

# Representation of functions on big data associated with directed graphs

Charles K. Chui\* and H. N. Mhaskar† and Xiaosheng Zhuang‡

## Abstract

This paper is an extension of the previous work of Chui, Filbir, and Mhaskar (Appl. Comput. Harm. Anal. 38 (3) 2015:489-509), not only from numeric data to include non-numeric data as in that paper, but also from undirected graphs to directed graphs (called digraphs, for simplicity). Besides theoretical development, this paper introduces effective mathematical tools in terms of certain data-dependent orthogonal systems for function representation and analysis directly on the digraphs. In addition, this paper also includes algorithmic development and discussion of various experimental results on such data-sets as CORA, Proposition, and Wiki-votes.

## 1 Introduction

In this section, we first give a very brief summary on the recent progress of the manifold and (undirected) graph approaches for processing high-dimensional (numeric) data, and then discuss the need for directed graphs (called digraphs). We will also discuss the need for processing non-numeric data associated with digraphs, by using the endocrine network of the human body as an example. The objective of this paper is to develop a theory, along with a demonstration of some methods and algorithms, for the representation of functions on non-numeric data for the digraph paradigm, based on a data-dependent orthogonal system, with associated filters, to be introduced in this paper. Since our approach is different from other studies in the literature, we will also give a toy example in this introduction section to illustrate the main idea of our approach. The organization of our presentation is outlined in the last paragraph of this section.

An earlier popular approach for processing high-dimensional numeric data is to consider that the data-set lies near or on some (unknown) lower-dimensional manifold and to apply such mathematical tools as manifold learning, manifold Laplacian, diffusion maps and diffusion wavelets, to extract the data geometry and other data structures for data analysis. The interested reader is referred to the special issue [3] of the journal, Applied and Computational Harmonic Analysis (ACHA), for some in-depth study in this research direction. In this regard, function approximation on such data-defined manifolds was also investigated in some depth (see, for example, [27, 11, 12, 30, 31]). On the other hand, since the discrete graph Laplacian well approximates the (continuous) manifold Laplacian (see [41] and the references therein), and since the subject of spectral graph theory (see [7]) has already been a well established research area, it was almost immediately clear to at least a handful of researchers that perhaps high-dimensional data could be understood, analyzed, and processed more fruitfully by associating the data directly with graphs, without embedding them into a lower-dimensional manifold. Indeed, many appealing aspects appearing in the analysis on the data-defined manifolds, such as the Hodge Laplacian, various properties of the Green kernel, and useful inequalities, can be analyzed extensively in the context of spectral graph theory (see, for example, [35, 43, 6, 26]). In addition, function approximation on graphs has also been discussed in the literature. For instance, it is shown in [13] that the solution of the wave equation corresponding to an edge-based Laplacian satisfies the property known as finite speed of wave propagation, and this, in turn, is equivalent to the Gaussian upper bound condition for small values of  $t$  [40, 11], so that the results in [27, 29] regarding function approximation

\*Department of Statistics, Stanford University, Stanford, CA 94305. The research of this author is supported by ARO Grant W911NF-15-1-0385. email: ckchui@stanford.edu.

†Department of Mathematics, California Institute of Technology, Pasadena, CA 91125; Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711. The research of this author is supported in part by ARO Grant W911NF-15-1-0385. email: hrushikesh.mhaskar@cgu.edu.

‡Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong. The research of this author is supported in part by the Research Grants Council of Hong Kong (Project No. CityU 11304414) and City University of Hong Kong (Project No.: 7200462 and 7004445). email: xzhuang7@cityu.edu.hk.

and harmonic analysis are directly applicable. It is noted, however, that the associated graphs mentioned above are undirected graphs.

For big data analysis and processing, the associated graphs of interest are often directed graphs (or digraphs). For instance, digraphs are used effectively to model social networks, technological networks, biological and biomedical networks, as well as information networks [34], with probably the most well-known example being the page-rank algorithm, where the nodes consist of urls of different web pages. Another example is the urban road network, where an intersection is represented by a vertex and a road section between adjacent intersections is denoted by an edge conforming to traffic flow direction on the road (e.g. [19]). The interested reader is referred to the special issue [36] of *Internet Mathematics* for an introduction to biological networks, with two papers [10, 8] dealing directly with the question of finding the correct digraph structures. For biomedical informatics, the most important big data are arguably those of the human physiology; namely, the data generated by the physical, mechanical, and bio-chemical functions of humans and of human organs and systems. In particular, among the human body network systems, the two major ones, being the endocrine and nervous systems, both of which can be viewed as digraphs. While it is perhaps easier to understand, even by the layman, that our nervous system is a complex network of nerves and cells that carry messages from the brain and spinal cord to various parts of the body, the endocrine network is far more complicated. A very brief description of the endocrine system is that it is made up of a network of glands that secrete various chemical signals, called hormones, that travel in our blood vessels to regulate our bodily functions, helping in the control of our growth and development, internal balance of our entire body system, body energy levels (or metabolism), reproduction, as well as response to stress and injury. We will return to this discussion after briefly introducing the concept of non-numeric data.

An advantage of the (undirected or directed) graph approach over the manifold approach is that both numeric and non-numeric data can be (either separately or together) dealt with directly. Non-numeric data are also called qualitative data or categorical data, since they are used to indicate the quality of an object from the observed data (usually by using a bar chart) and showing the various categories in which the object belongs from the data (usually by using a pie chart). The usual techniques for working with numerical data cannot be used for non-numeric data. For example, the diffusion matrix commonly used for representing numerical data as a graph cannot be constructed in the same way for non-numeric data.

Returning to the above discussion of the human body network systems, there are ten major glands that constitute the endocrine network, including the pituitary gland (also called the master gland, located at the base of the brain) that controls and regulates hormone secretion of the other glands. Among these major glands, the pair of adrenal glands, with one sitting atop each kidney, are essential for human life. However, until today, there is still no reliable method for acquiring and analyzing the adrenal hormone data. Blood and urine tests are still commonly used, at least for preliminary screening to establish the case. The acquired information is non-numeric, in that adrenal insufficiency is (usually) determined by 7 observations, namely: sodium level, potassium level, blood pressure, glucose level, aldosterone level, cortisol level and ACTH level, in terms of only 3 qualitative marks: low, high, and normal, with categorial classification depending on personal genetics and medical history. If necessary, ultrasound or X-ray imaging of the abdomen to view the adrenal glands to establish primary adrenal insufficiency (called Addison's disease), and perhaps followed by CT scan to view the size and shape of the pituitary gland, if adrenal insufficiency could be secondary. When non-numeric data are associated with (undirected or directed) graphs, the data are represented as information in the nodes of the graph. The nature of this information is not critical to the analysis, but is used only to determine the edge weights of the graph. In the second paragraph, we have already discussed the topic of approximation of functions on numeric data associated with graphs. Here we mention that representation of functions on non-numeric data has also caught some attention recently, for instance, by Smale and his collaborators [38], in introducing a mathematical foundation of molecular immunology, arguing that the study of peptide binding to some appropriate alleles can be thought of as a problem of approximating an unknown function on certain strings that represent the relevant alleles.

In the study of data associated with graphs, we note that digraphs are much less studied than undirected graphs in the literature. In fact, to the best of our knowledge, all current approaches to digraphs involve, in essence, the construction of an undirected graph that captures different features of the underlying digraph (see, for example, the recent surveys [28] by Malliaros and Vazirgiannis or [21] by Jia et. al.). For example, the Hodge Laplacian of a digraph is a symmetric matrix [26], and the weighted adjacency matrix of the graph Laplacian introduced by Chung [5] for a digraph is given by a symmetric matrix as well, although an asymmetric version, called dilaplacian, has been discussed by Li and Zhang in [25]. The concept of generalized Cheeger constant that plays an important role in graph partitioning algorithms, as introduced in [25] also utilizes a symmetrized version of the dilaplacian. In the current paper, we propose an alternative way to develop harmonic analysis on digraphs by extending the undirected graph approach from our earlier paper [4]. Our main idea is to represent a digraph by an asymmetric matrix  $W$

(such as the weighted adjacency matrix, the dilaplacian, etc.), and observe that if the singular values of  $W$  are all distinct, then  $W$  can be recuperated uniquely from the symmetric matrices (equivalently, weighted undirected graphs)  $WW^*$  and  $W^*W$ , or by their degree reduced forms. When the singular values are not distinct, the matrix is still a limit of matrices with distinct singular values. In other words, our viewpoint is that a digraph is a pair of undirected graphs. In this way, we can apply the well-known techniques for analysis on (undirected) graphs for developing an analysis on digraphs – in principle, showing in fact that analysis on digraphs is trivial, once it is developed for (undirected) graphs. Therefore, in this paper we will represent a digraph as two (undirected) graphs and apply the theory and methods developed in our paper [4].

To demonstrate this idea, let us first consider a toy example with the digraph shown in Figure 1. Here, the weighted adjacency matrix  $W$  is generated randomly but fixed throughout this example. In the general discussion of this paper, we will identify a digraph and its weighted adjacency matrix accordingly. If  $W$  is the (weighted) adjacency matrix, we apply a variant of the algorithm for hierarchical clustering described by Chaudhury and Dasgupta [2] for both  $WW^*$  and  $W^*W$ , where the Euclidean distance is replaced by the graph distance on these graphs. The resulting trees are shown in Figure 2. Although we do not show all the leaves of the two trees for the convenience of presentation, each of the trees corresponds to one node in the digraph  $W$ . Conversely, each node of  $W$  appears as a leaf on each of the two trees. Using the edge weights of  $W$ , we can easily construct a filtration for each of the trees as described in [4], so that each of these leaves is a sub-interval of  $[0, 1)$ . Suppose a node on  $W$  appears as the interval  $[a, b)$  on the tree corresponding to  $WW^*$  and as the interval  $[c, d)$  on the other tree. Then we will consider the node as the rectangle  $[a, b) \times [c, d) \subseteq [0, 1) \times [0, 1)$ , or according to convenience, any point on that rectangle. In particular, the digraph  $W$  can now be viewed as a set of rectangles in a partition of  $[0, 1) \times [0, 1)$ , where each horizontal stripe as well as each vertical stripe contains at least one point of  $W$ . This is illustrated in Figure 3. In the sequel,  $I^2$  will denote  $[0, 1) \times [0, 1)$ .

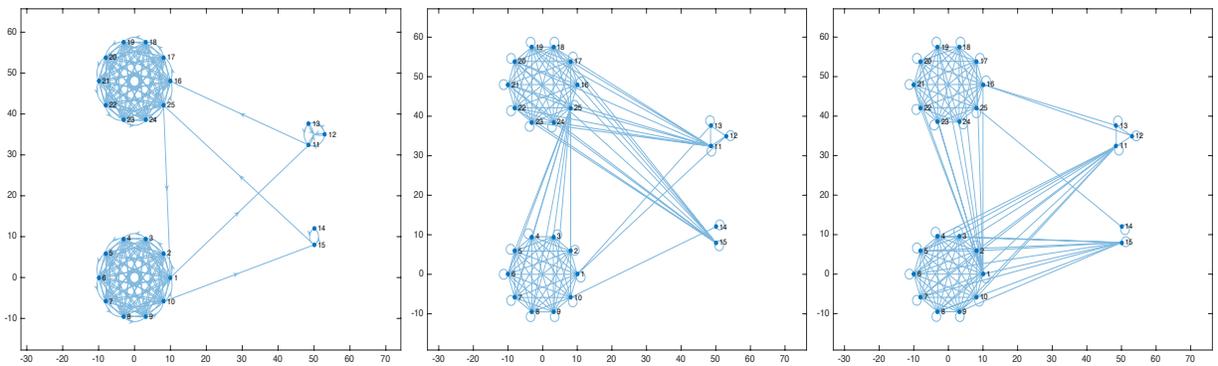


Figure 1: A simple, strongly connected digraph  $W$  (left), and the equivalent pair of graphs,  $WW^*$  (middle) and  $W^*W$  (right).

One major advantage of our approach is the following. It is observed in many examples, including the ones which we will study in this paper, that digraphs are typically highly disconnected. In contrast, the spectral theory for digraphs assumes a strongly connected (undirected) graph (e.g., [5, 25]). Our approach does not make any such assumption. We simply take the connected components of each of the two (undirected) graphs as the children of the root for the tree corresponding to the graph, and use clustering on each of these connected components, enabling us to use spectral clustering if necessary. On the other hand, every point on the square  $I^2$  does not correspond to a point on the digraph. Therefore, the theory of function approximation and harmonic analysis on digraphs in this paradigm must necessarily be totally data-driven, including data-dependent orthogonal systems and filters. In this paper, we will describe this theory in abstract.

