HAL
open science

Traitement du signal sur les complexes simpliciaux

Feng Ji, Giacomo Kahn, Wee Peng Tay

To cite this version:

Feng Ji, Giacomo Kahn, Wee Peng Tay. Traitement du signal sur les complexes simpliciaux. Extraction et Gestion des Connaissances (EGC), Jan 2021, Montpellier, France. hal-03077917

HAL Id: hal-03077917

https://hal.science/hal-03077917v1

Submitted on 16 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Traitement du signal sur les complexes simpliciaux

Feng Ji*, Giacomo Kahn*,**, Wee Peng Tay*

*Laboratoire INFINITUS, Nanyang Technological University, Singapour
jifeng@ntu.edu.sg et wptay@ntu.edu.sg
**Laboratoire DISP, Université Lumière Lyon 2
giacomo.kahn@univ-lyon2.fr

Résumé. Le traitement du signal sur les graphes (GSP) est une méthode récente ayant attiré une attention conséquente, aussi bien du côté théorique qu'appliqué, et qui se développe rapidement du fait de son utilisation en apprentissage profond. Les graphes sont souvent utilisés pour modéliser des relations simples, mais ils échouent à modéliser complètement des relations n-aires. Dans cet article, nous cherchons à dépasser cette limitation et présentons un cadre pour le traitement du signal sur des structures plus générales, les complexes simpliciaux, et à travers eux le cas général des hypergraphes. Nous expérimentons notre méthode en utilisant des exemples et sur des données réelles.

1 Introduction

De nombreux jeux de données sont générés *via* des réseaux de capteurs. Les capteurs sont placés dans un environnement et leur relevé à un instant donné constitue les données. En modélisant un tel réseau de capteurs par un graphe, deux capteurs sont reliés par une arête lorsqu'ils sont suffisamment *proches*, une condition qui s'appuie généralement sur des contraintes matérielles. Pour traiter les données générées par un tel réseau, il est alors important de prendre en compte la topographie de l'environnement. Une des méthodes incorporant des propriétés topologiques dans l'analyse de données est l'utilisation du traitement du signal sur les graphes (*graph signal processing*, GSP (Shuman et al. (2013); Sandryhaila et Moura (2013, 2014))). À partir de données récupérées sur des réseaux, par exemple des réseaux de capteurs ou des réseaux sociaux, GSP utilise des représentations de graphes et leurs spectres pour effectuer des tâches de traitement telles que la compression du signal, son échantillonnage ou le filtrage de certaines fréquences. Récemment, c'est *via* ses applications en apprentissage profond que le GSP a reçu beaucoup d'attention (Defferrard et al. (2016); Egilmez et al. (2017); Li et al. (2018)).

GSP, bien qu'étant un outil extrêmement efficace, a tout de même des limitations. Les données à traiter sont variées, et l'information topologique peut également être enrichie d'autres informations ou de connaissance. On peut également rencontrer des données ne pouvant pas être exprimées sous forme de graphes, tels des nuages de points ayant des propriétés géométriques en haute dimension (voir Figure 1) ou des réseaux complexes (Klamt et al. (2009); Flamm et al. (2015)). Il est nécessaire d'aller au delà des graphes pour modéliser ces interactions complexes.

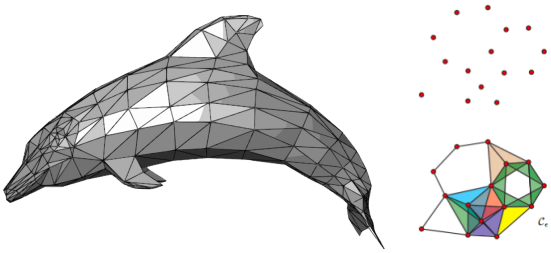


FIG. 1 – *À gauche une surface de dauphin approximée par un 2-complexe (de Wikimedia Commons), à droite un nuage de points (haut) et un complexe simplicial le couvrant (bas).*

Pour cela, nous nous intéressons aux complexes simpliciaux, une classe d'hypergraphes constituée de sommets, d'arêtes, de triangles, tétraèdres, et, en général, d-simplexes. Bien que le sujet soit relativement récent, plusieurs propositions ont été faites. Dans (Barbarossa et Tsitsvero (2016); Barbarossa et Sardellitti (2019)) les auteurs proposent une approche de traitement du signal sur des complexes simpliciaux de chaînes. Leur approche nécessite que chaque simplexe, y compris les arêtes, triangles, etc, soient munis d'un signal, et non uniquement les sommets. Dans (Puschel (2019); Wendler et Püschel (2019)), Puschël et Wendler proposent une approche de traitement du signal sur des semitreillis, en utilisant les opérations de *meet* et *join* pour décaler le signal. Ici aussi, chaque élément du treillis doit porter un signal, et pas uniquement les atomes ou co-atomes. Finalement, dans (Zhang et al. (2019)), les auteurs proposent une approche de traitement du signal plus générale, sur les hypergraphes. Comme pour notre approche, seuls les sommets portent un signal. Les hypergraphes sont représentés sous forme de tenseurs, et le signal est modifié pour avoir cette forme également. Leur méthode s'appuie ensuite sur de l'algèbre tensorielle. Nous avons fait le choix d'une modélisation plus simple des hypergraphes en introduisant la notion de *Laplacien généralisé*, et n'utilisons que de l'algèbre linéaire.

