Classification dans les graphes hétérogénes basée sur une représentation latente des noeuds

Yann Jacob, Ludovic Denoyer, Patrick Gallinari

Classification dans les graphes hétérogènes basée sur une représentation latente des nœuds

Yann Jacob — Ludovic Denoyer — Patrick Gallinari

Université Pierre et Marie Curie - LIP6
Boîte courrier 169, 4 place Jussieu 75252 PARIS cedex 05

RÉSUMÉ. *Les réseaux sociaux sont souvent composés de différents types de nœud. Apprendre et inférer sur ces réseaux hétérogènes est une tâche récente. Nous considérons la tâche d'étiquetage de nœuds dans les réseaux sociaux, où différents types de nœud doivent être étiquetés par différents jeux de catégories ou d'étiquettes. Nous proposons une nouvelle approche transductive qui apprend automatiquement à projeter les différents types de nœud dans un espace latent commun, cette représentation apprise étant utilisée ensuite pour classifier les différents éléments. Cette approche exploite l'idée que deux nœuds connectés dans un réseau social tendront à avoir des représentations latentes similaires peu importe leur type. Cette hypothèse nous permet d'apprendre les corrélations entre les catégories de nœuds de type différent, quand les méthodes de l'état de l'art traitent chaque type de nœud indépendamment. Nous avons testé ce modèle sur deux jeux de données et il obtient de bonnes performances.*

ABSTRACT. *Social networks are often composed of different types of nodes. Learning and performing inference on such heterogeneous networks is a recent task. We address the tasks of tagging of nodes in social networks, where the different types of nodes have to be labeled by different set of categories or tags. We propose a new transductive approach that automatically learns to project the different types of nodes onto a common latent space, this learned representation being then used for classifying the different elements. This framework exploits the idea that two nodes connected in a social network will tend to have a similar latent representation regardless of their type. This assumption allows us to learn correlations between the labels of nodes of different types, when state-of-the-art methods usually address each type of node separately. The model is tested on two datasets and shows good performance.*

MOTS-CLÉS : *Classification dans les graphes, Graphes hétérogènes, Réseaux sociaux*

KEYWORDS: *Graph-based classification,Heterogeneous networks,Social networks*

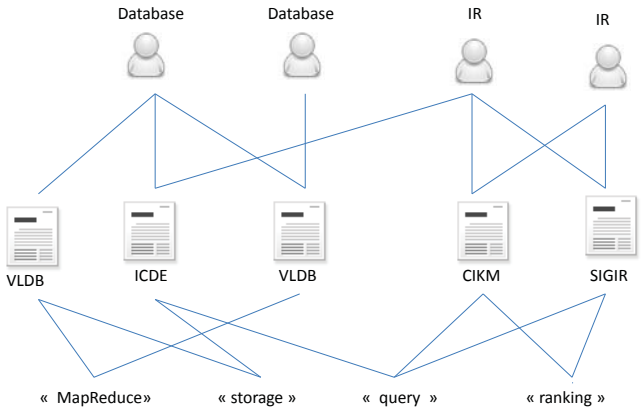


Figure 1. *Réseau bibliographique hétérogène : les auteurs sont connectés aux articles publiés. Les auteurs sont étiquetés par leur domaine, tandis que les articles le sont par leur conférence. Deux auteurs ou deux articles ne sont jamais connectés directement.*

1. Introduction

Les médias sociaux sur le web sont le plus souvent des réseaux hétérogènes complexes avec des nœuds et des relations entre nœuds de type différent, correspondant à différents objets, concepts et relations. Des exemples classiques sont les réseaux de contenu utilisateur comme Flickr où les nœuds sont des auteurs, photos, commentaires, et les relations des liens d'amitié ou de publication (entre un auteur et un contenu). Un autre exemple est un réseau bibliographique avec des nœuds correspondant aux auteurs, papiers, conférences, et des relations de citation, de coauteur, de présence d'un papier dans les actes de la conférence, etc. Les réseaux hétérogènes apparaissent fréquemment dans d'autres domaines comme les sites d'e-commerce avec des utilisateurs, objets, commentaires, recommendations comme sur TripAdvisor. La modélisation et l'analyse de tels réseaux complexes ont récemment été entamées pour la résolution de tâches génériques de fouille de données comme la classification (Angelova *et al.,* 2012), le clustering(Sun *et al.,* 2009), la prédiction de liens(Davis *et al.,* 2011, Wang *et al.,* 2011) ou l'analyse d'influence(Liu *et al.,* 2010).

Nous nous intéressons à la tâche de classification des nœuds dans les graphes hétérogènes composés de différents types de nœuds, chaque type étant associé avec son propre ensemble d'étiquettes. Par exemple dans un réseau bibliographique, les auteurs et les papiers peuvent être étiquetés respectivement par leur domaine de recherche et leur conférence (voir Figure 1), dans Flickr les utilisateurs et les photos peuvent être respectivement étiquetés par leurs groupes reflétant leurs domaines d'intérêt et par leurs tags visuels. Notre hypothèse est que les nœuds de type différent s'influencent entre eux, les étiquettes étant inter-dépendantes, modéliser cette dépendance est donc important pour classifier ou étiqueter les nœuds de manière précise et ce ne peut être

réalisé avec les formulations homogènes classiques. Par exemple, dans un réseau Flickr, les groupes d'un utilisateur sont reliés aux tags que l'utilisateur a utilisé et vice-versa.

