Signal Processing on Simplicial Complexes With Vertex Signals

Feng Ji, Giacomo Kahn, Wee Peng Tay

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 19, 2022, accepted April 10, 2022, date of publication April 13, 2022, date of current version April 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3167055

Signal Processing on Simplicial Complexes With Vertex Signals

FENG JI[1], GIACOMO KAHN[2], AND WEE PENG TAY[1], (Senior Member, IEEE)
[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798
[2]Univ Lyon, Univ Lyon 2, Université Claude Bernard Lyon 1, Université Jean Monnet Saint-Etienne, INSA Lyon, DISP-UR4570, 69500 Bron, France

Corresponding author: Feng Ji (jifeng@ntu.edu.sg)

This work was supported in part by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE2018-T2-2-019, and in part by Agency for Science, Technology and Research (A*STAR) under its RIE2020 Advanced Manufacturing and Engineering (AME) Industry Alignment Fund–Pre Positioning (IAF-PP) under Grant A19D6a0053.

ABSTRACT In classical graph signal processing (GSP), the underlying topological structures are restricted in terms of dimensionality. A graph or a 1-complex is a combinatorial object that models binary relations, which do not directly capture complex high arity relations. One possible high dimensional generalization of graphs is a simplicial complex. In this paper, we develop a signal processing framework on simplicial complexes with vertex signals, which recovers the traditional GSP theory. We introduce the concept of a generalized Laplacian, which allows us to embed a simplicial complex into a traditional graph and hence perform signal processing similar to traditional GSP. We show that the generalized Laplacian satisfies several desirable properties, same as the graph Laplacian. We propose a method to learn 2-complex structures and demonstrate how to perform signal processing by applying the generalized Laplacian on both synthetic and real data. We observe performance gains in our experiments when 2-complexes are used to model the data, compared to the traditional GSP approach of restricting to 1-complexes.

INDEX TERMS Graph Signal Processing, multidimensional data, simplicial complex.

I. INTRODUCTION

Over the heterogeneous landscape of data analytics, data come in various diverse forms incorporating simple binary relations to relations of high arity. One way to abstract relationships inherent in the data and to avoid insensitive metrics is to incorporate topological properties of the data (its *shape*) into data analytics. Graph signal processing (GSP) [1]–[8] is one of the approaches that allow this. From signals recorded on networks such as sensor or social networks, GSP uses the topology of the graph in performing sampling, translation and filtering of the signals. Recently, GSP-based graph convolution neural networks have also received much attention [9]–[12].

GSP, as useful a tool as it is, still has its limitations. The vast data landscape includes complex data, such as high dimensional manifolds, or point clouds possessing high dimensional geometric features (cf. Fig. 1). For example, 2D meshes can be used to approximate surfaces and high dimensional simplicial complexes [13] can be used to model discrete point clouds. Another example is in social networks such as Facebook, where an edge represents a friendship relation but higher arity edges can represent family links or the inclusion in the same groups. This model also works for group conversations or other user groups in social networks. In addition, some complex interactions cannot be fully grasped by reducing them to binary relations. This is the case in chemistry, where two molecules might interact only in the presence of a third that serves as catalyst [14], [15], or with datasets such as folksonomies, where data are ternary or quaternary relations [16]. Therefore, it is necessary to go beyond graphs to fully capture these more complex interaction mechanisms.

Moreover, in many applications, a graph learning procedure is involved based on information such as geometric distance, vertex feature similarity and graph signals [17]–[21]. Due to the lack of a definite meaning for edge connections, it is arguable whether a graph is the best geometric object for signal processing. For these reasons, there is a need for a framework that permits signal processing on such high dimensional geometric objects.

Simplicial complexes, as a high dimensional generalization of graphs, are independently found in many fields of computer science and mathematics. Simplicial complexes

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

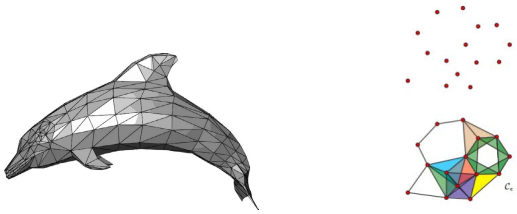

FIGURE 1. (a) Approximation of a 2D surface with a 2-complex from Wikimedia Commons. (b) Model of a point cloud (top), together with a simplicial complex built on it (bottom).

can be used to model and approximate any reasonable (topological) space according to the *simplicial approximation theorem* [13]. They are already used in topological data analysis [22], representation of surfaces in high dimensions, and modeling of complex networks [23]. In this paper, we extend the signal processing toolbox to include methods on simplicial complexes. Our goal is develop a GSP framework for signals defined on vertices but have high arity relations that can be modeled using simplicial complexes.

Despite the fact that the subject is relatively new, a few attempts have been made to develop signal processing methods over simplicial complexes [24]–[29]. In [24], [25] the authors develop a signal processing framework using a differential operator on simplicial chain complexes. They mainly consider complexes up to dimension two, even though their framework is more general and allows simplicial complexes of arbitrary dimension. Their framework considers flow data and signals associated with high dimensional simplexes such as edges and faces, and not only vertices. In [26], the authors study not a hypergraph framework, but a community aware type of GSP. This allows to highlight some groups of highly connected vertices in a way that is stronger than the Laplacian matrix. In [27], [28], the authors propose an approach on meet or join semi-lattices that uses lattice operators as the shift. Finally, in [29] the authors propose a framework on hypergraphs using tensor decomposition. Simplicial complexes are a class of hypergraph, and their framework also considers signals on vertices but their approach is different from ours in the following way. One of the main challenges is the mathematical representation of the hypergraph. When dealing with hypergraph representation, there are two main options. The first one, matrix-based representation, uses a modified adjacency matrix, while the second uses tensors to encode incidence in hyperedges. In [30], the authors claim that both approaches have shortcomings: tensor algebra is hard to understand, and this slows the process of using those representations in real life applications and industry, while for matrix-based representation, the difficulty lies in encoding a shift operator. In [29], they detail a complete framework for hypergraph signal processing using tensor decomposition, together with extensive examples of applications. We choose instead to focus on matrix-based representations of simplicial complexes, trying our hand at designing a shift operator.

In our paper, we propose a new signal processing framework for signals on vertices of simplicial complexes. Our approach makes full use of the geometric structure and strictly generalizes traditional GSP by introducing generalized Laplacians. Signal processing tasks can thus be performed similar to traditional GSP. Although for the later part of the paper, we focus on 2-complexes for illustration purposes, our approach can be easily generalized to higher order complexes. Throughout this paper, a weighted graph refers to a graph whose edge weights are positive.

The rest of the paper is organized as follows. We briefly recall fundamentals of simplicial complexes in Section II. In Section III, we introduce a general way to construct Laplacians on metric spaces, and then apply this approach to simplicial complexes. We focus on the special case of 2-complexes in Section IV and describe a procedure to construct 2-complex structures on a given graph. We present numerical results in Section V and conclude in Section VI. A preliminary version of this work was presented in [31]. In this paper, we present a more thorough theoretical discussion with rigorous proofs of all results and included further experiments.

In terms of common notations, Δ_n is used to denote a simplex, while X (possibly with a subscript) is for a simplicial complex. G (possibly with a subscript) is used to denote a graph. We use L (possibly with subscript) for a Laplacian or generalized Laplacian. We use boldface letters for graph signals and transformations, and a graph signal is usually denoted by x. We refer the readers to the main text for less common notations.

II. SIMPLICIAL COMPLEXES

In this section, we give a brief self-contained overview of the theory of simplicial complexes. We refer the interested reader to [13], [32] for more details.

Definition 1: The standard n-simplex (or dimension n simplex) Δ_n *is defined as*

$$\Delta_n = \left\{ (x_0, \ldots, x_n) \in \mathbb{R}_+^{n+1} \sum_{i=0}^{n} x_i = 1 \right\}.$$

Any topological space homeomorphic to the standard n-simplex is called an n-simplex. In Δ_n, if we require $k > 0$ coordinates being 0, we obtain an $(n-k)$-simplex, called a face.

A simplicial complex X *(see Fig. 2 for an example) is a set of simplexes such that any face from a simplex of X is also in X and the intersection for any two simplexes $\sigma_1, \sigma_2 \in X$ is a face of both σ_1 and σ_2. A simplex of X is called* maximal *if it is not the face of any other simplexes.*

We shall primarily focus on finite simplicial complexes, i.e., a finite set of simplexes. The dimension $\dim X$ of X is the largest dimension of a simplex in X. An m-dimensional simplicial complex is called an m-complex. A 1-complex is then a graph in the traditional sense. For each $m \geq 0$,

the subset of m-simplexes of X and their faces is called its *m-skeleton*, denoted by X^m.

Combinatorially, if we do not want to specify an exact homeomorphism of a n-simplex X with Δ_n, we may just represent it by $n+1$ labels. Therefore, its faces are just subsets of the labels. It is worth pointing out that according to the above definition, a simplicial complex is a set of topological spaces, each homeomorphic to a simplex and they are related to each other by face relations. However, it is possible to produce a concrete geometric object for each simplicial complex.

