Reconstruire l'invisible : GRIOT pour l'imputation des attributs dans les graphes par transport optimal

Richard Serrano*, Charlotte Laclau**, Baptiste Jeudy*, Christine Largeron*

*Laboratoire Hubert Curien, Saint-Étienne 42000, France
{prénom}.{nom}@univ-st-etienne.fr,
**Télécom Paris, Institut Polytechnique de Paris, France
{prénom}.{nom}@telecom-paris.fr

Résumé. Ces dernières années, l'apprentissage automatique pour les graphes attribués a progressé grâce aux réseaux de neurones pour graphes (GNN) (Kipf et Welling, 2017). Cependant, ces méthodes supposent que les attributs sont entièrement connus, ce qui est rarement le cas dans les graphes réels. Cet article explore le potentiel du transport optimal (TO) pour imputer les valeurs d'attributs manquantes dans des graphes attribués. Nous proposons une nouvelle fonction de perte multi-vues basée sur le TO, intégrant les attributs des nœuds et la structure du graphe. Cette fonction permet d'entraîner une architecture GNN capable d'imputer simultanément toutes les valeurs manquantes et, dans un context dynamique, d'imputer de nouveaux nœuds. Nous évaluons notre approche sur des données synthétiques et réelles. Les résultats montrent que notre méthode est compétitive avec l'état de l'art, et meilleure, en particulier sur les graphes faiblement homophiles.

1 Introduction

Les graphes attribués modélisent des relations complexes, mais souffrent souvent de données manquantes, compromettant leur analyse (Kossinets, 2006). Dans les réseaux sociaux, par exemple, certaines informations comme le genre ou l'âge peuvent être absentes, entraînant des valeurs manquantes dans les attributs des nœuds.

L'imputation des données manquantes est un défi important (Little et Rubin, 2019), influencé par les mécanismes des données manquantes, i.e. Masque Complètement Aléatoire (MCAR), Masque Aléatoire (MAR), et Masque Non Aléatoire (MNAR) (Schafer et Graham, 2002). Les méthodes d'imputation naïves ignorent la structure des graphes, limitant leur efficacité (Huisman, 2014). Les méthodes classiques reposent souvent sur l'homophilie des graphes (Rossi et al., 2022), mais les graphes réels peuvent être hétérophiles (Zheng et al., 2022). Notre méthode, basée sur le Transport Optimal (TO), impute les attributs manquants en tenant compte de la topologie, dans le but d'être robuste au mécanisme de données manquantes et à une faible homophilie. Le TO a prouvé son efficacité pour l'imputation de données tabulaires (Muzellec et al., 2020) et la prédiction de graphes (Brogat-Motte et al., 2022). Son utilisation se justifie

par sa capacité à mesurer, par la distance de Wasserstein, l'écart entre des distributions relatives aux attributs nœuds d'un graphe. Cependant, appliquer la distance de Wasserstein uniquement aux attributs néglige la topologie du graphe.

Contributions. ⋆ Nous proposons donc une nouvelle mesure de distance, *Multi-vues Wasserstein* (MultiW), prenant en compte plusieurs représentations du graphe (attributs, topologie, décomposition spectrale, etc.). Cette distance, plus flexible et rapide que FGW (Vayer et al., 2020) ou OTT (Kerdoncuff et al., 2022), est utilisée comme fonction de perte dans un modèle de réseau de neurones pour graphes (GNN) capable d'imputer les attributs manquants. Notre approche se distingue notamment par sa capacité à utiliser l'imputeur entraîné, en inférence sur de nouveaux nœuds, sans réentraînement, contrairement à FP (Rossi et al., 2022). Nos contributions incluent : (1) une fonction de perte multi-vues MultiW ; (2) un environnement d'imputation nommé GRIOT (**GR**aph **I**mputation with **O**ptimal **T**ransport) ; (3) une évaluation empirique exhaustive.

2 Travaux connexes

De nombreuses techniques ont été développées pour l'imputation des données manquantes (Van Buuren et Groothuis-Oudshoorn, 2011; Stekhoven et Bühlmann, 2011; Yoon et al., 2018), mais le domaine reste actif en raison de la complexité croissante des données.

