SEMI-SUPERVISED SPECTRAL CLUSTERING

Xiaoyi MAI[1], Romain COUILLET[12]

[1]CentraleSupélec, Université Paris-Saclay, France
[2] GIPSA-lab, University Grenoble-Alpes, France

ABSTRACT

In this article, we propose a semi-supervised version of spectral clustering, a widespread graph-based unsupervised learning method. The semi-supervised spectral clustering has the advantage of producing consistent classification of data with sufficiently large number of labelled or unlabelled data, unlike classical graph-based semi-supervised methods which are only consistent on labelled data. Theoretical arguments are provided to support the proposition of this novel approach, as well as empirical evidence to confirm the theoretical claims and demonstrate its superiority over other graph-based semi-supervised methods.

***Index Terms*—** semi-supervised learning, spectral clustering, graphs, consistency

1. INTRODUCTION

Semi-supervised learning aims to maximize the learning ability by using both labelled and unlabelled data, typically a small amount of labelled data with abundant unlabelled data. In many real-life machine learning tasks, the labeling process is much more expensive than the collection of unabelled data, it is thus of great practical value to study the semi-supervised approach. A successful semi-supervised learning method should be able to learn from both the labelled data and the underlying clusters in the unlabelled data. Combining these two types of information is conceptually more difficult than one-sided learning perspectives focused on labelled or unlabelled data. In this regard, some seemingly natural semi-supervised approaches may have unexpected outcomes.

A recent random matrix analysis [1] notably shows that in very high dimensions, a popular branch of graph-based semi-supervised learning methods (here referred to as Laplacian regularization) fails to exploit the clustering structure of unlabelled data. Also, the results in [2] state that the Laplacian regularization approach converges to a trivial solution in the limit of infinite data. As such, the Laplacian regularization methods tend to disregard additional unlabelled data which we refer to as being *inconsistent* with respect to the unlabelled dataset. As a consequence, given sufficient unlabelled data, Laplacian regularization will be outperformed by spectral clustering, an unsupervised method which learns from the data graph by optimizing the same objective as Laplacian regularization but whose consistency is proven in [3].

The graph-based semi-supervised learning method presented in this article is based on spectral clustering, in order to ensure its consistency with respect to unlabelled data. An original strategy is proposed to incorporate the labelled data in the learning process. The proposed method is easy to implement with similar cost to spectral clustering. To demonstrate its efficiency as a semi-supervised learning algorithm, we provide a mathematical discussion along with experimental confirmations, all showing that the semi-supervised approach yields better partitions of data points than spectral clustering as a result of exploiting labelled data.

The reminder of the article is organized as follows. We start by introducing preliminary notions in Section 2. The proposed semi-supervised spectral clustering method is presented in Section 3, along with theoretical justification and corroborating simulations. Finally, we end the article with concluding remarks in Section 4.

2. PRELIMINARIES

We present the common setting of graph-based methods in Section 2.1. The principles and consistency properties of spectral clustering, on which the proposed method is based, are briefly explained and discussed in Section 2.2, followed by a section on the Laplacian regularization methods.

2.1. Problem Setup

Graph-based methods represent data points as vertices in a graph with their similarities reflected by the edges. In this paper, we consider undirected and weighted graphs $G = (V, E)$ with set of vertices $V = \{x_1, \cdots, x_n\}$, which is also the dataset, and set of edges $E = \{(x_i, x_j, w_{ij}) | x_i, x_j \in V\}$, where (x_i, x_j, w_{ij}) denotes an edge of weight w_{ij} between x_i and x_j, and $w_{ii} = 0$. The degree d_i of a node x_i is defined as $d_i = \sum_j w_{ij}$. The adjacency matrix W of the graph is an $n \times n$ symmetric matrix with $W_{ij} = w_{ij}$, and the degree matrix $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $D_{ii} = d_i$. The Laplacian matrix L is defined as $L = D - W$. The symmetric normalized form of the Laplacian matrix is $L_s =$

$D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - W_s$ where $W_s = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ is the symmetric normalized adjacency matrix.

The objective of graph-based methods is to learn a data representation that is in accordance with the graph structure, meaning that data points connected with large weights have similar representations. This is usually achieved by minimizing a loss function:

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 = f^T L f \tag{1}$$

where f is the representation function. A normalized formulation of (1) writes as

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}\right)^2 = f^T L_s f. \tag{2}$$

It is easily seen that the minimization of (1) imposes data points connected with large weights to have close values on the representation function f, so does (2) for closely connected data points with similar degrees. In the following, we will refer to (1) and (2) as the graph smoothness penalty terms of f since their minimizations encourage f to have similar values on neighboring vertices in the graph.