Definition 2: The geometric realization *of a simplicial complex X is the topological space obtained by gluing simplexes with common faces.*

Example 1: Let X be a finite simplicial complex consisting of two types of simplexes E and V. Each simplex in E is 1-dimensional and V contains only 0 simplexes. The geometric realization of X is nothing but a graph with vertex set V and edge set E. More generally, for any simplicial complex X, the geometric realization of its 1-skeleton X^1 is a graph in the usual sense.

For convenience, we shall not distinguish a simplicial complex from its geometric realization when no confusion arises.

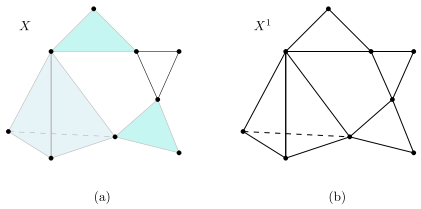

FIGURE 2. (a) X is (the geometric realization) of a 3-complex with a maximal 3-simplex, two maximal 2-simplexes and three maximal 1-simplexes. (b) X^1 is a connected graph with 15 edges and 9 vertices.

A useful notion is the *barycenter* of a simplex. Consider an n-simplex $X \cong \Delta^n$ in a Euclidean space such that

$$X = \left\{ \sum_{0\le i\le n} t_i v_i \sum_{0\le i\le n} t_i = 1,\ 0 \le t_i \le 1,\ \forall i \right\}, \qquad (1)$$

where $v_0, \ldots, v_n$ are affinely independent (i.e., $v_1 - v_0, \ldots, v_n - v_0$ are linearly independent). The barycenter (or the center of gravity) is the point

$$u = \frac{1}{n+1} \sum_{i=0}^{n} v_i. \qquad (2)$$

In this paper, we say that a simplicial complex X is weighted if X^1 is a weighted graph. Otherwise, we may make an unweighted simplicial complex become weighted by assigning length 1 to each 1-simplex. In this way, X^0 becomes a metric space. It is important to notice that we do not require weights associated with simplexes of dimension at least 2. In practice, such information might be harder to obtain. For example, on a 2D mesh, it is usually harder to find the area of each triangular component.

A more general notion than simplicial complex is that of a *hypergraph* [33], [34]. A hypergraph $H = (V, E)$ is a pair where V is a set of vertices and E is a set of non-empty subsets of V, called *hyperedges*. Each simplicial complex X might be viewed as a hypergraph $H_X = (X^0, E)$, where the vertex set is the 0-skeleton of X and a set of vertices of each simplex of dimension at least 1 is a hyperedge. Conversely, not all hypergraphs are simplicial complexes since any subset of a hyperedge might not be a hyperedge, violating the face condition of Definition 1. For each hypergraph H, the associated simplicial complex X_H whose vertices are spanned by hyperedges in E as well as all their proper subsets is the smallest simplicial complex containing H. Therefore, the signal processing framework developed in this paper can be directly applied to each hypergraph H, or more precisely, the associated simplicial complex X_H.

III. GENERALIZED LAPLACIAN

A. THE ABSTRACT CONSTRUCTION

Recall that graph signal processing relies heavily on the notion of a shift operator. A popular choice is the graph Laplacian. In this section, we generalize this notion for *simplicial complexes with vertex signals*.

Our approach is to embed the vertices of a given simplicial complex and their corresponding signals in a bigger graph, where the additional vertices and edges are designed to capture the high arity relationships expressed by the simplicial complex. The generalized Laplacian is then a combination of a linear transformation and the Laplacian of the bigger graph. In particular, we want this generalized Laplacian to have the usual properties of a graph Laplacian: symmetric positive semi-definite with the constant vector as an eigenvector with eigenvalue zero. We propose the following definition.

Definition 3: Let X be a finite simplicial complex. A generalized Laplacian consists of the following data: (A)

(A) a weighted, undirected graph $G_X = (V, E)$,
(B) a function $f : X^0 \to V$, and
(C) a linear transformation $\mathbf{T} : \mathbb{R}^{|X^0|} \to \mathbb{R}^{|V|}$,

such that the following holds: (a), nolistsep

(a) The function f is injective.
(b) For each $v \in X^0$ and $\mathbf{x} \in \mathbb{R}^{|X^0|}$, the $f(v)$-component of $\mathbf{T}(\mathbf{x})$ is the same as the v-component of $\mathbf{x}$.
(c) The sum of each row of $\mathbf{T}$ (written as a transformation matrix) is a constant.

Let L_{G_X} be the Laplacian of the weighted graph G_X. The generalized Laplacian associated with the data $(G_X, f, \mathbf{T})$ *is defined as*

$$L_{(G_X,f,T)} = \mathbf{T}' \circ L_{G_X} \circ \mathbf{T} : \mathbb{R}^{|X^0|} \to \mathbb{R}^{|X^0|},$$

where $\mathbf{T}'$ denotes the adjoint (transpose) of $\mathbf{T}$. We abbreviate $L_{(G_X,f,\mathbf{T})}$ by L_X if no confusion arises from the context.

Intuitively, we require that f is injective to ensure that f "embeds" X^0 in G_X so that we may perform the shift operation on G_X. Conditions (b) and (c) on $\mathbf{T}$ say that the signal on $v \in X^0$ is preserved at its image $f(v)$ in G_X, while signals on $V \backslash f(X)$ are formed from an averaging process.

Note that the traditional graph Laplacian for a 1-complex X is a special case of a generalized Laplacian with both f and $\mathbf{T}$ being the identity maps.

We further remark that Definition 3 can be applied to finite metric spaces with slight modifications.

Lemma 1: Suppose X is a simplicial complex and L_X is a generalized Laplacian as in Definition 3. Then, the following holds: (a)

(a) L_X is symmetric.
(b) L_X is positive semi-definite.
(c) Constant signals are in the 0-eigenspace of L_X. The 0-eigenspace E_0 of L_X is 1-dimensional if and only if G_X is connected.

Proof:

(a) L_X is symmetric because G_X is assumed to be undirected and hence L_{G_X} is symmetric.

(b) Similar to (a), L_X is positive semi-definite because L_{G_X} is positive semi-definite.

(c) As the sum of each row of $\mathbf{T}$ is a constant, if $\mathbf{x} \in \mathbb{R}^{|X^0|}$ is a constant vector, then so is $\mathbf{T}(\mathbf{x})$. Since constant vectors are in the 0-eigenspace of L_{G_X}, we have $L_{G_X} \circ \mathbf{T}(\mathbf{x}) = 0$ and hence $L_X(\mathbf{x}) = 0$. Therefore, the dimension of E_0 is at least 1. Now assume that $\mathbf{x}$ is in E_0. We have

$$0 = \langle \mathbf{x}, L_X(\mathbf{x}) \rangle = \langle \mathbf{T}(\mathbf{x}), L_{G_X} \circ \mathbf{T}(\mathbf{x}) \rangle.$$

Consequently, $\mathbf{T}(\mathbf{x})$ belongs to the 0-eigenspace of L_{G_X}, which is 1-dimensional if and only if G_X is connected. By Condition (b), the operator $\mathbf{T}$ is injective. Therefore, E_0 is 1-dimensional if and only if $\mathbf{T}(E_0)$ is 1-dimensional, which is in turn equivalent to G_X being connected as we just observed. ■

By Lemma 1, the generalized Laplacian L_X enjoys a few desired properties. In particular, being symmetric permits an orthonormal basis consisting of eigenvectors of L_X. Therefore, one can devise a Fourier theory analogous to traditional GSP. Moreover, as L_X is positive semi-definite, we may perform smoothness based learning. The constant vectors belonging to the 0-eigenspace is also desirable as it agrees with the intuition that "constant signals are smoothest".

On the other hand, Lemma 1 asserts that L_X is indeed very similar to the Laplacian of a graph even for higher dimensional complexes. The theory will be less useful if we are only able to produce weighted graph Laplacians, which we shall prove to be false in Lemma 2 after we propose a choice of $(G_X, f, \mathbf{T})$ below. To this point, we introduce the following notion.

Definition 4: We say that L_X is graph type *if all the diagonal entries of L_X are non-negative and all the off-diagonal entries are non-positive.*

B. GENERALIZED LAPLACIAN MADE EXPLICIT

1) THE CONSTRUCTION

We give an explicit construction of L_X together with a canonical choice of G_X, f and $\mathbf{T}$. For the simplest case, assume $X \cong \Delta_n$ is a weighted n-simplex where $n \geq 2$. We defer the case $n = 1$ towards the end of this subsection. For $n \geq 2$, X is homeomorphic to the standard n-simplex and its 1-skeleton X^1 is a weighted graph with edge weights given by the weight function $w(\cdot, \cdot)$. We label the vertices of X by $X^0 = \{v_0, \ldots, v_n\}$. The graph $G_X = (V, E)$ is constructed as follows: Let $V = \{v_0, \ldots, v_n, u\}$ with a single additional vertex u, which is taken to be the barycenter of X. There is no edge between v_i and v_j for any pair $0 \leq i \neq j \leq n$. On the other hand, there is an edge connecting v_i and u for each $0 \leq i \leq n$.

