Analyse statistique de données à structure de graphes, via des processus de diffusion de la chaleur.

Etienne Lasalle [1,2]

[1] *Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay, Rue Michel Magat Bâtiment 307, 91405 Orsay, France.*
[2] *DataShape, Inria Saclay.*
etienne.lasalle@universite-paris-saclay.fr

Résumé. Lorsqu'on travaille avec des graphes pondérés, on peut interpréter les poids des arêtes comme des conductivités thermiques: la chaleur se diffuse plus rapidement le long des arêtes ayant un poids important. Dès lors, on peut comparer les graphes en comparant leur répartition de chaleur au bout d'un temps de diffusion t. Ce paramètre t, jouant le rôle d'un paramètre d'échelle, doit être minutieusement choisi pour s'assurer des comparaisons pertinentes. A l'opposé de certains travaux considérant un temps de diffusion fixé arbitrairement ou choisi à partir des données, on se propose de prendre en compte tout le processus de diffusion de la chaleur. Pour cela, on définit des processus à valeurs réelles indexés par tous les temps de diffusion dans [0,T], concaténant toutes les comparaisons faites aux différentes échelles. Dans cet exposé, je présenterai ces processus de comparaisons, ainsi que leurs propriétés statistiques. Notamment, on montre qu'ils vérifient un théorème central limite fonctionnel et qu'ils admettent des approximations gaussiennes. En pratique, ces résultats permettent de construire des bandes de confiance asymptotiques ainsi que des tests à deux échantillons.

Mots-clés. graphes, processus empiriques, diffusion de la chaleur

1 Introduction

De part leurs capacités à modéliser de nombreuses situations et grâce aux moyens d'acquisition et de stockage de plus en plus développés, les données à structures de graphes reçoivent une attention croissante depuis ces dernières décennies. Une effort particulier a été fourni pour développer des notions de distance entre graphes. Cependant, ces notions de distances sont extrêmement dépendantes du contexte et des caractéristiques des graphes considérés. Plus particulièrement, différentes informations peuvent être disponibles, selon si les graphes sont orientés ou non, pondérés ou non, possèdent le même nombre de sommets ou pas. Une autre information clé peut être la connaissance d'une correspondance entre les sommets des graphes. Dans le cas où cette information est connue, de nombreuses métriques ont été développées et se basent sur des comparaisons

locales, à l'échelle des paires de nœuds. En revanche lorsque cette information manque, les graphes sont plutôt comparés à une échelle mésoscopique, voire macroscopique, à travers des résumés de leurs structures.

Dans ce travail, nous considérons des graphes non orientés, éventuellement pondérés et souhaitons être capables de gérer des graphes sans correspondance entre les sommets et éventuellement de tailles différentes. Pour cela, nous définissons deux processus de comparaisons basés sur la diffusion de la chaleur. Le premier suppose la correspondance entre les sommets connues, le second s'affranchit de ces hypothèses. Leur analyse nous mène au développement de méthodes bootstrap pour déterminer des bandes de confiances ainsi que des tests à deux échantillons. Nous présentons ces processus dans la section suivante.

2 Processus de comparaison

2.1 Notions de distance, via la diffusion de la chaleur

Considérons $\mathcal{G}_n$ l'ensemble des graphes pondérés de taille n partageant les mêmes sommets $\{1, \dots, n\}$ et tels que les poids des arêtes sont positifs. La diffusion de la chaleur sur un graph $G \in \mathcal{G}_n$ peut se caractériser par la matrice symétrique $K_t(G)$, de taille $n \times n$, appelée *noyau de la chaleur* au temps $t \geq 0$. Le coefficient (i, j) de $K_t(G)$ représente la quantité de chaleur au noeud i au temps t, lorsque seule une unité de chaleur se trouvait au nœud i au temps $t = 0$. Pour deux graphes G, G' dans $\mathcal{G}_n$, on définit leur *Heat Kernel Distance* (HKD) au temps t par

$$D_t((G, G')) = \|K_t(G) - K_t(G')\|_F \tag{1}$$

où $\|\cdot\|_F$ représente la norme de Frobenius.