2.2. Spectral Clustering

Evidently, the graph smoothness penalty term (1) is minimized when f has constant values on all data points, which is obviously a useless data representation for the partitioning. Similarly for the normalized Laplacian regularizer (2) with the trivial solution $f_i = \sqrt{d_i}$.

Therefore, the spectral clustering method states the problem as minimizing (1) or (2) in the subspace orthogonal to the trivial solution for a fixed norm of f. This is equivalent to finding the eigenvector associated with the second smallest eigenvalue for L (unnormalized spectral clustering) or L_s (normalized spectral clustering), as the trivial solutions are the eigenvectors of Laplacian matrices associated with the smallest eigenvalue 0. See [4] for more details.

The consistency properties of spectral clustering are investigated in [3], where the authors prove that under very mild assumptions, the clusters constructed by normalized spectral clustering converge to a limit clustering of the whole data space as the number of sampled data tends to infinity. It is also found that unnormalized spectral clustering is only consistent under strong additional assumptions, which are not always satisfied in real data.

2.3. Laplacian Regularization

The Laplacian regularization approach [5–7] is very similar to spectral clustering, in the sense that it searches the representation function f that minimizes the same graph smoothness penalty terms (1) and (2) as spectral clustering. The difference is that for Laplacian regularization, the representation function f has to also respect the labeling $y_{[l]}$ of the labelled data, which leads to imposing $f_{[l]} = y_{[l]}$[1] [5], where $f_{[l]}$ is the subset of f corresponding to the labelled data. The Laplacian regularization algorithms consist thus in minimizing (1) or (2) under the constraint $f_{[l]} = y_{[l]}$.

Although seemingly a perfectly natural way to learn the inherent clusters of the graph in a semi-supervised manner, the Laplacian regularization methods have some unexpected behaviors when dealing with large dimensional data, as demonstrated in [1] by random matrix arguments. The main issue is that for datasets of sufficiently large dimension and size, an increase in the number of unlabelled data has negligible contribution to the performance of Laplacian regularization algorithms, implying that the Laplacian regularization approach is inconsistent with respect to unlabelled data. Another theoretical work [2] shows that the Laplacian regularization methods yield flat solutions of $f_{[u]}$ ($f_{[u]}$ being the part of f corresponding to the unlabelled data) in the limit of infinite unlabelled data. Since a flat $f_{[u]}$ does not conform to the inherent clusters in the graph, it is conjectured that the same inconsistency problem of high dimensional data also occurs in small dimensions. We refer the reader to [2] and [1] for more details.

3. SEMI-SUPERVISED SPECTRAL CLUSTERING

Despite coming from the same idea of learning a data representation that is smooth on the graph by using the smoothness penalty terms (1) and (2), the Laplacian regularization approach does not share the same consistency properties as spectral clustering, as discussed in the above section. As an answer to this problem, we propose in this section a semi-supervised adaptation of the spectral clustering method. The idea is to preserve the data representation learned from spectral clustering while making use of labelled data, so that the proposed method is consistent with respect to unlabelled data as in the spectral clustering case. The challenge is to ensure that the use of labelled data improves the performance of the proposed algorithm, i.e., it should outperform spectral clustering on the same datasets.

Here are some notations that will be used in the following. Data points x_i are separated into k classes $\mathcal{C}_1, \cdots, \mathcal{C}_k$. For $a \in \{1, \cdots, k\}$, let n_a denote the number of data in class $\mathcal{C}_a$, with $n_{[l]a}$ labelled ones and $n_{[u]a}$ unlabelled ones. The total numbers of labelled and unlabelled data are $n_{[l]} = \sum_a n_{[l]a}$, $n_{[u]} = \sum_a n_{[u]a}$. The vector $j_a \in \mathbb{R}^n$ is defined as the indicator vector for data points in $\mathcal{C}_a$ with $[j_a]_i = 1$ if the data point x_i belongs to $\mathcal{C}_a$, otherwise $[j_a]_i = 0$. Similarly, $j_{[l]} \in \mathbb{R}^n$ (resp., $j_{[l]} \in \mathbb{R}^n$) is the indicator vectors for labelled (resp.,

