Une approche probabiliste pour l'identification de structures de communautés

Nacim Fateh Chikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles
Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex
{chikhi,rothenburger,aussenac}@irit.fr

Résumé. Dans cet article, nous valorisons et défendons l'idée que les modèles génératifs sont une approche prometteuse pour l'identification de structures de communautés (ISC). Nous proposons un nouveau modèle probabiliste pour l'identification de structures de communautés qui utilise le lissage afin de pallier le petit nombre de liens entre les nœuds. Notre modèle étant très sensible aux paramètres de lissage, nous proposons également une méthode basée sur la modularité pour leur estimation. Les résultats expérimentaux obtenus sur trois jeux de données montrent que notre modèle SPCE est largement meilleur que le modèle PHITS

1. Introduction

L'analyse de réseaux sociaux essaie entre autre d'extraire leur structuration en communautés distinctes (structure de communautés). Intuitivement, une communauté est un ensemble d'acteurs ayant plus de liens vers des acteurs dans cette communauté qu'avec des acteurs d'autres communautés. Nous pensons que les modèles génératifs, bien que peu utilisés dans ce cadre, ont plusieurs avantages pour l'identification automatique des structures de communautés (ISC). En premier lieu, ils nous permettent de prendre en compte le recouvrement entre communautés. En second lieu, ils peuvent analyser des graphes non-orientés mais aussi des graphes orientés. Enfin, ils sont basés sur l'existence d'une distribution de probabilités sous-jacente permettant d'expliquer les données observées.

De nombreuses techniques d'ISC ont été publiées. Les plus connues sont les méthodes de partitionnement de graphes (Shi et Malik, 1997), les approches basées sur la marche aléatoire (Pons et Latapy, 2006) ou sur l'optimisation de la modularité (Clauset et al., 2004). Pour un état de l'art plus exhaustif, on pourra se référer à Fortunato and Castellano (2008).

Pour notre part, nous avons opté pour l'utilisation des modèles génératifs. Cette classe de modèles utilisée pour plusieurs autres tâches d'analyse de données, n'a été que peu utilisée pour l'ISC. Une des rares exceptions est le modèle PHITS (Probabilistic HITS) proposé par Cohn et Chang (2000) pour l'analyse des citations ou des liens hypertextes entre documents. PHITS est un modèle proche de PLSA (Hofmann, 1999) qui a été proposé pour l'analyse de cooccurrences. L'idée de base de PHITS est que les liens dans un graphe peuvent être expliqués par un nombre restreint de variables cachées qui correspondent à la notion intuitive de communautés.

En s'inspirant du modèle PHITS, nous proposons le modèle SPCE (Smoothed Probabilistic Community Explorer) qui met en oeuvre une technique de lissage afin de surmonter la faible densité des graphes. Mais, le comportement de SPCE étant très dépendant des paramètres de lissage, nous avons également proposé une méthode permettant l'estimation de ces hyperparamètres.

Dans la section 2, nous présentons notre modèle pour l'identification de structures de communautés. Les résultats de l'évaluation sont présentés puis discutés dans la section 3. La section 4 permettra de conclure cet article.

2. Modèle probabiliste avec lissage pour l'identification de structures de communautés

Lors de la mise en œuvre de PHITS, nous avons constaté que les résultats obtenus étaient décevants. Nous pensons pouvoir affirmer que ce défaut est dû à l'utilisation de l'estimation du maximum de vraisemblance (EMV) pour l'apprentissage des paramètres. Dans cette section, nous décrivons le modèle génératif SPCE que nous proposons pour pallier ce défaut.

2.1. Processus génératif

Soit un graphe orienté G=(N,L) où N désigne l'ensemble des nœuds et L l'ensemble des liens (ou des arcs). Nous notons par N_N=|N| le nombre de nœuds et par N_L=|L| le nombre de liens. Soient S l'ensemble des nœuds source i.e. des sommets ayant au moins un lien sortant et D l'ensemble des nœuds destination i.e. des sommets ayant au moins un lien entrant. Notons également par N_C le nombre de communautés.

Le processus génératif des liens à partir des noeuds de S vers des noeuds de D est le suivant:

pour i = 1 à N_L

(1) sélectionner un noeud $s_i \sim$ Mult(1, P(S))

(2) choisir une communauté $c_i \sim$ Mult(1, P(C|S=s_i))

(3) générer un lien de s_i vers d_i où $d_i \sim$ Mult(1, P(D|C=c_i))

où Mult(n,p) est une distribution multinomiale de paramètres n et p.

