Transport Optimal pour les Signaux sur Graphes

Titouan VAYER[1], Laetitia CHAPEL[1], Rémi FLAMARY[2], Romain TAVENARD[3], Nicolas COURTY[1]

[1]Université Bretagne-Sud, CNRS, IRISA, F-56000 Vannes

[2]Université Côte d'Azur, CNRS, OCA Lagrange, F-06000 Nice

[3]Université de Rennes 2, CNRS, LETG, F-35000 Rennes

tvayer@irisa.fr,lchapel@irisa.fr,rflamary@unice.fr,
romain.tavenard@univ-rennes2.fr,ncourty@irisa.fr

Résumé – Ce papier considère le problème du calcul d'une distance entre deux signaux portés par des graphes. L'originalité de notre approche provient de l'interprétation probabiliste de cet objet, où le signal est modélisé comme une distribution de probabilités dans un espace métrique spécifique. Nous proposons une nouvelle distance basée sur la notion de **transport**. A la différence des distances de transport classiques comme la distance de Wasserstein ou de Gromov-Wasserstein, notre distance exploite simultanément l'information présente dans le signal et le graphe. Nous présentons dans ce papier certaines de ses propriétés, ainsi que son application dans des problèmes de calcul de barycentres.

Abstract – This work considers the problem of computing distances between structured objects such as undirected graphs, seen as probability distributions in a specific metric space. We consider **a new transportation distance** (*i.e.* that minimizes a total cost of transporting probability masses) that unveils the geometric nature of the structured objects space. Unlike Wasserstein or Gromov-Wasserstein metrics that focus solely and respectively on features (by considering a metric in the feature space) or structure (by seeing structure as a metric space), our new distance exploits jointly both information. After discussing some of its properties, we show results on its application to the computation of Fréchet means or barycenters of graphs.

1 Introduction

Le domaine du traitement du signal sur graphes (TSG) s'intéresse au traitement de données (signaux) supportées sur des structures (graphes) [12], les graphes représentant ici un lien ou une structure existant entre les données, alors que les signaux représentent des informations portées par les noeuds de ces graphes. De nombreuses instances de ce type de données existent dans le monde réel ; nous pouvons citer à titre d'exemples la modélisation de composés chimiques [9], la connectomique [10], ou bien encore les réseaux sociaux [21]. Alors que beaucoup d'approches classiques de traitement du signal (TS) présupposent la régularité et l'uniformité de l'échantillonnage du signal, la possibilité ici que les données soient disposées selon une structure irrégulière (le graphe) invite à repenser une bonne partie des outils existants. Les problèmes liés au TSG sont nombreux, par exemple l'extension des techniques classiques de TS pour le filtrage des données, l'échantillonnage de grands graphes, la détection de communautés ou bien l'apprentissage de graphes, dans la mesure où plusieurs instances sont disponibles et qu'il est possible de leur associer une étiquette. Plusieurs variantes de ce dernier problème existent, notamment si la tâche de classification porte plutôt sur les noeuds.

Dans la majorité des cas, les principaux outils mathématiques utilisés sont des extensions de l'analyse harmonique aux graphes, basée notamment sur les notions de Laplacien combinatoire des graphes, construits eux-mêmes à partir des matrices d'adjacence définissant la structure [12]. Basée sur les mêmes outils, l'extension de la notion de la convolution aux graphes a permis l'utilisation de réseaux de neurones profonds dans des tâches de classification [5], permettant d'apprendre de bout en bout la meilleure représentation du signal porté par les noeuds ainsi que le classifieur adapté à cette représentation.

