Comment rendre les GNN plus équitables pour la prédiction de liens ?

Manvi Choudhary, Antoine Gourru, Charlotte Laclau, Christine Largeron

Comment rendre les GNN plus équitables pour la prédiction de liens ?

M. Choudhary *, A. Gourru*, C. Laclau**, C. Largeron*

* Laboratoire Hubert Curien UMR5516, UJM-Saint-Etienne, CNRS, IOGS,
Université de Lyon, F-42023 St-Etienne, France
prenom.nom@univ-st-etienne.fr
**LTCI, Télécom Paris, Institut Polytechnique de Paris
charlotte.laclau@telecom-paris.fr

Résumé. L'équité algorithmique a suscité un grand intérêt dans la communauté de l'apprentissage automatique et plus récemment dans le domaine des données relationnelles représentées sous forme de graphe. Dans cet article, nous abordons le problème de l'apprentissage de représentations équitables des noeuds d'un graphe, en se concentrant plus spécifiquement sur l'équité dyadique pour la tâche de prédiction de liens dans des graphes attribués. Nous avons conçu un modèle qui, étant donné des paires de nœuds avec un attribut protégé/sensible, apprend une représentation basée sur le principe du Variationnal Information Bottleneck (Alemi et al., 2017) en utilisant un Graph Neural Network (GNN) comme encodeur. Le modèle proposé permet d'apprendre simultanément des plongements de nœuds non linéaires reflétant la structure du graphe, tout en contrôlant explicitement le niveau d'équité. Les expériences menées sur plusieurs jeux de données du monde réel ont confirmé la capacité de la méthode à maintenir une haute précision sur la tâche de prédiction de liens tout en réduisant significativement le biais.

1 Introduction

De nos jours, un nombre croissant de tâches sont exécutées ou assistées par des algorithmes d'apprentissage automatique ("machine learning", ou ML). Dans ce contexte, il est important de contrôler que les décisions prises ou assistées par ces algorithmes sont équitables. Prenons l'exemple de l'analyse automatique de demandes d'emploi. Dans ce cas, *équitable* peut avoir différentes significations. D'une part, on s'attend à ce que la recommandation faite par l'algorithme soit indépendante de certains attributs sensibles des candidats, par exemple, leur sexe ou leur origine ethnique ; ce type d'équité est appelé équité de groupe dans la littérature. D'autre part, nous voudrions également que la recommandation reste équitable d'un point de vue individuel, c'est-à-dire que deux candidats ayant des compétences similaires devraient obtenir une décision similaire.

Dans cet article, nous nous intéressons principalement à la notion d'équité dyadique (Li et al., 2021) ; ce qui signifie que nous nous attendons à ce que la probabilité d'un lien entre deux nœuds soit la même, qu'ils présentent ou pas la même valeur pour un attribut sensible.

Comment rendre les GNN plus équitables pour la prédiction de liens ?

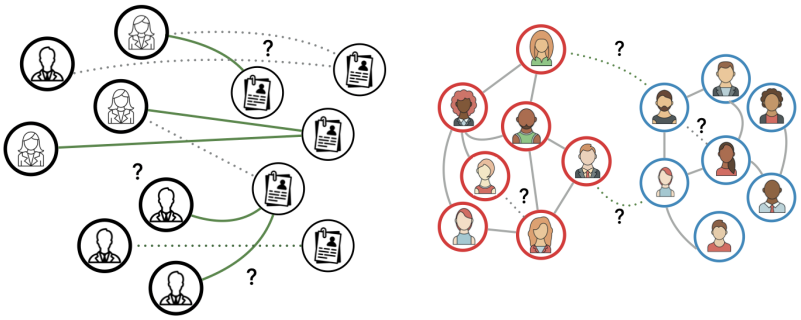


FIG. 1 – Exemples de prédiction de liens sensibles à la non-équité : recommandation d'emploi et réseaux sociaux polarisés

Par exemple, dans un réseau social de blogueurs où l'inclinaison politique peut être considérée comme un attribut protégé, nous aimerions, notamment, pour éviter la formation de bulles en ligne et la polarisation, que le modèle de prédiction ne favorise pas seulement les liens entre les personnes ayant la même idéologie politique, même si elles sont plus susceptibles d'être connectées dans un réseau homophilique, comme illustré en fig. 1. De même, la prédiction de lien étant largement utilisé dans des systèmes de recommandation, il est nécessaire que la prédiction soit indépendante du genre au moment de recommander une offre d'emploi à des personnes en recherche d'emploi.