Notre approche utilise la structure géométrique du complexe en utilisant un *Laplacien généralisé*, et généralise le GSP classique lorsqu'elle est limitée aux graphes. Elle peut également être appliquée directement aux cas général des hypergraphes en considérant un complexe simplicial contenant l'hypergraphe.

Cet article est organisé de manière classique : nous introduisons d'abord les fondamentaux nécessaires à la manipulation des complexes simpliciaux dans la Section 2. Ensuite, nous introduisons des notions de GSP en Section 3. Notre contributions commencent en Section 4 avec la définition du Laplacien généralisé. Nous étudions le cas spécial des 2-complexes dans la Section 5. Finalement, nous présentons des résultats expérimentaux en Section 6.

2 Complexes simpliciaux

Dans cette section, nous introduisons brièvement les notions de topologie algébrique qui nous seront utiles dans la compréhension de ce papier. Un lecteur curieux peut se référer à (Spanier (1989); Hatcher (2002)) pour de plus amples informations.

Définition 1 *Le simplexe standard de dimension n (ou n-simplexe standard), dénoté Δ_n, est l'ensemble* $\{x \in \mathbb{R}_+^{n+1} \mid x_0 + x_1 + \cdots + x_n = 1\}$.

Un simplexe est la généralisation d'un triangle à une dimension arbitraire. Tout espace topologique homéomorphe à un n-simplexe standard est appelé un n-simplexe. Nous donnons les exemples des simplexes des dimensions 0 à 3 en Figure 2. Dans Δ_n, si nous forçons k coordonnées à être nulles, obtenons un $(n-k)$-simplexe, appelé une *face* de Δ_n.

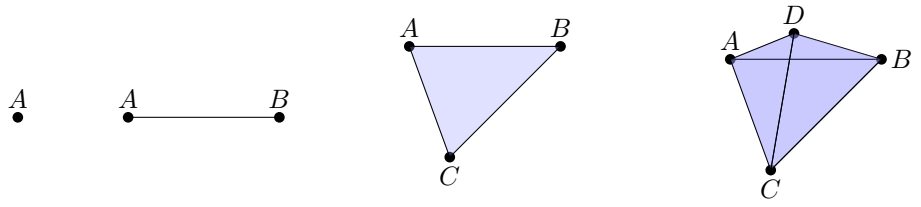


FIG. 2 – *De gauche à droite un 0-simplexe (un point), un 1-simplexe (un segment), un 2-simplexe (un triangle) et un 3-simplexe (un tetrahedron).*

Un *complexe simplicial* X (voir Figure 3) est un ensemble de simplexes tel que toute face d'un simplexe de X est aussi dans X, et l'intersection de deux simplexes σ_1 et σ_2 de X est aussi dans X. Un simplexe est *maximal* s'il n'est la face d'aucun autre simplexe. Nous ne nous intéressons qu'à des complexes simpliciaux finis, c'est-à-dire des ensembles finis de simplexes. Un d-complexe est un complexe dont la dimension – la plus haute dimension parmi ses simplexes – est d. Pour un entier positif m, la restriction d'un complexe simplicial X à ses simplexes de dimensions au plus m est le m-squelette du complexe et est notée X^m.

Pour notre utilisation dans cet article nous préférons une description combinatoire d'un n-simplexe sous la forme d'un ensemble de $n+1$ labels. De cette manière, les $faces$ d'un n-simplexe $\{0,\ldots,n\}$ sont simplement les sous-ensembles de $\{0,\ldots,n\}$. Un complexe simplicial X est alors la famille d'ensembles obtenue en prenant tous les sous-ensembles des simplexes maximaux de X. Il est toujours possible de créer un objet géométrique pour les complexes simpliciaux définis de cette manière. La *réalisation géométrique* d'un complexe simplicial X est l'espace topologique obtenu en collant les simplexes ayant des faces en commun.

Dans cet article, un complexe simplicial est pondéré si son 1-squelette est un graphe pondéré. S'il ne l'est pas, on considère que ses 1-simplexes (ses arêtes) ont un poids de 1.

