# Clustering on Multi-Layer Graphs via Subspace Analysis on Grassmann Manifolds

Xiaowen Dong, Pascal Frossard, Pierre Vandergheynst
Signal Processing Laboratories (LTS4/LTS2)
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
{xiaowen.dong,pascal.frossard,pierre.vandergheynst}@epfl.ch

Nikolai Nefedov
Signal and Information Processing Laboratory
Eidgenössische Technische Hochschule Zürich (ETH Zürich)
Zurich, Switzerland
nefedov@isi.ee.ethz.ch

***Abstract*—Relationships between entities in datasets are often of multiple types, which can naturally be modeled by a multi-layer graph; a common vertex set represents the entities and the edges on different layers capture different types of relationships between the entities. In this paper, we address the problem of analyzing multi-layer graphs and propose methods for clustering the vertices by efficiently merging the information provided by the multiple modalities. We propose to combine the characteristics of individual graph layers using tools from subspace analysis on a Grassmann manifold. The resulting combination can then be viewed as a low dimensional representation of the original data which preserves the most important information from diverse types of relationships between entities. We use this information in new clustering methods and test our algorithm on several synthetic and real world datasets to demonstrate its efficiency.**

***Index Terms*—Multi-layer graphs, subspace representation, Grassmann manifold, clustering.**

## I. Introduction

Graphs are powerful mathematical tools for modeling pairwise relationships among sets of entities. A graph traditionally captures a single form of relationships between entities. However, numerous emerging applications rely on different forms of such relationships, which can naturally be represented by a multi-layer graph whose layers share a common set of vertices but with different edge weights, depending on the type of information in each layer. Assuming that all the graph layers are informative, they are likely to provide complementary information and thus to offer richer information than any single layer taken in isolation. We thus expect that a proper combination of the information contained in the different graph layers leads to an improved understanding of the structure of the data.

In this paper, we consider a $M$-layer graph $G$ with individual graph layers $G_i = \{V, E_i\}$, $i = 1, \ldots, M$, where $V$ represents the common vertex set and $E_i$ represents the edge set in the $i$-th individual graph $G_i$, which corresponds to an adjacency matrix $W_i$. An example of a three-layer graph is shown in Fig. 1 (a) (we assume unitary edge weights). Clearly, different graph layers capture different types of relationships between the vertices. Our objective is to find a method that properly combines the information in these different layers. We first adopt a subspace representation for the information provided by the individual graph layers; by modeling each graph layer as a subspace on a Grassmann manifold, the problem of combining multiple graph layers is then transformed into the problem of efficiently merging different subspaces into a representative one on a Grassmann manifold. Specifically, we study the distances between the subspaces and develop a new framework where the overall distance between the representative subspace and the individual subspaces is minimized. The proposed method leads to a summarization of the information contained in the multiple graph layers, which reveals the intrinsic relationships between the vertices.

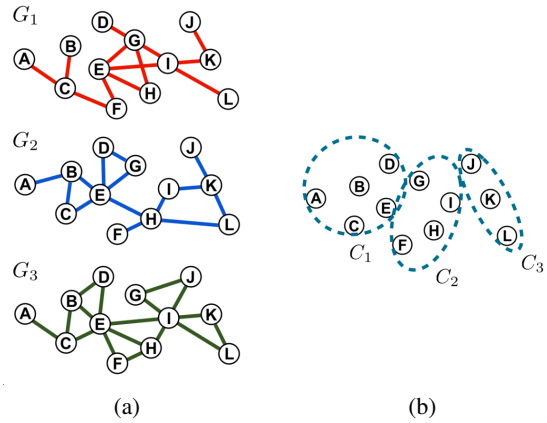


Fig. 1. (a) An illustration for a three-layer graph $G$, whose three layers $\{G_i\}_{i=1}^3$ share the same set of vertices but with different edges. (b) A potential unified clustering $\{C_k\}_{k=1}^3$ of the vertices based on the information provided by the three layers.

Finally, we study the clustering problem as an application of our framework: we want to find a unified clustering of the vertices (as illustrated in Fig. 1 (b)) by utilizing the representative subspace, which is better than the clustering achieved on any of the graph layers $G_i$ independently. We show that our clustering method achieves competitive performance compared to baseline and state-of-the-art techniques in various synthetic and real world datasets.

