Contrats doctoraux

en Intelligence Artificielle

Projet AHEAD

Artificial Intelligence for Health, Physics, Transportation and Defense

TITRE DE LA THESE EN FRANÇAIS

Le « Machine Learning » au service de la scénarisation d'attaques complexes à partir de journaux d'évènements

TITRE DE LA THESE EN ANGLAIS

Machine Learning for Complex attack scripting from logs

Co-financement et collaboration

Cette thèse est co-financée par le centre régional du Cnam Grand Est. Dans le cadre de son plan stratégique 2019-2023, ce dernier met la question des Transition(s), notamment numérique, au cœur de ses préoccupations d'investissements et d'actions. L'accompagnement à la transformation des organisations, en particulier dans le champ des vulnérabilités des systèmes d'information de celles-ci est une ligne d'action explicite. Hébergeur du LabS'DN, à l'initiative de nombreuses actions sur l'entrepreneuriat dans le contexte de l'économie de proximité et des TPE/PME-PMI, le centre montre une forte capacité d'innovation. Il a été récemment retenu pour porter un projet d'enseignement innovant de cybersécurité, totalement à distance. L'originalité de ce projet vise l'enseignement de la connaissance technique cyber de pointe et en particulier les capacités de remédiation en situation adverse. La plateforme propose pour cette raison la capacité de confronter des équipes attaquantes (« red team ») et de défense (« blue team »). L'un des points crucial pour le succès de ce projet sera la capacité de construire et diversifier des scenarii pédagogiques attaque/défense à des fins d'entrainement et d'enseignement. Les contributions de cette thèse visent à alimenter cette future plateforme.

Le Centre régional Grand Est collaborera dans le cadre de cette thèse avec les deux laboratoires du Cnam : CEDRIC (Centre d'étude et de recherche en informatique et communication) et SD (Laboratoire Sécurité Défense).

Contexte de la thèse

L'investigation numérique consiste à reconstituer le scenario d'attaque suite à un incident de sécurité informatique. Cette reconstitution s'effectue, par tâtonnement, au travers d'une analyse des traces émises dans les journaux d'évènements (logs) ; ce qui induit un traitement lourd compte tenu du volume, de la variété et de la variabilité des données ainsi que de leur forte dépendance du contexte. Elle est également non formalisée. Elle produit une chaine d'hypothèses beaucoup trop dense, un taux important d'explications erronées et une quantité importante de faux-positifs et faux-négatifs qui ne permet pas de caractériser correctement les incidents, les vulnérabilités ou les comportements malveillants. En outre, la quantité de donnée introduit une lenteur dans la démarche voire même jusqu'à rendre inopérantes les opérations d'un analyste non expérimenté. De même, l'analyste ne peut pas s'appuyer

totalement sur les outils d'aide à l'analyse de sécurité à base de Machine Learning (ML), du fait de leur incapacité à prendre en compte/combiner le contexte et les expertises des analystes.

La littérature propose une pléthore d'algorithmes à base de ML, qui se montre efficace pour la détection d'incidents de sécurité [HG,2019],[RAAH,2019],[CV,2009]. À notre connaissance, il n'existe pas de travaux de recherche proposant un choix d'algorithmes adapté à l'incident analysé tout en prenant en compte le contexte. La plupart des travaux de recherche en ML traitant des logs hétérogènes s'appuie sur des algorithmes prenant la structure des logs (syntaxe) ou les mots présents dans les logs tout en ignorant leur sens (sémantique) [AI,2017]. Cependant, ces algorithmes à base de ML sont des bons candidats pour guider l'humain lors de son investigation numérique dans les logs. Ils sont en effet efficaces pour explorer des bases de données massives. Ils sont en mesure d'apporter de nouvelles capacités d'analyse à condition qu'on les alimente de connaissances adéquates.

Objectif de la thèse

L'objectif de la thèse est de fournir un système d'aide à la décision fondé sur le Machine Learning. Ce système servira à l'explication d'incidents de sécurité en s'appuyant sur les algorithmes à base de ML qui exploiteraient les connaissances du domaine tel que les arbres d'attaque cachés dans les bases de connaissances CVE et qui prendraient en compte le contexte de l'entreprise.

Candidature

Le candidat à cette thèse doit avoir un Master 2 en informatique (ou équivalent). Il doit avoir des compétences dans au moins 2 des domaines suivants :

- Représentation des connaissances/modélisation conceptuelle des systèmes d'information
- Cybersécurité
- Machine Learning

Une expertise en R&D dans l'un de ces domaines seront un plus. Des compétences en programmation sont exigées.

Envoyer un CV et une lettre de motivation à veronique.legrand@lecnam.net, ilham.lammari@lecnam.net, jean-claude.bouly@lecnam.net

Bibliographie

[AI,2017] Arnaldo I., Cuesta-Infante A., Arun A., Lam M., Bassias C., Veeramachaneni K. (2017) Learning Representations for Log Data in Cybersecurity. In: Dolev S., Lodha S. (eds) Cyber Security Cryptography and Machine Learning. CSCML 2017. Lecture Notes in Computer Science, vol 10332. Springer, Cham

[HG,2019] Hermann, Geoffroy. « IA et cybersécurité : une boucle émergente de rétroactions », Revue Défense Nationale, vol. 821, no. 6, 2019, pp. 131-137.

[CV,2009] Varun Chandola, Aindam Banerjee, Vipin Kumar, « Anomaly detection: a survey. » ACM COMPUT. SURV., 2009, pp 15: 1-15: 58.

[RAAH,2019] Riyaz Ahamed Ariyaluran Habeeb, Fariza Nasaruddin, Abdullah Gani, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, Muhammad Imran, « Real-time big data processing for anomaly detection: A Survey », International Journal of Information Management, Volume 45, 2019, pp 289-307, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2018.08.006