The outline of this paper is as follows. In Section 2, we give examples of some of the algorithms used for hierarchical clustering in order to represent a digraph as a pair of filtrations. We will also discuss standard criteria to evaluate the quality of this clustering. These algorithms are tested in the case of three data-sets, an unlabelled data-set (Wiki-votes), a labeled data-set that is not hierarchically labeled (Proposition data), and a hierarchically labeled data-set (CORA). Each of these data-sets is non-numeric, and we make no effort to create a numeric data set out of them. The results are reported in Section 3. It is not our intention to investigate either the data-sets or the algorithms in their own right, but only to demonstrate that the choice of the algorithm can lead to a different structure for the digraph with vertices represented as elements of  $I^2$ . Therefore, unlike the study in classical harmonic analysis, not only the orthogonal system on the digraph, but also the very notion of smoothness and the

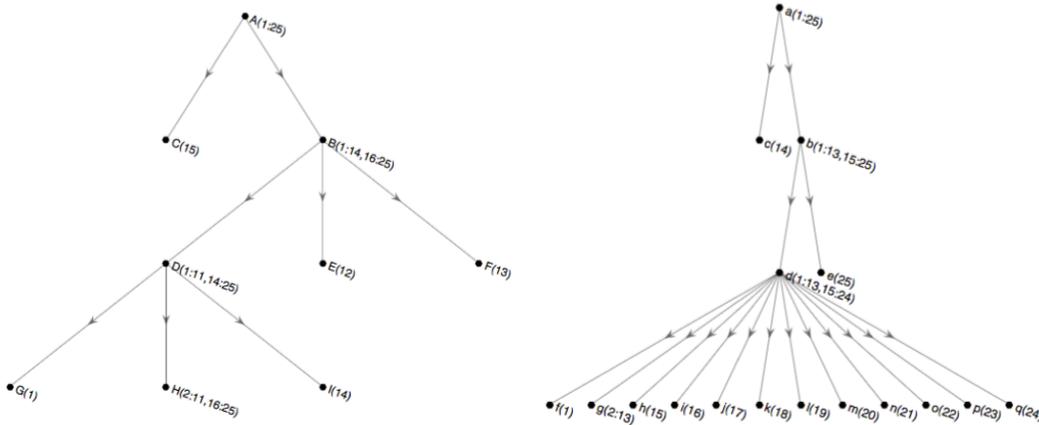


Figure 2: The trees corresponding to the example in Figure 1 using a variant of the algorithm in [2] with  $WW^*$  on left, and  $W^*W$  on right. Each node corresponds to either a cluster of nodes in  $W$  as indicated in the parenthesis, or a node itself, also indicated in parenthesis. Thus,  $q(24)$  is a leaf, corresponding to the node 24 in  $W$ ;  $H(2 : 11, 16 : 25)$  is a cluster with nodes 2–11 and 16–25. These nodes themselves are assumed to be the children of  $H$ .

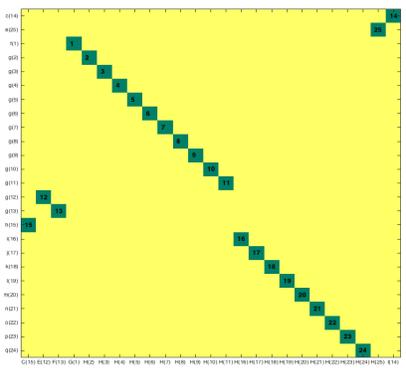


Figure 3: The digraph  $W$  of Figure 1 as rectangles in a partition of  $I^2 = [0, 1) \times [0, 1)$ , according to the trees in Figure 2.

various filters, must necessarily depend upon the data as well as the particular structure of the digraph obtained via the clustering algorithms. The theory of function approximation and harmonic analysis will be developed in Section 4.

## 2 Implementation and testing

Our paper [4] is motivated in part by the observation that a string as in [38] can be represented via arithmetic coding as a vertex on a tree. Our work is motivated also by the work of Coifman, Gavish, and Nadler [15, 16]. Their approach starts with clustering the vertices of the graph into several “folders”, followed by organizing these folders as another weighted graph, and repeating this process till only one folder remains. This organization generates a tree of sub-folders, sub-sub folders, and so on, till the lowest level that consists of the cluster-folders of the original vertices. In the paper [4], we have therefore assumed that the (undirected) graph has been converted to a tree using an appropriate clustering algorithm. In this section, we wish to extend this idea to digraphs. As explained in the introduction, a digraph can be viewed as a pair of undirected graphs. Clustering algorithms applied to each of these yield two corresponding trees, as well as a meaningful clustering of the digraph itself. The purpose of this section is to illustrate this concept using some concrete clustering algorithms and data sets.

After reviewing certain graph theory preliminaries, we review the algorithms we used for clustering (Sub-

section 2.1), as well as assessment tools for the quality of clustering (Sub-section 2.2).

## 2.1 Algorithms

For the convenience of the reader, we first review certain preliminaries about digraphs relevant to this paper.

A *digraph* is an ordered pair  $(V, W)$ , where  $V$  is a nonempty set, and  $W : V \times V \rightarrow [0, \infty)$ . Each element of  $V$  is called a *vertex* or a *node*. If  $u, v \in V$ , then there is an *edge from  $u$  to  $v$*  with weight  $W(u, v)$  if  $W(u, v) > 0$ . This fact is often denoted by  $(u, v) \in W$ . The digraph is *undirected* if  $W$  is symmetric. The term graph (or undirected graph) refers to an undirected digraph. A digraph  $(V, W)$  is a *tree* if there is a distinguished vertex  $u^*$  (*the root*) such that there is no edge from any other vertex to  $u^*$ , and for every  $v \in V \setminus \{u^*\}$ , there is a unique  $u$  such that  $(u, v) \in W$ . The vertex  $u$  is then called the *parent* of  $v$ , and  $v$  the *child* of  $u$ . We will follow the custom in computer science to treat the vertices as pointers to a record of information; for example, an entire file could be considered as the information stored in a vertex of some graph. For the convenience of exposition, we will often describe the vertex by the information it points to.

If  $V$  is a finite set, then  $W$  is represented by a matrix, called the *weighted adjacency matrix*. If the values of  $W$  are all in  $\{0, 1\}$  then  $W$  is called an *adjacency matrix*. In this section, we will assume  $V$  to be finite, and denote the transpose of  $W$  by  $W^*$ .

For a digraph  $(V, W)$ , the *underlying undirected graph* is given by  $(V, W_0)$ , where  $W_0 = (W + W^*)/2$ . If  $I$  is the identity matrix of the same size as  $W$ , the *extended graph*  $(V, W_e)$  with  $W_e = I + W$  is the same graph as  $(V, W)$  except for a new (or enhanced) self-loop inserted at each vertex. The *pre-symmetrized (ES) graph* (respectively, *post-symmetrized (OS) graph*) for  $(V, W)$  is defined by  $(V, W_{ES})$  (respectively,  $(V, W_{OS})$ ), where  $W_{ES} = W_e W_e^*$  and  $W_{OS} = W_e^* W_e$ . In the context of citation graphs, these have been called bibliographic coupling and co-citation graphs respectively [37].

If  $u, v \in V$ , then a *path from  $u$  to  $v$*  is an ordered set  $u_0 = u, u_1, \dots, u_n, u_{n+1} = v$  such that there is an edge between  $u_i$  and  $u_{i+1}$  for  $i = 0, \dots, n$ ; the weight of this path is  $\sum_{j=0}^n W(u_j, u_{j+1})$ . The *distance* from  $u$  to  $v$  is the minimum of the weights of all the paths from  $u$  to  $v$ , defined to be  $\infty$  if no such path exists. The distance matrix  $d_W$  is the matrix whose  $(u, v)$  entry is the distance from  $u$  to  $v$ .

A (undirected) graph is *connected* if for any  $u, v \in V$ , there exists a path from  $u$  to  $v$ . A digraph is *weakly connected* if the underlying undirected graph is connected. A *weak component* of a digraph is a subgraph whose vertices form a maximal weakly connected subset of the vertices of the original digraph. It is not difficult to show that if  $(V, W)$  is weakly connected, then the ES and OS graphs for  $(V, W)$  are connected (undirected) graphs.

Each of the algorithms we discuss below have the format described in Algorithm 1, which we will call Twin-tree construction algorithm (TWT).

---

**Algorithm 1 TWT:** A general top level description of the algorithms in this paper.

---

- a) Input a digraph  $(V, W)$ .
  - b) Let  $\{(V_j, W_j)\}_{j=1}^M$  be the weakly connected components of the extended graph  $(V, W_e)$ . With the digraph itself as the root, we construct two trees, with the ES (respectively, OS) graphs for  $(V_j, W_j)$ ,  $j = 1, \dots, M$  as leaves.
  - c) Taking each of the leaves above as roots, we construct subtrees by applying various hierarchical clustering algorithms with the connected graphs represented by these leaves.
  - d) The resulting trees are denoted by  $\mathcal{T}_{ES}^*$ , respectively,  $\mathcal{T}_{OS}^*$ . The symbol  $\mathcal{T}^*$  will denote either of these.
- 

In the sequel, we will describe our algorithms for a connected (undirected) graph; e.g., the leaf of the tree obtained in Step b of the algorithm TWT corresponding to the largest connected component of the corresponding undirected graph. Rather than complicating our notations, we will abuse the notation, and write  $G = (V, W)$  for this graph, keeping in mind that in practice, this is really one of the leaves generated at Step b of the algorithm in Figure 1. We denote the tree with root at  $G$ , resulting from the hierarchical clustering algorithm by  $\mathcal{T}_G$ .

Before describing the algorithms which we used extensively, we comment about some algorithms which we could not pursue vigorously.

The first algorithm to generate hierarchical tree structure from a connected graph is a variant of the algorithm described by Chaudhury and Dasgupta in [2]. This algorithm is developed primarily for clustering in high dimensional Euclidean spaces to achieve theoretically proven consistency results. In all the examples which we have studied in Section 3, the data is not numerical. Therefore, we replaced the Euclidean distance by the graph distance on  $G$ , as described in the introduction. However, we found this algorithm to be too slow for the examples.

The second one apparently highly cited algorithm is the Markov Clustering algorithm (MCL) [44, 39]. This is not a hierarchical clustering algorithm, and therefore, we did not pursue this further.

The other one is the MGL (Multiclass Ginzburg-Landau) algorithm described in the paper [14]. Since it is similar to the MBO algorithm and MBO outperforms MGL in most of the cases, we therefore use only the MBO algorithm.

We now describe a set of three algorithms (NHC, MLL, and MBO) which we used in our examples as follows.

We developed an unsupervised, hierarchical clustering algorithm based on an idea described briefly in [15, 16], that does not require the eigen-decomposition of a matrix. This variant is described in Algorithm 2. We will refer to this algorithm as NHC (Non-spectral Hierarchical Clustering) and note that the algorithm is easy to modify for semi-supervised learning by choosing the initial centers to include the labeled data points.