En ML, l'équité peut être prise en compte à différentes étapes du processus de décision : pendant le pré-traitement des données, en intégrant des contraintes ou des pénalités dans le modèle d'apprentissage lui-même, ou par un post-traitement qui débiaise directement la sortie du modèle (Mehrabi et al., 2021). En outre, l'importance majeure de l'apprentissage de représentation en ML ces dernières années a conduit aussi à l'apprentissage de représentation équitable, qui peut être considéré comme une étape entre le pré-traitement des données et l'entraînement du modèle lui-même. Par exemple, Zemel et al. (2013) ont proposé l'un des premiers algorithmes d'apprentissage de représentation pour la classification qui garantit à la fois l'équité individuelle et de groupe. Dans le même esprit, des techniques de régularisation adversariale (adversarial regularization) ont été introduites pour apprendre directement des représentations équitables (Madras et al. (2018)). Bien que ces solutions se soient avérées efficaces pour atténuer les biais algorithmiques potentiels, elles ont toutes été conçues pour des données indépendantes et identiquement distribuées. Or, dans cet article, nous nous intéressons aux données modélisées sous la forme de graphes qui sont devenus omniprésents pour décrire des structures complexes dans lesquelles cette propriété n'es pas vérifiée en général.

L'équité dans le contexte des données représentées par un graphe est un domaine de recherche émergent. À notre connaissance, la première contribution dans ce domaine a été proposée par Rahman et al. (2019), qui a étendu l'algorithme populaire Node2vec (Grover et Leskovec, 2016) en modifiant la procédure de marche aléatoire originale pour atteindre l'équité. Un inconvénient de cet algorithme est qu'il ne fonctionne pas dans le cas d'une forte dépendance de la structure communautaire du graphe à l'attribut sensible : en explorant le voisinage de chaque nœud, l'algorithme trouvera des nœuds ayant le même attribut sensible. Il en ré-

sultera une représentation biaisée vers un attribut particulier qui sera facilement identifié lors de l'étape de prédiction. De même, DeBayes (Buyl et Bie, 2020) adopte un modèle bayésien où l'information sensible est modélisée dans la distribution a priori. Buyl et Bie (2021) ont aussi proposé un cadre pour les modèles de représentation probabilistes, intégrant un terme de régularisation de l'équité basé sur la théorie de l'information. Dans Bose et Hamilton (2019), contrairement aux méthodes précédentes, les auteurs définissent un cadre adversarial générique permettant de filtrer a posteriori les informations sensibles des représentations. Cependant, cette approche ne peut garantir qu'une classification équitable des nœuds. Enfin, `FairGNN` (Dai et Wang, 2021) est une approche contrastive basée sur les Graph Neural Networks (GNN) qui se concentre sur la tâche de classification équitable des nœuds, traitant donc de l'équité au niveau individuel. D'autres méthodes Laclau et al. (2021); Li et al. (2021) entrent dans la catégorie des méthodes de prétraitement (débiaiser le graphe original) qui garantissent que l'entrée est sans biais mais qui ne sont pas capables de contrôler les biais potentiels qui peuvent apparaître lors de l'apprentissage des plongements des nœuds dans l'espace vectoriel.

Dans cet article, nous introduisons une approche nommée LEarning FAir Variational Embedding (`LEAVE`), capable d'inférer des représentations équitables des nœuds pour la tâche de prédiction des liens. `LEAVE` est un modèle qui ne repose pas sur un algorithme spécifique de représentation de nœuds comme beaucoup d'autres (Rahman et al., 2019; Buyl et Bie, 2020). De plus, l'architecture proposée pour l'apprentissage de représentation exploite l'aspect relationnel des graphes, à la fois pour apprendre les représentations et pour résoudre la tâche de prédiction de liens, tout en tenant compte des biais potentiels. De cette façon, notre méthode permet de contrôler explicitement l'équilibre entre la précision en prédiction de liens et la réduction du biais.

Notre contribution est triple : (1) nous proposons un modèle dérivé du principe du VIB (Variational Information Bottleneck, Alemi et al. (2017)), intégrant un double objectif : atténuer à la fois l'équité dyadique et le biais au niveau de la représentation, (2) nous implémentons ce modèle sur deux architectures différentes de GNN et, (3) nous démontrons expérimentalement son utilité sur plusieurs graphes du monde réel. Dans la section 2, nous définissons notre modèle `LEAVE` qui prend en compte simultanément la précision de la prédiction des liens et l'équité. Nous évaluons notre modèle sur différents jeux de données et présentons les résultats dans la section 3. Nous concluons ensuite en section 4.

2 Plongement variationnel équitable de noeuds

Nous considérons un réseau représenté par un graphe non orienté et non pondéré $G = (V, E)$, où V est l'ensemble de n sommets et E, l'ensemble de m arêtes observées. y est une fonction indicatrice des liens définie pour chaque paire de nœuds $(u, v) \in V \times V$ telle que $y_{u,v} = 1$ s'il existe un lien entre u et v et $y_{u,v} = 0$ sinon. En outre, nous supposons l'existence de A, un attribut sensible catégoriel qui attribue une valeur appartenant à $\{0, \cdots, l\}$ à chaque nœud du graphe. Par exemple, dans un réseau social où les arêtes représentent les interactions entre individus, A peut être le sexe ou le continent d'origine de chaque nœud. En outre, nous définissons s, un indicateur d'information sensible pour chaque paire de nœuds (u, v) tel que $s_{u,v} = 1$ si les nœuds u et v ont des valeurs d'attributs sensibles similaires, c'est-à-dire lorsque $A_u = A_v$ et, $s_{u,v} = 0$ s'ils n'ont pas la même valeur d'attribut sensible, c'est-à-dire lorsque $A_u \neq A_v$.