The edge weight $d(v_i, u)$ between v_i and u in G_X is computed as follows:

$$d(v_i, u) = \binom{n}{2}^{-1} \sum_{v_i \neq v_j \neq v_k \neq v_i} (v_j, v_k)_{v_i}, \tag{3}$$

where

$$(v_j, v_k)_{v_i} = \frac{1}{2}(w(v_i, v_j) + w(v_i, v_k) - w(v_j, v_k))$$

is the Gromov product [35], [36]. Illustrations for $X \cong \Delta_2$ and $X \cong \Delta_3$ are shown in Fig. 3. When $n = 2$, we recover pairwise distances in G_X between the vertices v_0, v_1 and v_2 in X, in the sense that for each pair of vertices v_i, v_j, we have $d(v_i, u) + d(v_j, u) = w(v_i, v_j)$.

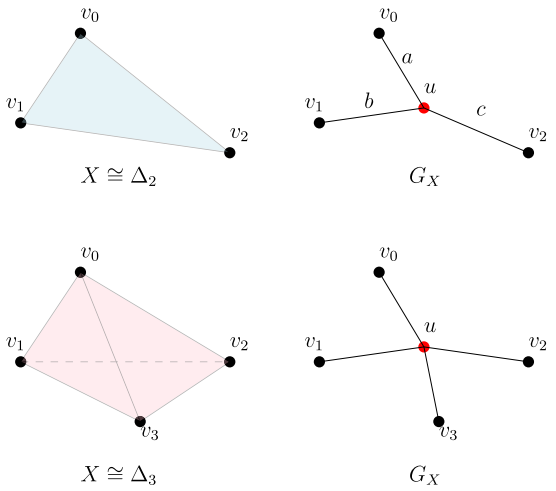

FIGURE 3. Graphical illustration of G_X for $X \cong \Delta_2$ and $X \cong \Delta_3$.

We have a canonical choice for $f : f(v_i) = v_i$ for $0 \leq i \leq n$. For $\mathbf{T}$, it is identity on each v_i-component for $0 \leq i \leq n$, while the average is assigned to the u-component. In matrix form,

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ 1/(n+1) & 1/(n+1) & \ldots & 1/(n+1) \end{bmatrix}.$$

It is straightforward to check that $(G_X, \mathbf{T}, f)$ satisfies the conditions of Definition 3. Thus, we have an associated generalized Laplacian L_X.

For a general finite simplicial complex X, we have a decomposition X^{max} as the subset of the maximal simplexes in X and the generalized Laplacian is defined as:

$$L_X = \sum_{\sigma \in X^{max}} L_\sigma, \tag{4}$$

where the summation is over all maximal simplexes of X with appropriate embedding of the vertex indices of σ in X.

We now consider the case where X itself is a graph. In this case, the maximal simplexes are just edges (i.e., Δ_1). For an edge $e = (v_1, v_2)$ with weight w, the associated graph G_e with a barycenter contains only 3 vertices, v_1, v_2 and an additional vertex u. Therefore, in this case, formula (3) no longer applies. On the other hand, to apply Definition 3, we choose G_X to be X itself and both f and $\mathbf{T}$ be the identity map. Therefore, we recover L_X as the usual graph Laplacian.

2) MOTIVATION AND INSIGHTS

To give motivation and insights into the above construction, we consider both theoretical (topological) and practical aspects.

First of all, we notice that for $X \cong \Delta_n$, it is topologically (homotopy) equivalent to a point [32], meaning that it can be continuously deformed to a point. Therefore, if we want to approximate X by a graph G_X that preserves this topological property, then G_X cannot contain a cycle and must be a tree. In addition, if we do not want to break the symmetry of the vertices, the most natural step to do so is to add one additional vertex (the barycenter) connected to every vertex in the original graph. The edge weights of G_X are chosen to approximate the metric of X.

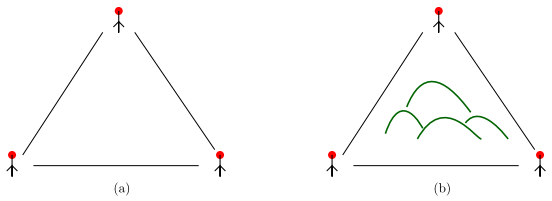

FIGURE 4. In (a), the sensors are located in an unblocked terrain; while in (b), the terrain is partially blocked.

Next, we describe a practical scenario to give a signal processing motivation. Consider the case (Fig. 4) of three sensors that are pairwise connected. If the area containing sensors are blocked, for example by geographic terrain such as mountains (Fig. 4(b)), it is reasonable to model the connections between the sensors by a graph with 3 edges. However, if the sensors are located in an unblocked terrain (Fig. 4(a)), it may no longer be reasonable to use a graph to model the sensor relationships, as there are additional paths from one sensor to another via the interior region. If f is a signal recorded by the sensors, an one-step cyclic shift in Fig. 4(a) is expected to have a stronger "smoothing" or "averaging" effect than in Fig. 4(a). The approach we propose indeed satisfies such a requirement as we effectively have an initial averaging step.

IV. 2-COMPLEXES

In this section, we focus on 2-complexes, which is sufficient for most applications. We first discuss the explicit construction in Section III-B in the context of 2-complexes and present spectral and shift invariance results. We then propose a method to learn 2-complex structures from graph data, and finally briefly indicate some signal processing tasks that can be achieved using our generalized Laplacian framework.

A. GENERALIZED LAPLACIANS FOR 2-COMPLEXES

For a weighted 2-simplex $X \cong \Delta_2$, assume that the edge weights are $w(v_0, v_1), w(v_0, v_2)$ and $w(v_1, v_2)$. Using the explicit construction proposed in Section III-B, the edge weights of G_X are (c.f. Fig. 3):

$$\begin{aligned} a &= (v_1, v_2)_{v_0} = (w(v_0, v_2) + w(v_0, v_1) - w(v_1, v_2))/2, \\ b &= (v_0, v_2)_{v_1} = (w(v_1, v_2) + w(v_0, v_1) - w(v_0, v_2))/2, \\ c &= (v_0, v_1)_{v_2} = (w(v_0, v_2) + w(v_1, v_2) - w(v_0, v_1))/2. \end{aligned}$$

If the edge weights in X satisfy the triangle inequality, then $a \geq 0, b \geq 0, c \geq 0$. Conversely, given $a \geq 0, b \geq 0, c \geq 0$, we are able to recover the edge weights by taking pairwise sums.

The generalized Laplacian L_X is then given by:

$$\begin{aligned} L_X &= \begin{bmatrix} 1 & 0 & 0 & 1/3 \\ 0 & 1 & 0 & 1/3 \\ 0 & 0 & 1 & 1/3 \end{bmatrix} \begin{bmatrix} a & 0 & 0 & -a \\ 0 & b & 0 & -b \\ 0 & 0 & c & -c \\ -a & -b & -c & a+b+c \end{bmatrix} \\ &\quad \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \\ &= \frac{1}{9} \begin{bmatrix} b+c+4a & c-2a-2b & b-2a-2c \\ c-2a-2b & a+c+4b & a-2b-2c \\ b-2a-2c & a-2b-2c & a+b+4c \end{bmatrix}. \end{aligned} \tag{5}$$

Definition 5: Define the shape constant γ_X *of* X *as*

$$\gamma_X = \min\{ \tfrac{5w(v_i,v_j) - w(v_i,v_k) - w(v_j,v_k)}{2} : \{i, j, k\} = \{0, 1, 2\}\}.$$

In general, γ_X can be negative. This happens when there is at least one very short edge in X. We use this observation to address an issue left over from the previous section. The following lemma shows that L_X is not necessarily of graph type (cf. Definition 4), i.e., the generalized Laplacian given in Definition 3 need not be a weighted graph Laplacian.

Lemma 2: Suppose $X \cong \Delta_2$ *is a 2-simplex. Then* L_X *in (5) is of graph type if and only if* $\gamma_X \geq 0$.

Proof: A direct computation shows that

$$-\gamma_X = \max\{c - 2a - 2b, b - 2a - 2c, a - 2b - 2c\}.$$

As the diagonal entries of L_X in (5) are all positive, it is of graph-type if and only if $-\gamma_X \leq 0$, i.e., $\gamma_X \geq 0$. ■

In the case of a 2-simplex, we may also give the following interpretation of L_X in (5) with the graph Laplacian L_{X^1} of the 1-skeleton X^1. Consider a graph signal $\mathbf{x} = (x_0, x_1, x_2)'$

on the vertices $\{v_0, v_1, v_2\}$. Let $\mathbf{y}$ be the first order difference $(x_2 - x_1, x_0 - x_2, x_1 - x_0)'$. By a direct computation, one observes that L_X is determined by

$$9\langle \mathbf{x}, L_X(\mathbf{x})\rangle = \langle \mathbf{y}, L_{X^1}(\mathbf{y})\rangle,$$

which says that L_X has similar effect as a second order difference. More specifically, it is proportional to the sum of the square of the second order differences in the signal components of $\mathbf{x}$. However, this point-of-view cannot be generalized beyond dimension 2.

From (4), if X is a general 2-complex, the generalized Laplacian L_X takes contribution from the generalized Laplacians of 2-simplexes computed as above and the usual edge Laplacians. We next study spectral properties of L_X. In particular, we want to compare L_X and L_{X^1} as the latter is well-studied. We write $A \preceq B$ if $B - A$ is positive semi-definite. Moreover, for each edge $e \in E$ of X^1, let k_e be the number of 2-simplexes in X containing e. Set $k_{\max} = \max_{e\in E} k_e$ and $k_{\min} = \min_{e\in E} k_e$.