Bien que beaucoup de travaux existent en classification dans les réseaux homogènes (i.e. composés d'un seul type de nœud), peu de tentatives ont été faites pour les réseaux hétérogènes. La plupart des travaux sur les graphes hétérogènes définissent une projection d'une formulation hétérogène du problème vers une formulation homogène pour ensuite pouvoir utiliser les techniques connues dans cette dernière formulation. Ils font souvent des hypothèses restrictives sur les données, comme utiliser le même jeu d'étiquettes pour tous les types de nœuds ou s'intéresser à des types particuliers de dépendances (chemins de longueur 2 entre nœuds de même type par exemple).

Nous proposons un nouvel algorithme pour apprendre à étiqueter les nœuds dans un réseau hétérogène qui n'a pas besoin de telles hypothèses. Cet algorithme opère dans le contexte transductif et dans le cadre de la régularisation. Il peut être utilisé pour étiqueter des nœuds de type différent avec des jeux d'étiquettes différents, dans un graphe de n'importe quelle structure. Il est capable d'apprendre les dépendances entre les jeux d'étiquettes associés aux différents types de nœuds, et d'inférer les étiquettes associées à un nœud par une fonction sur les propriétés globales du graphe et sur le voisinage local du nœud. Cet algorithme apprend une représentation latente des nœuds du réseau de façon à ce que tous les nœuds, peu importe leur type, partagent le même espace latent commun. Cette représentation sera utilisée ensuite pour inférer les catégories. De plus, la même formulation permet une modélisation naturelle des attributs des nœuds du graphe (information de contenu associé au nœud dans notre contexte), en considérant simplement ces attributs comme des nœuds. A notre connaissance, il s'agit du premier modèle capable de traiter le problème d'étiquetage dans un graphe hétérogène quelconque qui utilise les dépendances entre les étiquettes des nœuds de type différent.

Le papier est organisé comme suit : la Section 2 donne les notations, et introduit le contexte des modèles classiques d'étiquetage dans les graphes homogènes. La Section 3 décrit l'algorithme et une extension pour inclure l'information de contenu des nœuds dans notre modèle. La Section 4 présente des résultats expérimentaux sur trois jeux de données, et une analyse qualitative des représentations latentes apprises par le modèle. La Section 5 donne une synthèse de l'état de l'art des modèles d'étiquetage dans les graphes et des problèmes d'apprentissage automatique proches.

2. Contexte et notations

2.1. *Notations*

Un réseau hétérogène est modélisé comme un graphe pondéré non dirigé avec un type associé à chaque nœud.

Notons $\mathcal{T} = 1, 2, .., T$ l'ensemble des T types de nœud possibles. Les nœuds sont notés x_i, le nombre de nœud N et les arcs $w_{i,j} \in \mathbb{R}$, où $w_{i,j}$ est le poids de la relation entre x_i et x_j. $w_{i,j} = 0$ si il n'y a pas de lien entre x_i et x_j. Soit $t_i \in \mathcal{T}$ le type du nœud x_i, $\mathcal{Y}^t$ l'ensemble des catégories associées au type de nœud t, et C^t la cardinalité de $\mathcal{Y}^t$. Nous considérons un contexte transductif où le réseau est composé de $\ell < N$ nœuds étiquetés $(x_1, ..., x_\ell)$ avec $\forall i \in [1..\ell], y_i \in \mathbb{R}^{C^{t_i}}$, le composant j de y_i noté y_i^j est défini comme :

$$y_i^j = \begin{cases} +1 \text{ si } x_i \text{ appartient à la catégorie } j \text{ de } \mathcal{Y}^{t_i} \\ -1 \text{ sinon} \end{cases} .$$

2.2. *Modèle de propagation d'étiquette dans les graphes homogènes*

Où il n'y a qu'un seul type de nœud i.e. $T = 1$, les modèles transductifs classiques exploitent l'hypothèse de variété *deux nœuds connectés tendent à avoir les mêmes étiquettes*. Notons $\hat{y}_i$ le vecteur de scores pour les catégories prédites pour n'importe quel nœud x_i dans le réseau. L'apprentissage transductif dans les modèles de graphe donne souvent une fonction de la forme suivante (Abernethy *et al.,* 2008)(Zhou *et al.,* 2004)(Zhou *et al.,* 2005) :

$$(\hat{y_1}, ..., \hat{y_N}) = \underset{\tilde{y_1},...,\tilde{y_N}}{\operatorname{argmin}} \sum_{i=1}^{\ell} \Delta(\tilde{y_i}, y_i) + \lambda \sum_{i,j} w_{i,j} ||\tilde{y_i} - \tilde{y_j}||^2$$

où Δ correspond au coût de prédire les scores $\tilde{y_i}$ au lieu des catégories réelles y_i pour les nœuds étiquetés et $||\tilde{y_i} - \tilde{y_j}||^2$ est un terme de régularisation qui pousse les nœuds connectés à avoir les mêmes scores. λ est un hyper-paramètre qui correspond au compromis entre les erreurs de prédiction et la régularité voulue dans la structure du réseau. Plusieurs techniques peuvent être utilisées pour minimiser cette fonction comme un algorithme de descente de gradient stochastique, une méthode de descente de coordonnées ou une marche aléatoire.

2.2.1. *Extension aux graphes hétérogènes*

Cette formulation pour les graphes homogènes peut être étendue de plusieurs manières pour les graphes hétérogènes (Ji *et al.,* 2010, Hwang *et al.,* 2010, Angelova *et al.,* 2012). Cependant, de telles extensions ne permettent pas d'utiliser pleinement l'information présente dans le réseau hétérogène en particulier les corrélations entre les catégories de types de nœud différents.