Learning multi-view data have attracted a large amount of interest in the learning research communities, with representative techniques including unified matrix factorizations [1], [2], Canonical Correlation Analysis [3], co-training and co-regularization [4], [5], and graph regularization [6], to name a few. Compared to the related work, the main contributions of the paper are three-fold: First, we present a fundamental and intuitive mathematical framework for the learning problems on multi-layer graphs, which is somehow lacking in the literature. Second, we show the link between the projection distance on the Grassmann manifold and the empirical estimate of the Hilbert-Schmidt Independence Criterion (HSIC) [7]. Finally, we provide a simple yet competitive solution to the problem of clustering on multi-layer graphs using the construction of a proper representative layer.

## II. Preliminaries

### A. Layer subspace representation

In this section, we introduce a subspace representation for the information contained in a single graph. Such a representation is inspired by spectral clustering [8][9][10], which is a technique that studies the spectral properties of the graph Laplacian matrix for partitioning

the vertex set into several subsets. Without loss of generality, we consider a connected graph $G$ that has an adjacency matrix $W$. The degree matrix $D$ is defined as the diagonal matrix containing the degrees of the vertices along the diagonal. The normalized graph Laplacian matrix $L$ is then defined as: $L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$. The spectral clustering algorithm proposed in [9] solves the following trace minimization problem:

$$\min_{U \in \mathbb{R}^{n \times k}} tr(U'LU), \quad \text{s.t.} \quad U'U = I. \tag{1}$$

where $n$ is the number of vertices in the graph, and $k$ is the target number of clusters. The solution to the problem of Eq. (1) contains the first $k$ eigenvectors (which correspond to the $k$ smallest eigenvalues) as columns. The final clustering is then achieved by applying the $k$-means algorithm to the normalized row vectors of $U$.

As we can see, the low dimensional matrix $U$ consists of $k$ orthonormal eigenvectors as columns, and each row of $U$ (after normalization) can be viewed as the coordinates of the corresponding vertex in the $k$-dimensional space. Therefore, this defines a subspace representation of the original vertices in the spectral domain. This subspace representation contains the information on the connectivity of the vertices in the original graph, hence it can be used for clustering purposes. We adopt the same subspace representation for each layer of the multi-layer graph in our framework. Then, we gather the different subspaces on a Grassmann manifold, as explained below.

### B. Multi-layer manifold representation

By definition, a Grassmann manifold $\mathcal{G}(k, n)$ is the set of $k$-dimensional linear subspaces in $\mathbb{R}^n$, where each unique subspace is mapped to a unique point on the manifold. This provides a natural representation for our problem: the subspaces representing the individual graph layers can be considered as different points on the Grassmann manifold; it thus permits to use efficient tools to study the distances between points on the manifold, namely, distances between different subspaces.

Mathematically speaking, each point on $\mathcal{G}(k, n)$ can be represented by an orthonormal matrix $Y \in \mathbb{R}^{n \times k}$ whose columns span the corresponding $k$-dimensional subspace in $\mathbb{R}^n$; it is thus denoted as $span(Y)$. The distance between two points on the manifold, or between two subspaces $span(Y_1)$ and $span(Y_2)$, is then defined based on a set of principal angles $\{\theta_i\}_{i=1}^k$ between these subspaces [11]. These principal angles, which measure how the subspaces are geometrically close, are the fundamental measures used to define various distances on the Grassmann manifold, such as the Riemannian (geodesic) distance or the projection distance [12], [13]. In this paper, we use the projection distance as a metric, which is defined as: $d_{\text{proj}}(Y_1, Y_2) = (\sum_{i=1}^k \sin^2\theta_i)^{\frac{1}{2}}$, where $Y_1$ and $Y_2$ are the orthonormal matrices representing the two subspaces under comparison. The reason for choosing the projection distance is two-fold: (i) the projection distance is an unbiased definition since it uses all the principal angles; (ii) it can be interpreted using a one-to-one mapping that preserves distinctness: $span(Y) \rightarrow YY' \in \mathbb{R}^{n \times n}$. Note that the squared projection distance can be rewritten as:

$$d^2_{\text{proj}}(Y_1, Y_2) = \frac{1}{2}||Y_1{Y_1}' - Y_2{Y_2}'||^2_F = k - tr(Y_1{Y_1}'Y_2{Y_2}'). \tag{2}$$

Therefore, the projection distance can be related to the Frobenius norm of the difference between the mappings of two subspaces $span(Y_1)$ and $span(Y_2)$ in $\mathbb{R}^{n \times n}$. Moreover, the second equality of Eq. (2) provides an explicit way of computing the projection distance between two subspaces from their matrix representations $Y_1$ and $Y_2$. We are going to use it in developing the generic merging framework in the following section.