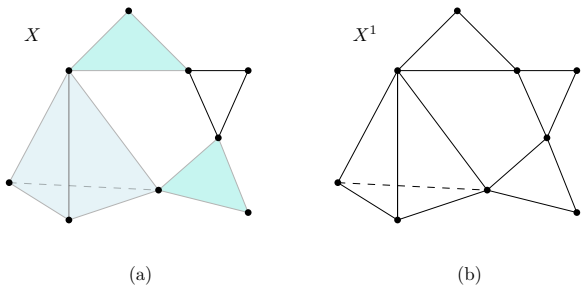


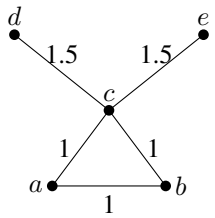
FIG. 3 – *(a) X est la réalisation géométrique d'un 3-complexe avec un 3-simplexe maximal, deux 2-simplexes maximaux et trois 1-simplexes maximaux. (b) X^1 est un graphe à 9 sommets et 15 arêtes.*

Les complexes simpliciaux sont une étape entre les graphes et les hypergraphes. Un hypergraphe $H = (V, E)$ est une paire où V est un ensemble de sommets et E un ensemble de sous-ensembles non vides de V, appelés hyperarêtes. Un complexe simplicial X peut être considéré comme un hypergraphe (X^0, E) où chaque simplexe de dimension supérieure à 0 est une hyperarête. On peut associer à tout hypergraphe H un complexe simplicial X_H dont les simplexes sont l'ensemble des sous-ensembles des hyperarêtes maximales de H. De cette manière, notre méthode de traitement du signal peut être appliquée à un hypergraphe H *via* son complexe simplicial associé X_H.

3 *Graph signal processing*

Nous invitons un lecteur intéressé à consulter (Stankovic et al. (2019)) pour une introduction en profondeur de GSP.

Soit $G = (V, E)$ un graphe non-dirigé à n sommets, muni d'une fonction de poids $w : E \mapsto \mathbb{R}_+$. Sa matrice d'adjacence, A, est une matrice carrée $n \times n$ ayant $w(u, v)$ en position u, v si (u, v) est une arête de E et 0 sinon. La matrice des degrés D de G est une matrice $n \times n$ où l'entrée (u, u) est égale au degré du sommet u (la somme des poids des arêtes qui lui sont adjacentes) et les autres entrées sont nulles. Finalement, la matrice Laplacienne L d'un graphe est donnée par $L = D - A$. Un graphe et ses matrices d'adjacence et Laplacienne sont donnés en Figure 4. Pour un graphe non-dirigé, ces matrices sont symétriques.



(a)

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1.5 & 1.5 \\ 0 & 0 & 1.5 & 0 & 0 \\ 0 & 0 & 1.5 & 0 & 0 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 5 & -1.5 & -1.5 \\ 0 & 0 & -1.5 & 1.5 & 0 \\ 0 & 0 & -1.5 & 0 & 1.5 \end{bmatrix}$$

(c)

FIG. 4 – *(a) Un graphe pondéré, (b) sa matrice d'adjacence et (c) sa matrice Laplacienne.*

Un signal sur un graphe à n sommets est obtenu en associant une valeur $x(v)$ à chaque sommet, comme illustré en Figure 5 (à gauche). Ce signal peut-être arrangé sous forme d'un vecteur colonne $\mathbf{x} = [x(0), x(1), \ldots, x(n-1)]^T$, ici $[0.7, 0.4, 0.2, -0.5, -0.8]^T$. Le graphe est alors considéré comme le domaine du signal. [1]

Là où l'échantillonnage ou le décalage d'un signal sur un axe temporel discret trouvent une solution immédiate (prendre une valeur sur deux, ou décaler d'une unité de temps), ce n'est pas le cas dans les graphes. La notion de "un sommet sur deux" n'admet pas toujours une réponse évidente dans le cas général. Il est cependant nécessaire pour les méthodes de traitement du signal sur les graphes de définir un opérateur de décalage ou de *shift*. Dans les faits, on utilise souvent la matrice d'adjacence ou la matrice Laplacienne. La Figure 5 donne un exemple de décalage du signal basé sur la matrice d'adjacence. Le signal est alors porté d'un sommet à ses

1. À noter qu'un signal sur un domaine temporel discrétisé est alors un signal sur un graphe de type chemin ou cycle.

FIG. 5 – *En utilisant la matrice d'adjacence comme opérateur de décalage, un signal est envoyé vers les sommets voisins. Un sommet reçoit donc la somme des signaux de ses voisins, pondéré par le poids des arêtes.*

sommets adjacents. Pour un signal $\mathbf{x}$, le signal décalé selon l'opérateur A (matrice d'adjacence) est alors $A\mathbf{x}$, ici $[0.6, 0.9, -0.85, 0.3, 0.3]^T$.