Comme illustration considérons le cas d'un graphe composé de deux types de nœud 1 et 2 avec les jeux d'étiquettes $\mathcal{Y}^1$ et $\mathcal{Y}^2$ — voir Figure 2. Il n'est pas possible de faire une propagation directe d'étiquettes de $\mathcal{Y}^1$ vers $\mathcal{Y}^2$ puisque les deux jeux d'étiquettes sont différents et correspondent à des sémantiques d'étiquettes différentes. Une solution proposée par (Angelova *et al.,* 2012) est équivalente à transformer le graphe hétérogène en deux graphes homogènes (avec les relations *plein* et *dotted*

fusionnées pour le type 1) comme illustré dans la Figure 2 puis à faire la propagation d'étiquettes sur les deux graphes homogènes séparément.

Ceci a deux inconvénients :

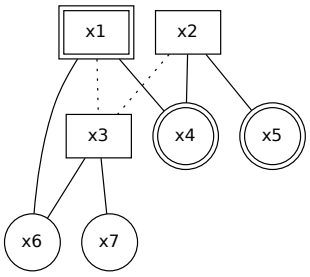
– Les dépendances entre les deux jeux d'étiquettes sont perdues, les étiquettes prédites pour un nœud n'étant dépendants que des étiquettes de ses voisins **du même type**. Par exemple, une relation comme *les utilisateurs étiquetés "vieux" sont liés aux images d'étiquette "fleurs"* ne peut pas être prise en compte par une telle décomposition. Plus il y a de types de nœud, plus la perte d'information est importante.

– Etendre cette idée aux graphes plus complexes peut ne pas être trivial, et les sémantique correspondantes peuvent devenir complexes. Considérons par exemple le cas où il y a plus de 2 types de nœud. Il y'a plusieurs manières de définir une telle transformation, les chemin liant deux nœuds de même type peuvent avoir différentes longueurs ou avec différents types de nœud sur le chemin. Que serait alors la sémantique d'un lien entre deux nœuds de même type ? Comment pondérer ces liens ? C'est pourquoi de telles extensions sont souvent limitées par des hypothèses trop simplificatrices sur la nature des graphes, les jeux d'étiquettes ou les relations. Un exemple simple est fourni dans la Figure 2, où deux types de relation sont présents dans le graphe initial (plein et pointillé). Différents graphes homogènes projetés peuvent être définis pour les nœuds de même type. La Figure 2 (en bas à gauche) illustre la projection utilisant un saut à deux bonds d'un nœud carré vers un nœud rond puis vers un nœud carré à nouveau. La figure 2 (en bas à droite) illustre une projection utilisant les liens directs entre les nœuds carrés. Les deux projections ont des sémantiques différentes.

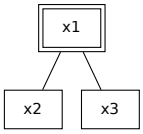
3. Modèle

3.1. *Apprentissage des représentations latentes des nœuds*

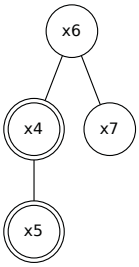
Nous proposons une solution qui ne repose pas sur une décomposition du problème initial en des problèmes d'apprentissage homogène et évite donc les limitations de ces dernières méthodes. Le but est de prendre directement en compte les dépendances entre les étiquettes des nœuds connectés, quelques soient leurs types. Nous introduisons d'abord le modèle où aucun attribut de contenu n'est associé aux nœuds du graphe puis nous décrirons dans la section 3.2 une extension naturelle pour incorporer le contenu. L'idée sous-jacente du modèle est la suivante : chaque nœud a une représentation latente dans l'espace vectoriel $\mathbb{R}^Z$ dont Z est la dimension. L'espace latent est commun à tous les types de nœud. Cette représentation latente définit une métrique dans l'espace $\mathbb{R}^Z$ telle que deux nœuds connectés tendront à avoir une représentation proche (*hypothèse de régularité*). Pour chaque type de nœud, une fonction de classification sera apprise simultanément avec la représentation latente, prenant en entrée la représentation latente d'un nœud et calculant en sortie ses étiquettes. Les nœuds de même type et de représentation latente proche tendront à avoir les mêmes étiquettes (hypothèse de régularité dans l'espace métrique) et les nœuds de types diffé-



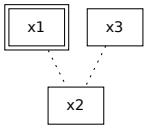
Réseau hétérogène inital avec différentes relations



Graphe homogène des nœuds de type 1 avec la relation *plein*



Graphe homogène des nœuds de type 2



Graphe homogène des nœuds de type 1 avec la relation *pointillé*

Figure 2. *Les nœuds carrés sont du type* 1 *et les ronds du type* 2*. Les nœuds doublement encadrés sont étiquetés. Transformation du réseau hétérogène (deux types de nœud) en deux réseaux homogènes. Différentes manières de transformer un réseau hétérogène en un réseau homogène existent (illustrées ici pour le type* 1*) : la transformation n'est pas triviale.*

rents avec des représentations proches s'influenceront entre eux (les corrélations entre des nœuds hétérogènes sont prises en compte). Dit autrement, à la fois les dépendances de graphe et les dépendances entre les étiquettes de types de nœud différents seront prises en compte à travers l'espace latent.

3.1.1. *Modèle transductif de classification*

Notons $z_i \in \mathbb{R}^Z$ la représentation latente du nœud x_i qui est un vecteur de taille Z.

Pour capturer la métrique du graphe dans l'espace latent, nous utilisons le coût suivant qui comporte l'hypothèse de régularité et force les nœuds liés à avoir des représentations similaires :

$$\sum_{i,j/w_{i,j}\neq 0} w_{i,j}||z_i - z_j||^2$$

Nous utilisons une norme L_2 dans l'espace latent, mais d'autres normes peuvent être utilisées. Ce terme est similaire au terme de régularité de la partie précédente si ce n'est qu'il est défini dans l'espace latent et non pas dans l'espace des étiquettes comme dans la partie précédente. La projection dans l'espace latent est apprise afin que les étiquettes puissent être prédites de la représentation latente pour chaque type de nœud. Pour cela, nous utilisons une fonction de classification notée f_θ^k pour chaque type de nœud k Cette fonction a en entrée la représentation d'un nœud et en sortie les étiquettes prédites. La f-fonction peut être apprise en minimisant un coût sur les données étiquetées comme suit :

$$\sum_{i=1}^{\ell} \Delta(f_\theta^{t_i}(z_i), y_i^x)$$

où $\Delta(f_\theta^{t_i}(z_i^x), y_i)$ est le coût de prédire les étiquettes $f_\theta^{t_i}(z_i)$ au lieu des étiquettes réelles y_i.