Comment rendre les GNN plus équitables pour la prédiction de liens ?

Nous nous concentrons sur la prédiction d'arête, où l'objectif est d'identifier les paramètres de $p(y_{u,v}|u,v)$. À cette fin, nous apprenons des représentations, ou plongements, équitables des nœuds dans un espace vectoriel de représentation à d dimensions, notées z, avec z_u la représentations du nœud u. Nous nous attendons à ce que ces représentations satisfassent les propriétés suivantes :

[P1] Les représentations doivent permettre de prédire l'existence d'arêtes entre les nœuds.
[P2] Les représentations doivent être indépendantes de l'attribut sensible A.

2.1 Fonction objectif de `LEAVE`

Notre objectif est d'apprendre des représentations de nœuds satisfaisant les propriétés [**P1**] et [**P2**]. Pour ce faire, nous considérons les deux problèmes simultanément. D'une part, nous voulons prédire correctement l'existence d'un lien entre deux nœuds et d'autre part, nous voulons échouer à prédire l'attribut sensible des nœuds. Nous concevons une fonction objectif permettant de rechercher un compromis entre la précision de la prédiction des arêtes et l'équité dyadique. Nous définissons la fonction objectif suivante à maximiser, qui étend l'approche du goulot d'information (Information Bottleneck, Tishby et al. (2000)) :

$$\mathcal{L}_{IB} = (1-\alpha)I((z_u, z_v), y_{uv}) - \alpha I((z_u, z_v), s_{uv}) - \beta I((z_u, z_v), (x_u, x_v)), \qquad (1)$$

où I est l'information mutuelle, x_u et x_v sont les représentations initiales des noeuds, $\alpha \in [0, 1]$ et $\beta \geq 0$ sont deux hyperparamètres du modèle. Les représentations initiales x_u et x_v sont issues d'un encodeur entraînés simultanément (voir le dernier paragraphe de la sous-section 2.2 au sujet de cette représentation initiale). Dans cet article, nous utilisons deux architectures GNN.

Décomposons cette fonction objectif : (1) Le premier terme évalue la dépendance entre les représentations Z et la vraie fonction de lien y. Ce terme permet de préserver l'information relationnelle en encodant la structure du graphe ; (2) le deuxième terme quantifie la dépendance entre les représentations Z et l'attribut sensible s, qu'on cherche à minimiser : c'est à dire qu'on construit des représentations pour "oublier" l'information sensible ; (3) le troisième terme évalue la dépendance entre les représentations z_u et z_v des nœuds u et v et leurs représentations initiales qui doivent être réduites pour compresser au maximum l'information.

De plus, l'information mutuelle permet d'incorporer une certaine forme d'incertitude/variance dans la fonction objectif, qui a plusieurs avantages en apprentissage de représentations (Oh et al., 2018; Gourru et al., 2020). Dans le contexte de l'apprentissage de représentations de graphes équitables, nous nous attendons, et montrons dans nos expériences, que l'apprentissage d'une mesure de variance peut réduire le biais des représentations. Notre intuition est que l'ajout d'une part d'aléatoire permet de prédire des liens *inattendus* pour les graphes qui démontrent de l'homophilie, c'est-à-dire de prédire des liens entre des nœuds ayant des attributs sensibles différents. En résumé, α contrôle le compromis entre **[P1]** et **[P2]**, lorsque $\alpha = 0$ le modèle se concentre uniquement sur la structure du graphe, alors que lorsque $\alpha = 1$ le modèle se concentre sur l'équité dyadique, indépendamment de la structure originale du graphe.

2.2 Borne et Optimisation de LEAVE

Les représentations probabilistes des noeuds sont apprises en maximisant l'équation (1). Cependant, le calcul de l'information mutuelle est généralement difficile, c'est pourquoi nous utilisons l'approximation variationnelle proposée par (Alemi et al., 2017), qui permet d'obtenir une borne inférieure de l'équation (1), qui devient :

$$\mathcal{L}_{VIB} = \mathbb{E}_{z_u \sim p(z_u|x_u), z_v \sim p(z_v|x_v)} \left[(1-\alpha)\mathbb{L}_{y_{uv}} - \alpha \mathbb{L}_{s_{uv}} \right] - \beta KL(p(z|x)||r(z)) \quad (2)$$

où $p(z|x)$ est la distribution conditionnelle latente de z dont les paramètres sont appris (l'encodeur), et $r(z)$ est un terme marginal qui est généralement fixé à une distribution gaussienne unitaire $\mathcal{N}(0, I)$. Dans cette équation, les deux premiers termes sont les log-vraisemblance négative par rapport aux liens observés y et aux attributs sensibles s. Enfin, le troisième terme correspond à l'aspect compressif du modèle et prend la forme d'une régularisation de Kullback-Leibler.

Nous présentons ci-dessous les détails du calcul de $\mathbb{L}_{y_{uv}}$ et de $\mathbb{L}_{s_{uv}}$ dans notre contexte, par rapport à y_{uv} et s_{uv}.