Une distance de transport pour le TSG. En contraste avec ces précédentes approches, nous proposons dans ce papier de modéliser les signaux sur graphes comme des distributions de probabilité existant dans un espace métrique donné. Nous dérivons alors une distance entre des distributions qui peut être potentiellement utilisée dans des applications du TSG. Le contexte est donc celui où plusieurs signaux sont disponibles, et doivent être comparés. Dans ce contexte, disposer d'une telle distance entre signaux présente de multiples intérêts : *i)* elle peut être utilisée dans la plupart des algorithmes d'apprentissage basés distances, comme par exemple une recherche de plus proches voisins, ou un algorithme des k-moyennes *ii)* a contrario des approches uniquement basées données/apprentissage qui construisent des représentations relativement à une tâche donnée, une distance ne dépend pas d'un quelconque ensemble d'apprentissage et nous renseigne sur la géométrie de l'espace des signaux sur graphes, et finalement *iii)* nous permet de considérer des objets géométriques tels que des interpolations géodésiques ou des barycentres. Ce travail constitue à notre connaissance

une des toutes premières approches pour définir une distance entre signaux sur graphes.

Définir une telle distance n'est cependant pas une tâche directe : même si on peut toujours comparer les valeurs du signal en utilisant une métrique standard comme ℓ_2, la comparaison de graphes nécessite une similarité basée sur la notion d'isométrie, dans la mesure où les noeuds ne sont pas ordonnés et directement comparables. Le transport optimal (TO) s'est avéré être un outil majeur pour la comparaison de distributions avec de nombreuses applications en apprentissage automatique (par exemple [8, 6, 3]), supporté par une théorie riche [20] et de nombreux progrès sur les algorithmes de résolution du problème sous-jacent [14]. Cependant, aucune formulation existante du TO n'intègre directement à la fois les données et l'information structurelle liant ces données. Il est à noter toutefois que des formulations alternatives du problème de transport originel existent. Notamment, suivant les travaux de Mémoli [11], Peyré *et al.* [15] ont proposé une manière de comparer des matrices de distances (et donc potentiellement la structure du graphe). Cette extension, appelée distance de Gromov-Wasserstein, a pour but de comparer deux espaces métriques, mais ne prend pas en compte les données portées par les noeuds. Des approches récentes visent à intégrer de la structure dans la distance de TO (ou distance de Wasserstein) par ajout d'une régularisation spécifique (voir par exemple [2] ou [6] qui contraignent la structure du plan de transport). Thorpe *et al.* ont proposé une distance [17], nommée TL^p, incluant un coût qui est une combinaison linéaire d'une distance entre données et entre éléments de structure, mais celle-ci nécessite une correspondance a priori entre les signaux (cas d'images ou de séries temporelles de tailles similaires). A l'inverse, nous proposons un cadre générique pour formuler la question d'une distance entre deux signaux portés par des graphes de tailles potentiellement différentes. La suite de cet article décrit ce formalisme (Section 3) avant de proposer quelques exemples d'utilisation de cette distance (Section 4).

Notations. Le simplex unité à n sommets et décrivant l'espace des histogrammes h est noté $\Sigma_n = \{h \in (\mathbb{R}_+^*)^n, \sum_{i=1}^n h_i = 1, \}$. On note $\otimes$ le produit tensoriel, *i.e.* pour un tenseur $L = (L_{i,j,k,l})$, $L \otimes B$ est une matrice $\left(\sum_{k,l} L_{i,j,k,l} B_{k,l}\right)_{i,j}$. $\langle . \rangle$ est le produit scalaire matriciel associé à la norme de Frobenius. Pour tout $x \in \Omega$, δ_x dénote la mesure de Dirac en x.

2 Signal sur graphe vu comme une mesure de probabilité

Dans ce papier, nous cherchons à comparer des signaux sur graphe, c'est-à-dire la combinaison d'un signal (porté par les noeuds) et d'une structure (relations entre les noeuds). Plus formellement, on considère un graphe non orienté comme un tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \ell_f, \ell_s)$ où $(\mathcal{V}, \mathcal{E})$ sont les ensembles de nœuds et arrêtes du graphe. $\ell_f : \mathcal{V} \to \Omega_f$ est une fonction qui associe à chaque nœud $v_i \in \mathcal{V}$ une valeur du signal $a_i \overset{\text{def}}{=} \ell_f(v_i)$. On