La fonction de coût de notre modèle combine ces deux coûts de classification et de régularisation :

$$L(z,\theta) = \sum_{i=1}^{\ell} \Delta(f_\theta^{t_i}(z_i), y_i) + \lambda \sum_{i,j/w_{i,j}\neq 0} w_{i,j}||z_i - z_j||^2.$$

La minimisation de cette fonction vise à trouver un compromis entre la régularité sur les représentation latentes des nœuds liés dans $\mathcal{Z}$ et sur la prédiction des étiquettes réelles dans $\mathcal{Y}_k$. Optimiser ce coût nous permet d'apprendre :

– La projection z_i de chaque nœud x_i dans l'espace latent.

– La fonction de classification f_θ^k pour chaque type de nœud k qui transforme un point dans l'espace latent en scores de catégories.

3.1.2. *Apprentissage*

L'apprentissage consiste à minimiser la fonction de coût. Différentes méthodes d'optimisation peuvent être utilisées comme des gradients batch ou minibatch. Nous avons utilisé une méthode de descente de gradient stochastique. L'apprentissage est détaillé dans l'algorithme 1.

L'algorithme choisit itérativement une paire de nœuds connectés puis fait une mise à jour sur les paramètres du modèle. Si un des nœuds est étiqueté, l'algorithme fait

Algorithm 1 Apprentissage avec descente de gradient stochastique

1: **procedure** APPRENTISSAGE(x, w, ϵ, λ)
2: **for** Un nombre d'itérations donné **do**
3: Choisir un (x_i, x_j) au hasard avec $w_{i,j} > 0$.
4: **if** $i \leq \ell$ **then** ▷ si x_i est étiqueté
5: $\theta \leftarrow \theta + \epsilon \nabla_\theta \Delta(f_\theta^{t_i}(z_i), y_i)$
6: $z_i \leftarrow z_i + \epsilon \nabla_{z_i} \Delta(f_\theta^{t_i}(z_i), y_i)$
7: **end if**
8: **if** $j \leq \ell$ **then** ▷ si x_j est étiqueté
9: $\theta \leftarrow \theta + \epsilon \nabla_\theta \Delta(f_\theta^{t_j}(z_j), y_j)$
10: $z_j \leftarrow z_j + \epsilon \nabla_{z_j} \Delta(f_\theta^{t_j}(z_j), y_j)$
11: **end if**
12: $z_i \leftarrow z_i + \epsilon \lambda w_{ij} \nabla_{z_i} ||z_i - z_j||^2$
13: $z_j \leftarrow z_j + \epsilon \lambda w_{ij} \nabla_{z_j} ||z_i - z_j||^2$
14: **end for**
15: **end procedure**

d'abord une mise à jour sur le premier terme de la fonction de coût. Cette mise à jour – lignes 5-6 et 9-10 – consiste à modifier successivement les paramètres de la fonction de classification θ et les paramètres des représentations latentes z_i et z_j de façon à minimiser le coût de classification. Puis, le modèle met à jour ses paramètres par rapport au terme de régularité de la fonction de coût – lignes 12-13. Ici, ϵ est le pas de gradient, et λ le compromis entre les termes de classification et de régularité.

Dans notre implémentation, nous avons utilisé une fonction hinge-loss pour Δ et un modèle linéaire pour f_θ^t :

$$f_\theta^{t,k}(z) = \langle \theta_{t,k}; z \rangle$$

$$\Delta(f_\theta^t(z), y) = \sum_{k=1} max(0; 1 - y^k f_\theta^{t,k}(z))$$

où y^k est le score désiré de la catégorie k pour le nœud x (-1 ou $+1$) et $f_\theta^{t,k}(z)$ est le score prédit de la catégorie k par le modèle.

3.2. *Modèle hétérogène et contenu*

La fonction objectif ne considère que la structure de graphe via sa matrice des poids et les étiquettes des nœuds. Parfois, des attributs sont associés aux nœuds (par exemple du contenu textuel pour des commentaires ou du contenu visuel pour des images). Soit un exemple où du contenu textuel est associé à certains types de nœuds (l'espace de description du contenu est le même pour un type de nœud donné). La formulation ci-dessus permet de modéliser facilement cette information en représentant chaque mot comme un nouveau nœud comme dans (Ji *et al.,* 2010) avec un lien

DBLP Sans Contenu	Noeuds	**Type**	**Nb. Noeuds**	**Nb. Etiquetés**	**Nb. Etiquettes différentes**
		Papier	14,376	14,376	20
		Auteur	14,475	4,057	4
	Arcs	**Type**	**Nb. Arcs**		
		Auteur→Papier	41,794		
DBLP Avec Contenu	Noeuds	Papier	14,376	14,376	20
		Auteur	14,475	4,057	4
		Mot	8,920	0	
	Arcs	Auteur→Papier	41,794		
		Papier→Mot	114,624		
Flickr	Noeuds	Photo	46,926	8,766	21
		Utilisateur	4,760	3,476	42
	Arcs	Utilisateur←Utilisateur	175,779		
		Utilisateur←Photo	46,926		

Tableau 1. *Statistiques sur les trois jeux de données*

Modèle	Type de nœud	Taille en entraînement		
		10%	30%	50%
Modèle homogène	Auteur	58.8	65.0	69.2
	Papier	28.2	32.7	34.8
Modèle hétérogène (sans Contenu)	Auteur	64.9	75	83
	Papier	30.4	34.5	37.2
Modèle hétérogène (avec Contenu)	Auteur	67	82	87.1
	Papier	30.6	35.6	37.9

Tableau 2. *Précision sur les données DBLP avec une représentation latente de taille 30.*

pondéré entre ce mot et le nœud le contenant. Le poids de la relation peut être défini par n'importe quelle mesure (par exemple un TF-IDF).