3.1 Transformée de Fourier sur un graphe

Certaines méthodes d'analyse de données utilisent une estimation du signal utilisant le domaine spectral. Dans cette section, nous montrons que l'on peut porter ces notions à un signal indexé par un graphe en utilisant le spectre de la matrice d'adjacence ou de la matrice Laplacienne. Dans cet article, nous privilégions la matrice Laplacienne pour les définitions et les applications.

Définition 2 *La* transformée de Fourier sur un graphe *d'un signal* $\mathbf{x}$ *est*

$$\mathbf{X} = \mathbf{U}^{-1}\mathbf{x}$$

où $\mathbf{X}$ *est un vecteur contenant les coefficients de la transformée de Fourier et* $\mathbf{U}$ *est la matrice dont les colonnes sont les vecteurs propres de la matrice Laplacienne.*

Les éléments de $\mathbf{X}$ sont notés $X(k)$, pour k dans $\{0, \ldots, n-1\}$. La matrice Laplacienne étant une matrice symétrique, $\mathbf{U}^{-1} = \mathbf{U}^T$. Chacun des éléments $X(k)$ de la transformée est une projection du signal sur le vecteur propre u_k . La transformée de Fourier sur un graphe est alors un ensemble de projections sur l'ensemble des vecteurs propres $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{n-1}$, qui servent de vecteurs orthonormés d'une base. La transformée de Fourier inverse est donnée par $x = \mathbf{U}\mathbf{X}$.

Grâce à ces deux opérations permettant de basculer du domaine du signal au domaine spectral (fréquentiel), nous pouvons effectuer diverses opérations : filtration des hautes fréquences en ne gardant que les fréquences associées aux valeurs propres les plus petites, compression du signal en ne gardant que certaines fréquences, évaluation de la régularité (*smoothness*) du signal, lissage du signal, etc.

4 Laplacien généralisé

Nous souhaitons généraliser le GSP, mais retrouver le GSP classique lorsque on utilise un graphe. Pour cela, nous introduisons la notion de Laplacien généralisé.

Définition 3 *Soit X un complexe simplicial fini de taille n. Soient un graphe pondéré, non-dirigé $G_X = (V, E)$, une fonction $f : X^0 \mapsto V$, une transformation linéaire $T : \mathbb{R}^{|X^0|} \mapsto \mathbb{R}^{|V|}$, tels que les conditions suivantes sont satisfaites :*

(a) f est injective,

(b) la composante $f(v)$ de $T(x)$ est la même que la composante v de x, pour tous les $v \in X^0$ et $x \in \mathbb{R}^{X^0}$,

(c) la somme de chaque ligne de la matrice de l'application linéaire T est une constante.

Soit L_{G_X} la matrice Laplacienne du graphe G_X. Le Laplacien généralisé *associé avec (G_X, f, T) est*

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

avec T' la matrice adjointe de T. Nous notons L_X pour $L_{(G_X,f,T)}$ si le contexte n'induit pas de confusion.

La condition sur l'injectivité de f nous assure que f "plonge" X dans G_X, de manière à ce que l'on puisse appliquer l'opérateur de *shift* directement sur G_X. Les conditions (b) et (c) sur T permettent qu'un signal sur $v \in X^0$ soit préservé par son image $f(v)$ dans G_X, tandis que les signaux dans $V \setminus f(X)$ seront formés par les valeurs moyennes.

Lemme 4 *1. L_X est symétrique,*

2. L_X est semi-définie positive,

3. les signaux constants sont dans l'espace propre associé à la valeur propre 0 de L_X. Ce 0-eigenspace est de dimension 1 si et seulement si G_X est connecté.

Preuve La preuve est disponible dans la version longue sur ArXiv (Ji et al. (2020)).

Ce lemme montre que L_X possède certaines propriétés que nous recherchons pour un opérateur de *shift*. Tout d'abord, sa symétrie nous assure que ses vecteurs propres sont une base orthonormée. Nous pouvons donc créer une théorie de Fourier analogue au GSP classique. Le fait que les vecteurs constants soient dans le 0-eigenspace nous permet également de faire un parallèle avec le GSP classique, car les signaux constants sont les plus lisses.

Le Lemme 4 nous assure que L_X est similaire au Laplacien d'un graphe. Cette approche ne serait pas utile si L_X était toujours le Laplacien d'un graphe. Ce n'est pas le cas : il arrive que L_X contienne des valeurs positives en dehors de sa diagonale (cf. Section 5).

4.1 Construction explicite

Dans cette section, nous donnons une construction de L_X en choisissant (G_X, f, T) pour un complexe simplicial. Un prérequis implicite est que nous souhaitons retrouver la matrice Laplacienne traditionnelle quand le complexe étudié est en fait un graphe.

