M2 Internship / PhD proposal

Generalization of Graph Neural Networks on Large Graphs

Graph Neural Network (GNN) [3] are state-of-the-art deep models that can perform a wide range of **Graph Machine Learning** (Graph ML) task on graph data, with many applications in chemistry, biology, or recommender systems, to name a few. Most GNN architectures are built by stacking layers of functions that perform *Message-Passing* (MP) along the edges of the graph, where each node receive messages from its neighbors and aggregate them to progressively build discriminative node representations.

As a core notion in ML, the **generalization** – the hability to generalize predictions from training data to test data – of GNNs is naturally an active area of research. In particular, for "large" graphs, moving away from combinatorial analyses toward statistical models of random graphs becomes necessary [1, 2]. Nevertheless, compared to actual applications of GNNs in Graph ML, many crucial components of GNNs (and Graph ML tasks in general) are relatively ignored by current analyses. In particular:

- *node features* are often ignored, as traditional models of random graphs focus on graph structures. Experiments on real data show however that node features are often as important as graph structure, with non-trivial interaction between the two;
- analyses are often restricted to *graph tasks*, classification or regression, where samples – whole graphs – are independent and identically distributed according to some model [2]. This allows to deploy the usual ML computations: Rademacher complexities, concentration of the empirical risk, and so on. However, on large graphs, most tasks are *node-tasks*, where classical ML notions need to be revisited: nodes are not independent, training and test nodes may live in the same graph, and so on.

In this internship, which may be potentially followed by a PhD thesis, we will explore new ways to compute generalization bounds for GNNs, by progressively incorporating new key elements into existing analyses. While the core of the proposal is theoretical, experiments on real and synthetica graphs will also be performed, depending on the candidate.

Infos. Location: IRISA, Rennes, France. Starting date in 2025. Funded by the MALAGA ERC Starting Grant: `https://nkeriven.github.io/malaga`

Contact. `nicolas.keriven@cnrs.fr`

References

[1] Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. Convergence and stability of graph convolutional networks on large random graphs. In *Advances in Neural Information and Processing Systems (NeurIPS)*, pages 1–26, 2020.

[2] Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. *Advances in Neural Information Processing Systems*, 35:1–63, 2022.

[3] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2020.