Learning graphs from observed graph signals

Description: Graph Signal Processing (GSP) is a new and emerging field at the intersection of Signal Processing, Graph Theory, and Machine Learning. GSP manifestates the generalization of standard Signal Processing tools, for example sampling, filtering, recovery, to signals recorded in complex environments. Such an environment comprises of multiple entities whose interrelations, or interactions, can be encoded in a graph and specifically in the links between its nodes. In more formal terms, a graph signal is a function defined on the nodes of a graph and can be represented as a vector with one component per graph node. In order to enjoy the promised benefits of GSP methods, the knowledge of the underlying graph is needed, which is a strong requirement for many real-world problems where the graph may be little or not at all known.

In this work we are going to study the *data-driven Graph Learning* problem where the objective is to use Machine Learning techniques to infer the underlying graph based on observed graph signals. In nature, this is an ill-posed problem since many graphs may be able to explain equally well the data. Therefore, the challenge is to introduce the right assumptions regarding the graph signals, the graph, and the interrelation between the two in order to solve tractable optimization problems to reach meaningful solutions. A survey of existing works is part of the mission of the project, as for target applications, those include physiological data such as fMRI, epidemiological data, or complex data from several other sources.

Topic keywords: graph signal processesing, graph theory, sparse coding, graph inference

Indicative references:

[1] Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., and Vandergheynst, P. (2013). “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains,” Signal Processing Magazine, vol. 30, no. 3, pp. 83–98.

[2] Dong, X., Thanou, D., Frossard, P., and Vandergheynst, P. (2016). “Learning Laplacian matrix in smooth graph signal representations,” Trans. Signal Processing, vol. 64, no. 23, pp. 6160–6173.

[3] Kalofolias, V., “How to learn a graph from smooth signals,” in Proc. of the conf. on Artificial Intelligence and Statistics, 2016, pp. 920–929.

[4] Dong, X., Thanou, D., and Frossard, P. (2018). "Learning Graphs from Data: A Signal Representation Perspective", arxiv preprint.

[5] Le Bars, B., Humbert, P., Oudre, L., amd Kalogeratos, A. (2019). “Learning Laplacian Matrix from Bandlimited Graph Signals”, IEEE International Conference on Acoustics, Speech, and Signal Processing.

Exploiting graph structure in diffusion control strategies

Description: The Dynamic Resource Allocation (DRA) has been proposed as a framework on which control strategies can be developed aiming to dynamically suppress a Susceptible-Infected-Susceptible (SIS) diffusion process [1,2]. The considered SIS process is a continuous-time Markov process that allows node recoveries and (re-)infections in a stochastic setting. The DRA strategy acts at a micro-level deciding exactly which nodes should receive the treatment resources. In the related work there has been proposed to develop score-based strategies where a deterministic criticality score is computed locally for each node providing a node ranking for deciding where to allocate the treatments, like LRIE [1] and Priority-Planning [2]. Roughly speaking, the first can be seen as a local approximation of the second approach. As such, it doesn't have a long term plan but it can be more appealing for cases where the network is partially known, and also adapt to changes in the environment (e.g. network changes).

The directions of work in this project can be:

- extending the modeling side behind LRIE by considering more general SIS-like epidemic models (i.e. allowing reinfections). Examples are models with intermediate incubation states or competitive scenarios.
- designing better performing greedy strategies that may combine structural properties of the network and compare that with meta-population models from the optimal control literature (i.e. considering a k-cluster structure at the top level, and a mixing model inside each cluster).
- go beyond the deterministic local scoring by incorporating Monte Carlo approximations of network evolution that can then be used as refined criticality scores. This can also help in cases of partial knowledge of the infections states of nodes.

The work will be based on existing work, therefore the students will need to review existing material, use code repositories, etc., and will be asked to generalize theoretically the methods to the aforementioned cases, run simulations, and finally highlight insights about the problem.

Topic keywords: epidemics, social interactions and behavior, diffusion control

Indicative references:

[1] Scaman, K., Kalogeratos, A., and Vayatis, N. (2015). "A Greedy Approach for Dynamic Control of Diffusion Processes in Networks". Proceedings of International Conference on Tools with Artificial Intelligence.
[2] Scaman, K., Kalogeratos, A., and Vayatis, N. (2016). "Suppressing Epidemics in Networks using Priority Planning". IEEE Transactions on Network Science and Engineering.
[3] Fekom M., Vayatis, N., and Kalogeratos, A. (2019). "Sequential Dynamic Resource Allocation for Epidemic Control". International Conference on Decision and Control.

Homogeneity testing in high dimensions with applications to graph sparsification and data clustering

Description: Testing whether a sample of observations is statistically homogeneous, is fundamental for assessing the complexity of the underlying distribution. Moreover, it can help in determining the model complexity (model selection) in machine learning applications. Nevertheless, it is unfortunate that, as most statistical hypothesis testing methods, homogeneity testing also suffers from being only effective in very few dimensions.

In this work we first aim to study hypothesis meta-tests, such as the one presented in [1], that devise ways to define and test a hypothesis over the result of multiple univariate unimodality tests [5]. The main idea is based on analyzing multiple histograms of pairwise similarities and then decide whether a high dimensional cloud of points forms one (null hypothesis, H_0) or more (alternative hypothesis, H_a).

We are specifically interested to measure the statistical power of such approaches, their scalability in the size of data, their adequacy to get easily recomputed (to get an updated result) when only small changes have taken place to the sample. Moreover, it is also interesting to try incorporating approaches related to histogram segmentation [3], k-modality testing, or other projection-based preprocessing. In terms of applications, we are planning to use such tests for kernel sparsification and/or data clustering.

Topic keywords: epidemics, social interactions and behavior, diffusion control

Indicative references:

[1] Kalogeratos, A. and Likas, A. (2012). "Dip-means: an incremental clustering method for estimating the number of clusters". NIPS.
[2] Tsapanos, N., Anastasios, T., Nikolaidis, N., and Pitas, I. (2015). "A distributed framework for trimmed kernel k-means clustering", Pattern recognition.
[3] Delon, J., Desolneux, A., Lisani, J.-L., and Petro, A.-B. (2007). "A non parametric approach for histogram segmentation". PAMI.
[4] Daskalakis, C., Diakonikolas, I., Servedio, R.A., Valiant, V., and Valiant, P. (2011). "Testing k-Modal Distributions: Optimal Algorithms via Reductions". arXiv preprint.
[5] Hartigan, J.A. and Hartigan, P. M. (1985). "The dip test of unimodality", Annals of Statistics.
[6] Siffer, A., Fouque, P.-A., Termier, A., and Largouët C. (2018). "Folding test. Are your data gathered?", KDD2018.
[7] Chronis P., Athanasiou, S., and Skiadopoulos, S. (2019). "Automatic clustering by detecting significant density dips in multiple dimensions", ICDM.

Contact:

Argyris Kalogeratos, kalogeratos@cmla.ens-cachan.fr (CMLA, ENS Paris-Saclay)