Commençons par le cas le plus simple, c'est-à-dire quand $X \cong \Delta_n$ est un n-simplexe pondéré (i.e., dont le 1-squelette est un graphe pondéré). Soit $d(\cdot, \cdot)$ la fonction de distance dans X. Les sommets de X sont notés $v_1, \ldots, v_{n+1}$. Le graphe $G_X = (V, E)$ est construit de la manière suivante (illustrée en Figure 6) : $V = \{v_1, \ldots, v_{n+1}, u\}$ est l'ensemble des sommets du simplexe avec un unique sommet additionnel u, que nous considérons comme le barycentre

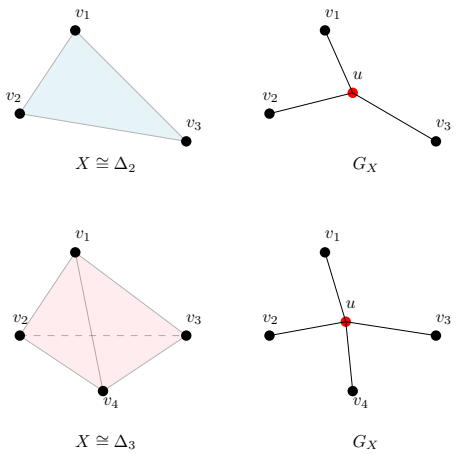


FIG. 6 – *Illustration de la forme de G_X lorsque $X \cong \Delta_2$ (haut) et $X \cong \Delta_3$ (bas).*

de X. Il n'y a pas d'arêtes entre les sommets v_i et v_j, pour i et j dans $\{1, \ldots, n+1\}$. En revanche, (v_i, u) est une arête pour tous les i dans $\{1, \ldots, n+1\}$.

Le poids $w(v_i, u)$ associé à l'arête (v_i, u) dans G_X est calculé de la manière suivante :

$$w(v_i, u) = \binom{n}{2}^{-1} \left(\sum_{v_i \neq v_j \neq v_k \neq v_i} (v_j, v_k)_{v_i} \right)$$

où $(v_j, v_k)_{v_i} = (d_X(v_i, v_j) + d_X(v_i, v_k) - d_X(v_j, v_k))/2$ est le produit de Gromov de v_j et v_k à v_x (Kapovich et Benakli (2002); Ji et al. (2019)). Quand $n = 1$, le poids à donner aux nouvelles arêtes est une version modifiée du produit de Gromov, qui donne $w(v_1, u) = (v_2, v_2)_{v_1} = d(v_1, v_2)$, ce qui nous permet de retrouver le Laplacien classique.

Cette construction mène à un choix canonique pour f, où chaque sommet est envoyé vers lui-même, c'est-à-dire $f(v_i) = v_i$. De plus, T est l'identité sur chaque composant v_i et la moyenne des signaux de $\{v_1, \ldots, v_{n+1}\}$ est associée à u. De cette manière, (G_X, f, T) satisfait les conditions de la définition 3 et donne lieu à un Laplacien généralisé L_x.

Dans un complexe simplicial X, nous pouvons obtenir une décomposition X^{max} de X comme étant l'ensemble des simplexes maximaux de X. De cette manière, le Laplacien généralisé d'un complexe simplicial X est

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

où la somme est sur les simplexes maximaux, en respectant les labels des sommets de σ dans X.

Une intuition de cette construction vient du fait qu'un simplexe $X \cong \Delta_n$ est homotopiquement équivalent à un point. Par conséquent, si on veut approcher X par un graphe G_X qui conserve cette propriété, il faut que G_X soit un arbre. De plus, si on ne veut pas casser la symétrie des sommets du graphe, il est naturel de considérer un sommet additionnel comme un

barycentre connecté à tous les sommets du simplexe d'origine. Les poids sur les arêtes de G_X sont choisis de manière à approcher la métrique de X.

5 Traitement du signal sur les 2-complexes

Dans cette section, nous nous intéressons principalement aux 2-complexes. Soit un 2-simplexe pondéré $X \cong \Delta_2$ dont les sommets sont $\{v_1, v_2, v_3\}$ et qui a pour fonction de poids $w(\cdot,\cdot)$. Selon les choix de G_X, f, T donnés en section précédente, les poids de G_x sont donc :

$$\begin{aligned} a &= d(v_1, u) = (v_2, v_3)_{v_1} = (w(v_1, v_3) + w(v_1, v_2) - w(v_2, v_3))/2, \\ b &= d(v_2, u) = (v_1, v_3)_{v_2} = (w(v_2, v_3) + w(v_1, v_2) - w(v_1, v_3))/2, \\ c &= d(v_3, u) = (v_1, v_2)_{v_3} = (w(v_1, v_3) + w(v_2, v_3) - w(v_1, v_2))/2. \end{aligned}$$

et le Laplacien généralisé associé

$$\frac{1}{9}\begin{bmatrix} b+c+4a & c-2a-2b & b-2a-2c \\ c-2a-2b & a+c+4b & a-2b-2c \\ b-2a-2c & a-2b-2c & a+b+4c \end{bmatrix}.$$

On note donc que pour certains X, L_X ne peut pas être le Laplacien d'un graphe : par exemple, si $c > 2a + 2b$, il y aura certaines entrées non-négatives en dehors de la diagonale.

