Big Data
National Research Programme

www.nrp75.ch
Wildhainweg 3, P.O. Box 8232, CH-3001 Berne

Swiss National Science Foundation
Programmes division
National Research Programmes
+41 (0)31 308 22 22
nrp75@snf.ch

# Project description for pre-proposal NRP "Big Data"

*The project description must fulfil the following criteria for successful submission:*

- *The project description is to be submitted in English, unless it can be shown that either German or French is intrinsically better suited to the research topic;*
- *Pre-Proposals must not exceed six pages, including the cover page;*
- *Pre-Proposals must be submitted using this form through mySNF (deadline: **13 January 2016**);*
- *Applicants must follow the structure given in this form and the instructions in the call.*

| | |
|---|---|
| **Responsible applicant**<br>Last name, first name | Vandergheynst, Pierre |
| **Further applicant(s)**<br>Last name, first name, institution | |
| **Project title** | New Clustering Approaches for the Big Data Era |

Please indicate to which research module your project belongs:

☒ **Moduel 1:** Computing and Information Technology

□ **Module 2:** Societal, Regulatory, and Educational Challenges

□ **Module 3:** Applications

Please list five publications from third parties (not yours) considered relevant as stepping stones for the proposed project:
1. Von Luxburg, U. A tutorial on spectral clustering. Stat. Comput. 17, 4 (2007), 395–416.
2. Shi, J. and Malik, J. Normalized Cuts and Image Segmentation, IEEE Trans. on PAMI, 22, 8 (2000).
3. Dhillon, I., Yuqiang G. and Kulis, B. Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. IEEE Trans on PAMI 29 (11): 1–14
4. Kannan, R., Vempala, S. and Vetta, A. On Clusterings : Good. Bad and Spectral. Journal of the ACM 51
5. Ng, A., Jordan, M. and Weiss, Y. On spectral clustering: Analysis and an algorithm. NIPS (2001)

Please list the most important and relevant publications of your team (not more than five):
1. Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. IEEE Signal Processing Magazine 30, 3 (2013), 83–98.
2. Tremblay, N., Puy, G., Borgnat, P., Gribonval, R., and Vandergheynst, P. Accelerated Spectral Clustering Using Graph Filtering of Random Signals. In 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016) (2016)
3.Puy, G., Tremblay, N., Gribonval, R., and Vandergheynst, P. Random sampling of bandlimited signals on graphs, http://arxiv.org/abs/1511.05118
4. Hammond, D.K., Vandergheynst, P. and  Gribonval,R. Wavelets on graphs via spectral graph theory, in Applied and Computational Harmonic Analysis, vol. 30, num. 2, p. 129-150, 2011

# 1 Research topic and objective of the project

Big Data has become an ubiquitous keyword in technology, describing the overwhelming availability of digital data in almost every field of human endeavour in Science, Technology and the Humanities. The use of Big Data is no more a promise: it already had a transformative action on industry, one whose pace is increasing under the action of new enablers such as digitisation or the Internet of Things.

A key technique in data analytics is clustering, the process of finding groups of similar entities within data. Clustering allows one to discover patterns but also to reduce dimensionality by coarsening data sets down to few relevant classes. One of the most used algorithm is Spectral Clustering (SC), an algorithm that relies on computing similarities among data points as a support to describe classes. However SC is computationally complex as we now recall.

In spectral clustering, we do not cluster the data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ directly. Instead, we first compute weights $w_{ij} \geqslant 0$ that models the similarity between pairs of data points $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. This gives rise to a graph $\mathcal{G}$ of $N$ nodes with adjacency matrix $\mathsf{W} = (w_{ij})_{1 \leqslant i,j \leqslant N} \in \mathbb{R}^{N \times N}$. Second, we calculate the Laplacian matrix $\mathsf{L} := \mathsf{D} - \mathsf{W}$ associated to $\mathsf{W}$, where $\mathsf{D}$ is the diagonal matrix with entries $D_{ii} = \sum_{1 \leqslant j \leqslant N} w_{ij}$. Third, we compute the eigenvectors $\mathsf{U}_k := (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) \in \mathbb{R}^{N \times k}$ associated to the first $k$ smallest eigenvalues of $\mathsf{L}$. Finally, we run $k$-means using the rows of $\mathsf{U}_k$ as features vectors to partition the $N$ data points into $k$ clusters.

