Offre de thèse/Thesis Offer

Title: Robustness and privacy of graph neural networks: homomorphic encryption and randomization

Keywords: Graph Neural Networks, Adversarial attacks, Graph classification, Robustness, Randomized algorithms, Functional encryption

Publication Date: 15 March 2019

Beginning Date: October-November 2019

Duration: 36 months

Contact: cedric.gouy-pailler@cea.fr and renaud.sirdey@cea.fr

Context

In various domains, graphs represent a useful representation for many types of data. Prominent examples entail behavioural analyses performed in cybersecurity or social network analysis. In the former, user internet behaviour can be observed by monitoring DNS requests, interpreted as successive steps of a random walker on a graph in which nodes represent domain names and edges represent population-level average behaviour. Therefore studying user behaviour can be done by analyzing the subgraph induced by specific user movements. In the latter, graph representations naturally emerge from user interactions. For example nodes can represent users, and any relation between two users (messages or common interests) can be interpreted as edges. Understanding and analyzing graph structures appear to be a key tool in many real-world applications. It is thus essential to find efficient and robust methods for tasks such as node or graph classification.

In the last decade, deep neural networks have reached an outstanding level of accuracy in numerous areas, such as image classification [KSH12] or object detection [RHGS15]. These models have also recently been formulated in the context of graph-structured data, and now play an important role in node classification and graph classification problems [SGT+09, DDS16, KW16, HYL17, WPC+19]. Such formulations are currently explored in many domains, jointly exploiting classical features, as well as graph-structured information. Successful applications have been developed in physics, in which graph neural networks are able to predict physical properties of molecules based on their molecular graphs [CBG+17]. Graph neural networks also have convincing applications in material sciences [XG18], structural forecasting [LYSL17], natural language processing [MT17], or communication optimization in multi-agent systems [SSF16].

Despite these powerful representation properties, recent issues have been demonstrated in deep learning-based approaches regarding their robustness to adversarial attacks (small perturbations of the input) and the resulting training data privacy (due to overfitting and over-parameterization posing threats on data privacy). Adversarial attacks [SZS+13] are small perturbations of an input that fools the results of classification for a network. Adversarial attacks raise questions of security and safety, and also responsibility in terms of law. Adversarial examples attacks against machine learning models have become a burning issue due to their efficiency, and the number of sensitive domains they could affect. Accordingly, both attacks and defenses are developed in a tight back-and-forth [GSS14, PMJ+16, PMG+17, DLT+18, SKC18]. Recently, the idea of using randomization in the learning process to ensure robustness against adversarial examples attacks have been successfully used [XWZ+17, MDST18, LCZH17, PMA+19]. In the context of graph neural networks, these issues are of primary importance for various reasons: first graph-structured data often bear much more information than classical tabular data. For example in social networks, tabular based approaches classicaly summarize neighbourhood information in a few variables (number of friends/degree, node/edge betweenness, ...), while the topological information structurally bear much more sensitive information about individuals. Second the complexity of information present in graphs makes the graph neural network approaches much more sensitive to attacks. Basically this comes from the fact that small perturbations in graph data consist in adding/deleting nodes, or modifying edges weight. While adversarial modifications in images and sounds might be globally noticeable, graph-based perturbations will be hard to detect. Coincidentally with the fact that graph neural network approaches have shown superior results in various domains, their robustness have been investigated and attacks have been developed in the context of node classification and graph classification [ZAG18, DLT+18, SWYL18].

In this context it is of primary importance to develop innovative approaches to ensure privacy and robustness of graph neural networks. Among possible approaches, lightweight approaches such as randomization will have to be adapted to these techniques. Depending on the criticity of the stage and needed privacy of data and models (learning or inference phases), randomization techniques should be complemented by homomorphic encryption approaches. With respect to privacy, a number of works have started to investigate how techniques for computing over encrypted data such as homomorphic cryptography (FHE) can be applied to the inference phase of deep neural network models with encouraging results when a clear-domain network is evaluated over an encrypted-domain input [BMMP18, CLM+19, CdWM+17, DGBL+16]. Yet there are a number of practically interesting extensions most notably with respect to GNN regarding specific optimizations that may render them more amenable to better FHE-execution performances. Also investigating the relevance and practicality of using these techniques during the learning phase of such models is of high practical interest. On top of privacy, the connection between cryptographic theory and techniques and counter-measures against the aforementioned adversarial attacks is an another important research topic which can be considered as part of this PhD subject. The goal of this thesis is to explore robustness and privacy of graph-neural network-based approaches, by considering solutions combining randomization and homomorphic encryption to ensure a satisfying compromise between performance, robustness and data privacy.