---

**Algorithm 2 NHC:** An unsupervised, hierarchical, eigen-decomposition free clustering algorithm.

---

- a) **Input:** undirected graph  $G = (V, W)$ .  $K = (k_1, k_2, \dots, k_L)$  with  $1 < k_1 < k_2 < \dots < k_L < N$ , where  $N$  is the number of vertices in the graph.
  - b) **Output:** tree structure of level  $0, 1, \dots, L + 1$ , where level 0 is the root ( $V$ ), level  $L + 1$  is the leaves of vertices, and in between are  $k_l$  clusters at level  $l$  for  $l = 1, \dots, L$ .
  - c) **Main Steps:**
    - 1: Initialization:  $\ell \leftarrow L$ ,  $V_0 \leftarrow V$ , and  $A_0 \leftarrow W$ .
    - 2: **while**  $\ell > 1$  **do**
    - 3:   compute graph distance matrix  $d_{A_0}$ .
    - 4:   **while** true **do**
    - 5:     randomly choose  $k = k_\ell$  vertices  $u_1, \dots, u_k$  from  $V_0$  as centers.
    - 6:     construct cluster  $C_j$  for  $j = 1, \dots, k$ :  $v \in V_0$  belongs to  $C_j$  if  $j = \operatorname{argmin}_{1 \leq i \leq k} d_{A_0}(u_i, v)$ .
    - 7:     update the centers: for each  $C_j$ , find a new center  $u \in C_j$  such that  $\sum_{v \in C_j} d_{A_0}(u, v)$  is minimal.
    - 8:     break if all centers remain the same.
    - 9:   **end while**
    - 10:   construct a new graph  $G_1 = (V_1, A_1)$  of  $k$  vertices by the adjacent matrix  $A_1$  of size  $k \times k$  as follows:  
 $A_1(i, j) = \sum_{i \in C_i, j \in C_j} A_0(i, j)$ ,  $1 \leq i, j \leq k$ .
    - 11:   update  $V_0 \leftarrow V_1$ ,  $A_0 \leftarrow A_1$ , and  $\ell \leftarrow \ell - 1$ .
    - 12: **end while**
- 

At the other end of the spectrum, we used an algorithm (Diffuse-Interphase Method) described by Garcia-Cardona, et. al. in [14] which in turn is a modification of the well known MBO algorithm based on a graph Laplacian. This method can be used for hierarchical clustering only if the class labels are also organized hierarchically. Otherwise, we use this algorithm for the primary clustering, and use the coarse-graining ideas in [22] to construct the remaining levels of the tree bottom-up in an unsupervised manner. We will refer to this algorithm as MBO.

In between the two, we used the algorithm described by Lafon and Lee in [22]. This is also a algorithm based on the graph Laplacian, but can be used both in the unsupervised setting (where the centers for clustering are chosen randomly) and in the semi-supervised setting (where the centers for clustering are chosen to be among the training data). We will refer to this algorithm as MLL (Modified Lafon-Lee).

In both of the MLL and MBO, we used the graph Laplacian. In our applications, it was not necessary to construct a diffusion matrix as in [14, 22]. We only need the adjacency matrix of the graph as an input. In each case, the tree  $\mathcal{T}_G$  has 4 levels. The root is at level 0 containing all vertices and the leaves are vertices at level 3. We cluster all vertices to  $k_2$  clusters at level 2 and cluster them further into  $k_1$  clusters at level 1 with  $(k_1, k_2)$  preassigned. Thus, a vertex  $v$  of  $\mathcal{T}_G$  at level 2 is a cluster of the vertices in the graph  $G$  which are children of  $v$  in  $\mathcal{T}_G$ , and similarly, a vertex  $u$  of  $\mathcal{T}_G$  at level 1 is a cluster comprising its children on the tree.

## 2.2 Quality of clustering

It is clear that any harmonic analysis/function approximation scheme based on tree polynomials would depend upon the tree itself or equivalently on the quality of clustering used to generate the same. The objective of this paper is only to illustrate the concepts, not to point out an optimal clustering algorithm. Therefore, rather than using the usual measurement of accuracy of classification for evaluating our experiments, we will use measurements for the quality of clustering at different levels.

As explained in the introduction, each node on the digraph  $W$  is interpreted as a rectangle contained in  $[0, 1) \times [0, 1)$ . The non-leaf nodes on the graphs would likewise be represented as rectangles as well, with each such node

being the union of rectangles corresponding to its children. These non-leaf nodes at different levels will be considered as clusters at that level.

We will use two measurements for the quality of clustering in digraphs in a hierarchical manner. To do so, we first make sure that the number of levels in the two trees corresponding to the digraph is the same. Suppose the tree  $\mathcal{T}_{W_{ES}}$  has  $L$  levels and the tree  $\mathcal{T}_{W_{OS}}$  has  $L' > L$  levels. Then we treat each node at level  $L$  in  $\mathcal{T}_{W_{ES}}$  as its own leftmost child, and continue this way until the tree  $\mathcal{T}_{W_{ES}}$  has  $L'$  levels as well. Equivalently, since all the nodes at level  $L$  in  $\mathcal{T}_{W_{ES}}$  are leaves, any cluster at a level  $> L$  is just a cluster according to  $\mathcal{T}_{W_{OS}}$ . A cluster at level  $\ell$  is then a rectangle in the partition of  $I^2$  corresponding to the trees truncated at level  $\ell$ .

For unsupervised learning, we will use a measurement called modularity metric as described in [28]. Various algorithms are recently designed to optimize this measure, for example, [20]. This metric is designed to measure the number of edges that lie within a cluster compared to the expected number of edges in a random digraph with the same in/out degree distribution. If  $k_i^{\text{in}}$ ,  $k_i^{\text{out}}$  represent the indegree, respectively the outdegree of node  $i$  in  $W$  (more precisely, the sum of weights on the incoming, respectively outgoing, edges at  $i$ ), we assume that in a random digraph with the same connectivity, an edge from  $i$  to  $j$  will exist with probability  $k_i^{\text{out}} k_j^{\text{in}} / m$ , where  $m$  is the total weight of the incoming/outgoing edges in the digraph; i.e., sum of the entries in  $W$ . Then the modularity metric introduced by Arenas, et. al. in [1] is defined by

$$\mathcal{M} = \frac{1}{m} \sum_{i,j} \left( W_{i,j} - \frac{k_i^{\text{out}} k_j^{\text{in}}}{m} \right) \delta(C_i, C_j), \quad (2.1)$$

where  $\delta(C_i, C_j)$  is 1 if the nodes  $i$  and  $j$  both belong to the same cluster  $C = C_i = C_j$ , and 0 otherwise. In our implementation of this metric, we will consider the nodes at each level of the trees as the clustering at that level.

In the semi-supervised setting, we used the  $F$ -measure described in [37]. If  $\{C_1, \dots, C_M\}$  are the obtained clusters in the digraph from certain clustering algorithm, and  $\{L_1, \dots, L_n\}$  is a partition of the nodes according to the (ground-truth) class labels (i.e.,  $L_j$  is the set of all nodes in  $W$  with the class label  $j$ ), then one defines

$$F(C_i) = 2 \max_{1 \leq j \leq n} \frac{|C_i \cap L_j|}{|C_i| + |L_j|}.$$

the (micro-averaged)  $F$ -measure is then defined by

$$\mathcal{F} = \frac{\sum_i |C_i| F(C_i)}{\sum_i |C_i|}. \quad (2.2)$$

In Sub-section 4.2, we use the confusion matrix (see (4.31)) to measure the approximation power of different algorithms using our framework.

### 3 Data sets and results

We present our results for the (1) CORA data set (2) Proposition data set, and the (3) Wiki-votes data set.

Each of these data sets contains only one large weakly connected component while others are of small size. In our exposition, we focus only on the largest weakly connected component as the leaf obtained in Step b of the algorithm in Figure 1. The same algorithms can be applied to the other weakly connected components. If these components are too small, we may treat their vertices as the children of the  $\mathcal{T}^*$ -vertices corresponding to these components.

By abuse of notation as before, let  $G = (V, W)$  be the subgraph with respect to the largest weakly connected component,  $W_{ES}$ ,  $W_{OS}$  be the ES (respectively, OS) graphs of  $(V, W)$ , and  $\mathcal{T}_{W_{ES}}$  (respectively,  $\mathcal{T}_{W_{OS}}$ ) be the resulting 4 level trees with  $G$  at its root.

For each of the methods (NHC, MLL, MBO), we randomly pick  $p\%$  of the data as training data (semi-supervised learning (SSL) while 0% means unsupervised learning (USL)) and perform the clustering algorithms. For the method MBO, we used 50 eigenvectors, the time step is 0.01, the stop criterion is  $10^{-3}$ , and the weight constant for the fidelity term is 50. For the algorithm MLL, we used 30 significant eigenvectors and ‘‘time parameter’’  $t = 1$ . We compute the modularity metric and  $F$ -measure for each of the levels as described in the introduction. In view of the random choices of centers in both unsupervised (USL) and semi-supervised (SSL) settings, we computed these measurements for each given  $p\%$  over 30 trials, and the modularity metric  $\mathcal{M}$  and  $F$ -measure  $\mathcal{F}$  are average over these 30 trials. Note that for the Wiki-votes data set, we can only compute the modularity metric.

### 3.1 The data set CORA

We worked with the CORA research paper classification data set downloaded from <https://people.cs.umass.edu/~mccallum/data.html>. The data set comprises a digraph with 225,026 publications as vertices, and edge from  $i$  to  $j$  means that paper  $i$  cited paper  $j$ . These publications are from several areas of computer science and information theory. The subject area of each publication is given at two levels; e.g., artificial intelligence/vision, artificial intelligence/agent, artificial intelligence/DNP. Out of the entire data set, only 28,135 are labeled. We considered only the subgraph whose vertices are from this labeled data set. Altogether there are 70 classes at the most refined level, which are subgrouped into 10 classes, yielding a hierarchically labeled data set. There are 4,070 weakly connected components, and 22,985 strongly connected components, most of which are singletons. Thus, the digraph is highly disconnected. The largest weakly connected component of  $G$  contains 23,567 vertices while other weakly connected components contain at most 12 vertices. For each of  $W_{ES}$  and  $W_{OS}$  from the largest weakly connected component, we cluster all vertices to 70 clusters at level 2 and then further cluster them to 10 clusters at level 1.

For this data set, we can perform both unsupervised learning and semi-supervised learning methods. The results are given in Table 1. From the table, in terms of the  $F$ -measure and modularity metric, semi-supervised method MBO performs better than the other two methods of NHC and MLL, especially when the size of the training data is small ( $\leq 30\%$ ). It is interesting to note that the best  $F$ -measure reported in [37] for this data set is 0.36 at level 2, while the the algorithm MBO applied with 40% training data yields a better  $F$ -measure both at levels 1 and 2.

$\mathcal{F}$	Trains (%)	0 (USL)	10	20	30	40	50	60	70	80	90
NHC	Level 2 (70)	0.10	0.13	0.17	0.23	0.31	0.41	0.51	0.62	0.74	0.87
	Level 1 (10)	0.15	0.49	0.49	0.54	0.59	0.65	0.71	0.77	0.84	0.92
MLL	Level 2 (70)	0.11	0.12	0.15	0.21	0.29	0.38	0.49	0.60	0.73	0.86
	Level 1 (10)	0.42	0.40	0.40	0.45	0.48	0.52	0.58	0.67	0.77	0.87
MBO	Level 2 (70)	N.A.	0.21	0.27	0.33	0.40	0.49	0.58	0.68	0.78	0.89
	Level 1 (10)	N.A.	0.54	0.56	0.61	0.61	0.66	0.73	0.79	0.86	0.93
$\mathcal{M}$	Trains (%)	0 (USL)	10	20	30	40	50	60	70	80	90
NHC	Level 2 (70)	0.34	0.32	0.32	0.32	0.34	0.36	0.40	0.45	0.49	0.55
	Level 1 (10)	0.35	0.30	0.33	0.34	0.36	0.38	0.42	0.46	0.50	0.55
LL	Level 2 (70)	0.23	0.23	0.24	0.26	0.29	0.32	0.37	0.42	0.47	0.55
	Level 1 (10)	0.22	0.24	0.23	0.25	0.28	0.31	0.36	0.40	0.46	0.53
MBO	Level 2 (70)	N.A.	0.51	0.48	0.45	0.44	0.44	0.45	0.48	0.51	0.56
	Level 1 (10)	N.A.	0.52	0.49	0.47	0.48	0.48	0.50	0.51	0.53	0.56

Table 1: CORA Data: the  $F$ -measure  $\mathcal{F}$  and modularity metric  $\mathcal{M}$  using NHC, MLL, and MBO for given 0% (USL method), 10%, 20%, ..., 90% training data at level 2 ( $k_2 = 70$  clusters) and level 1 ( $k_1 = 10$  clusters), respectively. All results are average over 30 trials.

### 3.2 The Proposition data set

This data set is described in detail in [45, 42]. The November 2012 California ballot contained 11 initiatives, or propositions, on a variety of issues, including state taxation, corrections, and food labelling among others. The data consist of Twitter posts related to initiatives, grouped according to different propositions. For each proposition, the data is a directed graph with edge from  $i$  to  $j$  if the tweet originated from user  $i$  to user  $j$ . The authors of [42] have assigned an evaluation of the emotion of the sender with each tweet, called sentimental values, thereby creating a real valued label. At level 1, we group the users into binary clusters by the sign of sentimental values (yes or no). At level 2, we divide these further into 10 finer groups according to the strength of the sentimental values (strongly disagree to strongly agree). In such a way, we can construct a hierarchical labelling and all the algorithms can be used here as well. We choose the largest data set (Prop 37). Its largest weakly connected component  $W$  contains 8,123 vertices (users) and 10,911 edges (tweet relations). We cluster  $W_{ES}$  and  $W_{OS}$  to 10 clusters at level 2 and then further cluster them to 2 clusters at level 1.