4. Expériences

4.1. *Données*

Les expériences ont été lancées sur deux jeux de données extraits de DBLP et Flickr. Pour le premier, la tâche est de la classification mono-étiquette, pour le second, de la classification multi-étiquette. Pour DBLP deux groupes d'expériences - avec et sans les mots de contenu- ont été lancés, ils sont respectivement notés *DBLP Sans Contenu* et *DBLP Avec Contenu* par la suite. Pour les expériences sur Flickr, aucune

Modèle	Type de nœud	Taille en entraînement 10%	30%	50%
Modèle homogène	Utilisateur	42	46.8	48.7
	Photo (Même auteur)	35.6	59.4	65.5
	Photo (Auteurs amis)	19.3	21.7	22
	Photo (Même auteur ou auteurs amis)	22.4	31.3	32.4
Modèle hétérogène	Utilisateur	42.9	45.7	49.1
	Photo	43.3	61.6	68.1

Tableau 3. *P@1 sur le corpus Flickr avec un espace latent de taille 200*

caractéristique n'existe pour les nœuds. Les deux jeux de données sont introduits ci-dessous.

– Le corpus **DBLP** est un réseau bibliographique composé d'auteurs et de papiers. Ce corpus a été publié dans (Sun *et al.,* 2009), où la tâche est de la classification hétérogène avec les auteurs et les papiers partageant les mêmes étiquettes (leur domaine de recherche). La tâche étudiée ici considère deux ensembles d'étiquettes : les auteurs sont étiquetés par leur domaine de recherche (4 domaines différents) tandis que les papiers le sont par la conférence dans laquelle ils ont été publiés (20 conférences). Les auteurs et les papiers sont connectés par une relation *auteur*. La version sans contenu est composée de deux types de nœud reliés de telle manière que le graphe est bipartite. La version avec contenu a trois types de nœud – auteur, papier et mot – où chaque papier est connecté au nœuds des mots de son titre – voir la Figure 1. La classification est mono-étiquette sur les papiers et les auteurs. Des statistiques générales sur le corpus sont données dans la Table 1.

– Le corpus **Flickr** est composé de photos et d'utilisateurs. Les étiquettes des photos correspondent aux différents tags tandis que les utilisateurs sont étiquetés par leurs groupes. La classification est multi-étiquette : les images et utilisateurs peuvent appartenir à plus d'une catégorie. Les photos sont reliées aux utilisateurs à travers une relation *authorship*, tandis que les utilisateurs sont reliés entre eux par une relation *friendship*. Des statistiques sont données dans la Table 1. Nous avons gardé les tags des images qui apparaissent dans au moins 500 images, et les groupes d'utilisateurs qui comportent au moins 500 membres résultant dans 21 tags possibles pour les photos et 41 groupes pour les auteurs.

4.2. *Modèles*

Nous avons comparé l'approche proposée notée **Modèle hétérogène** avec un modèle, noté **Modèle Homogène**, qui utilise des graphes homogènes multiples, un par type de nœud, pour représenter le réseau hétérogène. Le modèle homogène de l'état de l'art minimise le coût défini dans la section 2.2. Il correspond au modèle proposé dans (Denoyer *et al.,* 2010).

Pour le modèle homogène, des graphes homogènes ont été construits de la manière suivante :

– Pour le corpus DBLP, le graphe des auteurs est construit en connectant deux co-auteurs ; le graphe des papiers est construit en connectant deux papiers écrits par le même auteur.

– Pour Flickr, le graphe des utilisateurs est construit en connectant deux amis. Différents graphes de photos sont possibles : le graphe **Même auteur** est construit en connectant deux photos qui ont le même auteur, le graphe **Auteurs amis** est construit en connectant deux photos qui ont été publiées par des utilisateurs amis, le graphe **Même auteur ou auteurs amis** est construit est connectant deux photos qui ont été publiées par le même utilisateur ou des utilisateurs qui sont amis.

Notons que le modèle "Même auteur" où deux images sont liées à travers un nœud auteur intermédiaire, partage des similarités avec le modèle Graffiti (Angelova *et al.,* 2012) qui utilise une marche aléatoire avec un saut à deux bonds pour connecter des nœuds du même type liés par un chemin de longueur 2 passant par un nœud de type différent.

Les expériences ont été lancées avec 10 ensembles d'entraînement aléatoires différents et les performance rapportée est la moyenne sur les 10. Différentes configurations ont été testées :

– Les expériences ont été lancées avec différentes tailles en entraînement : 10%, 30%, 50%. La taille en entraînement réfère à **la proportion de nœuds étiquetés** utilisée pour l'ensemble d'entraînement. Par exemple dans DBLP Sans Contenu, une taille en entraînement de 10% signifie que 1,437 papiers sur 14,376 et 405 utilisateurs sur 4,050 étiquetés ont été mis dans l'ensemble d'entraînement. Notons que pour ce corpus, tous les papiers sont étiquetés alors que seulement 4,050 sur 14,475 auteurs le sont. Les nœuds non étiquetés apparaîtront seulement dans le terme de régularisation. L'évaluation n'est faite que sur les nœuds étiquetés durant l'entraînement.