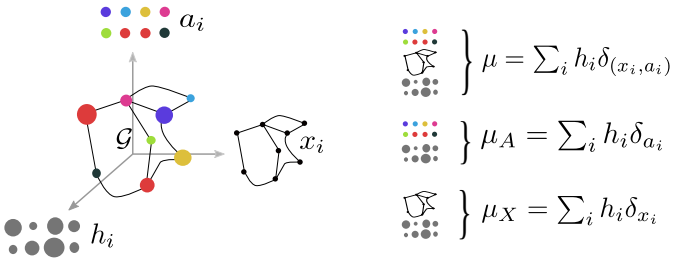

FIGURE 1 – (Figure de gauche) Signal sur graphe. $(a_i)_i$ est l'information du signal, $(x_i)_i$ sa structure et l'histogramme $(h_i)_i$ mesure une importance relative des nœuds. (Figure de droite) Le signal sur graphe associé est représenté par une mesure de probabilité μ sur l'espace produit signal/structure, avec comme marginales μ_X et μ_A.

appellera *signal* les valeurs $(a_i)_i$ sur les nœuds. De la même manière $\ell_s : \mathcal{V} \to \Omega_s$ est une fonction qui projette le nœud v_i dans un espace métrique (Ω_s, C) encodant les relations entre les nœuds du graphe $x_i \overset{\text{def}}{=} \ell_s(v_i)$. Cet espace est spécifique à chaque graphe et va donc changer lorsque la structure du graphe est modifiée. $C : \Omega_s \times \Omega_s \to \mathbb{R}_+$ est une application symétrique qui va mesurer la similarité entre deux nœuds. En pratique (Ω_s, C) est implicite et on travaille uniquement avec C. Dans la suite on utilisera C pour désigner la matrice carrée de similarité entre tous les nœuds du graphe $(C(i,k) = C(x_i, x_k))_{i,k}$. Nous utiliserons pour calculer cette matrice soit des distances de type plus court chemin ou des distances harmoniques [19].

Nous proposons d'enrichir la description d'un signal sur graphe ci-dessus par un histogramme qui va pondérer l'importance relative de chaque nœud. Pour un graphe de n nœuds, on associe donc à chaque nœud une pondération $(h_i)_i \in \Sigma_n$. Le signal sur graphe sera donc défini par le tuple $\mathcal{S} = (\mathcal{G}, h_\mathcal{G})$ où $\mathcal{G}$ est le graphe défini ci-dessus et $h_\mathcal{G}$ est la fonction qui associe la pondération à chaque nœud. Chaque graphe est donc une mesure de probabilité supportée par l'espace produit entre le signal et la structure du graphe $\mu = \sum_{i=1}^n h_i \delta_{(x_i,a_i)}$ (voir Figure 1). Si aucune information concernant l'importance des nœuds est disponible, un poids uniforme $h_i = \frac{1}{n}$ peut être utilisé pour décrire le signal sur graphe.

3 Distance Fused Gromov-Wasserstein

On cherche à définir une distance entre deux graphes $\mathcal{G}_1$ et $\mathcal{G}_2$, décrits respectivement par les distributions $\mu = \sum_{i=1}^n h_i \delta_{(x_i,a_i)}$ et $\nu = \sum_{i=1}^m g_j \delta_{(y_j,b_j)}$ où $h \in \Sigma_n$ et $g \in \Sigma_m$ sont des histogrammes. On supposera également $(x_i, a_i) \neq (x_j, a_j)$ pour $i \neq j$ (de même pour y_j et b_j).

Soit $\Pi(h, g)$ l'ensemble des couplages entre h et g, *i.e.* :

$$\Pi(h,g) = \{\pi \in \mathbb{R}_+^{n \times m} \text{ s.t. } \sum_{i=1}^n \pi_{i,j} = h_j \ , \sum_{j=1}^m \pi_{i,j} = g_i\},$$

où $\pi_{i,j}$ représente dans π la quantité de masse déplacée entre (x_i, a_i) et (y_j, b_j).

