# Sujet de thèse : Apprentissage de représentation de graphes pour la cybersécurité

Graph representation learning for cybersecurity

Direction :
- Directeur de thèse : Jamal Atif, Pr,PSL Université Paris Dauphine - LAMSADE (jamal.atif@dauphine.fr)
- Co-encadrant : Florian Yger, Mcf, PSL Université Paris Dauphine - LAMSADE (florian.yger@dauphine.fr)
- Co-encadrant : Fabrice Rossi, Pr, PSL Université Paris Dauphine - CEREMADE (rossi@ceremade.dauphine.f

## Contexte

Le domaine de la cybersécurité utilise abondamment les graphes dans nombre de ses applications. En effet, ce type d'objet peut

- décrire l'exécution et le fonctionnement d'un programme informatique, notamment grâce aux graphes de flot de contrôle ou graphe d'appel, avec les sous-ensembles des paquets standards utilisés.
- permettre de modéliser les interactions entre entités voire même la topologie d'un réseau.

Les applications qui en découlent vont de la découverte d'intrusions dans un réseau à la détection de nouveaux malwares. Ces applications ont en commun de mobiliser des graphes et des données structurées qui sont inadaptées à une utilisation directe des méthodes standards d'apprentissage automatique.

## Problématique

Jusqu'à récemment, le domaine de l'apprentissage automatique s'était surtout concentré sur le développement d'approches dédiées aux données euclidiennes. Mais ces progrès sont maintenant graduellement transférés à des données complexes, structurées [6] et plus généralement non-euclidiennes [3]. Les géométries mises en œuvre pour manipuler ces données impliquent des notions d'invariance complexes et souvent coûteuses en temps de calcul. Dans ce contexte, le choix de la représentation de telles données est fondamental. Une représentation bien choisie permettra ainsi d'induire les invariances nécessaires tout en permettant l'application d'algorithmes d'apprentissage classiques.

Par exemple, comparer deux graphes revient à trouver les structures équivalentes entre eux et peut se formuler de nombreuses façons (isomorphisme de graphe ou de sous-graphes, appariemment quadratique, chemin d'édition minimal,...). Dans tous les cas, un graphe subissant une permutation de ses sommets ne verra pas sa structure modifiée et devra être considéré comme similaire à sa version non permutée. Toute opération de comparaison de graphes devra donc être invariante aux permutations.
Ainsi, s'il existe de nombreuses façons de représenter un graphe (matrices d'adjacence, matrice d'incidence,...), rares sont celles qui induisent une invariance aux permutations. Il ne sera donc pas possible par exemple de comparer de manière pertinente deux graphes directement au moyen de leurs matrices d'adjacence. L'enjeu est donc de trouver et dans l'idéal d'apprendre une représentation pertinente de ces données. Ce sujet de thèse se situe à l'intersection de l'apprentissage automatique, des graphes et de la cybersécurité.

Dans un premier temps, nous nous concentrerons sur une application de détection de malwares. Dans ce cadre, nous pourrons étudier le jeu de données MalNet [5], récemment proposé pour lutter contre la dissémination des malwares sur Android. Ce jeu de données comporte 1.2 millions graphes d'appel de fonction extraits de malwares Android. Ces derniers se hiérarchisent en 47 classes et peuvent être utilisés pour déterminer la provenance d'autres graphes d'appels de fonctions. Comme illustré dans Figure 1, un graphe d'appel de fonction permet de représenter les appels possibles entre les sous-routines d'un programme. Les malwares sont par nature polymorphiques : ils tentent d'échapper aux outils de détection en opérant des modifications de leur code source. C'est pourquoi il est important d'en trouver des représentations plus robustes aux manipulations, comme celle du graphe d'appel de fonction. Tout l'enjeu est alors d'appliquer une démarche d'apprentissage artificiel sur de telles données structurées pour lesquelles les modèles classiques n'ont pas été pensés. À cela s'ajoute la difficulté du passage à l'échelle dans le cas du jeu de données MalNet.

Dans le cadre de cette thèse, nous nous proposons d'étudier la représentation de graphes [4] dans le cadre de la cybersécurité en privilégiant des approches d'apprentissage de métrique [1] et de plongement. Ces pistes seront à mettre en lien avec les approches à noyaux [7, 2] et les réseaux de neurones [9, 6, 8] proposés récemment dans la littérature.

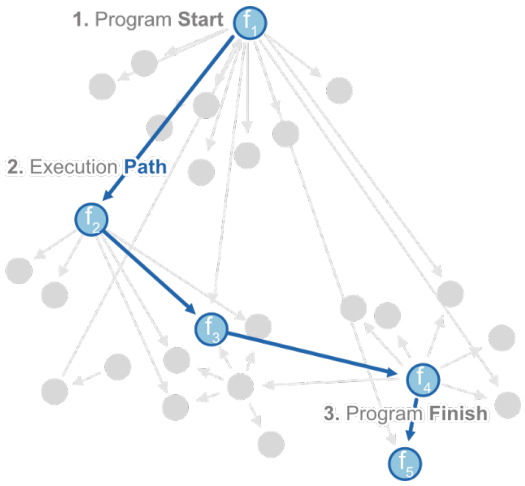


FIG. 1: Graphe d'appel de fonction (extrait de [5]). En gris apparaissent les appels entre sous-routines d'un programme et le chemin en bleu est une exécution possible du programme.

## Profil recherché

Nous recherchons des candidats et candidates très motivés, titulaires d'un excellent diplôme (master degree, école d'ingénieurs) dans les domaines des mathématiques appliquées ou de l'informatique. Les candidats et candidates doivent pouvoir démontrer d'un solide bagage en apprentissage artificiel et en théorie des graphes et doivent être à l'aise avec le développement informatique.

## Processus de sélection

Toute candidature doit comporter un CV détaillé, une lettre de motivation, les relevés de notes d'une ou deux années antérieures et un résumé du mémoire de master.

## Références

[1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A Survey on Metric Learning for Feature Vectors and Structured Data. 2013.

[2] Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray, and Bastian Rieck. Graph kernels : State-of-the-art and future challenges. *arXiv preprint arXiv :2011.03854*, 2020.

[3] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning : going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4) :18–42, 2017.

[4] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding : Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, 2018.

[5] Scott Freitas, Yuxiao Dong, Joshua Neil, and Duen Horng Chau. A large-scale database for graph representation learning. *NeurIPS Datasets and Benchmarks Track*, 2020.

[6] Martin Grohe. word2vec, node2vec, graph2vec, x2vec : Towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 1–16, 2020.

[7] Nils M Kriege, Fredrik D Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 2020.

[8] Andreas Loukas. How hard is to distinguish graphs with graph neural networks ? In *NeurIPS*, 2020.

[9] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Networks Learn. Syst.*, 2020.