Pour calculer $p\left(y_{uv}|z_u, z_v\right)$, nous adoptons l'approche contrastive, similairement à (Oh et al., 2018). Nous définissons les paires positives et négatives de noeuds de telle sorte qu'une paire (u, v) est une paire positive si $y_{uv} = 1$, c'est-à-dire si elles sont connectées, et une paire négative si $y_{uv} = 0$, c'est-à-dire si elles ne sont pas connectées. Ensuite, la probabilité qu'un exemple soit positif ou négatif en fonction de y_{uv} est donnée par :

$$p\left(y_{uv}|z_u, z_v\right) := \sigma\left(-a||z_u - z_v||_2 + b\right),$$

où a et b sont des paramètres entraînables, s.t. $c > 0$, $d \in \mathbb{R}$ et σ est la fonction sigmoïde $\sigma(t) = \frac{1}{1+e^{-t}}$. Après avoir défini la probabilité $p(y_{uv}|z_u, z_v)$, la vraisemblance contrastive correspondante est donnée par :

$$\mathbb{L}_{y_{uv}} = \begin{cases} \log p\left(y_{uv}|z_u, z_v\right), & \text{si } y_{uv} = 1 \\ \log(1 - p\left(y_{uv}|z_u, z_v\right)), & \text{si } y_{uv} = 0. \end{cases} \quad (3)$$

De la même manière, nous définissons les paires positives et négatives de nœuds par rapport à l'attribut sensible. Une paire positive à pour valeur $s_{uv} = 1$, c'est-à-dire que les nœuds de la paire possèdent le même attribut sensible, et inversement. Ensuite, la probabilité qu'un exemple soit positif ou négatif en fonction de s_{uv} est donnée par :

$$p\left(s_{uv}|z_u, z_v\right) := \sigma\left(-c||z_u - z_v||_2 + d\right).$$

La vraisemblance $\mathbb{L}_{s_{uv}}$ est définie similairement qu'en equation 3. En adoptant cette approche, nous minimisons la vraisemblance des observations s, et donc maximisons la perte d'un classifieur probabiliste prédisant la valeur de s. Le fonctionnement est ainsi similaire à une approche adversariale.

Détails de calcul Nous définissons $p(z|x)$ comme étant une gaussienne de dimension d avec une variance diagonale, comme cela a été fait dans (Oh et al., 2018).

Comment rendre les GNN plus équitables pour la prédiction de liens ?

$$z|x \sim \mathcal{N}(f(x), g(x)), \tag{4}$$

où f et g sont des perceptrons multicouches (MLP). Pour optimiser $\mathcal{L}_{min}$, nous reparamétrons $p(z|x)$ ((Kingma et Welling, 2014)), et approximons la vraisemblance avec un échantillonnage de Monte Carlo pour obtenir un flux lisse de gradients. Nous tirons $2k$ échantillons par paires et par époques :

$$z^{(k)} = f(x) + g(x) \odot \epsilon^{(k)}, \text{ avec } \epsilon^{(k)} \sim \mathcal{N}(0, 1) \tag{5}$$

Sur la représentation initiale des noeuds LEAVE repose principalement sur la représentation initiale des noeuds. Il peut s'agir de caractéristiques créées à la main, de vecteurs d'adjacence ou de représentations pré-entraînées. Dans ce travail, nous proposons d'utiliser une troisième fonction $x_u = h(u, G)$ dont les paramètres sont entraînés *simultanément*. L'utilisation de ce choix de modélisation rend le modèle général et indépendant de la modélisation initiale des nœuds. Dans nos expériences, nous avons utilisé des architectures GNN standard (Kipf et Welling, 2016; Veličković et al., 2018).

L'architecture globale est décrite en fig. 2.

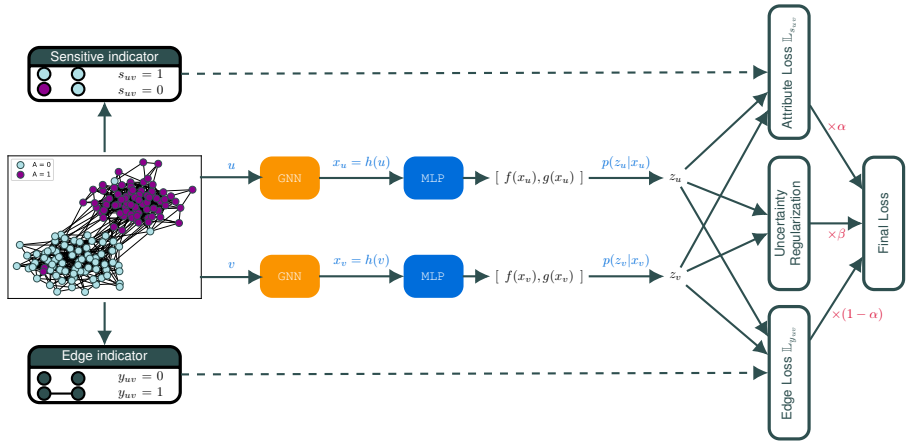


FIG. 2 – Illustration complète de l'architecture de LEAVE

3 Cadre expérimental et résultats

Dans cette section, nous présentons notre cadre expérimental et les résultats obtenus. Nous étudions également dans quelle mesure l'hyper-paramètre α nous permet de contrôler le compromis entre les propriétés **P1** et **P2**. Ces expériences sont menées sur plusieurs ensembles de données de référence dont les caractéristiques figurent dans le tableau 1.