For this data set, we can perform both unsupervised learning and semi-supervised learning methods. The results are given in Table 2. From the table, in terms of the  $F$ -measure, semi-supervised method MBO performs better than the other two methods of NHC and MLL while within unsupervised methods, MLL is better than NHC. In

terms of the modularity metric, NHC performs in general better than the other two methods especially when the size of the training data is small ( $\leq 60\%$ ).

$\mathcal{F}$	Trains (%)	0 (USL)	10	20	30	40	50	60	70	80	90
NHC	Level 2 (10)	0.15	0.11	0.15	0.25	0.31	0.39	0.48	0.60	0.73	0.86
	Level 1 (2)	0.65	0.55	0.49	0.47	0.47	0.49	0.58	0.64	0.73	0.83
MLL	Level 2 (10)	0.49	0.44	0.43	0.44	0.47	0.52	0.59	0.67	0.76	0.87
	Level 1 (2)	0.80	0.77	0.68	0.69	0.65	0.62	0.67	0.67	0.74	0.82
MBO	Level 2 (10)	N.A.	0.50	0.53	0.59	0.64	0.69	0.75	0.81	0.87	0.94
	Level 1 (2)	N.A.	0.84	0.86	0.86	0.86	0.86	0.88	0.88	0.90	0.91
$\mathcal{M}$	Trains (%)	0(USL)	10	20	30	40	50	60	70	80	90
NHC	Level 2 (10)	0.48	0.44	0.33	0.28	0.15	0.17	0.12	0.08	0.08	0.08
	Level 1 (2)	0.23	0.22	0.20	0.19	0.09	0.13	0.08	0.05	0.06	0.08
MLL	Level 2 (10)	0.09	0.07	0.06	0.05	0.04	0.06	0.04	0.04	0.07	0.09
	Level 1 (2)	0.08	0.04	0.03	0.02	0.01	0.04	0.03	0.03	0.05	0.07
MBO	Level 2 (10)	N.A.	0.28	0.25	0.17	0.15	0.13	0.11	0.09	0.10	0.09
	Level 1 (2)	N.A.	0.07	0.04	0.05	0.04	0.05	0.04	0.05	0.07	0.06

Table 2: Proposition Data: the  $F$ -measure  $\mathcal{F}$  and modularity metric  $\mathcal{M}$  using NHC, MLL, and MBO for given 0% (USL method), 10%, 20%, ..., 90% training data at level 2 ( $k_2 = 10$  clusters) and level 1 ( $k_1 = 2$  clusters), respectively. All results are average over 30 trials.

### 3.3 The Wiki-votes data set

The Wiki-votes data set [24, 23] is available from the Stanford large network data set collection at <https://snap.stanford.edu/data/wiki-Vote.html>. Per this website: “A small part of Wikipedia contributors are administrators, who are users with access to additional technical features that aid in maintenance. In order for a user to become an administrator a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship. Using the latest complete dump of Wikipedia page edit history (from January 3 2008) we extracted all administrator elections and vote history data. This gave us 2,794 elections with 103,663 total votes and 7,066 users participating in the elections (either casting a vote or being voted on). Out of these 1,235 elections resulted in a successful promotion, while 1,559 elections did not result in the promotion. About half of the votes in the dataset are by existing admins, while the other half comes from ordinary Wikipedia users. The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. Nodes in the network represent wikipedia users and a directed edge from node  $i$  to node  $j$  represents that user  $i$  voted on user  $j$ .”

The graph from the Wiki-votes data set has 7,115 vertices and 103,689 edges. It has 24 weakly connected components and 5,816 strongly connected components. The largest weakly connected component contains 7,066 vertices and others contains at most 3 vertices and can be viewed as singletons. Hence, we only consider the largest weakly connected component.

The Wiki-votes data set is unlabelled. Therefore, only the NHC and MLL algorithms can be used, and the performance can be only be tested using the modularity metric. We cluster all vertices to  $k_2$  clusters at level 2 and then further cluster them to  $k_1$  clusters at level 1. We choose  $k_2$  ranging from 4 to 11 and  $k_1$  ranging from 2 to 8 with step size 1. For each possible pair  $(k_1, k_2)$ , we compute the modularity metric from averaging over 30 trials. In terms of modularity, we found that the best modularity (see Table 3) is  $(\mathcal{M}_2, \mathcal{M}_1) = (0.040, 0.037)$  with respect to  $(k_2, k_1) = (6, 3)$  for NHC, where  $\mathcal{M}_2, \mathcal{M}_1$  are the modularity at level 2 and level 1, respectively. While it is  $(\mathcal{M}_2, \mathcal{M}_1) = (0.078, 0.074)$  with respect to  $(k_2, k_1) = (4, 3)$  for MLL.

## 4 Approximation and analysis of functions

To the best of our knowledge, representation of functions has been accomplished typically by a spectral decomposition of a graph Laplacian. We cite, for example, [18] for undirected graphs, which follows ideas developed in [32], and [17] for directed graphs. A more analytical treatment is typically based on embedding the digraph into

NHC( $\mathcal{M}_2, \mathcal{M}_1$ )								
$(k_1, k_2)$	4	5	6	7	8	9	10	11
2	(0.035,0.031)	(0.036,0.029)	(0.032,0.030)	(0.035,0.024)	(0.032,0.031)	(0.035,0.022)	(0.032,0.025)	(0.029,0.016)
3	(0.031,0.032)	(0.036,0.036)	(0.040,0.037)	(0.033,0.031)	(0.033,0.034)	(0.032,0.029)	(0.031,0.035)	(0.027,0.030)
4		(0.032,0.032)	(0.034,0.034)	(0.033,0.034)	(0.030,0.033)	(0.033,0.033)	(0.031,0.036)	(0.028,0.032)
5			(0.032,0.032)	(0.035,0.035)	(0.032,0.036)	(0.031,0.032)	(0.030,0.032)	(0.029,0.033)
6				(0.035,0.035)	(0.033,0.034)	(0.031,0.032)	(0.029,0.031)	(0.029,0.032)
7					(0.032,0.033)	(0.032,0.033)	(0.029,0.031)	(0.029,0.032)
8						(0.030,0.030)	(0.030,0.031)	(0.030,0.032)

MLL( $\mathcal{M}_2, \mathcal{M}_1$ )								
$(k_1, k_2)$	4	5	6	7	8	9	10	11
2	(0.078,0.058)	(0.056,0.052)	(0.048,0.048)	(0.046,0.045)	(0.041,0.036)	(0.036,0.038)	(0.032,0.042)	(0.028,0.042)
3	(0.078,0.074)	(0.056,0.060)	(0.047,0.050)	(0.046,0.048)	(0.040,0.044)	(0.036,0.040)	(0.032,0.042)	(0.030,0.040)
4		(0.056,0.061)	(0.048,0.049)	(0.045,0.046)	(0.040,0.046)	(0.035,0.040)	(0.033,0.045)	(0.031,0.039)
5			(0.047,0.049)	(0.046,0.050)	(0.040,0.044)	(0.037,0.043)	(0.032,0.041)	(0.029,0.041)
6				(0.046,0.049)	(0.040,0.041)	(0.036,0.041)	(0.032,0.036)	(0.030,0.034)
7					(0.040,0.041)	(0.036,0.038)	(0.033,0.036)	(0.029,0.033)
8						(0.036,0.039)	(0.032,0.035)	(0.030,0.033)

Table 3: Wiki-vote Data: the modularity metric pair  $(\mathcal{M}_2, \mathcal{M}_1)$  for the Wiki-Vote data set using GCN and LL unsupervised clustering ( $\mathcal{M}_2$  is the modularity at level 2 and  $\mathcal{M}_1$  is the modularity at level 1) for each  $(k_1, k_2) \in \{2, 3, \dots, 8\} \times \{4, 5, \dots, 11\}$ . All results are average over 30 trials.

a manifold [33, 29]. Our paper [4] gave a completely different approach that is free of any spectral decomposition, except for spectral clustering methods if used. The purpose of this section is to extend the results in [4] to the case of digraphs, represented by a pair of trees.

In this section, we assume that the trees corresponding to the digraph have been constructed using appropriate clustering algorithms. In greater abstraction, we will assume that the trees are both infinite. In practical terms, this allows us to add data to the digraph. We will then describe harmonic analysis on this infinite digraph. In Sub-Section 4.1, we will review the relevant ideas from [4]. They will be used in Sub-Section 4.2 to describe a very general harmonic analysis and function approximation paradigm.

## 4.1 Tree polynomials

Fundamental to our analysis is the notion of a filtration, as defined below.

**Definition 4.1** A *weighted tree* is a triplet  $(V, E, w)$ , where  $(V, E)$  is a tree and  $w : V \rightarrow (0, \infty)$  is a weight function. Let  $\mathbb{X} = (V, E, w)$  be a weighted tree, and  $v \in V$ . The tree with root  $v$  and leaves given by the children of  $v$  will be denoted by  $\mathbb{X}(v)$ . The tree  $\mathbb{X}(v)$  is called the **local (or conditional) filtration** at  $v$ , if the set  $\mathcal{L}_v$  of the children of  $v$  contains at least 2 elements and  $\sum_{u \in \mathcal{L}_v} w(u) = w(v)$ . The weighted tree  $\mathbb{X}$  is called a **filtration**, if  $w(\mathbf{r}^*) = 1$ ; and for each non-leaf vertex  $v \in V$ ,  $\mathbb{X}(v)$  is a local filtration at  $v$ .

It is possible to represent the vertices of a filtration as subintervals of  $[0, 1)$ . Let  $\mathbb{X} = (V, E, w)$  be a filtration. We associate the root  $\mathbf{r}^*$  with the unit interval  $[0, 1)$ . If the children of  $\mathbf{r}^*$  are  $v_1, \dots, v_M$ , we associate with each  $v_j$  the interval

$$\left[ \sum_{k=1}^{j-1} w(v_k), \sum_{k=1}^j w(v_k) \right).$$

Our assumption that  $\sum_{k=1}^M w(v_k) = w(\mathbf{r}^*) = 1$  implies that these intervals constitute a partition of  $[0, 1)$ . In a recursive manner, if the interval associated with a vertex  $v$  is  $[a, a + w(v))$ , and the children of  $v$  are  $u_1, \dots, u_K$ , then we associate with each  $u_j$ ,  $j = 1, \dots, K$ , the interval

$$\left[ a + \sum_{k=1}^{j-1} w(u_k), a + \sum_{k=1}^j w(u_k) \right).$$

Since  $\mathbb{X}(v)$  is a local filtration at  $v$ , the intervals associated with the  $u_j$ 's constitute a partition of the interval associated with  $v$ . In the sequel, we will refer to a vertex on the tree and the associated interval interchangeably. If  $S \subseteq [0, 1)$ , we denote by  $\chi_S$  the characteristic function of  $S$ ; i.e.,  $\chi_S(x) = 1$  if  $x \in S$ , and  $\chi_S(x) = 0$  if  $x \in [0, 1) \setminus S$ .

For a local filtration  $\mathbb{X}(v)$  at  $v$  associated with interval  $[a, a + w(v))$  for some  $a$ , let  $v_0, \dots, v_{m-1}$ ,  $m \geq 2$ , be the children of  $v$  in order. We wish to obtain the set of tree polynomials to have the same span as  $\{\chi_{v_j} : j = 0, \dots, m-1\}$ . In order to facilitate the construction of a consistently labeled system across the entire filtration, it

is convenient to substitute  $\chi_{v_0}$  by  $\chi_v$  in the list above. The resulting system is then defined by (4.2) as follows. Let  $p_j = w(v_j)$  for  $j = 0, \dots, m-1$  and

$$P_k := \sum_{j=0}^{k-1} p_j, \quad I_k := [a + P_k, a + P_{k+1}), \quad J_k := [a, a + P_k), \quad k = 1, \dots, m-1, \quad P_0 := 0, \quad P_{m+1} = w(v). \quad (4.1)$$

Define

$$\begin{aligned} \phi_0(x) &= \phi_0(v, x) = \chi_v, \\ \phi_k(x) &= \phi_k(v, x) = p_k \chi_{J_k}(x) - P_k \chi_{I_k}(x), \quad x \in [a, a + w(v)), \quad k = 1, \dots, m-1. \end{aligned} \quad (4.2)$$

Some of the important properties of these tree polynomials are listed in the following proposition ([4, Proposition 3.1, Proposition 3.2]).