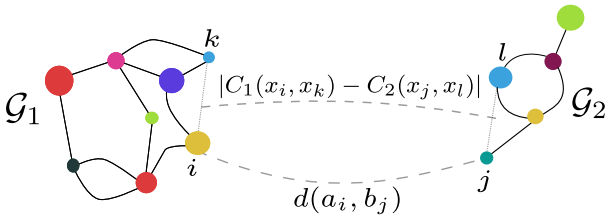

FIGURE 2 – Le coût E_q associé à la distance FGW pour un couplage π dépend à la fois de la similarité entre les signaux associés aux nœuds de chaque graphe $(d(a_i, b_j))_{i,j}$ et de la concordance entre les structures des graphes $(|C_1(x_i, x_k) - C_2(x_j, x_l)|)_{i,j,k,l}$.

$M_{AB} = (d(a_i, b_j))_{i,j}$ est une matrice de taille $n \times m$ contenant les distances croisées entre les valeurs des signaux. Les matrices de structure de graphe C_1 et C_2, encodent les informations de relation entre nœuds dans les graphes. Nous définissons une similarité entre nœuds de graphes différents à travers leurs relation pair-à-pair à l'aide du tenseur 4D $L(C_1, C_2)$:

$$L_{i,j,k,l}(C_1, C_2) = |C_1(i, k) - C_2(j, l)|.$$

Distance FGW On définit maintenant la distance basée transport optimal FGW pour "Fused Gromov-Wasserstein" pour un paramètre $\alpha \in [0, 1]$ comme

$$FGW_{q,\alpha}(\mu, \nu) = \min_{\pi \in \Pi(h,g)} E_q(M_{AB}, C_1, C_2, \pi) \qquad (1)$$

avec

$$E_q(M_{AB}, C_1, C_2, \pi) = \left\langle (1-\alpha) M_{AB}^q + \alpha L(C_1, C_2)^q \otimes \pi, \pi \right\rangle$$
$$= \sum_{i,j,k,l} (1-\alpha) d(a_i, b_j)^q + \alpha |C_1(i, k) - C_2(j, l)|^q \pi_{i,j} \pi_{k,l}$$

La distance FGW cherche un couplage π entre les nœuds des graphes qui minimise le coût E_q qui est une combinaison linéaire de $d(a_i, b_j)$ sur les valeurs des signaux et $|C_1(i, k) - C_2(j, l)|$ entre les paires de nœuds (cf Figure 2). Le paramètre α permet de trouver un compromis entre l'importance de l'information portée par le signal et celle portée par la structure du graphe. On cherchera donc à associer des nœuds qui sont similaires au sens de leur valeur et de leur relation aux autres dans le graphe. FGW utilise d pour mesurer la similarité entre les valeurs d'un signal, elle peut donc être utilisé sur des valeurs discrètes ou continues et également lorsque les nombre de nœuds sont différents.

Cette nouvelle distance FGW constitue une généralisation de la distance de Wasserstein [20] et Gromov-Wasserstein [11]. Il est possible de montrer que FGW possède bien les propriétés d'une métrique quand $q = 1$ et de demi-métrique pour $q > 1$ (nous renvoyons le lecteur à [18] pour le détail des preuves mathématiques). Elle peut être utilisée dans un large panel d'applications qui nécessitent une distance (k plus proches voisins, méthodes de plongement pour la visualisation, ensembles minimaux représentatifs, etc.). Une propriété intéressante est que l'inspection du couplage optimal π peut dans une certaine mesure être interprété, et magnifier certaines relations spécifiques entre deux signaux.

Le calcul de FGW est un problème d'optimisation quadratique en π sous contraintes linéaires d'égalité (contraintes de marginales). Il est possible de le résoudre de manière efficace à l'aide d'un algorithme de gradient conditionnel généralisé (Frank-Wolfe) comme proposé dans [6] pour le calcul de problèmes de transport optimal régularisés. Les détails de l'algorithme sont disponibles dans [18].

Barycentres Fused Gromov-Wasserstein Les barycentres à base de transport optimal ont de bonnes propriétés géométriques [1, 16], mais jusqu'à maintenant il n'était pas possible d'encoder à la fois des informations de signal et de graphe dans un même barycentre. Nous montrons que FGW permet de définir un barycentre de signaux sur graphes comme moyennes de Fréchet.

