Research Project

Using the graph structure to survive in high-dimensions: Graph arrangements, walks, and embeddings

Theme: Mathematics for Artificial Intelligence

Contact: Argyris Kalogeratos <argyris.kalogeratos@ens-paris-saclay.fr>

Responsible: A. Kalogeratos | Permanent researcher, co-ordinating the theme ML on Graphs

Institution: Centre Borelli, ENS Paris-Saclay – MLMDA research group (dir. Nicolas Vayatis)

Keywords: machine learning on graphs, combinatorial optimization, linear arrangement, cluster analysis, embeddings.

Context and motivation

The tremendous increase of information volume and variety has undoubtedly introduced a lot of challenges in managing and making sense out of such data by modern systems. Powerful computers, usually arranged in computational clusters, have helped to handle and processing the sheer volumes of data. Machine Learning and Artificial Intelligence is being clearly benefited from the advancements being brought by computer hardware and systems architecture to scale-up their solutions, with Deep Learning to be the warhorse of this story. At the same time, the availability of powerful computational resources allowed researchers to revisit problems that are considered small or typical with respect to their size. The aim is to evaluate how much has the increase in computational power helped us in solving better certain “difficult” problems. Not surprisingly, most of the time, what we see is that the existing methodological knowledge does not translate the orders of magnitude stronger computational resources to orders of magnitude better solutions. This is to remind that, what will always remain to be solved is the difficult mathematical/logical/algorithmic problems themselves that are in many cases naturally complex.

Graph-based modeling has proved to be a useful approach to deal with this complexity. Many aspects of the data can be encoded as a network of dependencies between different data objects or features, essentially as edges connecting the nodes of a graph. In some cases, this structure is inherent to the problem definition, as in social and transportation networks, while in others a graph topology can be inferred from observed data. The fact that graphs have been studied broadly in mathematics and computer science, provides rich theoretical results and effective algorithms to properly exploit them. Machine Learning on Graphs is a branch aiming to apply ML in a networked environment that can be represented by a graph model, as well as to exploit ML-oriented and graph-theoretic tools. Major recent advances include
Graph Neural Networks, Graph Embeddings (e.g. node-to-vec, deep-walk, etc.), or predicting and controlling interaction processes such as the spread of epidemics. This trend also fuels the hope that, using such tools, we will be able to attack more efficiently well-known hard combinatorial optimization problems (e.g. graph coloring, graph cut, etc.).

This research project is centered around the notion of graph linear arrangement. Let a graph be denoted by $G = \{V, E\}$, with $V$ being the set of $N$ vertices and $E \subseteq V \times V$ the set of edges. A linear arrangement is a bijection $\ell(G) : V \rightarrow 1,...,N$ that orders the nodes of a graph in a way that optimizes a considered objective function. What is measurable across a linear arrangement is the cost of splitting the graph in two parts, namely the left part and the right part, to a chosen cut position $p$. This cost is equal to the graph cut, i.e. the number of edges passing over that position connecting the two parts. There are several objective functions that one could optimize over an arrangement: typically one could be interested in finding the linear arrangement $\ell$ that has as small as possible max-cut (MaxLA), or as small as possible sum of cuts across the node arrangement (MinLA). Computationally, these are combinatorial optimization problems that are NP-hard for arbitrary graphs. However, there are useful approximations to work with, e.g. for MinLA using the Fiedler’s vector, which is the second smallest eigenvector of the Laplacian matrix $L = D - A$; here $A$ is the adjacency matrix of the graph and $D$ is the diagonal matrix containing the node degrees. The very interesting to observe here is that: what turns out to be a good approximation to the MinLA problem (i.e. the most important eigenvector of the Laplacian matrix) is at the same time a node embedding, which is a core tool originating from spectral graph theory, and is largely related to the graph cuts and spectral clustering. Conversely, this clearly suggest that the linear arrangement problem could offer a new viewpoint to study ML tasks related to the graph structure.

**Scientific objectives**

Having the previous points in mind, this project puts forward the following directions of work:

- Studying and developing efficient and tractable solvers for the graph linear arrangement problem, which is something the current literature lacks of.
- Devising greedy or randomized greedy approximations of the linear arrangement problem (or partial arrangements) can define novel ways of traversing through or walking around a graph. Note that random walks are widely-used for the extraction of node features in graphs, and consequently the computation of a node embedding in which the similarity between pairs of nodes can then be inferred. In that sense, the graph walks we aim to define based on linear arrangements have the potential to be plugged in such node feature extraction procedures.
There is a number of applications that can be tackled along this line of research. Examples are: data embedding, data clustering, graph exploration and visualization, graph signal processing; furthermore, there are certain resource allocation problems for network diffusion control, where gains can also be achieved via good quality graph linear arrangement solutions.

The project is essentially ambitious and can be transversal to a number of fields in mathematics and theoretical computer science. After establishing the theoretical framework, the expected output is the production of a number of articles aiming to high-quality conferences and journals. In addition, this research will be put in relation to the AI Chair of Centre Borelli, ENS Paris-Saclay (resp. Mathilde Mougeot), which could offer opportunities to apply our methodologies to real industrial application in collaboration with its partners.

**Indicative references**