5.1 Inférence d'un 2-complexe

Dans le cas général, le complexe simplicial peut ne pas être connu. Nous décrivons une approche, illustrée en Figure 7, qui permet d'approximer une structure de 2-complexe en utilisant le signal sur les sommets. Si X^1 n'est pas pondéré, les arêtes reçoivent un poids de 1.

Notre but étant de transformer certains triplés de sommets en 2-complexes, nous commençons par identifier les candidats potentiels. Si X^1 est connu, l'ensemble des candidats est l'ensemble des triplés de sommets reliés deux à deux dans X^1. Sinon, tous les triplés de sommets sont candidats.

On ordonne dans une file Q les triplés de candidats, de sorte à considérer en premier les triangles les plus petits, c'est-à-dire les triplés de sommets tels que leurs distances deux à deux sont les plus petites. On considère que ce sont ces triplés qui ont le plus de chance d'être, conceptuellement, des 2-simplexes. Si tous les poids sont égaux, Q permet de distribuer uniformément les triplés choisis sur la structure. On partitionne Q en p sous-ensembles $Q_i, i \in \{1, \ldots, p\}$ de taille uniforme.

Soit $X_0 = X^0 \cup X^1$. Pour chaque $1 \leq i \leq p$, on construit un 2-complexe X_i en ajoutant les simplexes de Q_i à X_{i-1}. On crée les Laplacien généralisés associés L_{X_i}. Ainsi, par exemple, si l'on fixe p à 10, X_0 sera le graphe de départ, X_1 verra les 10% de triangles les plus petits transformés en 2-simplexes, X_2 les 20%, et ainsi de suite.

On peut alors utiliser un des L_i pour approcher le "vrai" Laplacien. Cette étape, ainsi que la notion de "vrai" Laplacien, dépend beaucoup du contexte et du signal, et peut passer par la résolution d'un problème d'optimisation, comme explicité dans la section suivante.

Une fois qu'une structure est approchée et qu'un Laplacien est construit, on définit une théorie de Fourier de manière analogue au GSP classique.

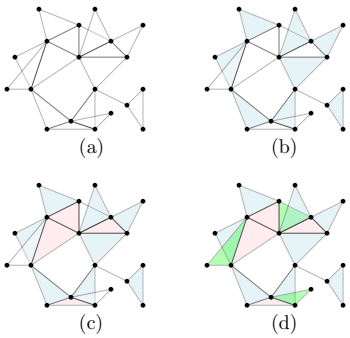


FIG. 7 – *Dans cet exemple, (a) montre un graphe, auquel on ajoute d'abord les 2-simplexes bleus (b), puis les 2-simplexes roses (c), et finalement les 2-simplexes verts dans (d), en suivant une file imaginaire Q.*

6 Résultats expérimentaux

6.1 Les graphes sont-ils toujours la meilleure structure ?

Soit G un graphe issu d'un réseau social (Klimt et Yang (2004) [2]) à 500 sommets et possédant 6815 triplés de sommets connectés deux à deux (triangles). Nous construisons un 2-complexe X en ajoutant aléatoirement des 2-simplexes pour certains triangles. De cette manière, le graphe $G = X^1$ est connu, mais pas X. Soit $B = \{f_1, \ldots, f_n\}$ une base propre de L_X ordonnée selon l'ordre croissant des valeurs propres associées. On génère aléatoirement un ensemble S_1 de signaux depuis le sous-espace vectoriel engendré par les $r_1\%$ premiers vecteurs de la base.

Pour inférer X depuis S_1, nous construisons des X_i et leurs Laplacien généralisés L_{X_i} comme décrit dans la section précédente, pour $i \in \{0, \ldots, 20\}$, avec $X_0 = X^1 = G$. Nous créons la matrice $V_{r_1,i}$ dont les colonnes sont les $r_1\%$ premières valeurs propres de L_{X_i}. Pour choisir le complexe simplicial X_b et son Laplacien généralisé, il faut alors résoudre le problème d'optimisation suivant :

$$b = \underset{i \in \{0,\ldots,20\}}{\arg\min} \sum_{f \in S_1} ||V_{r_1,i} V'_{r_1,i} f - f||_2^2.$$

Soit S_2 un ensemble de signaux créés à partir des $r_2\% \leq r_1\%$ premiers vecteurs de B, considérés comme un ensemble de signaux *compressibles*. Nous évaluons l'erreur err de compression du signal pour un Laplacien L_{X_b} comme étant $err = \sum_{f \in S_2} ||V_{r_2,b} V'_{r_2,b} f - f||_2$. Nous utilisons $L_{X_0} = L_G$ comme point de comparaison. En moyenne, l'erreur de compression est réduite de 33.2% pour $r_1 = r_2 = 30\%$ et 40.6% pour $r_1 = r_2 = 50\%$ par rapport à L_{X_0}.