**Proposition 4.1** *Let  $\mathbb{X}(v)$  be a local filtration, and  $\phi_j$ 's be defined as in (4.2). Let  $\ell, k$  be integers  $0 \leq \ell, k \leq m-1$ . Then*

(a)  $\phi_0 = \chi_v$  and for  $k = 1, \dots, m-1$ ,

$$\phi_k(x) = \begin{cases} p_k, & \text{if } x \in I_\ell, 1 \leq \ell \leq k-1, \\ -P_k, & \text{if } x \in I_k, \\ 0, & \text{if } x \in I_\ell, \ell \geq k+1; \end{cases} \quad (4.3)$$

(b)

$$\int_v \phi_k(x) \phi_\ell(x) dx = \begin{cases} w(v), & \text{if } k = \ell = 0, \\ p_k P_k P_{k+1}, & \text{if } k = \ell \geq 1, \\ 0, & \text{otherwise;} \end{cases} \quad (4.4)$$

(c)

$$\sum_{j=0}^{m-1} p_j \phi_k(a + P_j) \phi_\ell(a + P_j) = \int_v \phi_k(x) \phi_\ell(x) dx, \quad (4.5)$$

and for  $j = 0, \dots, m-1$ ,

$$1 + \sum_{k=1}^{m-1} \frac{\phi_k(a + P_j) \phi_k(a + P_\ell)}{p_k P_k P_{k+1}} = \begin{cases} 1/p_j, & \text{if } j = \ell, \\ 0, & \text{otherwise.} \end{cases} \quad (4.6)$$

(d) Let  $0 \leq k \leq m-1$  be an integer, and for  $x \in [a, a + w(v))$ ,

$$\tilde{\phi}_{k+1}(x) = \tilde{\phi}_{k+1}(v; x) = \begin{cases} 0, & \text{if } k = m-1, \\ (1 - P_{k+1}) \chi_{J_{k+1}}(x) - P_{k+1} \chi_{v \setminus J_{k+1}}(x), & \text{if } k \leq m-2. \end{cases}$$

Then

$$\int_v \phi_j(x) \tilde{\phi}_{k+1}(x) dx = 0, \quad j = 0, \dots, k, \quad (4.7)$$

and

$$\tilde{\Pi}_k := \tilde{\Pi}_k(v) := \text{span} \{\phi_0, \dots, \phi_k, \tilde{\phi}_{k+1}\} = \text{span} \{\chi_{I_0}, \dots, \chi_{I_k}, \chi_{v \setminus J_{k+1}}\}. \quad (4.8)$$

The orthogonal system on the whole tree is designed, so that in principle, when restricted to each local filtration, it should reduce to the local system for that filtration. To describe this in detail, we need to introduce the notion of Leave Left Out (LLO) enumeration. At each level  $L \geq 1$ , let (in this section only)

$$C_L = \{v : v \text{ is a vertex at level } L, v \text{ is not a left-most child of its parent}\},$$

$C_0 = \{\mathbf{r}^*\}$ . We associate  $M_0 = 0$  with the root  $\mathbf{r}^*$  at level 0. At each level  $L \geq 1$ , we enumerate the vertices in  $C_L$  by  $M_{L-1}, \dots, M_L - 1$ , left to right. This enumeration will be called the Leave Left Out (LLO) enumeration.

For example, the LLO enumeration of the vertices of the left tree in Figure 2 is  $\{A, B, E, F, H, I\}$ . We note that vertices which are left-most children of their parents are not numbered in this scheme.

Each vertex in  $V \setminus \{\mathbf{r}^*\}$  can also be enumerated among the children of its parent. For any  $v \in V$ , the children of  $v$  are enumerated left to right starting from 0, with 0 associated with the left-most child of  $v$ .

**Definition 4.2** For  $x \in [0, 1)$ , let  $\psi_0(x) = 1$ ; and for an integer  $n \geq 1$ , let  $u$  be the vertex corresponding to  $n$  under the LLO enumeration,  $v$  be the parent of  $u$ , and  $\ell$  be the enumeration of  $u$  as a child of  $v$ , so that  $\ell \geq 1$ . Set

$$\psi_n(x) = \phi_\ell(v, x). \quad (4.9)$$

The symbol  $\Pi_n$  denotes the span of  $\{\psi_0, \dots, \psi_n\}$ .

The following proposition [4, Theorem 4.2] summarizes some of the properties of the system  $\{\psi_n\}$ .

**Proposition 4.2** Let  $\mathbb{X}$  be a filtration,  $n, m \geq 0$  be integers,  $u = [a', b']$ ,  $u'$  be vertices corresponding to  $n$ ,  $m$  respectively in the LLO enumeration,  $v = [a, b)$  be the parent of  $u$ . Then

$$\int_0^1 \psi_n(x)\psi_m(x)dx = \begin{cases} 0, & \text{if } n \neq m, \\ 1, & \text{if } n = m = 0, \\ \aleph_n^{-1} := \frac{(b' - a')(a' - a)(b' - a)}{(b - a)^2}, & \text{if } n = m \neq 0. \end{cases} \quad (4.10)$$

If  $f : [0, 1) \rightarrow \mathbb{R}$  is any bounded and integrable function, then we define for integers  $k \geq 0$ ,  $n \geq 1$ ,

$$\hat{f}(k) = \aleph_k \int_0^1 f(t)\psi_k(t)dt, \quad s_n(f, x) = \sum_{k=0}^{n-1} \hat{f}(k)\psi_k(x). \quad (4.11)$$

A novelty of our system is that the partial sum operators  $\{s_n\}$  themselves are uniformly bounded, in contrast with the classical theory of Fourier series. Analogous to the summability methods in the theory of Fourier series, we define a more general version of these operators.

If  $\mathbf{h} = \{h_k\}_{k=0}^\infty$  is any sequence, we define

$$\sigma_n(\mathbf{h}, f) = \sum_{k=0}^n h_k \hat{f}(k)\psi_k, \quad n = 0, 1, \dots \quad (4.12)$$

We emphasize again that since the tree polynomials are piecewise constants, both the quantities  $s_n(f)$  and  $\sigma_n$  can be computed exactly as discrete sums, even though we find it convenient for theory and exposition purposes to write them as integral operators. Further details on this matter are given in [4].

We end this sub-section by enumerating some relevant properties of the operators  $\sigma_n$ . In the remainder of this sub-section, let  $\|\cdot\|$  denote the uniform norm on  $[0, 1)$ ,  $\mathcal{F}$  be the closure of  $\bigcup_{n=0}^\infty \Pi_n$  in this norm, and  $E_n(f) := \inf_{P \in \Pi_n} \|f - P\|$  be the degree of approximation of a function  $f$  defined on  $[0, 1)$ .

**Theorem 4.1** Let  $\mathbf{h} = \{h_k\}_{k=0}^\infty$  be a sequence of real numbers with

$$\mathcal{V}(\mathbf{h}) = \sup_{n \geq 0} \left\{ |h_n| + \sum_{k=0}^{n-1} |h_{k+1} - h_k| \right\} < \infty, \quad \lim_{N \rightarrow \infty} h_N = 0. \quad (4.13)$$

We have

$$\left\| \sum_{k=0}^n h_k \hat{f}(k)\psi_k \right\| \leq 3\mathcal{V}(\mathbf{h})\|f\|, \quad f \in \mathcal{F}, \quad n = 0, 1, \dots \quad (4.14)$$

In addition, if  $h_k = 1$  for  $0 \leq k \leq n$  for some integer  $n \geq 0$ , then for  $f \in \mathcal{F}$ ,

$$E_n(f) \leq \|f - \sigma_n(\mathbf{h}, f)\| \leq (1 + 3\mathcal{V}(\mathbf{h}))E_n(f). \quad (4.15)$$

## 4.2 Harmonic analysis on digraphs

In the case of a filtration, there is a one-to-one correspondence between functions on  $[0, 1)$  and functions on the vertices of the filtration, and it is a matter of convenience whether one thinks of the Lebesgue measure or a discrete measure defined on these vertices. When we represent a digraph using a pair of filtrations, this is no longer the case. It is seen already in the toy example in the introduction (cf. Figure 2), that the points on the digraph correspond only to 25 out of the 625 sub-rectangles of  $I^2$ . In the limiting case, as the number of nodes on the digraph tends to

infinity, it may or may not happen that the set of points on the square that correspond to the nodes on the digraph is a set of two dimensional Lebesgue measure 0. The structure of this set is hard to stipulate mathematically, since it depends upon the exact construction of filtrations, which in turn depends upon the particular digraph in question. Therefore, harmonic analysis and function approximation on digraphs in our paradigm is more delicate than the multivariate analogues of the univariate analysis in classical situations such as trigonometric series, splines, etc. It is not possible to give universal constructions in our paradigm. In this section, we outline an abstract, data-driven theory. In this section, we will use standard multivariate notation.

We denote the two filtrations corresponding to the given digraph by  $\mathbb{X}_1, \mathbb{X}_2$  respectively, and the set of points on the square  $I^2$  that correspond to the nodes on the digraph by  $\mathbb{G}$ . Thus,  $\mathbf{x} = (x_1, x_2) \in \mathbb{G}$  if and only if  $x_1$  corresponds to the same vertex on the digraph as a node on the filtration  $\mathbb{X}_1$  as the vertex corresponding to the point  $x_2$  as a node on the filtration  $\mathbb{X}_2$  (see Figure 3 for an example.) If  $w_j$  is the weight associated with  $x_j$  as a node on the filtration  $\mathbb{X}_j$ , then we associate the weight  $w_1 w_2$  with the point  $\mathbf{x} \in \mathbb{G}$ . The resulting measure will be denoted by  $\nu^*$ . We note that the measure  $\nu^*$  is a probability measure, but may well be singular with respect to the two dimensional Lebesgue measure on  $I^2$ . We will denote the uniform norm on  $\mathbb{G}$  by  $\|\cdot\|$ .

With an abuse of notation, we denote the **orthonormalized** system of tree polynomials on these by  $\psi_{n,1}, \psi_{n,2}$  respectively. Naturally, the tensor product tree polynomials

$$\psi_{\mathbf{k}}(\mathbf{x}) = \psi_{k_1,1}(x_1)\psi_{k_2,2}(x_2), \quad \mathbf{k} = (k_1, k_2) \in \mathbb{Z}_+^2, \quad \mathbf{x} = (x_1, x_2) \in I^2, \quad (4.16)$$

are an orthonormal basis for square integrable functions on  $I^2$ . However, many of these are possibly equal to 0 when restricted to  $\mathbb{G}$ . Let  $\Omega = \{\mathbf{k} : \psi_{\mathbf{k}}|_{\mathbb{G}} \neq 0\}$ . Since the tree polynomials  $\psi_{\mathbf{k}}$  are constants on rectangles in  $I^2$ , it is clear that

$$\int_{\mathbb{G}} \psi_{\mathbf{k}} \psi_{\mathbf{m}} d\nu^* = \int_{I^2} \psi_{\mathbf{k}}(\mathbf{x}) \psi_{\mathbf{m}}(\mathbf{x}) d\mathbf{x}, \quad \mathbf{k}, \mathbf{m} \in \Omega. \quad (4.17)$$

Thus,  $\{\psi_{\mathbf{k}}\}_{\mathbf{k} \in \Omega}$  is an orthonormal basis for  $L^2(\mathbb{G}, \nu^*)$ . The closure of the set  $\text{span}\{\psi_{\mathbf{k}} : \mathbf{k} \in \Omega\}$  in the uniform norm of  $\mathbb{G}$  will be denoted by  $\mathcal{F}$ , abusing again the notation from the case of single filtration.

We note that each of the tree polynomials is a piecewise constant function. The localization of these polynomials is illustrated in Figure 4 in the context of the toy example in the introduction, referring to the trees in Figure 2.