Pour cela, on cherche un signal sur graphe μ qui minimise une somme pondérée de distances FGW par rapport à un ensemble de signaux $(\mu_k)_k$ ayant des matrices de graphe $(C_k)_k$, des signaux $(B_k)_k$ et des histogrammes $(h_k)_k$. On suppose que les pondérations h du barycentre sont connues et fixées ; ainsi que le nombre de noeuds N du barycentre. Dns ce contexte pour un $N \in \mathbb{N}$ et $(\lambda_k)_k$ donné tel que $\sum_k \lambda_k = 1$, on cherche le signal $A = (a_i)_i$ et la matrice de graphe C du barycentre qui minimisent

$$\min_{\mu} \sum_k \lambda_k FGW_{q,\alpha}(\mu, \mu_k) \qquad (2)$$
$$= \min_{C \in \mathbb{R}^{N \times N},\ A \in \mathbb{R}^{N \times n}, (\pi_k)_k} \sum_k \lambda_k E_q(M_{AB_k}, C, C_k, \pi_k)$$

Ce problème est jointement convexe par rapport à C et A mais pas par rapport aux π_k. En pratique on le résoud avec un algorithme de descente alternée par bloc en mettant à jour C, A et π_k séquentiellement.

4 Illustrations des barycentres FGW

Barycentre de graphes Dans ce premier exemple, on utilise FGW pour estimer un barycentre de graphes. Pour cela on génère des graphes bruités ayant une forme de cercle ou de 8 avec un signal 1D suivant respectivement un sinus ou une variation linéaire. Le nombre de noeud est tiré aléatoirement entre 10 et 25 ainsi que du bruit additif sur le signal et des connections entre voisins de degré 2.

On calcule le barycentre de FGW entre 10 examples en prenant la distance des plus proches voisins pour les similarités entre les noeuds des graphes et la distance euclidienne entre les valeurs des signaux. On reconstruit un graph par seuillage sur la matrice de distance C du barycentre. Les résultats sont montrés dans la Figure 3 et illustrent la capacité de débruitage (signal et graphe) et de compression (en nombre de noeuds) de FGW. On voit clairement que non seulement la structure est préservée mais que le signal a également été lissé pour correspondre en moyenne à tous les autres examples. C'est à notre connaissance la seule méthode capable d'estimer des barycentres de signaux sur graphes ne partageant pas la même structures de graphe.

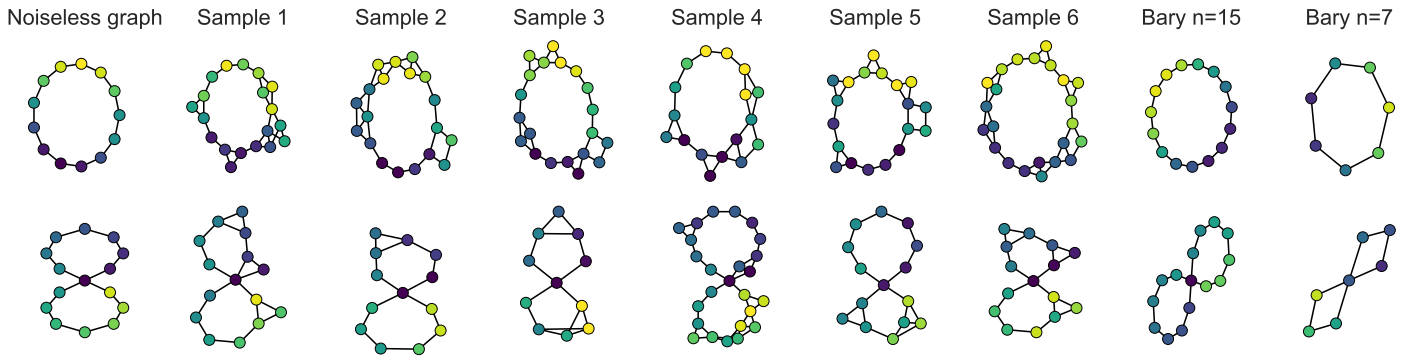

FIGURE 3 – Illustration de FGW pour les barycentres de graphes. La première colonne présente les graphes non bruités utilisés pour générer les exemples bruités qui constituent le jeu de données (colonnes 2 à 7). Les colonnes 8 et 9 présentent les barycentres pour chaque type de graphe, calculés avec différents nombres de noeuds. Un noeud bleu correspond à une valeur de signal proche de -1, jaune proche de 1.