Nous nous concentrons ici sur les valeurs manquantes dans les attributs des nœuds d'un graphe. Avec l'essor des réseaux de neurones pour graphes (GNNs), l'intérêt pour les graphes attribués complets a été ravivé. Des méthodes comme SAT (Chen et al., 2022), GCNMF (Taguchi et al., 2021), et PaGNN (Jiang et Zhang, 2020) ont tenté d'adapter les GNNs aux données manquantes, mais se sont concentrées davantage sur les performances des tâches à résoudre (telles que la classification, la prédiction de lien, etc.) que sur la qualité de l'imputation.

À ce jour, FP (Rossi et al., 2022) reste la méthode de référence pour l'imputation des attributs de nœuds manquants, bien qu'elle exige une forte homophilie des nœuds pour obtenir de bons résultats.

Notre approche se distingue en utilisant un imputeur GNN avec une fonction de perte MultiW qui prend en compte la topologie et les attributs des nœuds, rendant notre méthode particulièrement adaptée aux graphes avec faible homophilie et des mécanismes de données manquantes complexes.

3 Transport Optimal et fonction de perte MultiW

Nous définissons ici les notations, rappelons les concepts de Transport Optimal (TO) et de distance de Wasserstein, puis proposons une nouvelle fonction de perte, MultiW, pour l'imputation d'attributs sur les graphes.

Notations. Soit $\mathcal{G} = (\mathcal{V}, \mathcal{E}, F)$ un graphe non orienté où $\mathcal{V}$ est l'ensemble des n nœuds, $\mathcal{E}$ l'ensemble des arêtes représenté par la matrice d'adjacence A, et F la matrice $(n \times d)$, réelle ou binaire, des d attributs des nœuds. Un masque binaire Ω encode les valeurs observées

⋆. Code et article original (Serrano et al., 2024) publiés à ECML-PKDD2024 et annexe disponibles à : `github.com/RichardSrn/GRIOT`

(1) et manquantes (0) dans F. L'objectif est d'estimer $\hat{F}$ à partir de $\mathcal{G}$ et Ω, où $\hat{F}$ est une approximation de F^{gt}, les valeurs véritables.

Une *vue* ζ_j d'un graphe $\mathcal{G}$ est une collection de n vecteurs dans un espace de dimension z_i. Par exemple, F est une vue dans un espace de dimension d.

Transport Optimal et distance de Wasserstein. Le Transport Optimal (TO) trouve un plan de transport de coût minimal entre deux ensembles pondérés de points (X, w_1) et (Y, w_2) dans $\mathbb{R}^d$, où X et Y sont des ensembles de vecteurs, et w_1, w_2 leurs distributions de poids. Soit π le plan de transport, $M^{X,Y}$ la matrice de coût, et H l'entropie, la distance de Wasserstein est définie comme suit (détaillée dans l'article original (Serrano et al., 2024)) :

$$\mathcal{W}((X, w_1), (Y, w_2)) = \min_{\pi \in \Pi(w_1, w_2)} \langle M^{X,Y}, \pi \rangle_F + \varepsilon H(\pi) \tag{1}$$

Définition de la fonction de perte MultiW**.** Les graphes peuvent être représentés de diverses manières, i.e. observés selon différentes *vues*. Par exemple, la structure d'un graphe peut être encodée avec une matrice d'adjacence ou une matrice Laplacienne, chacune offrant une perspective différente. Pour exploiter ces différentes vues, nous avons conçu une fonction de perte qui exploite toutes ces vues simultanément.

Motivation pour le Transport Optimal. Le TO permet d'estimer la distance entre des distributions. Nous supposons que la distribution des valeurs imputées doit ressembler à celle des valeurs connues, relativement aux rôles que jouent les différents nœuds, ces derniers étant estimés en prenant en compte les différentes vues du graphe.

Définition générale. Considérons un graphe G avec n nœuds et q vues $\zeta = (\zeta_i)_{i \leq q}$ représentant G dans différents espaces. Pour deux sous-ensembles de nœuds aléatoires $\mathcal{V}^1$ et $\mathcal{V}^2$, et les vues respectives associées $(\zeta_i^1)_{i \leq q}$, $(\zeta_i^2)_{i \leq q}$, la fonction de perte MultiW est définie par :

$$\text{MultiW}_\alpha((\zeta_i^1)_{i \leq q}, (\zeta_i^2)_{i \leq q}) = \min_{\pi \in \Pi(w_1, w_2)} \langle \sum_{i=1}^{q} \alpha_i M^{\zeta_i^1, \zeta_i^2}, \pi \rangle_F + \varepsilon H(\pi) \tag{2}$$

où $(\alpha_i)_{i \leq q}$ sont les poids des vues.
Remarque : La complexité de MultiW croît linéairement avec le nombre de vues, ce qui la rend plus efficace que d'autres approches multi-vues comme l'Optimal Tensor Transport (OTT).