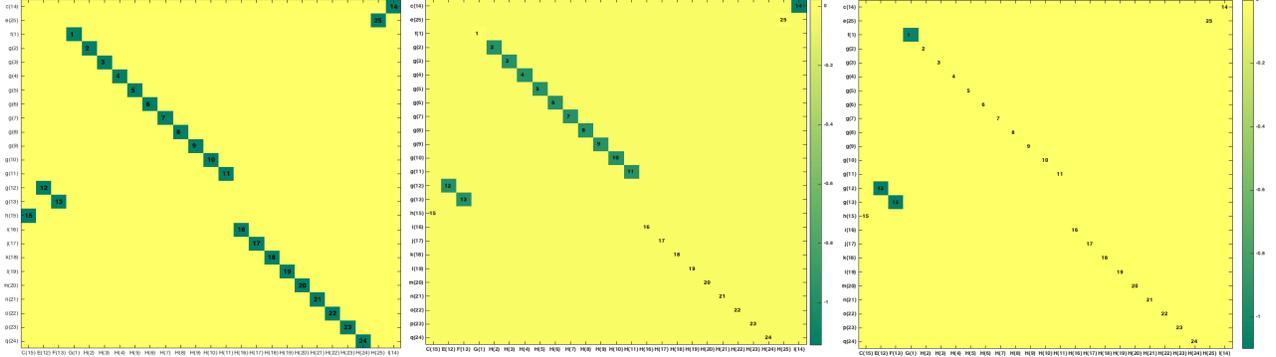


Figure 4: The localization of the orthogonal system on  $W$  is demonstrated by considering the tensor product tree polynomials for the toy example (see Figure 2). There are 6 polynomials  $\psi_{k_1,1}, k_1 = 1, \dots, 6$  from  $\mathcal{T}_{W_{ES}}$  and 14 polynomials  $\psi_{k_2,2}, k_2 = 1, \dots, 14$  from  $\mathcal{T}_{W_{OS}}$ . Left: tree vertex position of the tree on a square  $[0, 1] \times [0, 1]$ . Middle: tensor product polynomial on the tree w.r.t.  $\psi_{2,1}(x)\psi_{6,2}(y)$ . Right: tensor product polynomial on the tree w.r.t.  $\psi_{3,1}(x)\psi_{6,2}(y)$ .

Since each  $\psi_{\mathbf{k}}$  is bounded, we may define the Fourier coefficients, respectively, partial sums of  $f \in L^1(\mathbb{G}, \nu^*) \cap L^\infty(\mathbb{G}, \nu^*)$  by

$$\hat{f}(\mathbf{k}) = \int_{\mathbb{G}} f \psi_{\mathbf{k}} d\nu^*, \quad s_{\mathbf{m}}(f, \mathbf{x}) = \sum_{k_1=0}^{m_1} \sum_{k_2=0}^{m_2} \hat{f}(\mathbf{k}) \psi_{\mathbf{k}}(\mathbf{x}), \quad \mathbf{k}, \mathbf{m} \in \mathbb{Z}_+^2. \quad (4.18)$$

It is understood here that  $\hat{f}(\mathbf{k}) = 0$  if  $\mathbf{k} \in \mathbb{Z}_+^2 \setminus \Omega$ , so that the only nonzero summands in the definition of the partial sum correspond to  $\mathbf{k} \in \Omega$ .

For a (bi)-sequence  $h$ , we define

$$\Delta_1 h(\mathbf{k}) = h(k_1 + 1, k_2) - h(k_1, k_2), \quad \Delta_2 h(\mathbf{k}) = h(k_1, k_2 + 1) - h(k_1, k_2), \quad \Delta h = \Delta_1 \Delta_2 h = \Delta_2 \Delta_1 h, \quad (4.19)$$

and

$$\mathcal{V}(h) = \sup_{\mathbf{k} \in \mathbb{Z}_+^2} |h(\mathbf{k})| + \sup_{k_1 \in \mathbb{Z}_+} \sum_{k_2=0}^{\infty} |\Delta_2 h(\mathbf{k})| + \sup_{k_2 \in \mathbb{Z}_+} \sum_{k_1=0}^{\infty} |\Delta_1 h(\mathbf{k})| + \sum_{\mathbf{k} \in \mathbb{Z}_+^2} |\Delta h(\mathbf{k})|. \quad (4.20)$$

Denoting by  $E_1 h(\mathbf{k}) = h(k_1 + 1, k_2)$ ,  $E_2 h(\mathbf{k}) = h(k_1, k_2 + 1)$ , we note the following identity for future use. If  $h_1, h_2$  are sequences as above, then  $\Delta_j(h_1 h_2) = \Delta_j(h_1)E_j(h_2) + h_1 \Delta_j(h_2)$ ,  $j = 1, 2$ , and

$$\Delta(h_1 h_2) = \Delta(h_1)E_1(E_2 h_2) + \Delta_1 h_1 E_1(\Delta_2 h_2) + \Delta_1 h_2 E_2(\Delta_1 h_2) + h_1 \Delta(h_2). \quad (4.21)$$

In the sequel,  $a \lesssim b$  denotes that  $a \leq cb$  for a generic constant  $c$  that does not depend upon the target function and other obvious variables. The value of these generic constants may change at different occurrences, even within a single formula. By  $a \sim b$  we mean  $a \lesssim b$  and  $b \lesssim a$ . In particular,

$$\mathcal{V}(h_1 h_2) \lesssim \mathcal{V}(h_1) \mathcal{V}(h_2). \quad (4.22)$$

Using Theorem 4.1, it is not difficult to prove the following.

**Theorem 4.2** *Let  $h = \{h(\mathbf{k})\}_{\mathbf{k} \in \mathbb{Z}_+^2}$  be a finitely supported (bi)–sequence of real numbers. Then for  $f \in \mathcal{F}$ ,*

$$\left\| \sum_{\mathbf{k} \in \mathbb{Z}_+^2} h(\mathbf{k}) \hat{f}(\mathbf{k}) \psi_{\mathbf{k}} \right\| \lesssim \mathcal{V}(h) \|f\|. \quad (4.23)$$

In the classical theory of multivariate Fourier series, it is customary to define various notions of the degree of the polynomial: spherical, total, coordinatewise, hyperbolic cross, etc. One could do this in the context of tree polynomials on  $\mathbb{G}$  as well, but since the “frequencies” are limited to  $\Omega$ , it is convenient to define a more parsimonious notion by defining the analysis spaces first and defining the approximation spaces in terms of these.

**Definition 4.3** *A sequence of sequences  $\mathbf{g} = \{g_j : \Omega \rightarrow [0, 1]\}_{j=0}^{\infty}$  is called an **admissible partition of unity** on  $\Omega$  if  $g_0(\mathbf{0}) = 1$ , each  $g_j$  is supported on a finite set,*

$$\sum_{j=0}^{\infty} g_j(\mathbf{k}) = 1, \quad \mathbf{k} \in \Omega, \quad (4.24)$$

*and the following condition is satisfied: There exists an integer  $m^* = m^*(\mathbf{g}) \geq 0$  such that for  $j, j' \geq 0$ ,  $|j - j'| > m^*$ ,  $g_j$  and  $g_{j'}$  have disjoint supports; i.e.,  $g_j(\mathbf{k})g_{j'}(\mathbf{k}) = 0$  for all  $\mathbf{k} \in \Omega$ .*

In the remainder of this section, we will fix an admissible partition  $\mathbf{g}$  of unity. We set  $H_n(\mathbf{k}) = \sum_{j=0}^n g_j(\mathbf{k})$ ,  $\mathbf{k} \in \Omega$ , and define the class of multivariate tree polynomials of  $(\mathbf{g})$ –degree  $\leq n$  by

$$\mathbb{P}_n = \text{span} \{ \psi_{\mathbf{k}} : H_n(\mathbf{k}) > 0, \quad \mathbf{k} \in \Omega \}, \quad n \in \mathbb{Z}_+. \quad (4.25)$$

Since each  $g_j$  is finitely supported, so is each  $H_n$ .

As before, we define the degree of approximation of  $f \in L^\infty(\mathbb{G}, \nu^*)$  by

$$E_n(f) = \inf \{ \|f - P\| : P \in \mathbb{P}_n \}, \quad n = 0, 1, \dots. \quad (4.26)$$

It is convenient to extend this notation to  $n \in \mathbb{R}$  by setting  $E_n(f) = \|f\|$  if  $n < 0$  and  $E_n(f) = E_{\lfloor n \rfloor}(f)$  if  $n$  is not an integer.

Next, we define the reconstruction and analysis operators (with an abuse of notation) by

$$\begin{aligned} \sigma_n(f) &= \sum_{\mathbf{k} \in \Omega} H_n(\mathbf{k}) \hat{f}(\mathbf{k}) \psi_{\mathbf{k}}, \quad n = 0, 1, \dots, \\ \tau_j(f) &= \begin{cases} \sigma_0(f) = \hat{f}(\mathbf{0}), & \text{if } j = 0, \\ \sigma_j(f) - \sigma_{j-1}(f) = \sum_{\mathbf{k} \in \Omega} g_j(\mathbf{k}) \hat{f}(\mathbf{k}) \psi_{\mathbf{k}}, & \text{if } j = 1, 2, \dots. \end{cases} \end{aligned} \quad (4.27)$$

The following theorem lists some important properties of these operators.

**Theorem 4.3** Let  $\mathcal{V}(H_n) \lesssim 1$  for all  $n \geq 1$ , and  $f \in \mathcal{F}$ .

(a) We have

$$E_n(f) \leq \|f - \sigma_n(f)\| \lesssim E_{n-m^*}(f), \quad n = 0, 1, \dots \quad (4.28)$$

(b) We have

$$f = \sum_{j=0}^{\infty} \tau_j(f), \quad (4.29)$$

where the sum converges uniformly.

(c) We have

$$\int_{\mathbb{G}} |f(\mathbf{x})|^2 d\nu^*(\mathbf{x}) \sim \sum_{j=0}^{\infty} \int_{\mathbb{G}} |\tau_j(f)(\mathbf{x})|^2 d\nu^*(\mathbf{x}) \quad (4.30)$$

We digress to make some comments on the approximation power of our scheme in the context of classification problems. The value of the target function  $f$  for any leaf is its class label. For simplicity, let us consider a local filtration  $\mathbb{X}(v)$ , in which the majority of the leaves have a label 1, the others have a label 0. Assuming that both classes appear with equal probability, the value of the tree polynomial approximation to  $f$  at  $v$  is the expected value of the labels of the leaves. To view this as a class label, we need to round it to the nearest integer. This amounts to declaring that the label of the class at the level  $v$  is the same as that of the majority of the children of  $v$ . More generally, each cluster  $C_i$  in  $\{C_1, \dots, C_M\}$  obtained in the construction of the digraph is assigned the class label  $i_{j_0} = \operatorname{argmax}_j |C_i \cap L_j|$  by comparing to the ground-truth classes  $\{L_1, \dots, L_n\}$ . Then one defines a confusion matrix  $M$  of size  $n \times n$  by

$$M_{j,k} := \sum_{i_{j_0}=j} \frac{|C_{i_{j_0}} \cap L_k|}{|L_k|}, \quad j, k = 1, \dots, n, \quad (4.31)$$

where  $M_{j,k}$  is the  $(j, k)$ -entry of the matrix  $M$ . Note that the more  $M$  closes to the identity matrix, the better the classification result.

In Figures 5 and 6, we present the confusion matrices (as images) for the semi-supervised learning results in Sections 3.1 and 3.2 for datasets CORA and Proposition. We consider 70% of training data as input and run each clustering of the algorithms (NHC, MLL, MBO), which gives different clustering results. For each method, the confusion matrix  $M$  is averaged over 30 runs. From the images of Figures 5 and 6, we can see that for the CORA dataset, NHC has better classification (visual) results than the other two methods at both level 2 and level 1. For the Proposition dataset, MLL outperforms the other two methods at level 2 while the three methods performs more or less the same at level 1. Note that since the  $F$ -measure is computed differently from the class label assignment, the best  $F$ -measure might not correspond to the best confusion matrix, as measured by misclassification percentage.