Finalement, soit S_3 un signal créé de la même manière que S_2, mais pour lequel on va ajouter une anomalie à un sommet aléatoirement. La Figure 8 montre le spectre de L_{X_0} et d'une des L_{X_i} pour les hautes fréquences. En regardant le spectre des L_{X_b}, on voit plus facilement les variations dans les hautes fréquence lorsqu'une anomalie est introduite. Nous nous penchons plus en détail sur ce problème dans la section suivante.

2. `https://snap.stanford.edu/data/email-Enron.html`

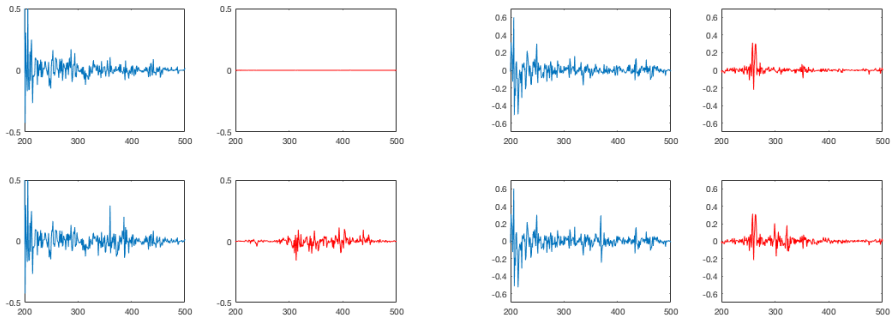


FIG. 8 – *La ligne du haut représente les hautes fréquences des signaux normaux pour L_{X_0} en bleu (notre point de comparaison avec l'existant sous la forme du GSP classique) et L_{X_b} en rouge. Sur la ligne du bas, on a introduit une anomalie dans le signal. Les spectres associés aux complexes montrent cette anomalie plus clairement que ceux associés au graphe.*

6.2 Détection d'anomalies

Le graphe utilisé dans cette section est construit à partir de stations météorologiques [3]. Le jeu de données identifie les positions des stations aux États-Unis d'Amérique, et nous en extrayons un graphe à 197 sommets, chaque station étant alors reliée à ses k voisins les plus proches. Ce graphe a 495 arêtes et 395 triangles. Comme précédemment, nous construisons les X_i et leur Laplacien généralisé pour $i \in \{0, \ldots, 20\}$ à partir de ce graphe.

Les signaux sont des relevés quotidiens de température de l'année 2013 [4]. Soit x un signal choisi parmi les signaux de l'année 2013, nous introduisons une anomalie en perturbant la valeur de x à un unique sommet (d'une quantité variable, servant de paramètre en Figure 9). Le signal résultant est noté x_a, x muni d'une anomalie. Nous souhaitons regarder les composants de haute fréquence de la transformée de Fourier de x_a, pour chacune des matrices L_{X_i}, $i \in \{0, \ldots, 20\}$. Notre protocole est le suivant : pour chaque instance, choisissons une date, et perturbons le signal pour obtenir x_a. À partir des signaux des trois jours précédents, nous obtenons un seuil pour les composants en hautes fréquences du signal. Nous déclarons que x_a est anormal si ses composants en hautes fréquence dépassent ce seuil.

Nous étudions les performances pour la détection d'une anomalie dans les cas suivants : (S1) en utilisant $L = L_{X_0}$, le Laplacien du graphe, (S2) la meilleure performance parmi les L_{X_i}, (S3) $L = L_{X_b}$ le Laplacien avec la meilleure performance en général ($b = 2$ dans notre cas), (S4) une anomalie est déclarée quand au moins $1/3$ des L_{X_i} la détectent.

Les résultats de ce protocole sont résumés en Figure 9. On peut voir qu'en général, on bénéficie de l'usage d'un complexe simplicial plutôt qu'un simple graphe. La meilleure structure dans notre cas est X_2, où environ 10% des triangles les plus "petits" sont ajoutés en tant que 2-complexes.