[1] This condition can be relaxed by adding a penalty term of the form $\|f_{[l]} - y_{[l]}\|^2$ to the optimization objective [8].

unlabelled) data points with $[j_{[l]}]_i = 1$ if x_i is labelled, otherwise $[j_a]_i = 0$. The same is understood with the indicator vectors $j_{[l]a}, j_{[u]a} \in \mathbb{R}^n$ for labelled and unlabelled data in $\mathcal{C}_a$. The operator $\mathcal{D}(v) = \mathcal{D}\{v_a\}_{a=1}^k$ is the diagonal matrix having $v_1, \cdots, v_k$ as its ordered diagonal elements.

3.1. Proposed Method

As explained in Section 2.2, spectral clustering consists in conducting first an eigendecomposition of the Laplacian matrix L or the normalized Laplacian matrix L_s, then selecting a few eigenvectors associated with the smallest eigenvalues except the first one to construct feature vectors of data, before the final partitioning step. Since the normalized spectral clustering algorithm is consistent [3], so is the learned data representation (i.e., the eigenvectors). In order to maintain the consistency property of normalized spectral clustering, the natural move is to base the semi-supervised learning algorithm on the eigenvectors of L_s.

In fact, there exist already such algorithms. Like spectral clustering, the manifold-based method [9] also uses the eigenvectors of Laplacian matrices $V = \{v_1, \cdots, v_m\}$ associated with the m smallest eigenvalues except the first one. Since eigenvectors with small eigenvalues are considered smooth on the graph, the manifold-base method constrains the representation function f to live in the subspace constructed by a certain number of eigenvectors with small eigenvalues, as a means to control the smoothness of f. This strategy can also be interpreted from a graph signal processing perspective [10, 11], where it is justified as limiting the "frequency" of f. The function $f = Va$ is determined by minimizing $\|f_{[l]} - y_{[l]}\|$, with a solved by the method of least squares.

The disadvantage of this method resides in the delicate choice of the number of selected eigenvectors. If the number is large, eigenvectors with relatively high eigenvalues are included, which induces a risk of hurting the smoothness of f; on the other hand, a decrease in the number of selected eigenvector will lead to greater loss of information contained in the discarded eigenvectors.

We propose here an algorithm based also on the eigenvectors of Laplacian matrices, but without the aforementioned disadvantage of the manifold-based method. In other terms, the algorithm should focus on the smoothest eigenvectors, and in the meantime, allow to leverage the whole Laplacian matrix. Obviously, a perfect data representations is composed of the class indicator vectors j_a. The consistency of normalized spectral clustering [3], as well as recent large dimensional arguments [12], implies that there exist informative eigenvectors of L_s which can be seen as a weighted sum of j_a plus noise. As we already know the class of labelled data, we can denoise the eigenvectors by replacing their labelled subsets with weighted sums of $j_{[l]a}$, the class indicator vectors for labelled data. In a subsequent step, the information on labelled points is propagated through the graph by multiplying the denoised eigenvectors with the normalized adjacency matrix W_s, similarly to the label propagation procedure of [6, 13]. The proposed semi-supervised spectral clustering is summarized in Algorithm 1.

Algorithm 1 Semi-supervised spectral clustering

1: **Input:** Normalized Laplacian matrix L_s. Number k of classes. Class indicator vectors $j_{[l]1}, \cdots, j_{[l]k}$ for labelled data.
2: **Output:** Classification of the unlabelled dataset.
3: Compute the $k+1$ eigenvectors $v_0, v_1, \cdots, v_k$ corresponding to the $k+1$ smallest eigenvalues of L_s.
4: For $a, b = 1, \cdots, k$, define $m_{ab} = v_a^T j_{[l]b}/n_{[l]b}$.
5: For $a = 1, \cdots, k$, compute $\hat{v}_a = \sum_{b=1}^k m_{ab} j_{[l]b} + \mathcal{D}(j_{[u]})v_a$ and $\tilde{v}_a = W_s\hat{v}_a/\|W_s\hat{v}_a\|$.
6: Let $\tilde{V} = [\tilde{v}_1, \cdots, \tilde{v}_k] \in \mathbb{R}^{n\times k}$.
7: For $i = 1, \cdots, n$, define $\phi_i \in \mathbb{R}^k$ as the feature vector for data point x_i and affect it with the i-th row of $\tilde{V}$.
8: Cluster $(\phi_i)_{i=1,\cdots,n}$ with k-means algorithm (or other clustering techniques) into $\mathcal{C}_1, \cdots, \mathcal{C}_k$.