– Le pas de gradient a été fixé à 0.1 – des pas plus petits ont été testés mais aboutissent à la même performance au prix d'un temps de convergence plus long.

– L'algorithme a été itéré jusque convergence – i.e. jusqu'à ce que les étiquettes des nœuds ne changent plus.

– Comme il y a beaucoup plus de liens entre utilisateurs qu'entre les utilisateurs et les photos dans les données Flickr, le poids des relations a été normalisé pour éviter que la relation $utilisateur \rightarrow utilisateur$ aie considérablement plus d'influence que la relation $utilisateur \rightarrow photo$. Le poids pour $utilisateur \rightarrow photo$ a été fixé à 1 et celui pour $utilisateur \rightarrow utilisateur$ à 0.1.

Pour l'évaluation, sur les données mono-étiquette (DBLP avec et sans contenu), la catégorie prédite est celle qui a le plus grand score. Nous utilisons la mesure de *précision* qui est le ratio entre le nombre de nœuds bien classifiés et le nombre de nœuds total. Pour les données multi-étiquettes, après avoir prédit les scores pour différentes catégories, il faut décider quel sous-ensemble de catégories assigner à chaque nœud

du réseau. Pour l'évaluation, nous avons utilisé la **Précision à 1 (P@1)** qui mesure le pourcentage de nœud où la catégorie avec le plus grand score est une catégorie réelle du nœud. Cela correspond à la capacité du modèle à prédire un bon score pour au moins une des étiquettes. La **Précision à k (P@k)** est la proportion d'étiquettes correctes parmi l'ensemble des k étiquettes avec les plus hauts scores prédits. Pour chaque nœud d'évaluation, k est artificiellement choisi pour être le nombre de catégories réelles. C'est une mesure optimiste de la capacité d'un modèle à ordonnancer correctement les k catégories réelles d'un nœud.

4.3. *Résultats*

4.3.1. *Hétérogène contre homogène*

Les Tableaux 2 et 3 présentent la performance obtenue sur les différents corpus pour les modèles hétérogène et homogène pour différentes tailles en entraînement. L'évaluation de ces modèles est faite séparément sur deux types de nœud – papier et auteur pour DBLP, utilisateur et photo pour Flickr. Globalement, le modèle hétérogène est capable d'obtenir de meilleurs performances que le modèle homogène. Par exemple, pour les données DBLP sans contenu, à une taille de 10% en entraînement, notre modèle obtient une précision de 64.9% sur les nœuds auteur et de 30.4% sur les nœuds papier tandis que le modèle homogène obtient 58.8% sur les auteurs et 28.2% sur les papiers. Quand le graphe est utilisé avec contenu, où les mots sont connectés aux papiers les contenant dans leur titre, le modèle hétérogène obtient toujours de meilleurs résultats et dépasse le modèle hétérogène sans contenu, montrant la capacité de notre modèle à gérer non seulement différents types de nœud, mais aussi l'information de contenu. L'augmentation de performance est importante pour les nœuds auteur, et beaucoup plus faible pour les nœuds papiers. Ceci est probablement dû au fait qu'il y a beaucoup moins d'auteurs étiquetés que de papiers, la tâche de classification sur les auteurs est donc plus dure et les modèles plus simples échouent. Le gain est également plus important quand la proportion de nœuds étiquetés est augmentée (par exemple un gain de 18% pour 50% de nœuds étiquetés). Ici le modèle hétérogène plus sophistiqué est capable de tirer parti de l'information additionnelle fournie par ces étiquettes.

Pour Flickr, nous comparons le modèle hétérogène avec différents graphes homogènes projetés pour les photos, comme décrit dans la section 2. Les performances pour $P@1$ sont données dans le Tableau 3. Le gain de performance est plus faible que dans DBLP, ce qui est probablement dû à la prééminence des relations "ami" et "même auteur" respectivement pour les utilisateurs et les photos. Dit autrement, pour le corpus Flickr, la plupart de l'information est présente dans ces deux types de relation. D'un autre coté, le modèle proposé nous permet d'obtenir la meilleure performance sans tester plusieurs projections de graphe et avec un temps de calcul moindre le nombre de liens dans le réseau hétérogène étant beaucoup plus faible que dans les graphes homogènes générés (Table 1 - dernière colonne).

Taille entraînement	Z	P@1/P@k Photo	P@1/P@k Utilisateur	Nb. Arcs
10%	Homogène	35.6/35.4	42/33.4	645,039
	20	38.4/33.8	**46.5/33.5**	222,705
	50	41.3/38	39.7/29.6	222,705
	100	**43.3**/40.4	39.8/30.5	222,705
	200	43.1/**41.7**	42.9/30.6	222,705
30%	Homogène	59.4/56.3	46.8/**36.2**	645,039
	20	48.8/46.4	**47.6**/33.7	222,705
	50	55/53.9	46.3/33.7	222,705
	100	60.4/58	45.6/33.5	222,705
	200	**61.6/58.2**	45.7/33.5	222,705
50%	Homogène	65.5/63.7	48.7/36.0	645,039
	20	50.3/48.7	**49.2**/30.5	222,705
	50	57.4/57.3	47.3/**36.4**	222,705
	100	67.6/**64.8**	48.3/35.1	222,705
	200	**68.1**/64.7	49.1/34.6	222,705