Lemma 3: Suppose X is a finite 2-complex with each edge of length 1. Then L_X given by the explicit construction in Section III-B satisfies

$$\frac{1}{6}\max\{*\}k_{\min}, 1 \cdot L_{X^1} \preceq L_X \preceq \frac{k_{\max}}{6} L_{X^1}.$$

Proof: We sketch the main idea of the proof. It suffices to show $L = L_X - \max\{\frac{k_{\min}}{6}, \frac{1}{6}\} \cdot L_{X^1}$ or $L = \frac{k_{\max}}{6} L_{X^1} - L_X$ is the Laplacian of a (possibly disconnected) graph. For this, one only needs to compute the off-diagonal entries of L and show that they are non-positive. It suffices to notice that each edge $e = (v_i, v_j)$ contributes -1 to the (i, j)-th entry of L_{X^1}; and $-k_e/6$ to the (i, j)-th entry of L_X. ∎

If X is a 2D-mesh (triangulation) of a compact 2-manifold, then $k_{\min} = 1$ and $k_{\max} = 2$. This is because at most two 2-simplexes can share a common edge and along the boundary each edge is contained in a single 2-simplex.

Recall that a filter F is *shift invariant* with respect to (w.r.t.) L_{X^1} if $F \circ L_{X^1} = L_{X^1} \circ F$. Notice it guarantees some other invariance properties such as permutation invariance. If the graph Laplacian L_{X^1} does not have repeated eigenvalues, then F is shift invariant if and only if $F = P(L_{X^1})$ for some polynomial P of degree at most $|X^0| - 1$. The shift invariant family is of particular interest and they are readily estimated as one only has to learn the polynomial coefficients. Due to this fact, L_X will be less interesting if it is shift invariant w.r.t. L_{X^1}, e.g., when X is a single 2-simplex with equal edge weights (more examples are shown in Fig. 5). However, this does not happen in general.

Proposition 1: Suppose X is a 2-complex with X^1 being a connected graph and such that the following conditions hold (illustrated in Fig. 6):

- *(a) In X, any two 2-simplexes are not connected by a direct edge.*
- *(b) In X, if a vertex v is not contained in any 2-simplex, then it is connected to at most one 2-simplex. There is at least one such vertex.*
- *(c) Each edge is contained in at most one 2-simplex.*

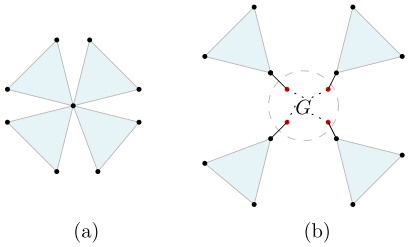

FIGURE 5. In (a), if all the edge weights are the same then $L_X = 1/3L_{X^1}$ is shift invariant w.r.t. L_{X^1}. However, in (b), as long as the 4 red vertices are contained in a graph G (at the center), then L_X is not shift invariant w.r.t. L_{X^1} by Proposition 1, even if we allow arbitrary positive edge weights.

Then the generalized Laplacian L_X given by the explicit construction in Section III-B is not shift invariant w.r.t. L_{X^1}.

Proof: See Apendix A. ∎

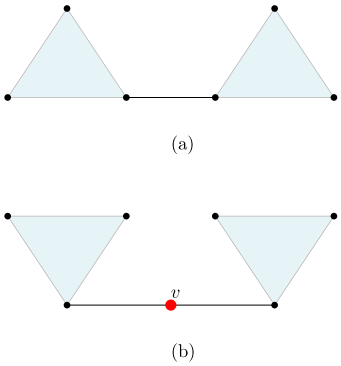

FIGURE 6. Illustration of the two situations disallowed by the first two conditions of Proposition 1.

B. CONSTRUCT 2-COMPLEXES

In this subsection, we present an approach to learn a 2-complex structure X given an undirected graph $G = (V, E)$ such that $X^1 = G$ and $X^0 = V$. If G is an unweighted graph, we assign weight 1 to each edge. Otherwise, if pairwise similarities of V are given, then we define the weight between two vertices to be the inverse of the similarity (i.e., we want two vertices to be closer if they are more similar). Therefore, without loss of generality, we may assume that $X^1 = G$ is weighted. We also assume that all edge weights in G are finite.

To construct a suitable 2-complex X that is consistent with G, we proceed as follows. We first identify the set C_{X^0} of all possible 2-simplexes. Depending on the problem, there are two main cases:

1) If $X^1 = G$ is given, then a triplet of vertices (u, v, w) belongs to C_{X^0} if and only if (u, v), (u, w) and (v, w) are all edges of G.
2) If only $X^0 = V$ is given, then we assume C_{X^0} contains any triplet of distinct vertices in V.

Given two non-negative numbers $r_1 \leq r_2$, we define $C_{X^0}(r_1, r_2)$ to be the subset of C_{X^0} consisting of triplets whose pairwise edge weights are within the interval $[r_1, r_2]$. Hence, we have the fundamental filtration

$\emptyset = C_{X^0}(0,0) \subset C_{X^0}(0,r) \subset C_{X^0}(0,r') \subset C_{X^0}(0,\infty) = C_{X^0}$ for $r \leq r'$. Next, we perform the following steps:

(a) Order all the 2-simplexes of C_{X^0} in a queue Q (illustrated in Fig. 7): (i),nolistsep
  (i) Choose $r_0 = 0 \leq r_1 \leq \ldots \leq r_m$ such that $C_{X^0} = C_{X^0}(0, r_m)$. A simplex in $C_{X^0}(0, r_i)$ is ordered before that in $C_{X^0}(0, r_{i+1}) \backslash C_{X^0}(0, r_i)$, i.e., triangles with small edge weights first.
  (ii) We order the 2-simplexes of $C_{X^0}(r_i, r_{i+1})$ in such a way that 2-simplexes sharing more edges are ordered later in the queue (with more details given below).

(b) Partition Q as a disjoint union $Q = \bigcup_{1 \leq i \leq p} Q_i$ such that their sizes are approximately uniform.

(c) Let X_0 be $X^0 \cup X^1$. For each $1 \leq i \leq p$, where p is a chosen upper limit, we construct a 2-complex X_i by adding the 2-simplexes of Q_i (and the associated edges) to X_{i-1}. We form the associated generalized Laplacians L_{X_i} using the procedure outlined in Section IV-A.

(d) Perform signal processing tasks using one of L_{X_i}. This step is problem dependent, which in particular relies on the given signal and usually involves an optimization step. We shall be more explicit in Section V-A.

For completeness, we propose one algorithm for Step (a)(ii):

(1) For each i, we randomly order all the 2-simplexes of $C_{X^0}(r_i, r_{i+1})$.

(2) We want to inductively re-order the members of $C_{X^0}(r_i, r_{i+1})$ from the initial ordering in Step IV-B above. Let s_1 be the first 2-simplex in $C_{X^0}(r_i, r_{i+1})$. Starting from $j = 1$, suppose we have already ordered $s_1, \ldots, s_j \in C_{X^0}(r_i, r_{i+1})$. They are fixed without further modifications. Search for the rest of the 2-simplexes of $C_{X^0}(r_i, r_{i+1})$. If a 2-simplex s is sharing a common edge with s_j, re-order $C_{X^0}(r_i, r_{i+1})$ by placing s at the end of $C_{X^0}(r_i, r_{i+1})$. Once all $s \in C_{X^0}(r_i, r_{i+1}) \backslash \{s_1, \ldots, s_j\}$ is gone through once, repeat the procedure by incrementing j, i.e., fixing $s_1, \ldots, s_{j+1}$ where s_{j+1} is the next 2-simplex in the ordering after s_j, and compare $s_k, k \geq j+2$ with s_{j+1}.

C. SIGNAL PROCESSING OVER 2-COMPLEXES

Given a graph $G = (V, E)$ and with L being one of the generalized Laplacians $L_{X_0}, \ldots, L_{X_p}$ constructed in Section IV-B, one may perform signal processing tasks, such as defining Fourier transform, frequency domain, sampling and filtering, similar to traditional GSP [1], as we now briefly recall. Let $L^2(V)$ be the space of signals on the vertex set V and $\{\mathbf{x}_i 0 \leq i \leq |V| - 1\}$ be an orthonormal eigenbasis of L (recall Lemma 11) corresponding to eigenvalues or *graph frequencies* $\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{|V|-1}$.