3.2. Theoretical Arguments

The objective of this section is to provide more technical details to the intuitive justification of the semi-supervised spectral clustering method given in Section 3.1 and to explain the success of this semi-supervised approach in leveraging labelled data to improve the performance of spectral clustering.

In common statistical modeling, the data in the same class $\mathcal{C}_a$ are considered to be drawn independently according to a certain probability measure P_a. Since w_{ij} measures the similarity between x_i, x_j, it is considered as a function of x_i, x_j, i.e., $w_{ij} = g(x_i, x_j)$ where g is a symmetric function with $g(x_i, x_j) = g(x_j, x_i)$. Therefore, for all pair of x_i, x_j belonging to the same combination of classes, $g(x_i, x_j)$ follow the same distribution. Otherwise speaking, the graph $G = (V, E)$ is statistically invariant to the exchange of data points in the same class. The adjacency matrix W is consequently written as

$$W = \mathbb{E}\{W\} + B = \sum_{a,b=1}^k q_{ab} j_a j_b^T + B$$

for some scalars q_{ab} with $q_{ab} = q_{ba}$, and B is a noise matrix with $\mathbb{E}\{B\} = 0$. And so is the normalized adjacency matrix W_s:

$$W_s = \mathbb{E}\{W_s\} + B' = \sum_{a,b=1}^k r_{ab} j_a j_b^T + B'. \qquad (3)$$

for some scalars r_{ab} with $r_{ab} = r_{ba}$, and B' is a noise matrix with $\mathbb{E}\{B'\} = 0$.

The key steps of Algorithm 1 are the steps 4 and 5, in which the class information of labelled data are employed to

modify the selected eigenvectors of L_s, before they are used as feature vectors for the final clustering step. In order to understand the purpose of these key steps and to demonstrate the performance gain achieved by them, we consider a selected eigenvector v_a of L_s with $\lambda_a v_a = L_s v_a$, where λ_a is the associated eigenvalue of v_a. Recalling that $L_s = I - W_s$, we have

$$W_s v_a = (I - L_s)v_a = (1 - \lambda_a)v_a, \tag{4}$$

v_a is thus an eigenvector of W_s with the corresponding eigenvalue $1 - \lambda_a$. Following the discussion above, the data points in the same class are interchangeable on W_s, it is easily deduced that v_a can be expressed as

$$v_a = \mathbb{E}\{v_a\} + \beta_a = \sum_{b=1}^{k} c_{ab} j_b + \beta_a. \tag{5}$$

for some scalars c_{ab}, and β_a is a noise vector with $\mathbb{E}\{\beta_a\} = 0$. Evidently, β_a is useless to the recovery of the underlying data clusters, due to the fact $\mathbb{E}\{\beta_a\} = 0$. It is thus reasonable to assume that it produces no information when propagated though the graph, i.e., $\tilde{\beta}_a = W_s \beta_a$ is also non-informative noise.

Since labelled data within the same class as unlabelled ones share the same statistical properties, the quantities m_{ab} defined in Algorithm 1 are estimators for c_{ab} in (5); these estimators are asymptotically consistent as the number of labelled data grows. As such, the vector $\hat{v}_a$ computed in the step 5 can be written as

$$\hat{v}_a = \sum_{b=1}^{k} c_{ab} j_b + \sum_{b=1}^{k} (m_{ab} - c_{ab}) j_{[l]b} + D(j_{[u]})\beta_a. \tag{6}$$

To put it simply, $\hat{v}_a$ is an estimation of $\mathbb{E}\{v_a\}$ plus the projection of β_a on the unlabelled points. If c_{ab} are satisfyingly estimated by m_{ab}, we have

$$\tilde{v}_a = W_s \hat{v}_a \simeq W_s \left[\mathbb{E}\{v_a\} + D(j_{[u]})\beta_a\right]. \tag{7}$$

Finally, we obtain from (4), (5) and (7) that

$$(1 - \lambda_a)v_a \simeq \tilde{v}_a + W_s D(j_{[l]})\beta_a. \tag{8}$$

As discussed before, the term $L_s D(j_{[l]})\beta_a$ only introduces additional noise, v_a is thus less informative than $\tilde{v}_a$ as features of data points, suggesting that the semi-supervised spectral clustering method has a superior learning ability than spectral clustering.