3. `https://www.ncdc.noaa.gov/data-access/land-based-station-data/station-metadata`
4. `ftp://ftp.ncdc.noaa.gov/pub/data/gsod`

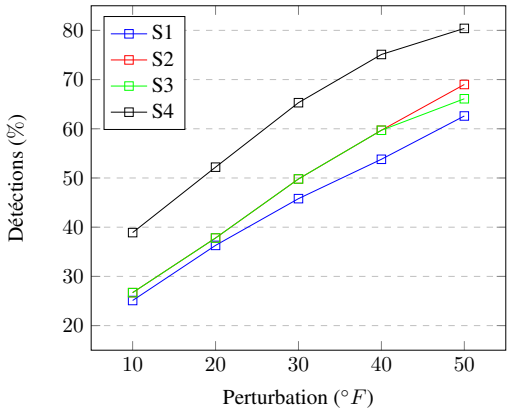


FIG. 9 – *Performances de la détection d'une anomalie sur le réseau de stations météorologiques aux USA.*

7 Conclusion

Nous donnons une méthode de construction d'un Laplacien généralisé sur un espace qui n'est pas forcément un graphe. De là, le traitement d'un signal découle d'une manière similaire à celle de GSP. En testant notre approche sur des données synthétiques et réelles, nous observons qu'il y a un avantage à travailler sur des structures en hautes dimensions plutôt que de simples graphes.

Références

Barbarossa, S. et S. Sardellitti (2019). Topological signal processing over simplicial complexes. *arXiv preprint arXiv :1907.11577*.

Barbarossa, S. et M. Tsitsvero (2016). An introduction to hypergraph signal processing. In *EEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 6425–6429.

Defferrard, M., X. Bresson, et P. Vandergheynst (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Inform. Process. Syst.*, USA, pp. 3844–3852.

Egilmez, H. E., E. Pavez, et A. Ortega (2017). Graph learning from data under laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing 11*(6), 825–841.

Flamm, C., B. M. Stadler, et P. F. Stadler (2015). Generalized topologies : hypergraphs, chemical reactions, and biological evolution. In *Advances in Mathematical Chemistry and Applications*, pp. 300–328. Elsevier.

Hatcher, A. (2002). *Algebraic Topology*. Cambridge University Press.

Ji, F., G. Kahn, et W. P. Tay (2020). Signal processing on simplicial complexes. *arXiv preprint arXiv :2004.02392*.

Ji, F., W. Tang, et W. P. Tay (2019). On the properties of Gromov matrices and their applications in network inference. *IEEE Trans. Signal Process. 67*(10), 2624 – 2638.

Kapovich, I. et N. Benakli (2002). Boundaries of hyperbolic groups. *arXiv preprint math/0202286*.

Klamt, S., U.-U. Haus, et F. Theis (2009). Hypergraphs and cellular networks. *PLoS computational biology 5*(5), e1000385.

Klimt, B. et Y. Yang (2004). Introducing the enron corpus. In *CEAS*.

Li, R., S. Wang, F. Zhu, et J. Huang (2018). Adaptive graph convolutional neural networks. In *Thirty-second AAAI conference on artificial intelligence*.

Puschel, M. (2019). A discrete signal processing framework for meet/join lattices with applications to hypergraphs and trees. In *IEEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 5371–5375.

Sandryhaila, A. et J. M. F. Moura (2013). Discrete signal processing on graphs. *IEEE Trans. Signal Process. 61*(7), 1644–1656.

Sandryhaila, A. et J. M. F. Moura (2014). Big data analysis with signal processing on graphs : Representation and processing of massive data sets with irregular structure. *IEEE Signal Process. Mag. 31*(5), 80–90.

Shuman, D. I., S. K. Narang, P. Frossard, A. Ortega, et P. Vandergheynst (2013). The emerging field of signal processing on graphs : Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag. 30*(3), 83–98.

Spanier, E. H. (1989). *Algebraic topology*. Springer Science & Business Media.

Stankovic, L., D. Mandic, M. Dakovic, M. Brajovic, B. Scalzo, et T. Constantinides (2019). Graph signal processing–part I : Graphs, graph spectra, and spectral clustering. *arXiv preprint arXiv :1907.03467*.

Wendler, C. et M. Püschel (2019). Sampling signals on meet/join lattices. In *Proc. Global Conference on Signal and Information Processing (GlobalSIP)*.

Zhang, S., Z. Ding, et S. Cui (2019). Introducing hypergraph signal processing : theoretical foundation and practical applications. *arXiv preprint arXiv :1907.09203*.

Summary

Theoretical development and applications of graph signal processing (GSP) have attracted much attention. In classical GSP, the underlying structures are restricted in terms of dimensionality. A graph is a combinatorial object that models binary relations, and it does not directly model complex n-ary relations. A possible high-dimensional generalisaiton of graphs are simplicial complexes. They are a step between the constrained case of graphs and the general case of hypergraphs. In this paper, we develop a signal processing framework on simplicial complexes, such that we recover the traditional GSP theory when restricted to signals on graphs. We show some possible applications of our framework with high dimensional sensor networks.