(a) Fourier transform: For a signal $\mathbf{x} \in L^2(V)$, its *Fourier transform* is given by

$$\widehat{\mathbf{x}}(i) = \langle \mathbf{x}, \mathbf{x}_i \rangle,$$

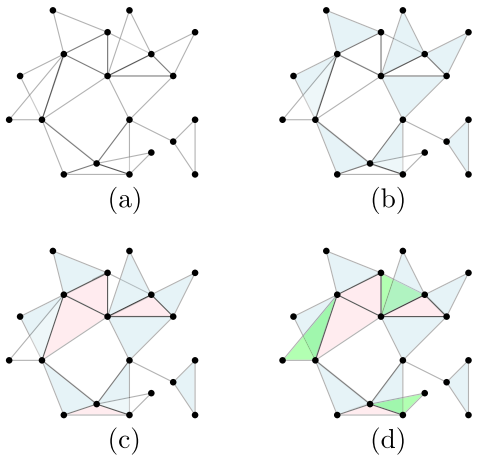

FIGURE 7. In this example, (a) shows the graph G. The blue 2-simplexes in (b) are (randomly) ordered first in Q. After which, we have the pink 2-simplexes in (c). Finally, the green 2-simplexes are ordered last in Q.

for $0 \leq i \leq |V| - 1$. The inverse transformation is given by:

$$\mathbf{x} = \sum_{0 \leq i \leq |V|-1} \widehat{\mathbf{x}}(i) \mathbf{x}_i.$$

(b) Bandlimit and bandpass filters: Suppose B is a subset of $\{0, \ldots, |V| - 1\}$. A signal $\mathbf{x} \in L^2(V)$ has *bandlimit* B if $\widehat{\mathbf{x}}(i) = \langle \mathbf{x}, \mathbf{x}_i \rangle = 0$ for $i \notin B$. The bandpass filter associated with B is given by $\mathbf{x} \mapsto \sum_{i \in B} \widehat{\mathbf{x}}(i) \mathbf{x}_i$. For denoising and data-compression, one may consider bandpass filters associated with B consisting of indices; while for anomaly detection, one may instead choose B containing large indices.

(c) Downsampling: If a signal $\mathbf{x} \in L^2(V)$ is bandlimited with B of small size, we can recover $\mathbf{x}$ by sampling signal values at a carefully chosen subset $V' \subset V$ of size $|V'| = |B|$ (V' is called a uniqueness set [37]). This is called *downsampling*.

(d) Convolution: A convolution is a generalization of bandpass filters. A convolution kernel is a signal $\mathbf{z} \in L^2(V)$. The associated convolution filter $\mathbf{x} \mapsto \mathbf{z} * \mathbf{x}$ is defined by requiring $\widehat{\mathbf{z} * \mathbf{x}}(i) = \widehat{\mathbf{z}}(i) \widehat{\mathbf{x}}(i)$ for $0 \leq i \leq |V| - 1$.

(e) Shift invariant filters: A filter F is *shift invariant* w.r.t. L if $F \circ L = L \circ F$. Suppose L does not have repeated eigenvalues. Then a shift invariant filter F is a polynomial of L.

V. NUMERICAL RESULTS

In this section, we perform numerical experiments using either synthetic and real data. Each experiment or dataset is associated with a graph G. We apply the 2-complex learning procedure in Section IV-B to obtain the generalized Laplacians $L_{X_0}, \ldots, L_{X_p}$. Our objective is to provide insights into the properties of these generalized Laplacians and which achieve the best performance metric. In particular, the baseline comparison is with the usual graph Laplacian L_{X_0}. In terms of complexity, once the Laplacians are obtained,

signal processing with any of $L_{X_i}, 0 \leq i \leq p$ is of the same complexity.

A. STRUCTURE LEARNING

We consider the Enron email graph[1] G with $n = 500$ vertices and 6815 pair-wisely connected triplets [38].

1) PROPERTIES OF GENERALIZED LAPLACIANS

We first provide some insights into the properties of the generalized Laplacians constructed as in Section IV-B. Let these be L_{X_i} for $0 \leq i \leq p = 20$ where X_i is the corresponding 2-complex and $X_0 = G$. The spectrum of the generalized Laplacians L_{X_i} for $0 \leq i \leq 20$ are plotted in Fig. 8. In our case, the number of pair-wisely connected triplets (6815) is small as compared to the size of the graph (500), as the number of ways choosing 3 vertices among n vertices is of order $\mathcal{O}(n^3)$. As a consequence, the triplets for our G do not share many common edges. Therefore, as i grows, entries of L_{X_i} tend to have smaller magnitude thus yielding a shift of the spectrum towards 0, as we observed in Fig. 8 (c.f. Lemma 3). However, the situation can reverse if a graph is densely connected with a large amount of pair-wisely connected triplets. Moreover, we observe that the spectrum pattern more or less stabilize beyond $i = 10$, which suggests that we may choose smaller p as the spectrum stratify the eigenbasis according to smoothness.

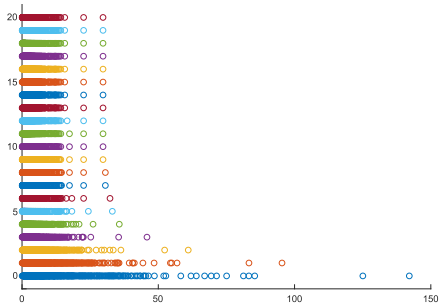

FIGURE 8. The plot shows the eigenvalue distribution (the horizontal axis) of $L_{X_0} = L_G$ to $L_{X_{20}}$ from bottom-to-top, for the Enron email graph. The spectrum tends to shift to the left, which indicate a "more connected" structure.

For further investigation of the relations among the generalized Laplacians L_{X_i}, we perform the following experiments. For each $0 \leq i \leq 20$, we generate graph signals by using the first $r = 30\%$ of the eigenvectors of L_{X_i} with random coefficients drawn uniformly from [0, 1]. For each such signal $\mathbf{f}_i$ and $0 \leq j \neq i \leq 20$, we "compress" it w.r.t. L_{X_j} as $\mathbf{f}_{i,j} = V_{r,j}V'_{r,j}\mathbf{f}_i$ and compute the error $\epsilon_{i,j} = \|\mathbf{f}_i - \mathbf{f}_{i,j}\|$. We average over different instances of $\mathbf{f}_i$ for each $0 \leq i, j \leq 20$. The heatmap for the resulting average $\epsilon_{i,j}$ is shown in Fig. 9. We see that it is more or less symmetric. For small i, j, $\epsilon_{i,j}$ is large even if $|i - j|$ is small. However, for large i, j, $\epsilon_{i,j}$ is always relatively small.

[1] https://snap.stanford.edu/data/email-Enron.html

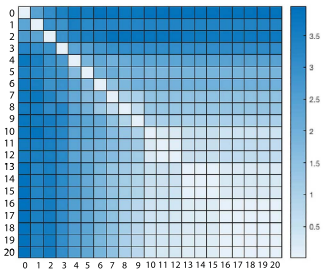

FIGURE 9. Heatmap for average $\epsilon_{i,j}$ as given by the (i, j)-th box.

2) PERFORMANCE EVALUATION

We construct a 2-complex X by randomly adding 2-simplexes for pair-wisely connected triplets in G. We suppose that $G = X^1$ is observed and X is unobserved. Let $B = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be an orthonormal eigenbasis of the generalized Laplacian L_X as constructed in Section IV-A, arranged according to increasing order of their associated eigenvalues. We randomly generate a set S_1 of signals from the span of the first $r_1 \in (0, 1]$ proportion of the eigenbasis B. Let $V_{r_1,i}$ be the matrix whose columns are the orthonormal eigenvectors corresponding to the first r_1 proportion of the eigenvalues of L_{X_i}. Then the estimated simplicial complex X_b and its Laplacian L_{X_b} is obtained by solving the optimization problem:

$$b = \arg min_{0 \leq i \leq p=20} \sum_{\mathbf{f} \in S_1} \left\| V_{r_1,i} V'_{r_1,i} \mathbf{f} - \mathbf{f} \right\|_2^2 . \tag{6}$$

We generate a set S_2 from the first $r_2 \leq r_1$ of the base signals in B, considered as a set of *compressible signals*. We want to estimate the signal compression error of the estimated Laplacian L_{X_b} as:

$$\epsilon = \sum_{\mathbf{f} \in S_2} \left\| V_{r_2,b} V'_{r_2,b} \mathbf{f} - \mathbf{f} \right\|_2 .$$

For comparison, we perform the same estimation on the usual graph Laplacian L_{X_0}, for which we do not consider high dimensional structures. On average for different choices of X, as compared to using L_{X_0}, the reduction in compression error with L_{X_b} is summarized in Table 1.

TABLE 1. Reduction in compression error.

$r_1 = r_2$	30%	50%
Error reduction	33.2%	50%

Finally, we introduce anomalies to signals in a new set S_3 (with $r_1 = 2r_2 = 50\%$) by perturbing the signal value at one random vertex. We perform spectrum analysis of the anomalous signals using both L_{X_b} and L_{X_0}. Two typical examples of the spectral plots are shown in Fig. 10 (note that we plot only the high graph frequencies). We see that using L_{X_b} (red), the anomalous behavior is more easily detected by inspecting high frequency portions.