3.3. Experimentation

In this section, we provide simulations conducted on synthetic and real datasets. For simplicity, we focus here on binary tasks. The synthetic data are generated from a Gaussian mixture model, i.e., $x_i \in \mathcal{C}_a \Leftrightarrow x_i \sim \mathcal{N}(\mu_a, C_a)$. The real datasets tested come from the MNIST database [14], a standard database of hand-written digits.

To guarantee the generality of the experimental results, we use simple settings. The similarities w_{ij} are computed with a Gaussian kernel function: $w_{ij} = \exp(-\|x_i - x_j\|^2/p)$ where p is the dimension of data vectors x_i. All possible numbers of selected eigenvectors are tested for the manifold-based method in order to find the best performance[2]. Even though performance gains are observed when using multiple eigenvectors for the proposed spectral clustering algorithm, notably on MNIST datasets, we report only results with one eigenvector.

The first purpose of the experimentation is to verify empirically that the proposed semi-supervised spectral clustering algorithm learns indeed from both labelled and unlabelled data, by showing that it surpasses the unsupervised spectral clustering on the same datasets, as theoretically claimed in Section 1. To illustrate this point, the accuracy[3] curves as a function of the ratio of labelled data $n_{[l]}/n$ are given in Figure 1 and Figure 2, where semi-supervised spectral clustering is shown to have a growing performance gain over spectral clustering as the ratio of labelled data increases.

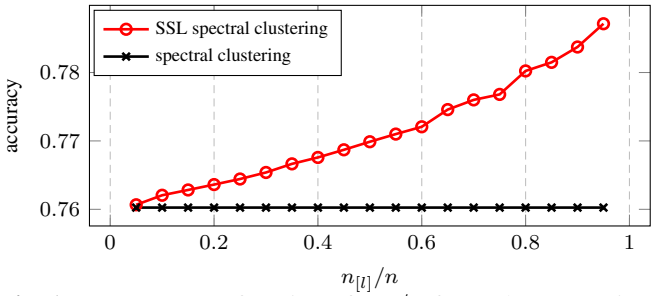

Fig. 1. Accuracy as a function of $n_{[l]}/n$ for 2-class Gaussian data vectors of dimension $p = 200$ with $\|\mu_1 - \mu_2\| = 2$ and $C_1 = C_2 = I_p$, $n = 600$, $n_{[l]1} = n_{[l]2}$, $n_{[u]1} = n_{[u]2}$. Averaged over 500 iterations.

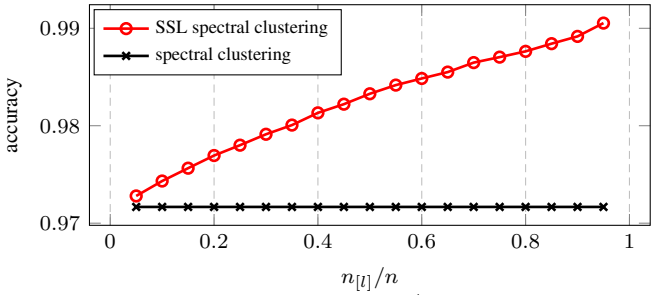

Fig. 2. Accuracy as a function of $n_{[l]}/n$ for 2-class MNIST data (6,9), $n = 600$, $n_{[l]1} = n_{[l]2}$, $n_{[u]1} = n_{[u]2}$. Averaged over 500 iterations.

After confirming the proposed method as a valid semi-supervised approach, we now move to compare it with the

[2]Thus the provided manifold performances can be seen as oracle ones

[3]The accuracy refers to the proportion of correctly classified unlabelled data.

two most classic graph-based semi-supervised techniques: the Laplacian regularization method [5] and the manifold-based method [9]. Figure 3 and Figure 4 show that unlike Laplacian regularization, the semi-supervised clustering algorithm and the manifold-based method both benefit from an increasing number of unlabelled data, with the manifold-based method consistently outperformed by the proposed approach. It should be mentioned that the extremely poor performances displayed by the Laplacian approach are due to the very small $n_{[l]}$ of the tested datasets.

