

DeepSight[5] propose les données collectées par les anti-virus Symantec. On peut ainsi citer des plateformes plus complètes comme RecordedFuture, Threat Connect (qui propose une plateforme générique à laquelle on peut connecter plusieurs sources externes), Malware Information Sharing Platform (MISP), McAfee Threat Intelligence Exchange, FireEye iSIGHT, RSA NetWitness, Anubis Networks Cyberfeed, etc.

A notre connaissance l'utilisation de l'intelligence artificielle pour l'analyse et la classification de la donnée de sécurité est très peu explorée dans le domaine de la recherche CTI. On peut noter quelques tentatives d'utiliser les algorithmes d'apprentissage automatisés dans la découverte et la classification des vulnérabilités [12] mais le lien avec les plateformes CTI reste très marginal. Une offre commerciale Recorded Future [13] propose d'appliquer les algorithmes de ML pour la prédiction des menaces selon la même méthodologie qu'on a proposé dans [3]. D'autres articles de presse récents insistent sur l'importance d'utiliser l'intelligence artificielle pour la prédiction des menaces. Avec cette thèse nous proposons d'étendre l'utilisation de cette technologie à la classification, l'analyse, le regroupement et la structuration des données de sécurité. Le but est de proposer une couche intermédiaire entre les sources de données et les plateformes de visualisation de haut niveau.

3. Objectifs de la thèse

La thèse aura pour objectif principal de proposer des solutions innovantes basées sur les technologies d'intelligence artificielle afin de pallier au manque de structuration des données dans les systèmes CTI actuels. Cet objectif global devra être bâti autour des objectifs fondateurs suivants :

- Répertorier tous les types de données utilisées dans les plateformes CTI et les classifier (en ontologie éventuellement).
- Identifier les besoins d'automatisation dans les plateformes CTI.
- Identifier les contraintes de confidentialité et de droit d'accès par rapport aux données collectées, échangées et traitées dans les plateformes CTI .
- Comparer et sélectionner les algorithmes de traitement de texte, de catégorisation de contenu et de groupement de données. Proposer des améliorations pour adapter les algorithmes d'intelligence artificielle aux besoins de sécurité et de confidentialité.
- Développer une plateforme de classification, d'analyse, de regroupement et de structuration de données selon les standards existants
- Concevoir des méthodes de protection de données qui d'une part fourniront le niveau de confidentialité souhaité et d'autre part seront adaptées au bon fonctionnement et à l'efficacité des algorithmes d'analyse et de traitement de données (les mécanismes d'apprentissage automatique).
- Étudier la faisabilité des nouvelles méthodes d'aprentissages en terme d'efficacité, de performance et de sécurité sur des données de sécurité réelles.
- Proposer de nouvelles applications pour les CTI.
- Transférer les prototypes obtenus aux unités de production et de développement.

Le prototype final devra atteindre le niveau de maturité TRL7[6] afin qu'il puisse être testé en mode productif au sein de SAP ou des Clients de SAP.

---

[5] DeepSight https://www.symantec.com/content/dam/symantec/docs/data-sheets/deepsight-intelligence-overview-en.pdf

[6] https://fr.wikipedia.org/wiki/Technology_readiness_level

4. Méthodologie de Travail

L'objectif de cette thèse étant de concevoir des solutions d'organisation et de traitement de données de sécurité en utilisant des outils sophistiqués comme l'apprentissage automatique, le candidat devra d'abord faire une étude bibliographique sur les méthodes qui existent et évaluera les principales faiblesses de ces solutions existantes. Grâce à cette étude, l'étudiant pourra identifier les besoins essentiels pour la conception de nouvelles solutions de traitement de données en terme d'efficacité, de performance et de confidentialité. L'étudiant pourra ensuite définir l'architecture de la nouvelle plateforme d'analyse de données en suivant une approche modulaire qui tiendra compte de l'hétérogénéité des types et sources de données et des besoins déjà identifiés. Les outils d'intelligence artificielle utilisées seront basés sur les algorithmes d'apprentissage machine ainsi que d'apprentissage profond (deep learning) comme les réseaux de neurones. L'étudiant étudiera également les différentes solutions de contrôle d'accès et de protection de données comme notamment les techniques d'anonymisation ou de "differential privacy" pour permettre de ne pas divulguer des informations à des entités non autorisées. Une fois la plateforme développée, le candidat débutera ensuite l'évaluation des nouveaux modules en utilisant des données réelles selon un certain nombre de critères (ou de métriques) comme la fréquence d'analyse, le taux des faux positifs, la taille des données, le nombre de sources de données.