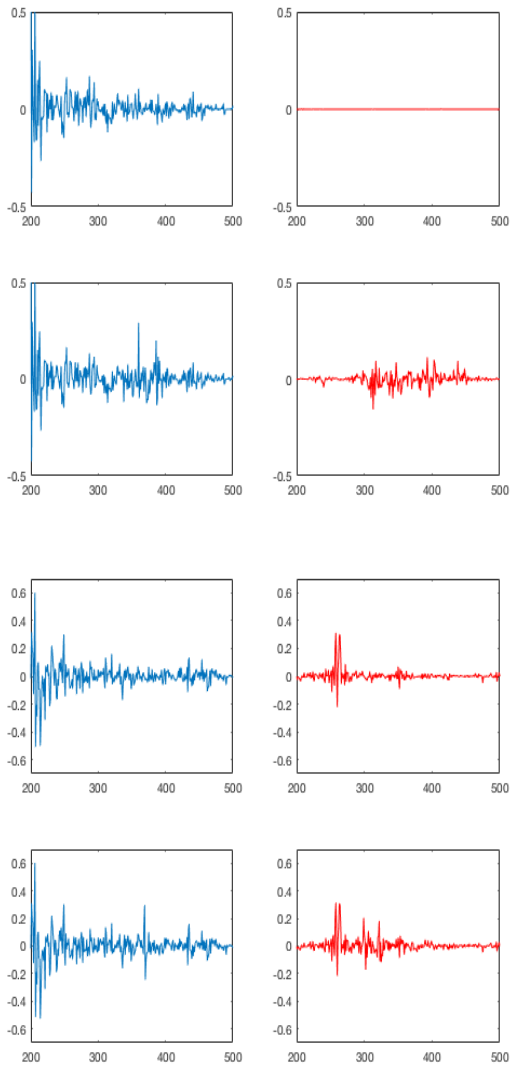

FIGURE 10. In the plots, the horizontal axis is for the graph frequency indices and the vertical axis is for the graph frequency coefficients. We have two sets of high frequency component plots, each consists of a group of four images. The top row of each set shows the plot for normal signals and the bottom row shows the plot for abnormal signals. For each set, the left figures (blue) are plots for L_{X_0} and the right figures (red) are the plots for L_{X_b}. The anomalous behavior is more visible for L_{X_b}. Note that the case on the first set is ideal in the sense that L_{X_b} happens to be Laplacian that used to generate the signal.

B. ANOMALY DETECTION

The graph used in this subsection is a weather station network in the United States with 197 vertices.[2] The graph G for the weather station network is constructed using the k-nearest neighbor algorithm based on the geographical locations of the weather stations. The graph G has 495 edges and 395 pair-wisely connected triplets.

As in Section IV-B, we construct X_i and L_{X_i} for $0 \leq i \leq p = 20$. We perform anomaly detection with real dataset. The signals on G are daily temperatures recorded over the year 2013.[3] For a daily temperature graph signal $\mathbf{x}$, we introduce anomaly to $\mathbf{x}$ by randomly perturbing the value of $\mathbf{x}$ at a single vertex. The resulting signal is denoted by $\mathbf{x}_a$.

As mentioned in Section V-A, we may look at the high frequency components of the Fourier transform of $\mathbf{x}_a$, decomposed w.r.t. L_{X_i} for $0 \leq i \leq p$, to detect if there is an anomaly. The experiment details are given as follows. We fix numbers $0 < r, \tau < 1$ and let $L \in \{L_{X_i} 0 \leq i \leq p\}$. For each instance, we randomly choose a date, and let the temperature signals on the 4 consecutive days starting from the chosen date be $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and $\mathbf{x}$, respectively. The signal $\mathbf{x}_a$ is the perturbed version of $\mathbf{x}$. Suppose that we only observe the normal signals $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and the anomalous signal $\mathbf{x}_a$. We perform graph Fourier transform w.r.t. L on $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and $\mathbf{x}_a$ to obtain $\widehat{\mathbf{x}}_1, \widehat{\mathbf{x}}_2, \widehat{\mathbf{x}}_3$ and $\widehat{\mathbf{x}}_a$, respectively. Define

$$\alpha = \max_{j=1,2,3;197r\leq k\leq 197} |\widehat{\mathbf{x}}_j(k)|,$$
$$\beta = \max_{197r\leq k\leq 197} |\widehat{\mathbf{x}}_a(k)|$$

as a measure of magnitude of the high frequency components of the signals. We declare that x_a is abnormal if $\beta/\alpha > 1+\tau$.

We run experiments with $r = 0.2$ and two choices of $\tau = 0.05, 0.2$. We average over 5000 instances for each set of parameters. We are interested in the performance in terms of percentage of successful detections under the following scenarios: S1

- S1 $L = L_{X_0}$, the usual graph Laplacian for all levels of perturbation.
- S2 The best performance among L_{X_i} for each level of perturbation.
- S3 Anomaly is declared when at least a third of the generalized Laplacians L_{X_i}, $0 \leq i \leq p$ say so.

One should note that when we set perturbation at 0, we obtain the false positive rate.

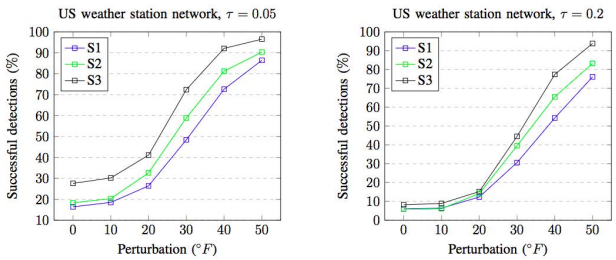

FIGURE 11. Performance of anomaly detection on the US weather station network.

The results are summarized in Fig. 11. For both cases of $\tau = 0.05, 0.2$, the best simplicial structure corresponding to scenario V-B, is X_6 when approximately 30% of pair-wisely connected triplets are added as 2-simplexes. It has a consistent overall performance. We see that in general, we do gain benefits by working with a simplicial complex instead of a graph in anomaly detection. There is little difference in the false positive rate between scenarios V-B and V-B.

[2]http://www.ncdc.noaa.gov/data-access/land-based-station-data/station-metadata

[3]ftp://ftp.ncdc.noaa.gov/pub/data/gsod

Moreover, by aggregating observations from different Laplacians L_{X_i} together, we have a much better rate of successful detection for large perturbations. For $\tau = 0.05$, the false positive rate is relatively large, while for $\tau = 0.2$, there is no significant loss in false positive rate.

C. NOISY LABEL CORRECTION

In this experiment, we consider noisy label correction with the following experimental setup. On a graph G with n vertices, suppose every vertex v belongs to a class among k classes, and thus has a class label $\mathbf{x}(v) \in \{1, \ldots, k\}$. We assume that a certain percentage of the labels are corrupted by noise. Our objective is to recover the true label. We view the vector of labels $\mathbf{x} = (\mathbf{x}(v))_{v \in V}$ as a graph signal.

An approach is to apply a convolution filter to the noisy labels. More specifically, let $\mathbf{x}_b$ be the graph signal of noisy labels and L be a shift operator. Fix a number $0 < r < 1$ and a scaling factor $0 \leq s < 1$. We first find the Fourier transform $\widehat{\mathbf{x}}_b$ of $\mathbf{x}_b$ w.r.t. L. To denoise, we scale down $\widehat{\mathbf{x}}_b(i)$, $rn \leq i \leq n$ by the factor s to obtain $\mathbf{y}_b$. To obtain the recovered label, we round off the inverse transform of $\mathbf{y}_b$ to obtain labels in $\{1, \ldots, k\}$.

The purpose of this paper is to investigate the gain or loss of using simplicial complexes over traditional graphs. Therefore, we apply the same set of parameters for different choices of L. As in Sections V-A and V-B, we construct X_i and L_{X_i} for $0 \leq i \leq p = 10$.

The graphs we consider are citation graphs: Citeseer (2120 vertices, 3679 edges and 1084 triangles) and Cora (2485 vertices, 5069 edges and 1558 triangles) [10], [39]. We briefly recall that for both graphs each vertex represents a document, and the label is the category of the document. The edges are citations links, forming the citation graphs.

For each experiment, we add white Gaussian noise to randomly selected 60% of the labels, with a constant signal-to-noise ratio (SNR) in dB. As a consequence, the observed labels might be non-integer values deviating from true labels. As we add relatively large noise, rounding-off gives wrong labels in most cases. We perform denoising with $r = 0.01$ and $s = 0.9$, as suggested by sample frequency plot of $\mathbf{x}$ and $\mathbf{x}_b$ shown in Fig. 12. For each set of parameters, we average over 500 experiments.

In terms of the reducing the amount of erroneous labels, different generalized Laplacians L_{X_i}, $0 \leq i \leq p$ perform without noticeable difference. Therefore, for each instance, we determine the L_{X_i} yielding the largest amount of correct labels. For each i, we can estimate the fraction of instances for which L_{X_i} is the best. The results are summarized in Tables 2 and 3. We boldface the three top performers across each row of the tables. In addition, the top performer is highlighted in blue.

From the results, we observe that L_{X_1} has the best overall performance consistently, which is when approximately 10% of pair-wisely connected triplets are added as 2-simplexes. Moreover, $L_{X_0}, L_{X_1}, L_{X_3}$ have consistent overall best performance. As a general trend however, the performance drops if a large amount of 2-simplexes are added, e.g., in X_9 and X_{10}.