## III. Framework for merging subspaces and clustering

We now present the proposed framework for combining multiple graph layers by merging multiple subspaces, as well as its application to clustering on multi-layer graphs.

### A. Layer merging framework

Equipped with the subspace representation for individual graphs and with a distance measure to compare different subspaces on the Grassmann manifold, we present now our generic framework for merging the information from multiple graph layers. Our philosophy is to find a representative subspace $span(U)$ that is close to all the individual subspaces $span(U_i)$; at the same time the representation $U$ should preserve the vertex connectivity in each graph layer. For notational convenience, in the rest of the paper we simply refer to the matrix representations $U$ and respectively $U_i$ as the corresponding subspaces, unless indicated differently.

We can generalize the squared projection distance defined in Eq. (2) to the case of multiple subspaces, by defining the squared projection distance between the target representative subspace $U$ and the $M$ individual subspaces $\{U_i\}_{i=1}^M$ as the sum of squared projection distances between $U$ and each individual subspace given by $U_i$:

$$d^2_{\text{proj}}(U, \{U_i\}_{i=1}^M) = \sum_{i=1}^M d^2_p(U, U_i) = kM - \sum_{i=1}^M tr(UU'U_iU_i'). \tag{3}$$

The minimization of the distance measure in Eq. (3) enforces the representative subspace $U$ to be close to all the individual subspaces $\{U_i\}_{i=1}^M$ in terms of the projection distance on the Grassmann manifold. At the same time, we want $U$ to preserve the vertex connectivity in each graph layer. This can be achieved by minimizing the Laplacian quadratic form evaluated on the columns of $U$, as also indicated by the objective function in Eq. (1) for spectral clustering. Therefore, we finally propose to merge multiple subspaces by solving the following optimization problem that combines Eq. (1) and Eq. (3):

$$\begin{aligned} \min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^M tr(U'L_iU) + \alpha[kM - \sum_{i=1}^M tr(UU'U_iU_i')], \\ \text{s.t.} \quad U'U = I, \end{aligned} \tag{4}$$

where $L_i$ and $U_i$ are the graph Laplacian and the subspace representation for $G_i$, respectively. The regularization parameter $\alpha$ trades-off the two terms in the objective function. By ignoring the constant terms and rearranging the trace form in the second term of the objective function, Eq. (4) can be rewritten as

$$\min_{U \in \mathbb{R}^{n \times k}} tr[U'(\sum_{i=1}^M L_i - \alpha \sum_{i=1}^M U_iU_i')U], \quad \text{s.t.} \quad U'U = I. \tag{5}$$

This is the same trace minimization problem as in Eq. (1), but with a "modified" Laplacian: $L_{\text{mod}} = \sum_{i=1}^M L_i - \alpha \sum_{i=1}^M U_iU_i'$. Therefore, the solution to the problem of Eq. (5) is given by the first $k$ eigenvectors of the modified Laplacian $L_{\text{mod}}$, which can be computed using efficient algorithms for eigenvalue problems [14], [15].

### B. Discussion of the merging framework

Interestingly, the choice of projection distance as a similarity measure between subspaces can be well justified from information-theoretic and statistical learning points of view. First, Hamm et al. [16] have shown that the squared projection distance is consistent

with the Kullback-Leibler (K-L) divergence [17] by taking a probabilistic viewpoint of the subspace representation. Second, we show that it is also consistent with the Hilbert-Schmidt Independence Criterion (HSIC) [7], which measures the statistical dependence between two random variables. Specifically, given $K_{\mathcal{X}_1}, K_{\mathcal{X}_2} \in \mathbb{R}^{n\times n}$ that are the centered Gram matrices of some kernel functions defined over two random variables $\mathcal{X}_1$ and $\mathcal{X}_2$, the empirical estimate of HSIC is given by: $d_{\text{HSIC}}(\mathcal{X}_1, \mathcal{X}_2) = tr(K_{\mathcal{X}_1} K_{\mathcal{X}_2})$. In our case, we can consider the rows of the individual subspace representations $U_i$ and $U_j$ as two particular sets of sample points in $\mathbb{R}^k$, which are drawn from two probability distributions governed by the information on vertex connectivity in $G_i$ and $G_j$, respectively. In other words, the sets of rows of $U_i$ and $U_j$ can be seen as realizations of two random variables $\mathcal{X}_i$ and $\mathcal{X}_j$. Therefore, we can define the Gram matrices of linear kernels on $\mathcal{X}_i$ and $\mathcal{X}_j$ as $K_{\mathcal{X}_i} = (U_i{'})'(U_i{'}) = U_i U_i{'}$ and $K_{\mathcal{X}_j} = (U_j{'})'(U_j{'}) = U_j U_j{'}$. We see that:

$$d_{\text{HSIC}}(\mathcal{X}_i, \mathcal{X}_j) = tr(K_{\mathcal{X}_i} K_{\mathcal{X}_j}) = tr(U_i U_i{'} U_j U_j{'}) = k - d^2_{\text{proj}}(U_i, U_j).$$

This shows that the projection distance between subspaces $U_i$ and $U_j$ can be interpreted as the negative dependence between $\mathcal{X}_i$ and $\mathcal{X}_j$, which reflects the information provided by the two individual graph layers $G_i$ and $G_j$. In other words, the smaller the projection distance, the larger the dependence (or consistency) between $G_i$ and $G_j$. Therefore, from both information-theoretic and statistical learning points of view, the smaller the projection distance between two subspace representations $U_i$ and $U_j$, the more similar the information in the respective graphs that they represent. As a result, the representative subspace (the solution $U$ to the problem of Eq. (4)) can be considered as a subspace representation that "summarizes" the information from the individual graph layers, and at the same time captures the intrinsic relationships between the vertices in the graph.

### C. Clustering on multi-layer graphs

The merging framework proposed in the previous section leads to a natural solution to the clustering problem on multi-layer graphs. In more details, similarly to the spectral embedding matrix in the spectral clustering algorithm, which is a subspace representation for one individual graph, our merging framework provides a representative subspace that contains the information from multiple graph layers. Using this representative subspace, we can then follow the same steps as spectral clustering [9] to achieve the final clustering of the vertices. The proposed clustering algorithm is summarized in Algorithm 1.

**Algorithm 1** Spectral Clustering on Multi-Layer graphs (**SC-ML**)

**Input:**
$\{W_i\}_{i=1}^M$: $n \times n$ weighted adjacency matrices of individual graph layers $\{G_i\}_{i=1}^M$
$k$: target number of clusters
$\alpha$: regularization parameter
Compute the normalized Laplacian matrix $L_i$ and the subspace representation $U_i$ for each $G_i$.
Compute the modified Laplacian matrix $L_{\text{mod}} = \sum_{i=1}^M L_i - \alpha \sum_{i=1}^M U_i U_i{'}$.
Compute $U \in \mathbb{R}^{n\times k}$ that is the matrix containing the first $k$ eigenvectors $u_1, \ldots, u_k$ of $L_{\text{mod}}$. Normalize each row of $U$ to get $U_{\text{norm}}$.
Let $y_j \in \mathbb{R}^k$ $(j = 1, \ldots, n)$ be the $j$-th row of $U_{\text{norm}}$.
Cluster $y_j$ in $\mathbb{R}^k$ into $C_1, \ldots, C_k$ using the $k$-means algorithm.
**Output:**
$C_1, \ldots, C_k$: The cluster assignment

## IV. Experimental results

We now study the performance of our clustering algorithm on one synthetic and two real world datasets with multi-layer graph representations. The first dataset is a synthetic dataset, where we have three point clouds in $\mathbb{R}^2$ forming the English letters "N", "R" and "C" (shown in Fig. 2). Each point cloud is generated from a five-component Gaussian mixture model, where each component represents a class of 500 points with specific color. A 5-nearest neighbor graph is then constructed for each point cloud by assigning the weight of the edges as the reciprocal of the Euclidean distance between vertex pairs. This gives us a 3-layer graph of 2500 vertices. The second dataset contains the mobile phone data of 136 users living and working in the Lake Léman region in Switzerland [18], which are grouped into eight clusters by their email affiliations. Considering the users as vertices in the graph, we construct three graphs by measuring the proximities between these users in terms of GPS locations, Bluetooth scanning activities and phone communication. For GPS locations and Bluetooth scans, we measure how many times two users are sufficiently close geographically and how many times two users' devices have detected the same bluetooth devices. For phone communication, we measure the number of calls between any pair of two users. This forms a 3-layer graph of 136 vertices. The third dataset is a subset of the Cora bibliographic dataset[1], which contains 292 research papers from three different fields. Considering papers as vertices in the graph, we construct the first two graphs by measuring the similarities among the title and the abstract of these papers, in terms of cosine similarities using the *Term Frequency-Inverse Document Frequency (TF-IDF)* [19] weighting scheme. We add a third graph that reflects the citation relationships among the papers. This results in a 3-layer graph of 292 vertices.