Dans Polblogs (Adamic et Glance, 2005), les nœuds représentent les blogs et les sommets représentent les hyperliens entre deux blogs. Chaque blog est accompagné de la tendance politique que nous considérons comme l'attribut sensible (binaire). Pour LastFM (Rozemberczki

TAB. 1 – Caractéristiques des graphes. #groups : nombre de modalités de l'attribut sensible, r : *fair* mixing coefficient, S : attribut sensible

Dataset	$\|V\|$	$\|E\|$	mc	densité	S	#groups
Polblogs (Adamic et Glance, 2005)	$1,490$	$19,090$	0.81	$2e^{-2}$	party	2
LastFM (Rozemberczki et Sarkar, 2020)	$7,624$	$27,806$	0.86	$1e^{-3}$	country	16
Facebook-P (Rozemberczki et al., 2021)	$22,470$	$171,002$	0.82	$7e^{-4}$	page-type	4

et Sarkar, 2020), les noeuds représentent les utilisateurs et les arêtes représentent les relations mutuelles de "followers" entre les utilisateurs. L'attribut sensible est le pays de l'utilisateur. La distribution de ces modalités est très déséquilibrée, c'est pourquoi nous ne considérons que les modalités avec plus de 150 nœuds, ce qui donne au final 11 modalités pour l'attributs sensible. Dans Facebook-P (Rozemberczki et al., 2021), les nœuds correspondent aux pages Facebook et une arête aux "likes" mutuels entre les pages. Le type de page est traité comme l'attribut sensible.

Pour évaluer le biais potentiel déjà présent dans les graphes utilisés, nous proposons de regarder de plus près le coefficient de mélange ("mixing coefficient") par rapport à l'attribut protégé. Le coefficient de mélange, dans l'analyse des réseaux sociaux, permet d'évaluer la tendance des nœuds à se connecter avec d'autres ayant des attributs similaires (Newman, 2003). Le mélange assortatif est lié à l'homophilie et ce coefficient s'est révélé être un indicateur fort de la ségrégation dans les réseaux en ligne et hors ligne (Hofstra et al., 2017). Nous l'adaptons dans le contexte de l'équité dyadique, et il est désigné par mc dans le tableau 1. Le coefficient mc se situe dans $[-1, 1]$, où 1 correspond au cas parfaitement assortatif, c'est-à-dire que les nœuds ayant la même valeur pour A se connectent exclusivement entre eux ; -1 correspond au cas dissortatif, c'est-à-dire que les nœuds se connectent uniquement avec les nœuds ayant une valeur différente pour A. Dans notre contexte, un graphe *équitable* a un coefficient de mélange de 0. Dans ce qui suit, nous utiliserons également ce critère pour réduire la recherche sur grille des hyperparamètres. Pour les trois jeux de données, nous constatons que mc est élevé, LastFM étant la structure la plus biaisée à cet égard. Une implication directe de cette remarque est que l'on peut s'attendre à ce que, dans ces cas, l'imposition d'une contrainte d'équité dans le processus d'apprentissage entraîne une baisse de la précision de la prédiction de liens.

3.1 Mesures d'évaluation

Biais de représentation (RB) Proposée à l'origine par (Bose et Hamilton, 2019), puis formalisée par (Buyl et Bie, 2020), la mesure RB évalue le biais des représentations de nœuds, en considérant l'attribut protégé A comme variable cible. Étant donné les vecteurs de représentation des nœuds en entrée, le RB calcule la moyenne pondérée des scores AUC un-contre-un obtenus à partir de la sortie $\mathbb{P}_h(a, z_v)$ d'un classificateur h entraîné à prédire l'attribut protégé A. En définissant par $V_a = \{v|A_v = a\}$ l'ensemble des nœuds prenant la valeur a pour l'attribut protégé A, on rappelle que $A_u \in \{0, .., \ell\}$le score RB est donné par

$$RB = \sum_{a=0}^{\ell} \frac{\ell}{|V_a|} \text{AUC}(\{\mathbb{P}_h(a, z_v)|\forall v \in V_a\}).$$

RB $\in [0, 1]$ et est idéalement proche de l'aléatoire (0.5), ce qui signifie que le classifieur entrainé à partir des représentations de nœuds effectue une prédiction aléatoire pour l'attribut sensible. Il convient de noter que, par rapport aux autres métriques, RB n'est pas un critère d'équité dyadique car il se concentre sur chaque nœud individuellement, et n'est donc pas suffisant pour évaluer la prédiction équitable des liens.