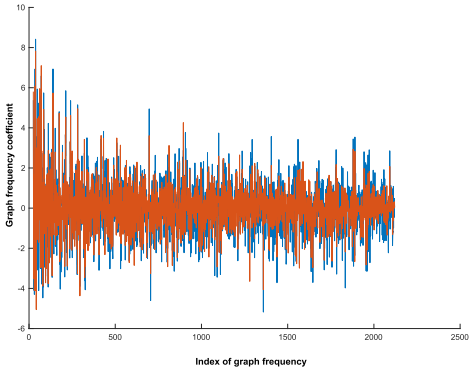

FIGURE 12. A sample frequency plot of the actual labels $\mathbf{x}$ (red) and noisy labels $\mathbf{x}_b$ (blue) in the graph frequency index range $\approx 0.01n$ to n. We observe fluctuations for both $\mathbf{x}$ and $\mathbf{x}_b$, while the amplitudes of x is smaller in general.

TABLE 2. Proportion of instances as best performer in citeseer.

SNR	L_{X_0}	L_{X_1}	L_{X_2}	L_{X_3}	L_{X_4}	L_{X_5}
12	**0.127**	**0.126**	0.084	**0.105**	0.103	0.076
10	**0.118**	**0.131**	0.079	0.094	**0.114**	0.087
8	**0.127**	**0.136**	0.083	0.093	0.082	0.092
6	**0.124**	**0.135**	0.081	0.086	0.064	0.102
4	0.102	**0.124**	0.080	**0.121**	0.087	0.088
SNR	L_{X_6}	L_{X_7}	L_{X_8}	L_{X_9}	$L_{X_{10}}$	
12	0.086	0.103	0.096	0.046	0.046	
10	0.101	0.099	0.094	0.041	0.041	
8	0.094	0.096	**0.102**	0.047	0.048	
6	0.089	**0.111**	0.099	0.055	0.055	
4	**0.104**	0.090	0.096	0.054	0.054	

TABLE 3. Proportion of instances as best performer in cora.

SNR	L_{X_0}	L_{X_1}	L_{X_2}	L_{X_3}	L_{X_4}	L_{X_5}
12	**0.134**	**0.172**	0.099	**0.121**	0.083	0.084
10	**0.115**	**0.133**	0.101	0.085	0.103	**0.129**
8	**0.124**	0.093	**0.132**	**0.153**	0.089	0.078
6	**0.136**	**0.127**	0.099	**0.104**	0.097	0.094
4	**0.136**	**0.139**	0.094	**0.106**	0.093	0.094
SNR	L_{X_6}	L_{X_7}	L_{X_8}	L_{X_9}	$L_{X_{10}}$	
12	0.050	0.087	0.060	0.051	0.060	
10	0.077	0.041	0.051	0.070	0.091	
8	0.042	0.059	0.076	0.089	0.065	
6	0.076	0.078	0.055	0.063	0.068	
4	0.077	0.061	0.057	0.059	0.085	

VI. CONCLUSION

In this paper, we have proposed a signal processing framework for signals on simplicial complexes. To do so, we introduced a generalized Laplacian that allows us to embed high order simplicial complexes in a graph. We showed that this generalized Laplacian includes the traditional graph Laplacian as a special case and enjoys several properties same as the graph Laplacian. In particular, it admits an eigenbasis,

which allows us to perform signal processing in a way similar to traditional GSP. We test the framework with both synthetic and real datasets, and observe that performance gains are achieved by working with simplicial complexes and our generalized Laplacian.

The simplicity of our framework allows natural adaptation of various GSP formulations and approaches to simplicial complexes. An interesting future research direction is to explore various such adaptations and generalizations, including continuous filter estimation and data driven end-to-end learning on simplicial complexes.

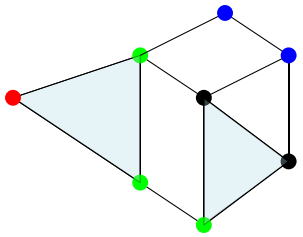

FIGURE 13. In this example, both blue vertices are 1-interior and the red vertex is 2-interior. The black and green vertices are neither of these. For the parameters in Lemma 55 $m_1 = 2$, $m_2 = 2$ and $m_3 = 1$. Moreover, in Lemma 66, $m_4 = 3$ counts the three green vertices.

APPENDIX I. NON SHIFT INVARIANCE

In this appendix, we assume X is a 2-complex and discuss conditions that ensure L_X given by the explicit construction in Section III-B is not shift invariant w.r.t. L_{X^1}. We are mainly interested in geometric conditions, which can be observed directly from the shape of X. As a corollary, we prove Proposition 1. For convenience, we introduce the following notion.

Definition 6: If a matrix M is the Laplacian of a weighted graph G, then we say M is of graph type G. We say that a 2-complex X has distinctive 2-simplexes *if (a) either $L_X - L_{X^1}$ or $L_{X^1} - L_X$ is of graph type G; and (b) if an edge belongs to a 2-simplex of X, then it also exists in G.*

Recall $k_{\max}$ and $k_{\min}$ as introduced before Lemma 3.

Lemma 4: X has distinctive 2-simplexes if any of the following holds:

(a) $k_{\max} \leq 1$, *i.e., each edge is contained in at most one 2-simplex.*
(b) $k_{\max} \leq 2$ *and all the edges have weight* 1.
(c) $k_{\min} \geq 4$ *and all the edges have weight* 1.

Proof: For claim (a), from Lemma 1(c), any constant vector is in the kernel of $L = L_{X^1} - L_X$, and the sum of each row of L is 0. If (v_i, v_j) is an edge of X not contained in any 2-simplex, then the (i, j)-th entry of L is 0. This is because the (i, j)-th entries of L_{X^1} and L_X are both negative of the weight of (v_i, v_j). It suffices to show that if (v_i, v_j), $i \neq j$, is any edge contained in a 2-simplex, then the (i, j)-th entry of L is negative. Let $a > 0$ be the weight of (v_i, v_j) and $b > 0, c > 0$ be the weights of the other two edges of the 2-simplex containing (v_i, v_j). A direct calculation shows that the (i, j)-th entry of L is $-(13a + b + c)/18 < 0$.

Claims (b) and (c) can be shown by the same argument by considering $L_{X^1} - L_X$ and $L_X - L_{X^1}$ respectively. ∎

Assume for the rest of this appendix that X has distinctive 2-simplexes. We want to study common eigenvectors of both L_{X^1} and L_X. To this end, we divide the discussion into two parts: for such an eigenvector, whether the corresponding eigenvalues are the same or different.

Definition 7: We say that a vertex v is 1-interior if it is not contained in any 2-simplex and 2-interior if each edge containing v belongs to a 2-simplex (see Fig. 13 for an example).

Lemma 5: (a) Let K be the vector space spanned by common eigenvectors of L_X and L_{X^1} that share the same corresponding eigenvalues w.r.t. L_{X^1} and L_X. Then K is a subspace of $\ker(L_X - L_{X^1})$.

(b) Let m_1 be the number of 1-interior vertices of X, m_2 be the number of connected components of the smallest complex containing all the 2-simplexes of X, and m_3 be the number of such components containing some 2-interior vertices. Then $\dim K \leq m_1 + m_2 - m_3$.

Proof:

(a) As we assume that X has distinctive 2-simplexes, $L_{X^1} - L_X = L_G$ or $-L_G$ for some graph G whose positive edge weights are supported on 2-simplexes of X. Therefore, if w is a common eigenvector of L_X and L_{X^1} with the same eigenvalue, then $L_G(w) = 0$, i.e., $w \in \ker(L_G)$. As $\ker(L_G)$ is a vector space, K as spanned by these v's is also contained in $\ker(L_G)$.

(b) Notice that the dimension of $\ker(L_G)$ is the same as the number of connected components of G. More specifically, each signal in $\ker(L_G)$ is constant on each component. The set of connected components of G consists of the following: (1) each 1-interior vertex of X is an isolated component of G, and (2) a collection of connected 2-simplexes. They are of sizes m_1 and m_2 respectively. Suppose a component C of the second type contains a 2-interior vertex v and w is a common eigenvector of L_X and L_{X^1} with the same eigenvalue $\lambda > 0$. Therefore, $\mathbf{w} \in \ker(L_G)$. As v is 2-interior and w is constant on C, for any neighbor v' of v in X, $w(v) = w(v')$. As a consequence, $L_X(w)(v) = \sum_{v'} w'(v, v')(\mathbf{w}(v) - \mathbf{w}(v')) = 0$, where the sum is over all neighbors v' of v and $w'(\cdot, \cdot)$ are the weights derived from (5). On the other hand, $L_X(w)(v) = \lambda w(v)$, and hence $w(v) = 0$. Furthermore, w is 0 on all of C as w is constant on C. Hence, the vectors of K vanish on such a C. Therefore, $\dim K \leq m_1 + m_2 - m_3$. ∎

Now we consider common eigenvectors of L_{X^1} and L_X with different eigenvalues.

Lemma 6: Suppose w is a common eigenvector of L_{X^1} and L_X with different eigenvalues. Then

(a) w is 0 *at 1-interior vertices of X.*
(b) If a vertex v belongs to a 2-simplex and any 1-interior neighbor of v is not connected to any other vertex belonging to a 2-simplex, then w is 0 *at v. Denote*

the number of such vertices by m_4 (see Fig. 13 for an example).

Proof: Suppose the eigenvalues of $\mathbf{w}$ are $\lambda_1 \neq \lambda_2$.