Moyenne de séries temporelles. Nous nous plaçons dans le contexte de séries temporelles. Nous prenons comme exemple une classe du jeu de données *Trace* venant de l'archive UEA [4]. Nous calculons la moyenne des exemples de cette classe échantillonnée en 20 instants et nous comparons notre méthode aux algorithmes DBA [13] et soft-DTW [7] (Figure 4). La distance ℓ_2 entre les temps des échantillons est utilisée pour l'information structurelle, ainsi que pour comparer les valeurs des séries. Les approches DTW sont invariantes aux distortions temporelles et échouent à capturer la "forme" des séries, alors que FGW produit une moyenne visuellement plus satisfaisante. De plus, on constate que les instants supports de cette série ne sont pas répartis uniformément, montrant la capacité de la méthode à densifier les instants d'échantillonnages pour capturer des changements abruptes du signal.

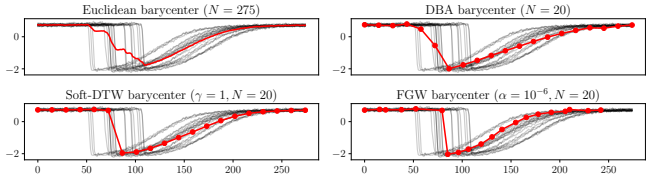

FIGURE 4 – Moyenne de séries temporelles correspondant au jeu de données *Trace*

Références

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2) :904–924, 2011.

[2] David Alvarez-Melis, Tommi S. Jaakkola, and Stefanie Jegelka. Structured Optimal Transport. In *AISTATS*, 2018.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223, Sydney, Australia, 06–11 Aug 2017.

[4] Antony Bagnall, Jason Lines, Williams Vickers, and Eamonn Keogh. The uea & ucr time series classification repository. `www.timeseriesclassification.com`.

[5] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning : going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4) :18–42, 2017.

[6] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEETPAMI*, 39(9) :1853–1865, 2017.

[7] Marco Cuturi and Mathieu Blondel. Soft-DTW : a differentiable loss function for time-series. In *Proceedings of the ICML*, volume 70, pages 894–903, Sydney, Australia, 06–11 Aug 2017.

[8] G. Huang, C. Guo, M. Kusner, Y. Sun, F. Sha, and K. Weinberger. Supervised word mover's distance. In *NIPS*, pages 4862–4870, 2016.

[9] Nils M. Kriege, Pierre-Louis Giscard, and Richard C. Wilson. On valid optimal assignment kernels and applications to graph classification. *CoRR*, abs/1606.01141, 2016.

[10] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. Distance metric learning using graph convolutional networks : Application to functional brain networks. In *MICCAI*, pages 469–477, 2017.

[11] Facundo Memoli. Gromov wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, pages 1–71, 2011.

[12] Antonio Ortega, Pascal Frossard, Jelena Kovačević, Jose MF Moura, and Pierre Vandergheynst. Graph signal processing : Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5) :808–828, 2018.

[13] François Petitjean, Alain Ketterlin, and Pierre GanÃ§arski. A global averaging method for dynamic time warping, with applications to clustering. *Elsevier Pattern Recognition*, 44(3) :678–693, 2011.

[14] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *arXiv preprint arXiv :1803.00567*, 2018.

[15] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *ICML*, New-York, 2016.

[16] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *ICML*, pages 2664–2672, 2016.

[17] Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K. Rohde, and Dejan Slepčev. A transportation l^p distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59(2) :187–210, Oct 2017.

[18] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data. *arXiv preprint arXiv :1805.09114*, 2018.

[19] Saurabh Verma and Zhi-Li Zhang. Hunt for the unique, stable, sparse and fast feature learning on graphs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*, pages 88–98. Curran Associates, Inc., 2017.

[20] Cédric Villani. *Optimal Transport : Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, 2009 edition, 2008.

[21] Pinar Yanardag and S.V.N. Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1365–1374, 2015.