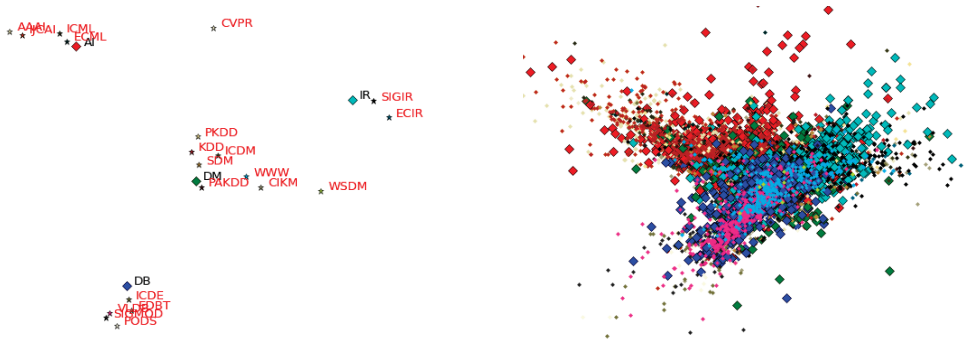
Tableau 4. *P@1 et P@k sur Flickr en fonction de la taille de représentation Z par rapport au modèle homogène.*

4.3.2. *Influence de la taille de représentation*

Examinons maintenant comment la performance varie en fonction de la taille de l'espace latent Z. Pour le corpus DBLP, utiliser un espace de taille 30 donne le plus souvent les meilleurs résultats pour toutes les tailles en entraînement. Quand l'espace latent est trop petit, le modèle est incapable de trouver une bonne représentation, quand il est trop grand, le modèle surapprend, résultant en une perte de précision. Sur le corpus Flickr, les résultats sont illustrés dans le Tableau 4. Un espace latent de taille 200 donne de bonnes performances pour les utilisateurs et les photos. En restreignant la taille de l'espace latent à 20, le modèle semble se concentrer sur les utilisateurs plus que sur les photos. Ceci est probablement dû au fait que les utilisateurs sont connectés à travers des relations directes, alors que les photos sont liées par des chemins au moins de longueur 2 : quand le modèle n'a pas de capacité de representation assez forte, il tend à faire de la propagations sur les nœuds connectés et à ignorer la propagation sur les chemins indirects.

4.3.3. *Résultats qualitatifs*

Pour visualiser les représentations latentes, nous avons lancé une ACP sur ces représentations pour les articles et les auteurs sur les données DBLP pour les nœuds de test. Une projection sur les deux premières dimensions propres de l'ACP est illustrée dans la Figure 3. La figure de droite montre tous les nœuds étiquetés dans l'espace latent (après réduction de dimension par l'ACP) - différentes formes correspondent à différents types de nœud (les formes en gros diamant correspondent aux auteurs, les petits diamants aux conférences), et différents couleurs correspondent à différentes étiquettes. La projection montre que le modèle proposé est capable de placer les nœuds



(a) Centroïdes des nœuds en test pour les catégories (b) Les nœuds en test

Figure 3. *ACP sur le corpus DBLP - projections pour 10% en entraînement et une couche latente de taille 30.*

avec les mêmes étiquettes dans la même partie de l'espace latent. La partie gauche de la figure montre la projection moyenne de la représentation latente pour chaque étiquette. Les quatre diamants correspondent aux centroïdes des quatre domaines de recherche des auteurs, tandis que les vingt petites étoiles correspondent aux centroïdes des papiers publiés à chacun des vingt conférences. On peut voir que, pour chaque domaine de recherche, le modèle est capable d'apprendre une représentation latente proche des conférences reliées à ce domaine, montrant la capacité du modèle à apprendre les corrélations entre les étiquettes de nœuds de types différents.

5. Etat de l'art

La classification des nœuds dans les graphes a motivé beaucoup de travaux dans la dernière décénnie et différents modèles ont été proposés. Il y'a deux familles principales de modèles : (i) les modèles itératifs – *Iterative Classification* (Sen *et al.,* 2008) et ses variantes comme SICA (Maes *et al.,* 2009) , *Gibbs Sampling* (Macskassy *et al.,* 2003), ou *Stacked Learning* (Kou, 2007), qui utilisent des classifieurs prenant en entrée les attributs du nœud et les étiquettes de ses voisins. Le processus de classification est répété itérativement pour chaque nœud de manière séquentielle (un nœud après l'autre) ou en parrallèle (tous les nœuds en même temps) jusque convergence. (ii) Les modèles régularisés semi-supervisés et transductifs – (Zhou *et al.,* 2005), (Belkin *et al.,* 2006) et (Zhou *et al.,* 2004) – qui sont basés sur la minimisation d'une fonction objectif qui encourage les nœuds connectés à avoir les mêmes étiquettes. Des extensions des modèles régularisés capables de prendre en compte l'information de contenu ont aussi été developpées avec des applications à l'étiquettage de réseaux sociaux (Denoyer *et al.,* 2010) et à la détection de spam (Abernethy *et al.,* 2008). Tous ces modèles ont été developpés pour des graphes homogènes et reposent sur l'idée que des nœuds voisins dans le graphe doivent partager les mêmes propriétés. La fouille de réseaux hétérogènes est un domaine plus récent où différentes tâches ont été abordées : classification de nœuds ((Ji *et al.,* 2010, Hwang *et al.,* 2010, Angelova *et al.,* 2012)), prédiction de liens ((Davis *et al.,* 2011, Wang *et al.,* 2011)), analyse d'influence ((Liu *et al.,*