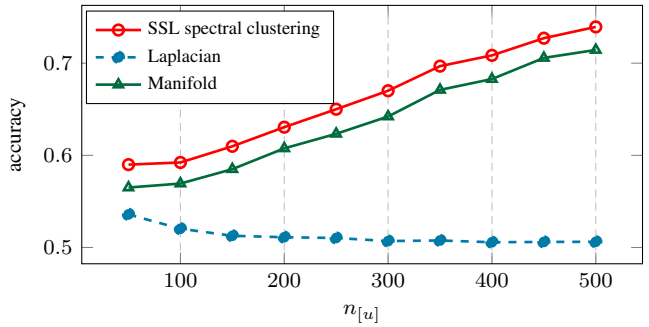

Fig. 3. Accuracy as a function of $n_{[u]}$ for 2-class Gaussian data vectors of dimension $p = 200$ with $\|\mu_1 - \mu_2\| = 2$ and $C_1 = C_2 = I_p$, $n_{[l]1} = n_{[l]2} = 5$, $n_{[u]1} = n_{[u]2}$. Averaged over 500 iterations.

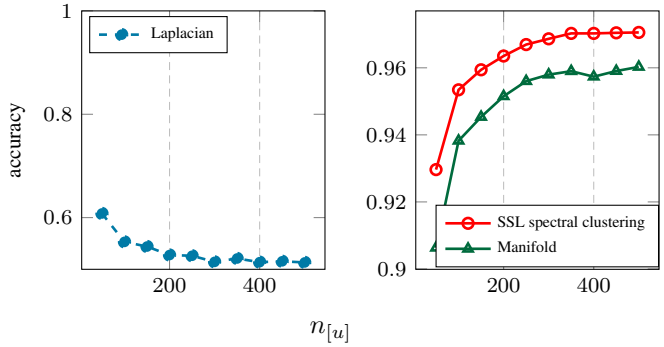

Fig. 4. Accuracy as a function of $n_{[u]}$ for 2-class MNIST data (6,9), $n_{[l]1} = n_{[l]2} = 2$, $n_{[u]1} = n_{[u]2}$. Averaged over 500 iterations.

4. CONCLUDING REMARKS

Mixing labelled and unlabelled data makes the design of semi-supervised techniques an interesting but difficult question. Some common semi-supervised methods, such as Laplacian regularization, do not really learn from both types of data. This article proposes a semi-supervised spectral clustering method which is consistent with respect to both labelled and unlabelled data, with significantly better performance over other classical graph-based semi-supervised approaches.

The semi-supervised spectral clustering algorithm presented in this article contains some new critical steps, entirely different from existing graph-based semi-supervised strategies. A more systematic analysis, as a follow up of [1, 15], on the asymptotic performances of the proposed method, is however needed to *quantitatively* evaluate further the gains as well as the limitations of the approach, which will be the subject of forthcoming investigations.

5. REFERENCES

[1] Xiaoyi Mai and Romain Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *arXiv preprint arXiv:1711.03404*, 2017.

[2] Boaz Nadler, Nathan Srebro, and Xueyuan Zhou, "Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data," in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2009, pp. 1330–1338.

[3] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet, "Consistency of spectral clustering," *The Annals of Statistics*, pp. 555–586, 2008.

[4] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[5] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al., "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, 2003, vol. 3, pp. 912–919.

[6] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.

[7] Olivier Delalleau, Yoshua Bengio, and Nicolas Le Roux, "Efficient non-parametric function induction in semi-supervised learning.," in *AISTATS*, 2005, vol. 27, p. 100.

[8] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, *Semi-supervised learning*, MIT press, 2006.

[9] Mikhail Belkin and Partha Niyogi, "Using manifold stucture for partially labeled classification," in *Advances in neural information processing systems*, 2003, pp. 953–960.

[10] Akshay Gadde, Aamir Anis, and Antonio Ortega, "Active semi-supervised learning using sampling theory for graph signals," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 492–501.

[11] Aamir Anis, Aly El Gamal, Salman Avestimehr, and Antonio Ortega, "Asymptotic justification of bandlimited interpolation of graph signals for semi-supervised learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5461–5465.

[12] Romain Couillet, Florent Benaych-Georges, et al., "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[13] Xiaojin Zhu and Zoubin Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep., Citeseer, 2002.

[14] Yann LeCun, Corinna Cortes, and Christopher JC Burges, "The mnist database of handwritten digits," 1998.

[15] Romain Couillet and Florent Benaych-Georges, "Kernel spectral clustering of large dimensional data," *arXiv preprint arXiv:1510.03547*, 2015.