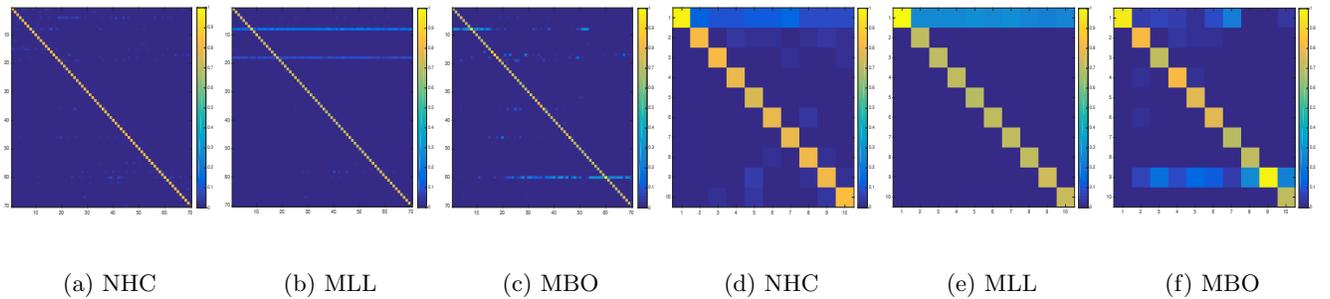


Figure 5: Confusion matrices (images) for CORA dataset. Left 3 (Level 2: 70 classes). Right 3 (Level 1: 10 classes)

We resume the main discussion with a definition of the smoothness classes in terms of the degree of approximation as in [4]. Let  $0 < \rho \leq \infty$ ,  $\gamma > 0$ , and  $\mathbf{a} = \{a_n\}_{n=0}^{\infty}$  be a sequence of real numbers. We define

$$\|\mathbf{a}\|_{\rho, \gamma} := \begin{cases} \left\{ \sum_{n=0}^{\infty} 2^{n\gamma\rho} |a_n|^\rho \right\}^{1/\rho}, & \text{if } 0 < \rho < \infty, \\ \sup_{n \geq 0} 2^{n\gamma} |a_n|, & \text{if } \rho = \infty. \end{cases} \quad (4.32)$$

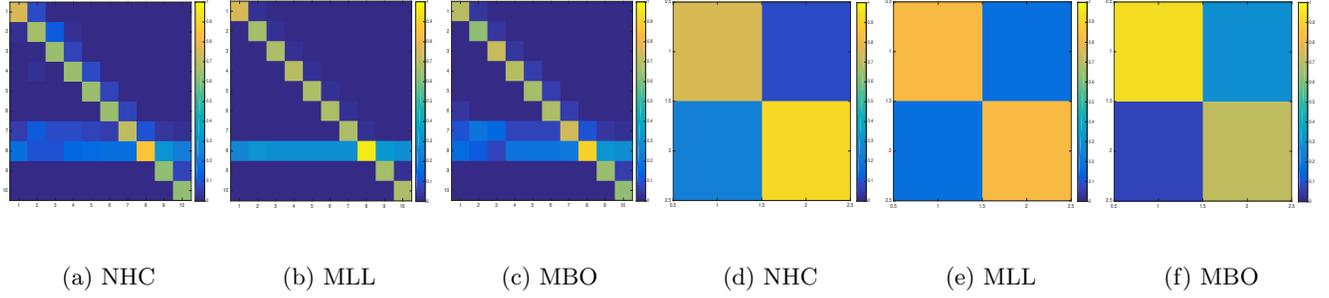


Figure 6: Confusion matrices (images) for Proposition dataset. Left 3 (Level 2: 10 classes). Right 3 (Level 1: 2 classes)

The space of sequences  $\mathbf{a}$  for which  $\|\mathbf{a}\|_{\rho,\gamma} < \infty$  will be denoted by  $\mathbf{b}_{\rho,\gamma}$ . The smoothness class  $B_{\rho,\gamma}$  is defined by

$$B_{\rho,\gamma} := \{f \in \mathcal{F} : \{E_{2^n}(f)\}_{n=0}^\infty \in \mathbf{b}_{\rho,\gamma}\}. \quad (4.33)$$

In order to develop both function approximation estimates and a wavelet-like characterization of smoothness spaces using the atoms  $\tau_j(f)$ , we need an appropriate notion of derivatives. In the case of classical theory of multivariate Fourier series, this is typically done via multipliers; e.g., denoting by  $\hat{f}(\mathbf{k})$  the trigonometric Fourier coefficient, the mixed partial derivative of  $f$  is a function  $g$  with  $\hat{g}(\mathbf{k}) = k_1 k_2 \hat{f}(\mathbf{k})$ , and a spherical partial derivative is a function  $g$  with  $\hat{g}(\mathbf{k}) = (|\mathbf{k}|^2 + 1)^{1/2} \hat{f}(\mathbf{k})$ . We have adopted a similar strategy also for tree polynomials in [4]. In the current context, since we don't know the structure of  $\Omega$ , it is not feasible to define a derivative by means of a fixed multiplier sequence. The following Definition 4.4 gives our substitute for a derivative of order  $r$ .

For any subset  $S \subseteq \mathbb{Z}_+^2$ , we define

$$\mathcal{E}(S) = \{\mathbf{k} \in \mathbb{Z}_+^2 : \mathbf{m} \in S \text{ for some } \mathbf{m} \text{ with } |\mathbf{k} - \mathbf{m}|_\infty \leq 2\}.$$

By the restriction of a sequence  $h$  to  $S$ , we mean the sequence whose value at  $\mathbf{k} \in S$  is  $h(\mathbf{k})$ , and 0 otherwise. The sequence  $h^{[-1]}$  is defined by  $h^{[-1]}(\mathbf{k}) = (h(\mathbf{k}))^{-1}$  if  $h(\mathbf{k}) \neq 0$  and  $h^{[-1]}(\mathbf{k}) = 0$  otherwise.

**Definition 4.4** *Let  $r \geq 1$  be an integer. A sequence  $\mu$  is called a ( $\mathbf{g}$ -)multiplier sequence of order  $r$  if  $\mu(\mathbf{k}) > 0$  for every  $\mathbf{k} \in \mathcal{E}(\Omega)$ , and for every integer  $j \geq 0$ , if  $\mu_j$  is the sequence  $\mu(\mathbf{k})$  restricted to  $\text{supp}(g_j)$ , then*

$$\mathcal{V}(\mu_j) \sim (\mathcal{V}(\mu_j^{[-1]}))^{-1} \sim 2^{jr} \quad (4.34)$$

*The derivative of  $f : \mathbb{G} \rightarrow \mathbb{R}$  (in the sense of  $\mu$  and  $\mathbf{g}$ ) is a function  $\mathcal{D}(f) : \mathbb{G} \rightarrow \mathbb{R}$  such that  $\widehat{\mathcal{D}(f)}(\mathbf{k}) = \mu(\mathbf{k})\hat{f}(\mathbf{k})$ ,  $\mathbf{k} \in \mathbb{Z}_+^2$ , if such a function exists.*

In the rest of this paper, we will fix  $r$  and  $\mu$ .

The  $K$ -functional in the theory of function approximation on  $\mathbb{G}$  is now defined for  $f \in \mathcal{F}$  by

$$K(f, \delta) = \inf\{\|f - g\| + \delta^r \|\mathcal{D}g\| : g, \mathcal{D}g \in \mathcal{F}\}. \quad (4.35)$$

The analogue of [4, Theorem 5.2, Theorem 5.3] is the following.

**Theorem 4.4** *Let  $0 < \rho \leq \infty$  and  $\gamma > 0$ . Let  $\mathcal{V}(H_n) \lesssim 1$  for all  $n \geq 1$ , and  $f \in \mathcal{F}$ . The following statements are equivalent.*

1.  $f \in B_{\rho,\gamma}$ .
2.  $\{\|\tau_j(f)\|\}_{j=0}^\infty \in \mathbf{b}_{\rho,\gamma}$ .
3.  $\{\|f - \sigma_n(f)\|\}_{j=0}^\infty \in \mathbf{b}_{\rho,\gamma}$ .
4.  $\{K(f, 2^{-j})\} \in \mathbf{b}_{\rho,\gamma}$ .

## 5 Proof of the theorems in Section 4.2.

In the sequel, we find it convenient to extend the sequences on  $\Omega$  to sequences on  $\mathbb{Z}_+^2$  by setting their values to be 0 outside  $\Omega$ . This will facilitate the use of summation indices and relationships such as (4.21) and (4.22) provided we keep in mind that the validity of (4.24) is assumed only on  $\Omega$ . Similarly, functions on  $\mathbb{G}$  are assumed to be extended to  $I^2$  by setting them equal to 0 outside of  $\mathbb{G}$ .

PROOF OF THEOREM 4.2. In this proof, let

$$s_{(n_1, n_2)}(f)(x_1, x_2) = \sum_{k_1=0}^{n_1} \sum_{k_2=0}^{n_2} \hat{f}(k_1, k_2) \psi_{k_1, 1}(x_1) \psi_{k_2, 2}(x_2), \quad n_1, n_2 = 0, 1, \dots,$$

where  $\hat{f}(k_1, k_2) = 0$  if  $(k_1, k_2) \notin \Omega$ . Then [4, Theorem 5.1] regarding the uniform boundedness of the Fourier partial sums in the univariate case leads to

$$\|s_{\mathbf{k}}(f)\| \lesssim 1, \quad \mathbf{k} \in \mathbb{Z}_+^2. \quad (5.1)$$

Next, we observe that  $\hat{f}(\mathbf{k})\psi_{\mathbf{k}}(\mathbf{x}) = \Delta s_{\mathbf{k}}(f)(\mathbf{x})$  for all  $\mathbf{k} \in \mathbb{Z}_+^2$ . A summation by parts shows that

$$\sum_{\mathbf{k} \in \mathbb{Z}_+^2} h(\mathbf{k}) \hat{f}(\mathbf{k}) \psi_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{k}} \Delta h(k_1 - 1, k_2 - 1) s_{\mathbf{k}}(f)(\mathbf{x}),$$

where  $h(k_1, k_2) = 0$  if either  $k_1$  or  $k_2$  is negative. The estimate (4.23) follows from (5.1) and the definition of  $\mathcal{V}(h)$ .  $\square$

PROOF OF THEOREM 4.3. The first estimate in (4.28) is obvious from the definition. The second estimate is also obvious in light of Theorem 4.2 and the definition if  $n < m^*$ . In the remainder of this proof, let  $n \geq m^*$ , and

$$\Omega_n = \{\mathbf{k} : H_n(\mathbf{k}) \neq 0\}, \quad n = 0, 1, \dots$$

If  $\mathbf{k} \in \Omega_{n-m^*}$ , then there exists  $j \leq n - m^*$  such that  $g_j(\mathbf{k}) \neq 0$ . Since  $g_j(\mathbf{k})g_{j'}(\mathbf{k}) = 0$  if  $|j - j'| > m^*$ , this implies that  $g_{j'}(\mathbf{k}) = 0$  for all  $j' > n$ . Consequently, (4.24) shows that  $H_n(\mathbf{k}) = \sum_{j=0}^n g_j(\mathbf{k}) = \sum_{j \in \Omega} g_j(\mathbf{k}) = 1$ . So, if  $P = \sum_{\mathbf{k} \in \Omega_{n-m^*}} \hat{P}(\mathbf{k})\psi_{\mathbf{k}} \in \mathbb{P}_{n-m^*}$ , then for all  $\mathbf{x} \in \mathbb{G}$ ,

$$\sigma_n(P)(\mathbf{x}) = \sum_{\mathbf{k}} H_n(\mathbf{k}) \hat{P}(\mathbf{k}) \psi_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{k} \in \Omega_{n-m^*}} H_n(\mathbf{k}) \hat{P}(\mathbf{k}) \psi_{\mathbf{k}}(\mathbf{x}) = \sum_{\mathbf{k} \in \Omega_{n-m^*}} \hat{P}(\mathbf{k}) \psi_{\mathbf{k}}(\mathbf{x}) = P(\mathbf{x}).$$

Therefore, Theorem 4.2 leads to

$$\|f - \sigma_n(f)\| \leq \|f - P\| + \|\sigma_n(f - P)\| \lesssim \|f - P\|,$$

and hence, to the second estimate of (4.28).

The second estimate in (4.28) can be rewritten in the form

$$\left| f - \sum_{j=0}^n \tau_j(f) \right| = \|f - \sigma_n(f)\| \lesssim E_{n-m^*}(f).$$

Since  $f \in \mathcal{F}$ ,  $E_{n-m^*}(f) \rightarrow 0$  as  $n \rightarrow \infty$ , this is equivalent to (4.29) in the sense of uniform convergence. This proves part (b).