Dans le cas de GRIOT, lorsque MultiW est utilisée pour estimer la distance avec les attributs ($\hat{F}$) et la structure (P), on calcule $M_\alpha = (1 - \alpha) M^{\hat{F}_1, \hat{F}_2} + \alpha M^{P_1, P_2}$.

4 Imputation des attributs, GRIOT et MultiW

Nous introduisons GRIOT, illustrée en Figure 1, une architecture qui entraîne un imputeur en optimisant la fonction de perte MultiW pour imputer les attributs manquants. Le pseudo-code ainsi que le code source sont disponibles en ligne*.

Entrée. GRIOT prend comme entrée un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, une matrice d'attributs F, et un masque des attributs manquants Ω pour reconstruire $\hat{F}$. Les attributs manquants de F sont initialisés par des valeurs aléatoires, tirées de lois normales, suivant les paramètres (μ_j, σ_j) (j de 1 à d), moyenne et écart-type du j-ième attribut.

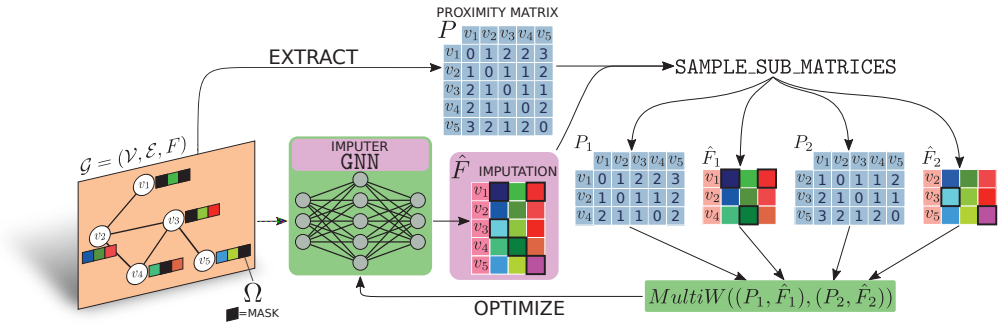

FIG. 1 – *Architecture de l'environnement* GRIOT. *Étant donné un graphe* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *avec des attributs de nœuds* F *et un masque de données manquantes* Ω *en entrée*, GRIOT *a deux composants : l'imputeur GNN et la fonction de perte* MultiW *pour l'optimiser. En sortie : la matrice des attributs imputés* $\hat{F}$ *et l'imputeur* GNN *entraîné.*

Imputeur GCN. L'imputeur est un réseau de neurones convolutif pour graphes (GCN) (Kipf et Welling, 2017) qui utilise la matrice $\hat{F}$ et A pour produire F^{imp}, la matrice des attributs imputés. Le masque Ω permet de remplacer uniquement les attributs manquants.

Entraînement. Pour l'entraînement du GCN, deux sous-ensembles de nœuds $\mathcal{V}^1$, $\mathcal{V}^2$ sont tirés aléatoirement de $\mathcal{V}$, puis MultiW est calculé entre les vues des nœuds de $\mathcal{V}^1$ et $\mathcal{V}^2$ (voir section 3). La perte est calculée pour n_p couples de sous-ensembles de nœuds :

$$loss = \sum_{i=1}^{n_p} \text{MultiW}((\hat{F}_i^1, P_i^1), (\hat{F}_i^2, P_i^2)).$$

Cette opération est répétée sur `epochs` itérations, recalculant l'imputation à chaque époque.

Sortie. La sortie de `GRIOT` inclut la matrice imputée $\hat{F}$ et l'imputeur GNN entraîné, permettant l'imputation d'attributs pour de nouveaux nœuds, dans le cas de graphes dynamiques.

Optimisation. Notre méthode permet une imputation simultanée de tous les attributs via un imputeur GNN unique, contrairement aux approches séquentielles (e.g., Round-Robin) peu efficaces pour les données de haute dimension.