We readily observe that for large datasets the above algorithm has a huge computational and memory cost. The main bottleneck is the computation of the first $k$ eigenvectors of $\mathsf{L}$. The computational cost of $k$-means also increases with the size of the datasets.

Our first main objective is to formulate an alternative clustering algorithm, whose complexity is much lower than SC, but that provably recovers the solution SC would have computed with no comptutational restriction. Our second objective is to formulate a dynamic version of the aforementioned algorithm that will be able to deal with time-dependent data sets and only update the clustering solution based on limited computations. Ultimately, we are seeking a provably correct approximation of SC to streaming data. Our third objective will be to extend the traditional notion of clustering to multimodal datasets, without explicitly defining a similarity measure among items.

# 2 State of research

In [3] we have shown that one can estimate the result of spectral clustering by extracting randomized information via filtering on graphs to generate features, thereby completely bypassing the SVD step. This technique works by performing a controlled estimation $\tilde{D}_{ij}$ of the spectral clustering distance $D_{ij}$ *without* partially diagonalizing the Laplacian: by fast filtering a few random vectors, thereby making an important connection with the new domain of signal processing on graphs [2] - one of the key expertise of the applicant. We have shown using results from random embedding theory that it is sufficient to compute $\mathcal{O}(\log N)$ such feature maps to guarantee a precise estimation. This naturally leads to a great reduction in computational complexity. However, for very large graphs, one still has to cluster a fairly large feature matrix of size of the order of $\log(N) \times N$. One of our goals will be to break to $\mathcal{O}(N)$ scaling bottleneck using recent results on sampling smooth scalar fields on graphs.

# 3 Research plan: approaches and methods

## 3.1 Towards dynamic spectral clustering

We propose a novel approach to considerably reduce the computational and storage requirements for Spectral Clustering. Our proposition builds on the SVD-less approach described above. One has to think of the feature maps computed by low-pass filtering random signals as smoothly varying scalar fields over the vertex set. The low pass filter induces strong correlations among samples of these fields, which means they can be heavily downsampled without loss of information. Ultimately, we expect to show that only collecting $\mathcal{O}(k \log k)$ samples will be enough to fully characterize these features maps. This means all the cluster information would now be contained in a much small matrix of size $\log(N) \times k \log(k)$. Once the $k \log(k)$ sampled vertices have been assigned clusters, these initial assignments $\boldsymbol{c}_i^r$ are used as observations to recover the cluster membership of all unobserved vertices. This can be done by solving a convex optimization problem that will penalize for smooth cluster membership maps while enforcing coherence with the observations :

$$\min_{\boldsymbol{x} \in \mathbb{R}^N} \|\mathsf{M}\boldsymbol{x} - \boldsymbol{c}_i^r\|_2^2 + \lambda \, \boldsymbol{x}^\intercal g(\mathsf{L})\boldsymbol{x}, \tag{1}$$

where $\mathsf{M}$ is the subsampling operator and $\boldsymbol{x}^\intercal g(\mathsf{L})\boldsymbol{x}$ is a generalized Dirichlet semi-norm enforcing graph smoothness. In a different context, we have recently proved [1] that this class of optimization problem can robustly recover smooth signals from compressed measurements. By interpreting cluster assignments as smooth signals, we plan to adapt this theory to precisely characterize the robustness of accelerated spectral clustering.

Even though (1) leads to a tractable numerical scheme, it may still be complex to solve for very large graphs. We therefore plan to study numerical methods for approximating the solution of this problems in a Big Data setting. A first approach consists in replacing (1) by iterations of a diffusion process, followed by a pointwise non-linearity to filter out information leakage. Cluster assignments are then replaced by $k$ competing diffusion processes, and a vertex is assigned to the first diffusion front that hits it. Another approach of interest, inspired by large scale machine learning, would be to formulate a stochastic gradient descent version of (1).