To prove part (c), we observe that since  $\nu^*$  is a probability measure, the system  $\{\psi_{\mathbf{k}}\}$  is a fundamental system for the  $L^2(\mathbb{G}, \nu^*)$  closure of  $\mathcal{F}$ . Therefore, for  $f \in \mathcal{F}$ , Parseval identity holds, and we obtain

$$\int_{\mathbb{G}} |f(\mathbf{x})|^2 d\nu^*(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}_+^2} |\hat{f}(\mathbf{k})|^2 = \sum_{j=0}^{\infty} \sum_{\mathbf{k} \in \mathbb{Z}_+^2} g_j(\mathbf{k}) |\hat{f}(\mathbf{k})|^2. \quad (5.2)$$

Since  $g_j(\mathbf{k})^2 \leq g_j(\mathbf{k})$  for all  $j$  and  $\mathbf{k}$ , this shows using Parseval identity again that

$$\sum_{j=0}^{\infty} \int_{\mathbb{G}} |\tau_j(f)(\mathbf{x})|^2 d\nu^*(\mathbf{x}) = \sum_{j=0}^{\infty} \sum_{\mathbf{k} \in \mathbb{Z}_+^2} g_j(\mathbf{k})^2 |\hat{f}(\mathbf{k})|^2 \leq \int_{\mathbb{G}} |f(\mathbf{x})|^2 d\nu^*(\mathbf{x}). \quad (5.3)$$

Since  $g_j(\mathbf{k})g_m(\mathbf{k}) = 0$  if  $|j - m| > m^*$ , it is easy to verify using the definitions that

$$\int_{\mathbb{G}} \tau_j(f)(\mathbf{x})\tau_m(f)(\mathbf{x})d\nu^*(\mathbf{x}) = 0, \quad |j - m| > m^*.$$

Using (4.29), we see that

$$\int_{\mathbb{G}} |f(\mathbf{x})|^2 d\nu^*(\mathbf{x}) = \sum_{j=0}^{\infty} \sum_{m=0}^{\infty} \int_{\mathbb{G}} \tau_j(f)(\mathbf{x})\tau_m(f)(\mathbf{x})d\nu^*(\mathbf{x}) = \sum_{j=0}^{\infty} \sum_{m=\max(j-m^*, 0)}^{j+m^*} \int_{\mathbb{G}} \tau_j(f)(\mathbf{x})\tau_m(f)(\mathbf{x})d\nu^*(\mathbf{x}).$$

An application of Schwarz inequality and Parseval theorem then lead to

$$\int_{\mathbb{G}} |f(\mathbf{x})|^2 d\nu^*(\mathbf{x}) \lesssim \sum_{j=0}^{\infty} \int_{\mathbb{G}} |\tau_j(f)(\mathbf{x})|^2 d\nu^*(\mathbf{x}).$$

Together with (5.3), this completes the proof of (4.30).  $\square$

**Theorem 5.1** (a) For  $n \geq 0$  and  $P \in \mathbb{P}_n$ ,

$$\|\mathcal{D}P\| \lesssim 2^{nr} \|P\|. \quad (5.4)$$

(b) If  $f \in \mathcal{F}$  and  $\mathcal{D}^r f \in \mathcal{F}$ , then for  $n \geq 0$ ,

$$E_n(f) \lesssim 2^{-nr} \|\mathcal{D}f\|. \quad (5.5)$$

PROOF. We observe first that for  $j = 0, 1, \dots$  and  $\mathbf{k} \in \mathbb{Z}_+^2$ ,  $\widehat{\tau_j(f)}(\mathbf{k}) = g_j(\mathbf{k})\hat{f}(\mathbf{k})$ . Hence, Theorem 4.2 and the conditions (4.34) together imply that

$$\|\mathcal{D}\tau_j(f)\| = \left\| \sum_{\mathbf{k}} \mu(\mathbf{k})g_j(\mathbf{k})\hat{f}(\mathbf{k})\psi_{\mathbf{k}} \right\| = \left\| \sum_{\mathbf{k}} \mu_j(\mathbf{k})g_j(\mathbf{k})\hat{f}(\mathbf{k})\psi_{\mathbf{k}} \right\| \lesssim 2^{jr} \|\tau_j(f)\| \lesssim 2^{jr} \|f\|, \quad (5.6)$$

and similarly,

$$\|\tau_j(f)\| = \left\| \sum_{\mathbf{k}} \mu_j^{[-1]}(\mathbf{k})\mu_j(\mathbf{k})g_j(\mathbf{k})\hat{f}(\mathbf{k})\psi_{\mathbf{k}} \right\| \lesssim 2^{-jr} \left\| \sum_{\mathbf{k}} g_j(\mathbf{k})\mu(\mathbf{k})\hat{f}(\mathbf{k})\psi_{\mathbf{k}} \right\| = 2^{-jr} \|\tau_j(\mathcal{D}f)\| \lesssim 2^{-jr} \|\mathcal{D}f\|. \quad (5.7)$$

To prove part (a), we observe in view of (5.6) that

$$\|\mathcal{D}P\| = \|\mathcal{D}\sigma_{n+m^*}(P)\| = \left\| \sum_{j=0}^{n+m^*} \mathcal{D}\tau_j(P) \right\| \lesssim \sum_{j=0}^{n+m^*} 2^{jr} \|P\| \lesssim 2^{nr} \|P\|.$$

This proves (5.4).

To prove part (b), we use (4.29) and (5.7) to deduce that

$$E_n(f) \leq \left\| \sum_{j=n+1}^{\infty} \tau_j(f) \right\| \leq \sum_{j=n+1}^{\infty} \|\tau_j(f)\| \lesssim \|\mathcal{D}f\| \sum_{j=n+1}^{\infty} 2^{-jr} \lesssim 2^{-nr} \|\mathcal{D}f\|.$$

This proves (5.5).  $\square$

PROOF OF THEOREM 4.4. Using Theorem 5.1, the proof of Theorem 4.4 follows standard arguments. For the equivalence of the items 1, 2, 3, these arguments are exactly the same as those in [32, Theorem 4] (with different notation). The equivalence of items 1 and 4 is shown using the arguments in [9, Theorem 9.1 in Section 7.9, also Chapter 6.7]. We omit the details.  $\square$

## 6 Acknowledgments

We thank Professors Percus and Hunter at Claremont Graduate University and Claremont McKenna College respectively for many useful discussions as well as their help in securing the Proposition data set, which was sent to us by Dr. Linhong Zhu at USC Information Sciences Institute in Marina Del Ray, California. We thank Dr. Garcia-Cardona for giving us a C code for the algorithm MBO.

## References

- [1] A. Arenas, J. Duch, A. Fernández, and S. Gómez. Size reduction of complex networks preserving modularity. *New Journal of Physics*, 9(6):176, 2007.
- [2] K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, pages 343–351, 2010.
- [3] C. K. Chui and D. L. Donoho. Special issue: Diffusion maps and wavelets. *Appl. and Comput. Harm. Anal.*, 21(1), 2006.
- [4] C. K. Chui, F. Filbir, and H. N. Mhaskar. Representation of functions on big data: graphs and trees. *Applied and Computational Harmonic Analysis*, 38(3):489–509, 2015.
- [5] F. Chung. Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.
- [6] F. Chung. The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences*, 104(50):19735–19740, 2007.
- [7] F. R. K. Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.
- [8] J. J. Crofts and D. J. Higham. Googling the brain: Discovering hierarchical and asymmetric network structures, with applications in neuroscience. *Internet Mathematics*, 7(4):233–254, 2011.
- [9] R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.
- [10] M. Elberfeld, V. Bafna, I. Gamzu, A. Medvedovsky, D. Segev, D. Silverbush, U. Zwick, and R. Sharan. On the approximability of reachability-preserving network orientations. *Internet Mathematics*, 7(4):209–232, 2011.
- [11] F. Filbir and H. N. Mhaskar. A quadrature formula for diffusion polynomials corresponding to a generalized heat kernel. *Journal of Fourier Analysis and Applications*, 16(5):629–657, 2010.
- [12] F. Filbir and H. N. Mhaskar. Marcinkiewicz-Zygmund measures on manifolds. *Journal of Complexity*, 27(6):568–596, 2011.
- [13] J. Friedman and J.-P. Tillich. Wave equations for graphs and the edge-based laplacian. *Pacific Journal of Mathematics*, 216(2):229–266, 2004.
- [14] C. Garcia-Cardona, E. Merkurjev, A. L. Bertozzi, A. Flenner, and A. G. Percus. Fast multiclass segmentation using diffuse interface methods on graphs. Technical report, DTIC Document, 2013.
- [15] M. Gavish and R. R. Coifman. Sampling, denoising and compression of matrices by coherent matrix organization. *Applied and Computational Harmonic Analysis*, 33(3):354–369, 2012.
- [16] M. Gavish, B. Nadler, and R. R. Coifman. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 367–374, 2010.
- [17] G. A. Gidelew. *Topics in harmonic analysis on combinatorial graphs*. PhD thesis, Drexel University, 2014.
- [18] D. K. Hammond, P. Vandergheynst, and R. Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.
- [19] X. Han, Y. Chen, J. Shi, and Z. He. An extended cell transmission model based on digraph for urban traffic road network. In *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, pages 558–563. IEEE, 2012.
- [20] H. Hu, T. Laurent, M. A. Porter, and A. L. Bertozzi. A method based on total variation for network modularity optimization using the mbo scheme. *SIAM Journal on Applied Mathematics*, 73(6):2224–2246, 2013.

- [21] H. Jia, S. Ding, X. Xu, and R. Nie. The latest research progress on spectral clustering. *Neural Computing and Applications*, 24(7-8):1477–1486, 2014.
- [22] S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(9):1393–1403, 2006.
- [23] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.
- [24] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1361–1370. ACM, 2010.
- [25] Y. Li and Z.-L. Zhang. Digraph laplacian and the degree of asymmetry. *Internet Mathematics*, 8(4):381–401, 2012.
- [26] L.-H. Lim. Hodge laplacians on graphs. In *Geometry and Topology in Statistical Inference, Proceedings of Symposia in Applied Mathematics*, volume 73. Amer. Math. Soc., 2015.
- [27] M. Maggioni and H. N. Mhaskar. Diffusion polynomial frames on metric measure spaces. *Applied and Computational Harmonic Analysis*, 24(3):329–353, 2008.
- [28] F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95–142, 2013.
- [29] H. N. Mhaskar. A unified framework for harmonic analysis of functions on directed graphs and changing data. <http://arxiv.org/abs/1604.06835>.
- [30] H. N. Mhaskar. Eignets for function approximation on manifolds. *Applied and Computational Harmonic Analysis*, 29(1):63–87, 2010.
- [31] H. N. Mhaskar. A generalized diffusion frame for parsimonious representation of functions on data defined manifolds. *Neural Networks*, 24(4):345–359, 2011.
- [32] H. N. Mhaskar and J. Prestin. Polynomial frames: a fast tour. *Approximation Theory XI: Gatlinburg*, pages 101–132, 2004.
- [33] S. Mousazadeh and I. Cohen. Embedding and function extension on directed graph. *Signal Processing*, 111:137–149, 2015.
- [34] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.
- [35] M. M. H. Pang. Heat kernels of graphs. *Journal of the London Mathematical Society*, 2(1):50–64, 1993.
- [36] N. Przulj. Introduction to the special issue on biological networks. *Internet Mathematics*, 7(4):207–208, 2011.
- [37] V. Satuluri and S. Parthasarathy. Symmetrizations for clustering directed graphs. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 343–354. ACM, 2011.
- [38] W.-J. Shen, H.-S. Wong, Q.-W. Xiao, X. Guo, and S. Smale. Towards a mathematical foundation of immunology and amino acid chains. *arXiv preprint arXiv:1205.6031*, 2012.
- [39] Y.-K. Shih, S. Kim, Y. Ruan, J. Cheng, A. Gattani, T. Shi, and S. Parthasarathy. Component detection in directed networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1729–1738. ACM, 2014.
- [40] A. Sikora. Riesz transform, Gaussian bounds and the method of wave equation. *Mathematische Zeitschrift*, 247(3):643–662, 2004.
- [41] A. Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.
- [42] L. M. Smith, L. Zhu, K. Lerman, and Z. Kozareva. The role of social media in the discussion of controversial topics. In *Social Computing (SocialCom), 2013 International Conference on*, pages 236–243. IEEE, 2013.
- [43] C. Smulders. Heat kernels on graphs. Technical report, Technische Universiteit Eindhoven, 2004.
- [44] S. M. Van Dongen. *Graph clustering by flow simulation*. PhD thesis, University of Utrecht, 2001.
- [45] L. Zhu, A. Galstyan, J. Cheng, and K. Lerman. Tripartite graph clustering for dynamic sentiment analysis on social media. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1531–1542. ACM, 2014.