We adopt three baseline algorithms as well as a state-of-the-art technique, namely the co-regularization approach (denoted as **SC-CoR**) introduced in [5], in our clustering performance comparison. The three baseline comparative algorithms work as follows:

- **SC-Single**: Spectral Clustering [9] applied on a single graph layer, where the graph is chosen to be the one that leads to the best clustering results.
- **SC-Sum**: Spectral clustering applied on a global matrix $W$ that is the summation of the normalized adjacency matrices of the individual layers: $W = \sum_{i=1}^M D_i^{-\frac{1}{2}} W_i D_i^{-\frac{1}{2}}$.
- **SC-KSum**: Spectral clustering applied on the summation $K$ of the spectral kernels [1] of the adjacency matrices: $K = \sum_{i=1}^M K_i$ with $K_i = \sum_{m=1}^d u_{im} u_{im}{'}$, where $n$ is the number of vertices, $d \ll n$ is the number of eigenvectors used in the definition of the spectral kernels $K_i$, and $u_{im}$ represents the $m$-th eigenvector of the Laplacian $L_i$ for the graph $G_i$.

In the implementation of **SC-ML** and **SC-CoR**, we choose the value of the corresponding parameters $\alpha$ through multiple empirical trials and report the best clustering performance. More details about the selection of the parameters are provided in [20].

We evaluate the performance of the different clustering algorithms with three different criteria, namely *Purity*, *Normalized Mutual Information (NMI)* and *Rand Index (RI)* [19]. The results are summarized in Table I (a), (b) and (c) for the synthetic, NRC and Cora dataset, respectively. It is clear that **SC-ML** and **SC-CoR** generally outperform the baseline approaches for the three datasets. Note that

[1]Available online at "http://people.cs.umass.edu/~mccallum/data.html" under category "Cora Research Paper Classification".

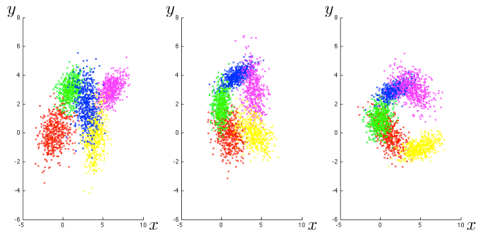

Fig. 2. Three five-class point clouds in $\mathbb{R}^2$ forming English letters "N", "R" and "C".

**SC-CoR** also uses the projection distance, however it involves a different alternating optimization scheme that optimizes, at each step, one subspace representation, while fixing the others. This differs from **SC-ML** mainly in the following aspects. First, **SC-CoR** requires a sensible initialization; then, it does not guarantee that all the subspace representations converge to one point on the Grassmann manifold[2]. In contrast, **SC-ML** directly finds a single representation without the need for alternating optimization steps and careful initializations. These are possibly the reasons why **SC-ML** performs slightly better than **SC-CoR** for the synthetic and NRC datasets. Second, from a computational point of view, the iterative nature of **SC-CoR** requires solving an eigenvalue problem for $MN$ times, where $M$ and $N$ are the number of individual graphs and the number of iterations needed for the algorithm to converge, respectively. In contrast, **SC-ML** needs to solve an eigenvalue problem only once.

## V. Conclusions

In this paper, we provide a framework for analyzing information in multi-layer graphs and for clustering vertices of graphs in rich datasets. Our generic approach is based on the transformation of the information contained in the individual graph layers into subspaces on a Grassmann manifold. The estimation of a representative subspace can then be essentially considered as the problem of finding a good summarization of multiple subspaces using distance analysis on the Grassmann manifold. Our framework is well motivated by analyses from information-theoretic and statistical learning points of views. We show that it can be applied to the clustering problem on multi-layer graphs and that it provides an efficient solution that is competitive with the state-of-the-art techniques.