Analyse de Complexité. Soit d le nombre d'attributs, q le nombre de vues, n la taille des graphes $\mathcal{V}_1$ et $\mathcal{V}_2$, `n_epochs` et `n_pairs` sont des hyperparamètres (détails en annexe), alors, la complexité temporelle de `GRIOT` est :$O(d \times q \times n^2 \times \texttt{n\_epochs} \times \texttt{n\_pairs})$.

5 Analyse expérimentale

Nous comparons ici `GRIOT` à des approches de l'état de l'art selon plusieurs scénarios : deux types de masquage d'attributs (MCAR et MNAR), différents pourcentages de données manquantes (20%, 50%, 80%), et divers niveaux d'homophilie. Nous évaluons la qualité des valeurs imputées et leur impact sur la classification des nœuds. Nous cherchons à savoir (1) si `GRIOT` est plus performant que les méthodes de l'état de l'art, (2) si la stratégie de masquage influence l'imputation, et (3) comment se comporte `GRIOT` dans un contexte dynamique.

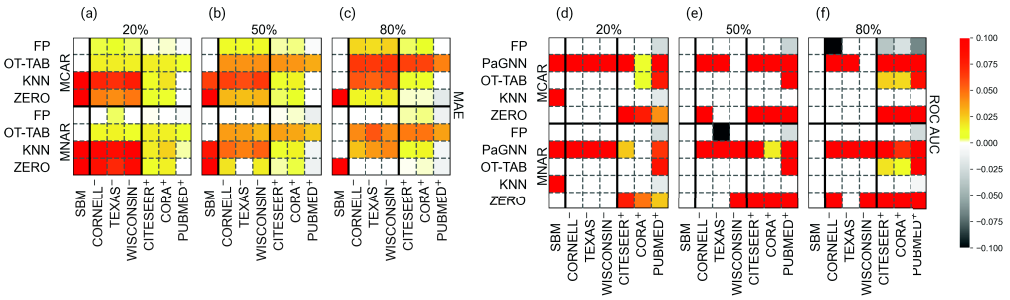

FIG. 2 – *Comparaison de* `GRIOT` *aux méthodes de référence : (a,b,c) MAE, (d,e,f) ROC AUC sur plusieurs jeux de données et mécanismes de données manquantes (MCAR et MNAR). Les carrés colorés indiquent des améliorations significatives de* `GRIOT`.

5.1 Protocole expérimental

Modèles de référence. Nous comparons avec des méthodes naïves comme remplacer les valeurs manquantes par 0 (dans un contexte d'attributs épars), ZERO, et l'imputation par la moyenne des attributs des nœuds voisins, KNN. Les modèles de l'état de l'art incluent OT-TAB (Muzellec et al., 2020), Feature Propagation (FP) (Rossi et al., 2022) pour l'imputation et `PaGNN` (Jiang et Zhang, 2020) pour la classification seule.

Jeux de données. Nous évaluons sur des graphes synthétiques et réels avec divers niveaux d'homophilie (Table 4 en annexe). Nous utilisons un Modèle de Bloc Stochastique (SBM) (homophilie variable), trois graphes WebKB (Craven et al., 1998), Cornell, Texas et Wisconsin, de faible homophilie et trois graphes de citation Planetoid (Yang et al., 2016), CiteSeer, Cora et PubMed, de forte homophilie.

Masquage des données. Deux stratégies de masquage : MCAR (aléatoire uniforme) et MNAR (aléatoire dépendant des données et de la structure du graphe). Le pourcentage de données manquantes varie entre 20%, 50%, et 80%.

Métriques d'évaluation. La qualité de l'imputation est mesurée par l'Erreur Absolue Moyenne (MAE). L'impact sur la classification des nœuds est évalué par le score ROC-AUC.

Architecture de l'imputeur. L'imputeur est un GNN qui prend en entrée $\hat{F}$ et $\mathcal{E}$. L'architecture comprend 2 couches GCN et 1 couche linéaire, avec dropout de 50%. Les paramètres sont optimisés avec Adam (Kingma et Ba, 2014), avec un taux d'apprentissage de 0.01.

Architecture du classifieur. Le classifieur est un GNN qui prend $(\mathcal{E}, \hat{F})$ en entrée et classe les nœuds dans k classes possibles. Il comporte 2 couches `Cheb` (Defferrard et al., 2016) et 1 couche linéaire, dont les hyper-paramètres sont optimisés par validation croisée.