(a) Let v be a 1-interior vertex. Then $L_X(w)(v) = L_{X^1}(w)(v)$ as the 1-hop neighborhood of v in X and X^1 are identical. This implies that $\lambda_1 w(v) = \lambda_2 w(v)$. This is possible only if $w(v) = 0$ as $\lambda_1 \neq \lambda_2$.

(b) Let v' be a 1-interior neighbor of v. By (a), $w(v') = 0$. As v' is not connected to any other vertex belonging to a 2-simplex, w is 0 at the neighbors of v' except at v again by (a). Hence $0 = L_{X^1}(w)(v') = aw(v)$, where a is the positive edge weight between v and v'. This forces $w(v) = 0$ and proves (b). ■

Now, we are ready to state and prove the main result of this section.

Theorem 1: Suppose X is a 2-complex, X^1 is a connected graph, $|X^0| = n$ and $m_1, \ldots, m_4$ and K are as defined in Lemma 5 and Lemma 6. If $\dim K \leq m_1 + m_4 < n$, then there does not exist any orthonormal basis consisting of common eigenvectors of both L_{X^1} and L_X. In particular, this holds if $m_2 \leq m_3 + m_4$ and $m_1 + m_4 < n$.

Proof: Suppose on the contrary that $\mathbf{W} = \{w_1, \ldots, w_n\}$ is an orthonormal basis consisting of common eigenvectors of L_{X^1} and L_X. There are at most $\dim K$ vectors of $\mathbf{W}$ each sharing the same corresponding eigenvalue w.r.t. L_{X^1} and L_X. Without loss of generality, assume they are $\{w_1, \ldots, w_{\dim K}\}$ and let w_1 be the constant vector $(1/\sqrt{n}, \ldots, 1/\sqrt{n})'$. Moreover, by re-indexing, we further assume that the first $m_1 + m_4$ indices correspond to the set S of 1-interior vertices and vertices satisfying Lemma 6(b).

By abuse of notation, write $\mathbf{W}$ for the $n \times n$ matrix whose i-th column is w_i. As the columns of $\mathbf{W}$ form an orthonormal basis, so do the rows of $\mathbf{W}$. On the other hand, by Lemma 6, only the leading $(m_1 + m_4) \times \dim K$ block $\mathbf{W}_1$ of the first $m_1 + m_4$ rows of $\mathbf{W}$ can contain non-zero entries. Hence, the rows of $\mathbf{W}_1$ forms an orthonormal system. This shows that $m_1 + m_4 \leq \dim K$.

We claim that $m_1 + m_4 \neq \dim K$. For otherwise, $\mathbf{W}_1$ is a $\dim K \times \dim K$ matrix with orthonormal rows. Hence, the columns of $\mathbf{W}_1$ also forms an orthonormal system. However, this is impossible as the norm of the first column of $\mathbf{W}_1$ is $\dim K/n < 1$.

Therefore, we have shown that $m_1 + m_4 < \dim K$ with the existence of $\mathbf{W}$. This contradicts the assumption that $\dim K \leq m_1 + m_4$. Furthermore, the condition $m_2 \leq m_3 + m_4$ implies that $\dim K \leq m_1 + m_4$ by Lemma 5(b). ■

As a corollary, we can prove Proposition 1 by counting. First of all, by condition (c) in Proposition 1, X has distinctive 2-simplexes. In order to show L_X is not shift invariant w.r.t. L_{X^1}, we want to prove that they cannot have a common orthonormal eigenbasis. By Theorem 1, it suffices to show that $m_2 \leq m_4$ under the assumptions of Theorem 1. Let C be a union of 2-simplexes contributing 1 to m_2 in X. In C, there is at least one vertex v_C connected to a 1-interior point for otherwise, we can either add another 2-simplex to enlarge C or X contains no 1-interior point, which contradicts condition (b) in Proposition 1. Moreover, v_C cannot be shared by another connected union of 2-simplexes by condition (a) in Proposition 1. In conclusion, $C \mapsto v_C$ is a one-one map and hence $m_2 \leq m_4$, and Proposition 1 follows.

REFERENCES

[1] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[2] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.

[3] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, Sep. 2014.

[4] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, Dec. 2016.

[5] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 825–841, Sep. 2017.

[6] F. Grassi, A. Loukas, N. Perraudin, and B. Ricaud, "A time-vertex signal processing framework: Scalable processing and meaningful representations for time-series on graphs," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 817–829, Feb. 2018.

[7] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[8] F. Ji and W. P. Tay, "A Hilbert space theory of generalized graph signal processing," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6188–6203, Dec. 2019.

[9] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3844–3852.

[10] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[11] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[12] F. Ji, J. Yang, Q. Zhang, and W. P. Tay, "GFCN: A new graph convolutional network based on parallel flows," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 3332–3336.

[13] E. H. Spanier, *Algebraic Topology*. New York, NY, USA: Springer, 1989.

[14] S. Klamt, U.-U. Haus, and F. Theis, "Hypergraphs and cellular networks," *PLoS Comput. Biol.*, vol. 5, no. 5, May 2009, Art. no. e1000385.

[15] C. Flamm, B. M. Stadler, and P. F. Stadler, "Generalized topologies: hypergraphs, chemical reactions, and biological evolution," in *Advances in Mathematical Chemistry and Applications*. Amsterdam, The Netherlands: Elsevier, 2015, pp. 300–328.

[16] S. Lohmann and P. Díaz, "Representing and visualizing folksonomies as graphs: A reference model," in *Proc. Int. Work. Conf. Adv. Vis. Interfaces*, 2012, pp. 729–732.

[17] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[18] X. Dong, D. Thanou, M. Rabbat, and P. Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 44–63, May 2019.

[19] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, May 2019.

[20] M. Ramezani-Mayiami, M. Hajimirsadeghi, K. Skretting, R. S. Blum, and H. V. Poor, "Graph topology learning and signal recovery via Bayesian inference," in *Proc. IEEE Data Sci. Workshop (DSW)*, Jun. 2019, pp. 52–56.

[21] F. Ji, W. Tang, W. P. Tay, and E. K. P. Chong, "Network topology inference using information cascades with limited statistical knowledge," *Inf. Inference, J. IMA*, vol. 9, no. 2, pp. 327–360, Jun. 2020.

[22] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc.*, vol. 46, no. 2, pp. 255–308, Apr. 2009.

[23] O. T. Courtney and G. Bianconi, "Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 93, no. 6, Jun. 2016, Art. no. 062311.
[24] S. Barbarossa and S. Sardellitti, "Topological signal processing over simplicial complexes," 2019, *arXiv:1907.11577*.
[25] S. Barbarossa and M. Tsitsvero, "An introduction to hypergraph signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6425–6429.
[26] M. Petrovic, R. Liegeois, T. A. W. Bolton, and D. Van De Ville, "Community-aware graph signal processing," 2020, *arXiv:2008.10375*.
[27] M. Puschel, "A discrete signal processing framework for meet/join lattices with applications to hypergraphs and trees," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5371–5375.
[28] C. Wendler and M. Püschel, "Sampling signals on meet/join lattices," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.
[29] S. Zhang, Z. Ding, and S. Cui, "Introducing hypergraph signal processing: Theoretical foundation and practical applications," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 639–660, Jan. 2020.
[30] S. Zhang, Z. Ding, and S. Cui, "Introducing hypergraph signal processing: Theoretical foundation and practical applications," 2019, *arXiv:1907.09203*.
[31] F. Ji, G. Kahn, and W. P. Tay, "Traitement du signal sur les complexes simpliciaux," in *Proc. Extraction et Gestion des Connaissances*, Montpellier, France, Jan. 2021.
[32] A. Hatcher, *Algebraic Topology*. Cambridge, U.K.: Cambridge Univ. Press, 2002.
[33] C. Berge, *Graphs and Hypergraphs*. Amsterdam, The Netherlands: Elsevier, 1973.
[34] X. Ouvrard, "Hypergraphs: An introduction and review," 2020, *arXiv:2002.05014*.
[35] I. Kapovich and N. Benakli, "Boundaries of hyperbolic groups," 2002, *arXiv:math/0202286*.
[36] F. Ji, W. Tang, and W. P. Tay, "On the properties of gromov matrices and their applications in network inference," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2624–2638, May 2019.
[37] I. Pesenson, "Sampling in Paley–Wiener spaces on combinatorial graphs," *Trans. Amer. Math. Soc.*, vol. 360, no. 10, pp. 5603–5627, 2008.
[38] B. Klimt and Y. Yang, "Introducing the Enron corpus," in *Proc. CEAS*, 2004.
[39] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI Mag.*, vol. 29, no. 3, p. 93, 2008.

FENG JI received the B.S. and Ph.D. degrees in mathematics from the National University of Singapore, in 2008 and 2013, respectively. He is currently a Senior Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

GIACOMO KAHN received the Ph.D. degree in computer science from Université Clermont Auvergne, France, in 2018. From October 2018 to August 2019, he was a Researcher and a Teaching Assistant (Ater) at LIFO and Université d'Orléans. In August 2019, he was a Research Fellow at Nanyang Technological University, Singapore. He is currently an Associate Professor at Université Lumière Lyon 2. He teaches at IUT Lumière and work at the DISP Laboratory for research.

WEE PENG TAY (Senior Member, IEEE) received the B.S. degree in electrical engineering and mathematics and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2002, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.