Effet disparate (DI et IDI) évalue le biais au niveau de la prédiction d'arête (Laclau et al., 2021) et est calculé comme le rapport entre les probabilités de prédire une arête entre deux nœuds sachant que ces nœuds possèdent le même attribut sensible ou ont des attributs sensibles différents. Formellement, nous avons :

$$\text{DI} = \frac{P(\hat{y}_{uv} = 1 | A_u \neq A_v)}{P(\hat{y}_{uv} = 1 | A_u = A_v)}, \text{et IDI} = \frac{1}{\text{DI}}.$$

Par souci de facilité de lecture, nous présentons l'inverse de cette valeur dans nos expériences qu'on note IDI. Comme le RB, plus la valeur IDI est grande, plus la probabilité de prédire un lien entre noeud de même attribut est grande par rapport à des noeuds d'attributs différents (ce qui correspond à la situation le plus souvent observée).

Aire sous la courbe ROC (AUC) mesure la performance en prédiction de lien et se situe dans $[0, 1]$, 1 étant la valeur optimale.

3.2 Protocole d'évaluation

Nous entraînons les modèles sur 70% des arêtes observées, et utilisons les 10 % et 20% restants pour la validation et le test, respectivement. Pour chaque ensemble, nous générons aléatoirement un nombre égal d'arêtes négatives, c'est à dire qu'on sous échantillonne les liens non observés, conformément aux travaux de l'état de l'art.

Nous implémentons notre modèle avec deux architectures basées sur un GNN comme fonction d'encodage $h(.)$, à savoir un Graph Convolutional Network (`GCN`) (Kipf et Welling, 2016) et un Graph Attention Network (`GAT`) (Veličković et al., 2018). Pour chaque architecture, nous évaluons `LEAVE` et réalisons une étude d'ablation pour mettre en évidence l'impact des différents termes composant la vraisemblance. `GAT` et `GCN` correspondent au cas où α et β sont tous deux fixés à 0, et $z = x$ est la sortie de l'encodeur sans aspect probabiliste. Ces modèles se concentrent sur les performances en prédiction de liens (AUC élevée). `LEAVE-GAT`$_{woVIB}$ (resp. `LEAVE-GCN`$_{woVIB}$) sont des versions dans lesquelles $\alpha \neq 0$ mais $z = x$ (z déterministe et $\beta = 0$). Ces modèles se concentrent sur la recherche d'un compromis performance en prédiction de lien - équité (IDI faible). `LEAVE-GAT` (resp. `LEAVE-GCN`) sont les modèles incluant à la fois le régularisateur d'équité dyadique ($\alpha \neq 0$) et le terme d'incertitude ($\beta \neq 0$), la aussi ajusté grâce à l'ensemble de validation. Nous espérons ici renforcer l'équité, tant des représentations (RB) que de la prédiction de liens (IDI).

Nous comparons avec la méthode Fair I-Projection Regularizer `FIPR` (Buyl et Bie, 2021) qui s'appuie sur un terme de régularisation qui encourage l'équité dyadique. Nous implémentons ce régularisateur sur les deux architectures basées sur le GNN et désignons par `GAT-FIPR` et `GCN-FIPR` , les méthodes de référence correspondantes. Cette approche s'est avérée

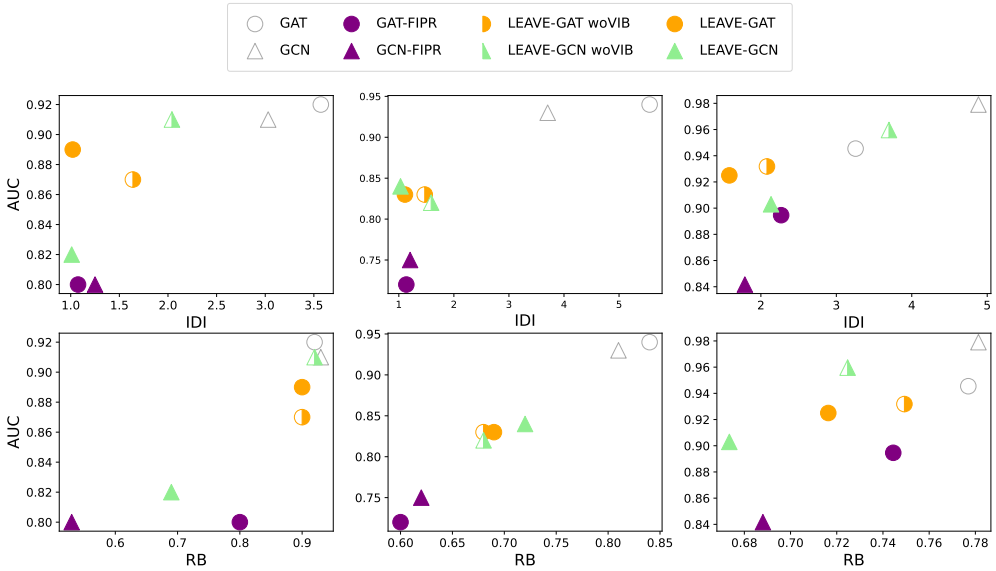


FIG. 3 – Résultats sur les réseaux Polblogs (à gauche), LastFM (au milieu) et Facebook (à droite) - La ligne supérieure représente l'AUC par rapport à IDI et la ligne inférieure l'AUC par rapport au RB.

jusque là plus performante que tous les travaux existants tels que (Buyl et Bie, 2020; Rahman et al., 2019; Bose et Hamilton, 2019).