Toutes les innovations et les solutions de recherche proposées à chaque étape de la thèse seront publiées dans les conférences internationales et les journaux de recherche. Nous nous proposons de viser une liste non exhaustive de conférences spécialisées dans la cyber sécurité comme USENIX, OWASP, SANS, PETS, ACM CCS, IEEE Symposium on Security and Privacy, ESORICS, etc.

Le travail s'effectuera dans le laboratoire de recherche SAP ainsi qu'à EURECOM. En fonction du type de taches à effectuer par le candidat, celui-ci sera à SAP pour la partie contact avec les utilisateurs, développement, prototypage et évaluation des nouvelles solutions proposées et à EURECOM pour la partie conception de ces nouvelles solutions d'apprentissage dédiées aux données de sécurité, des solutions de protections de données selon les niveaux de confidentialité, la rédaction des papiers et toutes activités scientifiques.

5. Bibliographie

[1] Trabelsi S., Plate H., Abida A., Ben Aoun M., Zouaoui A., Missaoui C., Gharbi S. and Ayari A. Monitoring Software Vulnerabilities through Social Networks Analysis . In Proceedings of the 12th International Conference on Security and Cryptography - Volume 1: SECRYPT, (ICETE 2015) ISBN 978-989-758-117-5, pages 236-242. DOI: 10.5220/0005538602360242

[2] Sauerwein, C., Sillaber, C., Mussmann, A., & Breu, R. (2017). Threat Intelligence Sharing Platforms: An Exploratory Study of Software Vendors and Research Perspectives.

[3] Slim Trabelsi, Skander Ben Mahmoud, and Anis Zouaoui. 2016. Predictive Model for Exploit Kit based Attacks. In Proceedings of the 13th International Joint Conference on e-Business and Telecommunications (ICETE 2016)

[4] Sarah Brown, Joep Gommers, and Oscar Serrano. 2015. From Cyber Security Information Sharing to Threat Management. In Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security (WISCS '15). ACM, New York, NY, USA, 43-49

[5] Gordon, L. A., Loeb, M. P., Lucyshyn, W., & Zhou, L. (2015). The impact of information sharing on cybersecurity underinvestment: a real options perspective. Journal of Accounting and Public Policy, 34(5), 509-519.

[6] Dandurand, L., Serrano, O.: Towards improved cyber security information sharing. In: 5th International Conference on Cyber Conflict (CyCon), pp. 1–16. IEEE, Los Almitos (2013)

[7] Structured Threat Information eXpression (STIX™) 1.x Archive Website:A structured language for cyber threat intelligence https://stixproject.github.io/

[8] Barnum, S. (2012). Standardizing cyber threat intelligence information with the Structured Threat Information eXpression (STIX). MITRE Corporation, 11, 1-22.

[9] Incident Object Description Exchange Format (IODEF) RFC 5070 https://tools.ietf.org/html/rfc5070

[10] Cyber Observable eXpression (CybOX™) Archive Website https://cyboxproject.github.io/

[11] Burger, E. W., Goodman, M. D., Kampanakis, P., & Zhu, K. A. (2014, November). Taxonomy model for cyber threat intelligence information exchange technologies. In Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security (pp. 51-60). ACM.

[12] Yamaguchi, F., Lindner, F., & Rieck, K. (2011, August). Vulnerability extrapolation: assisted discovery of vulnerabilities using machine learning. In Proceedings of the 5th USENIX conference on Offensive technologies (pp. 13-13). USENIX Association.

[13] Staffan Truvé, 4 Ways Machine Learning Is Powering Smarter Threat Intelligence, White Paper https://go.recordedfuture.com/hubfs/white-papers/machine-learning.pdf

[14] C. Van Rompay, R. Molva, M. Önen A leakage abuse attack against multi-user searchable encryption , , PETS 2017, Minneapolis, USA