CEA background in these fields

CEA LIST has been a key leader in fully homomorphic encryption techniques `https://github.com/CEA-LIST/Cingulata`. In the context of FHE, machine learning applications appear as a killer application. Many key advances have yet to be considered to fully address machine learning applications using FHE technologies. Next technological barriers depend on the computational cost of the considered stage (training or inference) but the main approaches are: first to limit operators used in graph neural networks such that FHE associated computational cost is kept reasonable. Second FHE can be viewed as a building block, which could be activated in specific parts of the pipeline to ensure model or data privacy.

CEA LIST is also very active in the field of randomization algorithms to ensure data privacy and robustness to adversarial attacks. Past works include PhD thesis of Anne Morvan and Rafael Pinot.

Expected work

- Experimental study of state-of-the-art attacks.
- Theoretical approach of defenses.
- Corresponding implementation, and experimentation of defenses.

Required profile

- You are currently in the final year of engineering school or in M2 at the university with specialization in computer science and/or statistics.
- You have a strong background in applied mathematics/computer science (probability, statistics, graph theory).
- You have good programming skills (Python/R, torch/tensorflow, C++).
- Academic research interests you, but also applications to concrete problems.

Conditions
The doctoral will take place at CEA LIST in Saclay, where you will work with Renaud Sirdey and Cédric Gouy-Pailler. Access to the CEA is based on a daily bus network service covering the entire Paris region (several buses from Paris in particular).

References

[BMMP18] Florian Bourse, Michele Minelli, Matthias Minihold, and Pascal Paillier. Fast homomorphic evaluation of deep discretized neural networks. In

Hovav Shacham and Alexandra Boldyreva, editors, *Advances in Cryptology – CRYPTO 2018*, pages 483–512, Cham, 2018. Springer International Publishing.

[CBG+17] Connor W. Coley, Regina Barzilay, William H. Green, Tommi S. Jaakkola, and Klavs F. Jensen. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *Journal of Chemical Information and Modeling*, 57(8):1757–1772, August 2017.

[CdWM+17] Hervé Chabanne, Amaury de Wargny, Jonathan Milgram, Constance Morel, and Emmanuel Prouff. Privacy-preserving classification on deep neural network. *IACR Cryptology ePrint Archive*, 2017:35, 2017.

[CLM+19] Hervé Chabanne, Roch Lescuyer, Jonathan Milgram, Constance Morel, and Emmanuel Prouff. Recognition over encrypted faces. In Éric Renault, Selma Boumerdassi, and Samia Bouzefrane, editors, *Mobile, Secure, and Programmable Networking*, pages 174–191, Cham, 2019. Springer International Publishing.

[DDS16] Hanjun Dai, Bo Dai, and Le Song. Discriminative embeddings of latent variable models for structured data. *CoRR*, abs/1603.05629, 2016.

[DGBL+16] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 201–210. JMLR.org, 2016.

[DLT+18] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. 2018.

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[HYL17] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *CoRR*, abs/1706.02216, 2017.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[KW16] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[LCZH17] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. *CoRR*, abs/1712.00673, 2017.

[LYSL17] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. *arXiv:1707.01926 [cs, stat]*, July 2017.

[MDST18] Seyed-Mohsen Moosavi-Dezfooli, Ashish Shrivastava, and Oncel Tuzel. Divide, denoise, and defend against adversarial attacks. *CoRR*, abs/1802.06806, 2018.

[MT17] Diego Marcheggiani and Ivan Titov. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. *arXiv:1703.04826 [cs]*, March 2017.

[PMA+19] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization: The case of the Exponential family. *arXiv:1902.01148 [cs, stat]*, February 2019.

[PMG+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 506–519, New York, NY, USA, 2017. ACM.

[PMJ+16] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[SGT+09] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[SKC18] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *ArXiv e-prints*, May 2018.

[SSF16] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning Multiagent Communication with Backpropagation. *arXiv:1605.07736 [cs]*, May 2016.

[SWYL18] Lichao Sun, Ji Wang, Philip S. Yu, and Bo Li. Adversarial Attack and Defense on Graph Data: A Survey. *arXiv:1812.10528 [cs]*, December 2018.

[SZS+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[WPC+19] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A Comprehensive Survey on Graph Neural Networks. *arXiv:1901.00596 [cs, stat]*, January 2019.

[XG18] Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14):145301, April 2018.

[XWZ+17] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. *CoRR*, abs/1711.01991, 2017.

[ZAG18] D. Zügner, A. Akbarnejad, and S. Günnemann. Adversarial Attacks on Neural Networks for Graph Data. *ArXiv e-prints*, May 2018.