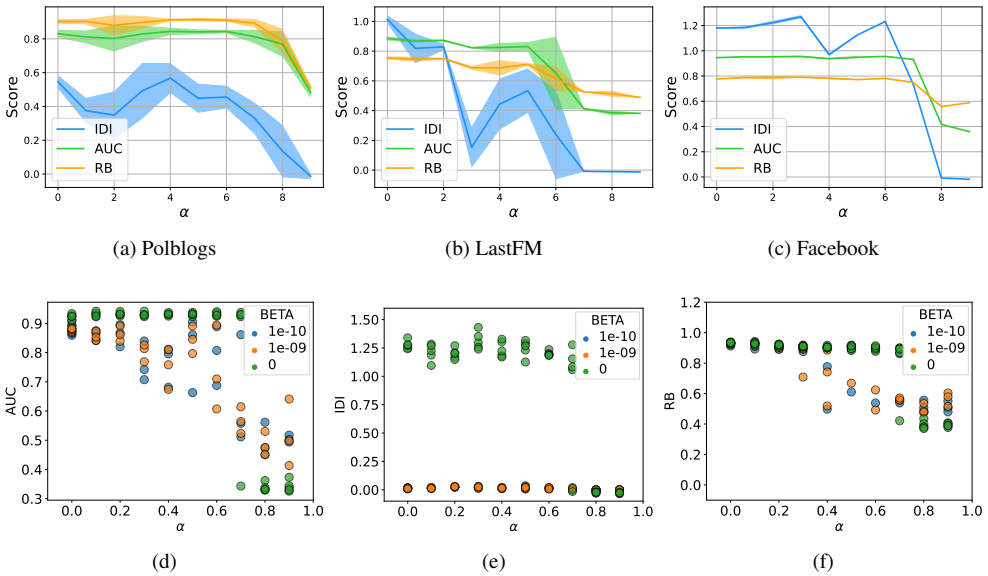
Pour les architectures `GNN`, nous réglons le nombre de couches, le nombre d'unités cachées par couche, et la dimension de représentation par recherche par grille. Nous avons également ajusté α pour le modèle sans le principe VIB, et ajusté β pour les modèles basés sur VIB. Enfin, pour `FIPR` , nous avons également ajouté le poids de régularisation dans l'ensemble des hyperparamètres. Comme l'équité consiste souvent à trouver un bon compromis utilité-équité, nous utilisons la moyenne harmonique entre l'AUC et le IDI pour sélectionner le meilleur ensemble d'hyperparamètres pour toutes les méthodes.

3.3 Resultats

Résultats en prédiction de liens La figure 3 présente les résultats obtenus sur les jeux de données. Le modèle optimal devrait être situé dans le coin supérieur gauche (AUC élevée et IDI/RB faible). Sans surprise, sur Polblogs et LastFM, `GAT` et `GCN` obtiennent de meilleures AUC, mais les pires IDI et RB, puisqu'ils ne prennent pas en compte les contraintes d'équité. Nous pouvons également noter que `GAT` est plus enclin au traitement inéquitable que `GCN`.

D'un point de vue global : en termes d'IDI, nos modèles permettent d'obtenir une AUC plus élevée à IDI égal, par rapport aux modèles FIPR et, en termes de RB, ils permettent un compromis entre FIPR (RB faible mais perte importante d'AUC) et les modèles non régularisés (AUC élevée mais RB élevé). On obtient donc avec `LEAVE` un classifieur équitable qui conserve de bonnes performances sur la tâche de prédiction de liens, mais avec des représen-

Comment rendre les GNN plus équitables pour la prédiction de liens ?



(a) Polblogs (b) LastFM (c) Facebook

(d) (e) (f)

FIG. 4 – Première ligne : impact de α avec $\texttt{LEAVE-GAT}_{wo_{VIB}}$ pour les trois repères (a à c) ; deuxième ligne : impact de la combinaison de α et β pour LEAVE-GAT sur Polblogs (d à f). Chaque point correspond aux résultats obtenus avec une graine aléatoire différente. L'IDI est en échelle logarithmique, donc équitable si proche de 0.

tations qui restent partiellement séparables. L'intégration de la variance, dans LEAVE versus $\texttt{LEAVE}_{wo_{VIB}}$, améliore globalement l'équité (à l'exception du RB sur LastFM pour GCN) tout en maintenant, voire en améliorant l'AUC. Ainsi, l'introduction d'une dimension probabiliste nous permet de prédire des liens "inattendus" mais probables.