Pour comparer des graphes sans correspondance entre les sommets et potentiellement de tailles différentes, on travaille dans $\mathcal{G}^n$, l'ensemble des graphes de taille *au plus* n. On définit alors une nouvelle notion de distance, en utilisant la théorie de la persistance étendue, un outil provenant de l'analyse topologique des données (voir Cohen-Steiner et al. (2009); Edelsbrunner and Harer (2010); Oudot (2015) pour une introduction complète). Pour un graphe G de taille n avec pour sommets $V = \{1, \dots, n\}$, Hu et al. (2014) définissent la Heat Kernel Signature (HKS) au temps t par la fonction $h_t(G) : V \to \mathbb{R}$, correspondant à la diagonal du noyau de la chaleur $K_t(G)$. Ainsi, $h_t(G)(i)$ correspond à la chaleur restante au nœud i lorsqu'au temps initial l'unité de chaleur était placé en i. Une fois les graphes de $\mathcal{G}^n$ munis de leur HKS, on peut définir leur diagrammes de persistance $Dg : G \to Dg(G, h_t(G))$ en suivant l'approche de Carrière et al. (2020). Les diagrammes de persistance sont des multi-ensembles de points de $\mathbb{R}^2$. Ils encode la manière dont les composantes topologiques, comme les composantes connexes ou les trous, évoluent au sein des familles de sous-graphes associés aux sous-niveaux et sur-niveaux des HKS. Les

diagrammes de persistances peuvent ensuite être comparés via la distance Bottleneck d_B, et cela en s'affranchissant de toute hypothèse sur la taille des graphes ou l'identification de sommets. Ainsi, on peut définir la Heat Persistence Distance (HPD) au temps t entre deux graphes G, G' dans $\mathcal{G}^n$ par

$$H_t((G, G')) = d_B(Dg(G, h_t(G)), Dg(G', h_t(G'))). \tag{2}$$

A la fois pour la HKD et la HPD, le choix d'un paramètre t pertinent et informatif est crucial. Cependant, en l'absence d'information *a priori* sur le problème, ce choix semble compliqué à effectuer. Ainsi, nous proposons d'adopter un point de vue processus, et de considérer l'ensemble des distances pour des t dans $[0, T]$.

2.2 Empirical Processes

Soit P une mesure de probabilité sur $\mathcal{G}_n \times \mathcal{G}_n$ et considérons un N-échantillon $((G_1, G'_1), \ldots, (G_N, G'_N))$ tiré sous P. On s'intéresse alors aux propriétés statistiques du processus empirique associé à la HKD :

$$\left\{ \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} D_t((G_i, G'_i)) - \mathbb{E}_P \left[D_t((G, G')) \right] \right), \ t \in [0, T] \right\} \tag{3}$$

De même,pour un N-échantillon $((G_1, G'_1), \ldots, (G_N, G'_N))$ tiré sous une mesure de probabilité P sur $\mathcal{G}^n \times \mathcal{G}^n$, on étudie le processus empirique associé à la HPD :

$$\left\{ \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} H_t((G_i, G'_i)) - \mathbb{E}_P \left[H_t((G, G')) \right] \right), \ t \in [0, T] \right\} \tag{4}$$

Grâce au caractère lipschitzien des fonctions $t \to D_t((G, G'))$ et $t \to H_t((G, G'))$, nous sommes en mesure de prouver la convergence faible des ces processus vers des processus gaussiens indexés par $[0, T]$. Ce résultat assure la validité asymptotique des méthodes bootstrap construisant des bandes de confiance contenant $t \to \mathbb{E}_P \left[D_t((G, G')) \right]$ ou $t \to \mathbb{E}_P \left[H_t((G, G')) \right]$ (voir Kosorok (2008)), ou construisant des tests à deux échantillons (voir Van Der Vaart and Wellner (1996)).

De plus, on montre que la vitesse de convergence des processus HKD et HPD vers les processus gaussiens est indépendante de la dimension du problème, *i.e.* la taille des graphes n.

Bibliographie

Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M., and Umeda, Y. (2020). Perslay: a neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pages 2786–2796. PMLR.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2009). Extending persistence using poincaré and lefschetz duality. *Foundations of Computational Mathematics*, 9(1):79–103.

Edelsbrunner, H. and Harer, J. (2010). *Computational topology: an introduction*. American Mathematical Soc.

Hu, N., Rustamov, R. M., and Guibas, L. (2014). Stable and informative spectral signatures for graph matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2305–2312.

Kosorok, M. R. (2008). Introduction to empirical processes. *Introduction to Empirical Processes and Semiparametric Inference*.

Oudot, S. Y. (2015). *Persistence theory: from quiver representations to data analysis*, volume 209. American Mathematical Society Providence.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*. Springer.