5.2 Analyse des résultats

Valeurs imputées. La Figure 2 (a,b,c) montre la MAE pour toutes les méthodes. `GRIOT` obtient de meilleurs résultats, surtout pour les graphes hétérophiles. Les différences de performances avec FP augmentent avec le pourcentage de données manquantes.

Classification des nœuds. La Figure 2 (d,e,f) montre les différences de score AUC. Bien que `GRIOT` soit supérieur pour l'imputation, cela ne se traduit pas systématiquement par

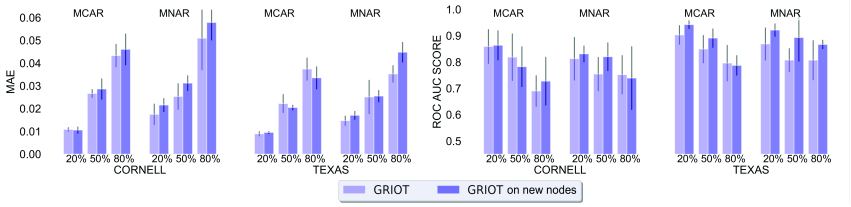

FIG. 3 – *Performance de* `GRIOT` *sur données dynamiques : MAE (gauche) et AUC (droite).*

	(a) sous-performe	(b) similaire	(c) sur-performe
MAE	6%	26%	**68%**
ROC	8%	56%	**37%**

TAB. 1 – *Pourcentage de fois où* `GRIOT` *(a) sous-performe, (b) est similaire (non significatif), ou (c) sur-performe des méthodes de référence solides, en moyenne sur tous les scénarios.*

une amélioration significative de la classification des nœuds. Mais `GRIOT` surpasse toujours `PaGNN` et OT-TAB avec 80% de données manquantes sur les graphes homophiles.

Alignement des objectifs. L'analyse de α (voir Table 5 en annexe) montre que la structure est plus cruciale pour l'imputation sur les graphes hétérophiles et moins pour la classification.

Cas d'un graphe dynamique. `GRIOT` permet d'imputer les valeurs manquantes pour de nouveaux nœuds sans réentraînement ; ceci correspond au cas d'un graphe dynamique. Les résultats pour les nœuds inconnus sont similaires aux précédents, avec une légère augmentation de la MAE dans les cas extrêmes (80% de données manquantes), mais sans baisse significative de l'AUC (Figure 3).

Complexité temporelle. `GRIOT` est comparable à OT-TAB en temps d'exécution (voir Table 3 en annexe). `GRIOT` peut être plus efficace que FP dans des environnements dynamiques avec des nœuds ajoutés après l'entraînement.

Résumé des résultats. La comparaison entre `GRIOT` et les méthodes de l'état de l'art (Tableau 1) montre que `GRIOT` améliore la reconstruction des données dans 68% des cas et la classification des nœuds dans 37% des cas. Malgré une performance similaire après imputation avec la méthode FP pour la classification des nœuds, `GRIOT` se distingue particulièrement dans des graphes dynamiques, tels que les réseaux sociaux, où le nombre de nouveaux utilisateurs augmente continuellement.

Nos résultats apportent une perspective novatrice sur les relations contre-intuitives entre la qualité des valeurs imputées et la performance de classification des nœuds, offrant des pistes pour l'amélioration des stratégies d'imputation dans diverses applications.

6 Conclusion et perspectives

Nous proposons `GRIOT`, un environnement utilisant la théorie du transport optimal et la fonction de perte MultiW pour l'imputation d'attributs manquants dans les graphes attribués. Ses atouts incluent la prise en charge de multiples représentations, une imputation paralléli-

sée efficace, une utilisation du TO qui supporte le passage à l'échelle, et la réutilisation d'un imputeur préalablement entraîné pour de nouveaux nœuds. GRIOT est applicable au-delà des tâches spécifiques comme la classification des nœuds.

Les expériences sur des jeux de données synthétiques et réels avec divers schémas de données manquantes montrent que GRIOT est compétitif, surtout sur des graphes de faible homophilie. Pour les graphes hétérophiles, une architecture d'imputeur moins dépendante de l'hypothèse d'homophilie, telle que les *Graph Convolutional Transformers* (Dwivedi et Bresson, 2020), pourrait être explorée. Bien que nos essais avec cette architecture n'aient pas encore surpassé le GCN, des recherches futures porteront sur l'effet de différentes vues pour diverses tâches, au-delà de la classification des nœuds.