Impacte d'α et β Enfin, la figure 4 montre l'impact d' α et de β. Nous présentons le logarithme de l'IDI, ce qui signifie que le modèle est équitable s'il est proche de 0. Comme prévu, lorsque α augmente, l'AUC, le RB et l'IDI diminuent, confirmant une amélioration de l'équité au détriment des performances en prédiction de liens. Néanmoins, l'évolution de IDI n'est pas monotone, et nous observons des régions d'augmentation de cette mesure, mais aussi des régions où l'IDI diminue sans perturber l'AUC (la solution idéale). Enfin, les scores RB et AUC évoluent conjointement, suggérant que les performances sont, sur ces données et avec ces encodeurs, fonction de la séparabilité des représentations de l'attribut sensible. Une contrainte plus forte sur les représentations est donc nécessaire pour débiaiser ces architectures. En ce qui concerne l'impact de β, l'ajout du terme d'incertitude attaché à β permet d'obtenir un compromis plus lisse et plus intéressant entre une représentation exacte et une représentation équitable. Pour DI, nous observons que l'ajout du second terme de régularisation est en fait suffisant pour atteindre l'équité de ce point de vue (log de IDI proche de 0 lorsque $\alpha = 0$).

4 Conclusion

Nous avons abordé le problème de l'apprentissage de représentations équitables des nœuds avec LEAVE, un modèle de bout en bout optimisant une fonction qui prend en compte simultanément la tâche de prédiction de liens et la contrainte d'équité dyadique. Notre modèle permet de contrôler explicitement le compromis entre la capture de la structure relationnelle du graphe et la réduction du biais potentiel pour la prédiction des liens. Nos résultats expérimentaux confirment que les représentations produites par LEAVE peuvent être utilisées efficacement pour la prédiction équitable des arêtes. Les perspectives de recherche futures sont nombreuses. Par exemple, nous aimerions explorer la possibilité d'étendre notre modèle au cas des attributs sensibles continus et aux graphes pondérés.

Références

Adamic, L. A. et N. Glance (2005). The political blogosphere and the 2004 us election : divided they blog. In *International Workshop on Link discovery*, pp. 36–43.

Alemi, A., I. Fischer, J. Dillon, et K. Murphy (2017). Deep variational information bottleneck. *International Conference on Learning Representations,*.

Bose, A. et W. Hamilton (2019). Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*, pp. 715–724.

Buyl, M. et T. D. Bie (2020). Debayes : a bayesian method for debiasing network embeddings. In *International Conference on Machine Learning*, pp. 2537–2546.

Buyl, M. et T. D. Bie (2021). The kl-divergence between a graph model and its fair i-projection as a fairness regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 351–366. Springer.

Dai, E. et S. Wang (2021). Say no to the discrimination : Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 680–688.

Gourru, A., J. Velcin, et J. Jacques (2020). Gaussian embedding of linked documents from a pretrained semantic space. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3912–3918.

Grover, A. et J. Leskovec (2016). node2vec : Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864.

Hofstra, B., R. Corten, F. Van Tubergen, et N. B. Ellison (2017). Sources of segregation in social networks : A novel approach using facebook. *American Sociological Review 82*(3).

Kingma, D. P. et M. Welling (2014). Auto-encoding variational bayes. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kipf, T. N. et M. Welling (2016). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.

Laclau, C., I. Redko, M. Choudhary, et C. Largeron (2021). All of the fairness for edge prediction with optimal transport. In *International Conference on Artificial Intelligence and*

Statistics, pp. 1774–1782. PMLR.

Li, P., Y. Wang, H. Zhao, P. Hong, et H. Liu (2021). On dyadic fairness : Exploring and mitigating bias in graph connections. In *International Conference on Learning Representations*.

Madras, D., E. Creager, T. Pitassi, et R. Zemel (2018). Learning adversarially fair and transferable representations. In *ICML*, pp. 3384–3393. PMLR.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, et A. Galstyan (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv. 54*(6).

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E 67*(2).

Oh, S. J., K. P. Murphy, J. Pan, J. Roth, F. Schroff, et A. C. Gallagher (2018). Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*.

Rahman, T. A., B. Surma, M. Backes, et Y. Zhang (2019). Fairwalk : Towards fair graph embedding. In *Proceedings of the Twenty-Seventh International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3289–3295.

Rozemberczki, B., C. Allen, et R. Sarkar (2021). Multi-scale attributed node embedding. *Journal of Complex Networks 9*(2), cnab014.

Rozemberczki, B. et R. Sarkar (2020). Characteristic functions on graphs : Birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1325–1334.

Tishby, N., F. C. Pereira, et W. Bialek (2000). The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pp. 368–377.

Veličković, P., G. Cucurull, A. Casanova, A. Romero, P. Liò, et Y. Bengio (2018). Graph attention networks. In *International Conference on Learning Representations*.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, et C. Dwork (2013). Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR.

Summary

Although algorithmic fairness has recently raised a great deal of interest in the machine learning community, the number of contributions specific to graph data remains scarce. In this paper, we address the problem of fair representation learning for graph data with a focus on the notion of dyadic fairness in the context of edge prediction for attributed graphs. We designed a model that, given pairs of nodes along with a protected attribute, learns individual representation based on the variational information bottleneck principle (Alemi et al., 2017). The proposed model allows us to simultaneously learn non-linear node embeddings reflecting the graph structure, while explicitly controlling the level of fairness. Experiments carried out on several real-world datasets confirmed the capacity of the proposed method both to maintain high accuracy on the edge prediction task while significantly reducing bias.