## Acknowledgement

This work has been partly supported by Nokia Research Center (NRC) Lausanne, and the EDGAR project funded by Hasler Foundation, Switzerland.

## References

[1] W. Tang, Z. Lu, and I. Dhillon, "Clustering with Multiple Graphs," in *International Conference on Data Mining*, 2009.

[2] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-View Clustering via Joint Nonnegative Matrix Factorization," in *SIAM International Conference on Data Mining*, 2013.

[3] M. B. Blaschko and C. H. Lampert, "Correlational Spectral Clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] A. Kumar and H. Daumé III, "A Co-training Approach for Multi-view Spectral Clustering," *International Conference on Machine Learning (ICML)*, 2011.

[2]In [5], the authors have also proposed a "centroid-based co-regularization approach" that introduces a consensus representation. However, such a representation is still computed via an alternating optimization scheme, which needs a sensible initialization and keeps the same iterative nature.

TABLE I
Performance comparison of different clustering algorithms on (a) the synthetic dateset, (b) the NRC dataset, and (c) the Cora dataset.

| | SC-Single | SC-Sum | SC-KSum | SC-CoR | SC-ML |
|---|---|---|---|---|---|
| *Purity* | 0.8580 | 0.9732 | 0.9772 | 0.9784 | 0.9828 |
| *NMI* | 0.7266 | 0.9180 | 0.9280 | 0.9278 | 0.9414 |
| *RI* | 0.9018 | 0.9791 | 0.9821 | 0.9830 | 0.9864 |

(a)

| | SC-Single | SC-Sum | SC-KSum | SC-CoR | SC-ML |
|---|---|---|---|---|---|
| *Purity* | 0.5147 | 0.5441 | 0.5368 | 0.5809 | 0.6103 |
| *NMI* | 0.3133 | 0.3609 | 0.3410 | 0.4056 | 0.4156 |
| *RI* | 0.7326 | 0.7706 | 0.7666 | 0.7878 | 0.7929 |

(b)

| | SC-Single | SC-Sum | SC-KSum | SC-CoR | SC-ML |
|---|---|---|---|---|---|
| *Purity* | 0.9555 | 0.9692 | 0.9521 | 0.9829 | 0.9829 |
| *NMI* | 0.8314 | 0.8656 | 0.8117 | 0.9304 | 0.9175 |
| *RI* | 0.9426 | 0.9600 | 0.9383 | 0.9776 | 0.9775 |

(c)

[5] A. Kumar, P. Rai, and H. Daumé III, "Co-regularized Multi-view Spectral Clustering," *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[6] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, "Clustering with Multi-Layer Graphs: A Spectral Perspective," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5820–5831, Nov 2012.

[7] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf, "Measuring Statistical Dependence with Hilbert-Schmidt Norms," in *International conference on Algorithmic Learning Theory*, 2005.

[8] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, Aug 2000.

[9] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[10] U. von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec 2007.

[11] G. H. Golub and C. F. V. Loan, "Matrix Computations (3rd Edition)," *Johns Hopkins University Press*, 1996.

[12] A. Edelman, T. A. Arias, and S. T. Smith, "The Geometry of Algorithms with Orthogonality Constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.

[13] J. Hamm and D. Lee, "Grassmann Discriminant Analysis: a Unifying View on Subspace-Based Learning," *International Conference on Machine Learning (ICML)*, 2008.

[14] D. C. Sorensen, "Implicit Application of Polynomial Filters in a k-Step Arnoldi Method," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 1, pp. 357–385, Jan 1992.

[15] R. B. Lehoucq and D. C. Sorensen, "Deflation Techniques for an Implicitly Restarted Arnoldi Iteration," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 4, pp. 789–821, Oct 1996.

[16] J. Hamm and D. Lee, "Extended Grassmann Kernels for Subspace-Based Learning," *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[17] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar 1951.

[18] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila, "Towards rich mobile phone datasets: Lausanne data collection campaign," in *International Conference on Pervasive Services*, 2010.

[19] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," *Cambridge University Press*, 2008.

[20] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, "Clustering with Multi-Layer Graphs via Subspace Analysis on Grassmann Manifold," *arXiv:1303.2221 [cs.LG]*, 2013.