2010)), etc. La classification de nœuds, objet de ce papier, a été récemment abordée dans quelques papiers. Le travail de Ji et al. dans (Ji *et al.,* 2010) repose sur la transformation du problème de classification hétérogène en un problème de classification multi-relationnel où il y'a un type de relation associé à chaque paire de type de nœud. Par exemple dans un réseau hétérogène d'auteurs et de papiers comme DBLP, ce modèle définira trois types de relations : auteur-auteur, papier-papier et auteur-papier. (Hwang *et al.,* 2010) est aussi basé sur l'idée de transformer un réseau hétérogène en un réseau homogène – mais ici la propagation d'étiquette est faite séparément sur chaque sous-réseau homogène, quand dans (Ji *et al.,* 2010) la propagation est faite sur tout le graphe multi-relationnel. Ces deux derniers modèles sont limités aux réseaux où les différents types de nœud ont le même ensemble de catégories. A notre connaissance, le seul modèle existant abordant la classification de réseaux hétérogènes avec différents ensembles d'étiquettes par type de nœud est (Angelova *et al.,* 2012). Cet algorithme est basé sur une marche aléatoire, où en plus des sauts simples aux nœuds voisins qui peuvent être de n'importe quel type, un saut à deux bonds entre nœuds de même type est permis. Les nœuds de même type peuvent être connectés par un chemin de taille 2 où le nœud intermédiaire est de type différent. Les étiquettes se propagent entre les nœuds de même type. Les chemins joignant des nœuds de même type et de longueur supérieure à 2 sont ignorés. Comparé à cette approche, notre modèle est capable de prendre en compte les corrélations entre les étiquettes de nœuds connectés de type différent, ce qui n'est pas le cas dans leur modèle de marche aléatoire. Notre modèle n'est pas limité aux sauts à deux bonds mais peut prendre en compte les corrélations entre les étiquettes de nœuds distants. Dans la communauté de l'apprentissage, certains modèles basés sur la régularisation dans un espace de représentation latent ont été developpés – (Weston *et al.,* 2008) par exemple – mais n'abordent pas des tâches de classification de nœuds dans les graphes comme celle présentée ici.

6. Conclusion

Nous avons proposé un nouveau modèle capable d'étiqueter les nœuds dans des réseaux hétérogènes où les nœuds sont de types différents, chaque type correspondant à un ensemble particulier de catégories possibles. A l'opposé des modèles existants, notre modèle est capable d'utiliser les corrélations entre les étiquettes des nœuds connectés de types différents, et donc d'utiliser la structure complexe du graphe pour bien étiqueter. Notre algorithme est basé sur l'idée de calculer une représentation des nœuds dans un espace latent commun à tous les types de nœuds, et de supposer que deux nœuds connectés tendront à avoir une représentation proche. Les étiquettes sont ensuite déduites de ces représentations. Nous avons fait des expériences sur deux corpus montrant la capacité de notre modèle à dépasser les approches classiques et nous avons présenté une analyse qualitative montrant que le modèle présenté dans ce papier capture correctement les interdépendances entre les étiquettes des différents types de nœud.

Remerciements

Ce travail a été partiellement financé par le projet DIFAC (FUI 12).

7. Bibliographie

Abernethy J., Chapelle O., Castillo C., « WITCH : A New Approach to Web Spam Detection », *In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb*, 2008.

Angelova R., Kasneci G., Weikum G., « Graffiti : graph-based classification in heterogeneous networks », *World Wide Web*, vol. 15, n° 2, p. 139-170, 2012.

Belkin M., Niyogi P., Sindhwani V., « Manifold Regularization : A Geometric Framework for Learning from Labeled and Unlabeled Examples », *J. Mach. Learn. Res.*, vol. 7, p. 2399-2434, December, 2006.

Davis D., Lichtenwalter R., Chawla N., « Multi-Relational Link Prediction in Heterogeneous Information Networks », *ASONAM*, 2011.

Denoyer L., Gallinari P., « A Ranking Based Model for Automatic Image Annotation in a Social Network », *ICWSM*, 2010.

Hwang T., Kuang R., « A heterogeneous label propagation algorithm for disease gene discovery », *SDM*, p. 12, 2010.

Ji M., Sun Y., Danilevsky M., Han J., Gao J., « Graph regularized transductive classification on heterogeneous information networks », *ECML PKDD*, vol. 0053, Springer, p. 570-586, 2010.

Kou Z., « Stacked graphical models for efficient inference in markov random fields », *In Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.

Liu L., Tang J., Han J., Jiang M., « Mining topic-level influence in heterogeneous networks », *CIKM*, 2010.

Macskassy S. A., Provost F., « A Simple Relational Classifier », *Proceedings of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-2003*, p. 64-76, 2003.

Maes F., Peters S., Denoyer L., Gallinari P., « Simulated Iterative Classification A New Learning Procedure for Graph Labeling », *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases Part II*, vol. II, p. 47-62, 2009.

Sen P., Namata G., Bilgic M., Getoor L., Gallagher B., Eliassi-Rad T., « Collective Classification in Network Data », *AI Magazine*, vol. 29, n° 3, p. 93-106, 2008.

Sun Y., Yu Y., Han J., « Ranking-based clustering of heterogeneous information networks with star network schema », *KDD*, p. 797-806, 2009.

Wang C., Raina R., Fong D., Zhou D., Han J., Badros G., « Learning Relevance from Heterogeneous Social Network and Its Application in Online Targeting », *Knowledge Creation Diffusion Utilization*, 2011.

Weston J., Ratle F., Collobert R., « Deep learning via semi-supervised embedding », *ICML*, p. 1168-1175, 2008.

Zhou D., Bousquet O., Lal T. N., Weston J., Schölkopf B., « Learning with Local and Global Consistency », *in* , S. Thrun, , L. Saul, , B. Schölkopf (eds), *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA, 2004.

Zhou D., Huang J., Schölkopf B., « Learning from labeled and unlabeled data on a directed graph », *Proceedings of the 22nd international conference on Machine learning*, ICML '05, ACM, New York, NY, USA, p. 1036-1043, 2005.