The compressed clustering approach depicted above would be efficient for static data, but if the dataset is changing with time the naive approach that would consist in re-computing cluster assignments would clearly be tremendously complex. Fortunately, our technique lends itself to an interesting strategy that allows us to tackle dynamic clustering. Indeed, imagine the graph to be clustered slowly changes with time (a fraction of connections changes) and let $C_t$ be the cluster assignments computed at time $t$. The random filtering features computed at $t$ carry some information pertinent for $C_{t+1}$ since the whole graph did not change. It is therefore plausible that only a fraction of these projections ought to be re-computed to evaluate the new assignments. We therefore propose to study, theoretically and numerically, the problem of recovery with smoothly time-varying graphs with partly updated random filtering features and $C_t$ as constraints to the recovery of $C_{t+1}$. On the theoretical front, we plan to formulate an equivalent of the Davis-Kahan theorem where the control between the original and the perturbed subspaces is in terms of the embedding of those subspaces via random filtering. Interestingly, the Davis-Kahan theorem is a statement about the incoherence of subspaces and this is precisely the quantity that governs the quality of embedding by random filtering.

### 3.2 A latent space model for multimodal graphs

The second problem we would like to tackle in this project is how to do data analytics in the presence of wildly different modalities. Classical clustering is performed on feature vectors extracted from a single modality (think text, images or any other source). However it becomes increasingly important to analyse jointly different sources of information and mine for multi-modal patterns. A good example is medical records, constituted of text (doctor reports), images but also increasingly of signals collected by wearables. Finding patterns in such cross-modal datasets could reveal information that up to now has been ignored in the light of mono-modal analytics.

Formally, we describe a multimodal dataset via feature vectors $f_i$ associated to data items. Multimodality entails $f_i \in \mathbb{R}^{d_i}$, i.e the feature dimension depends on the item and it is not trivial to measure the affinity among features via a distance function. We will therefore take a different route by thinking of the graph among features as a latent model for the data. In this case, any item $i \in \{1, ..., N\}$ is mapped to a latent variable $x_i \in [0, 1]$, irrespective of the inital dimension. That mapping cannot be constructed explicitely but will be adjusted in a data driven manner. To that end, we will use a metric on the latent space (for instance the simple euclidean metric) and recover the latent variables by imposing that the scalar field $x_i$ is smooth over the graph so induced :

$$\arg\min_{x \in \mathcal{B}} x^t L x,$$

where $\mathcal{B}$ is a set of constraints imposed to avoid a trivial solution.

## 4 Timeframe and milestones

We propose two sub-proposals that will share a lot of cross development and knowledge on graph-based data analytics as well as similar numerical methods. We therefore foresee two PhD students, each for a standard duration of 36 months. The timeframe and milestones for both sub-projects is as follows :

## 5 Expected benefit and possible application of results

We expect that, if successful, this project can have a strong impact on the field of data analytics by providing new and efficient means of mining large datasets for the most common pattern typically sought: clusters. Moreover, beyond the sheer volume of data, this project also plans to tackle the issue of dynamic data, where streams of features must be processed to recover time-varying cluster information. The expected benefit of this first research thrust is to be able to scale up data analytics based on clustering to big data sets using principled, well-motivated methodologies.

| **Project 1** | **Project 2** |
|---|---|
| milestone 1, month 12: *theoretical formulation and numerical experiments on compressive clustering* | |
| milestone 2, month 18: *numerical optimization and validation of compressive clustering over large graphs* | milestone 1, month 18: *theoretical formulation of the latent space model, first implementation* |
| milestone 3, month 24: *theoretical formulation of dynamic clustering* | milestone 2, month 24: *validation of multimodal clustering on various datasets* |
| milestone 4, month 30: *numerical optimization and validation of dynamic clustering on large datasets* | milestone 3, month 30: *numerical optimization for large scale datasets* |
| month 36: end of thesis | month 36: end of thesis |

Figure 1: Timeframe and Milestones

Finally the potential to deal with multimodal datasets mixing texts, images, speech or any other form of data can also have far reaching impact in any application field where mining digital data is important. The expected benefit here is to open up new realms for data analytics where data sources can be mixed arbitrarily and patterns are defined in a cross-modal way.

## References

[1] Puy, G., Tremblay, N., Gribonval, R., and Vandergheynst, P. Random sampling of bandlimited signals on graphs. Tech. rep., 2015.

[2] Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. *IEEE Signal Processing Magazine 30*, 3 (2013), 83–98.

[3] Tremblay, N., Puy, G., Borgnat, P., Gribonval, R., and Vandergheynst, P. Accelerated Spectral Clustering Using Graph Filtering of Random Signals. In *41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)* (2016).