Remerciements Ce travail a été financé par une subvention publique de l'Agence Nationale de la Recherche (ANR) dans le cadre du plan d'investissement "France 2030", avec la référence EUR MANUTECH SLEIGHT - ANR-17-EURE-0026.

Références

Brogat-Motte, L., R. Flamary, C. Brouard, J. Rousu, et F. d'Alché-Buc (2022). Learning to predict graphs with Fused Gromov-Wasserstein barycenters. In *ICML*, pp. 2321–2335.

Chen, X., S. Chen, J. Yao, H. Zheng, Y. Zhang, et I. W. Tsang (2022). Learning on attribute-missing graphs. *IEEE PAMI 44*(2), 740–757.

Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, et S. Slattery (1998). Learning to extract symbolic knowledge from the World Wide Web. *AAAI/IAAI 3*(3.6), 2.

Defferrard, M., X. Bresson, et P. Vandergheynst (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems 29*.

Dwivedi, V. P. et X. Bresson (2020). A generalization of transformer networks to graphs. *CoRR abs/2012.09699*.

Huisman, M. (2014). Imputation of Missing Network Data : Some Simple Procedures. In *Encyclopedia of Social Network Analysis and Mining*, pp. 707–715.

Jiang, B. et Z. Zhang (2020). Incomplete graph representation and learning via partial graph neural networks. *arXiv preprint arXiv :2003.10130*.

Kerdoncuff, T., R. Emonet, M. Perrot, et M. Sebban (2022). Optimal tensor transport. In *AAAI Conference on Artificial Intelligence*, Volume 36, pp. 7124–7132.

Kingma, D. P. et J. Ba (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

Kipf, T. N. et M. Welling (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 247–268.

Little, R. J. et D. B. Rubin (2019). *Statistical analysis with missing data*, Volume 793. John Wiley & Sons.

Muzellec, B., J. Josse, C. Boyer, et M. Cuturi (2020). Missing data imputation using optimal transport. In *ICML*, pp. 7130–7140.

Rossi, E., H. Kenlay, M. I. Gorinova, B. P. Chamberlain, X. Dong, et M. M. Bronstein (2022). On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. In *Learning on Graphs Conference*.

Schafer, J. L. et J. W. Graham (2002). Missing data : our view of the state of the art. *Psychological methods 7*(2), 147.

Serrano, R., C. Laclau, B. Jeudy, et C. Largeron (2024). Reconstructing the unseen : Griot for attributed graph imputation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 269–286. Springer.

Stekhoven, D. J. et P. Bühlmann (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics 28*(1), 112–118.

Taguchi, H., X. Liu, et T. Murata (2021). Graph convolutional networks for graphs containing missing features. *Future Generation Computer Systems 117*, 155–168.

Van Buuren, S. et K. Groothuis-Oudshoorn (2011). mice : Multivariate imputation by chained equations in r. *Journal of statistical software*.

Vayer, T., L. Chapel, R. Flamary, R. Tavenard, et N. Courty (2020). Fused Gromov-Wasserstein distance for structured objects : Theoretical foundations and mathematical properties. *Algorithms*.

Yang, Z., W. Cohen, et R. Salakhudinov (2016). Revisiting semi-supervised learning with graph embeddings. In *ICML*, pp. 40–48.

Yoon, J., J. Jordon, et M. van der Schaar (2018). GAIN : Missing data imputation using generative adversarial nets. In *Proceedings of ICML*, Volume 80, pp. 5689–5698.

Zheng, X., Y. Liu, S. Pan, M. Zhang, D. Jin, et P. S. Yu (2022). Graph neural networks for graphs with heterophily : A survey. *arXiv preprint arXiv :2202.07082*.

Summary

In recent years, machine learning in managing attributed graphs has experienced significant growth thanks to the Graph Neural Networks (GNN) (Kipf et Welling, 2017). However, these methods assume fully known attributes, which is often unrealistic. This paper explores the potential of optimal transport (OT) to impute missing attribute values on graphs. We propose a new multi-view OT loss function, integrating node attributes and topological structure. This loss is used to train a graph convolutional network (GCN) architecture capable of imputing all missing values simultaneously. We evaluate our approach with experiments on synthetic data and real-world graphs. The results show that our method is competitive with the state-of-the-art, especially on weakly homophilic graphs.