La Figure 1 décrit le modèle graphique correspondant à cette procédure. La variable c dans ce modèle est une variable cachée. On trouvera plus de détails sur ce type de représentation dans (Bishop, 2007).

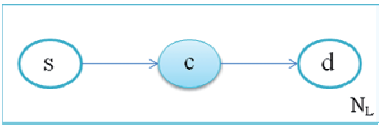

FIG. 1 – *Modèle graphique de SPCE*

2.2. Apprentissage des paramètres

L'inconvénient majeur de PHITS est l'utilisation de l'EMV pour l'apprentissage des paramètres. L'EMV donne effectivement des résultats fiables lorsque la taille de l'échantillon est grande par rapport au nombre de paramètres (Agresti, 2007). Dans notre cas, cette condition etant difficilement vérifiée nous remplaçons l'EMV par l'estimateur du Maximum a posteriori (MAP). L'idée est d'utiliser des distributions a priori sur les paramètres afin d'obtenir un effet de lissage. Nous avons utilisé des a priori de Dirichlet sur les paramètres

car ce sont des conjugués de la distribution multinomiale permettant de simplifier grandement la procédure d'estimation des paramètres. Les étapes EM de l'algorithme SPCE sont les suivantes :

$$E-step: \qquad P(c_i = k \mid s_i, d_i, \gamma', \phi') = \frac{\gamma'_{k s_i} \phi'_{d_i k}}{\sum_{t=1}^{N_C} \gamma'_{t s_i} \phi'_{d_i t}}$$

$$M-step: \qquad \gamma_{k s_i} = \frac{\alpha - 1 + \sum_{j=1}^{N_L} A_{s_i d_j} P(c = k \mid s_i, d_j, \gamma', \phi')}{N_C(\alpha - 1) + \sum_{t=1}^{N_C} \sum_{j=1}^{N_L} A_{s_i d_j} P(c = t \mid s_i, d_j, \gamma', \phi')}$$

$$\phi_{d_j k} = \frac{\beta - 1 + \sum_{i=1}^{N_L} A_{s_i d_j} P(c = k \mid s_i, d_j, \gamma', \phi')}{N_N(\beta - 1) + \sum_{i=1}^{N_L} A_{s_i d_i} P(c = k \mid s_i, d_i, \gamma', \phi')}$$

Où A est la matrice d'adjacence du graphe, $\gamma_{k s_i} = P(C = k \mid S = s_i)$, $\phi_{d_j k} = P(D = d_j \mid C = k)$, γ' et ϕ' sont les estimations actuelles des paramètres, γ et ϕ sont les nouvelles estimations des paramètres.

Dans l'étape M, α et β jouent le rôle de pseudo-comptes permettant à SPCE de prendre en compte la faible densité du graphe. Ainsi, lorsque $\alpha = \beta = 1$, SPCE est équivalent à PHITS. Ceci est évident car dans un pareil cas, les a priori de Dirichlet sont uniformes et les distributions a posteriori ne dépendent donc que de la vraisemblance. Afin d'éviter des valeurs de probabilité incohérentes, nous imposons $\alpha \geq 1$ et $\beta \geq 1$. En outre, dans le cas où $\alpha = \beta = 2$, cela revient au lissage de Laplace (ajout de 1) utilisé en recherche d'information.

2.3. Estimation des paramètres de lissage

Nous constatons que les paramètres de lissage (α et β) influent considérablement sur le comportement de SPCE. Nous proposons donc une méthode pour l'estimation de ces hyperparamètres basée sur la modularité. La modularité est une fonction de qualité récemment proposée par Newman (2006). Elle est définie formellement par : $Q = \sum_{i,j \in V} \left[\frac{A_{ij}}{2m} - \frac{d_i d_j}{4m^2} \right] \delta(c_i, c_j)$ où A est la matrice d'adjacence, d_i est le degré du noeud i, m est le nombre total de liens, et δ est la fonction delta de Kronecker.

Notons que cette définition concerne les graphes non-orientés et qu'elle a été adaptée aux graphes orientés par (Leicht et Newman, 2008).

Les valeurs possibles pour la modularité sont dans l'intervalle]-1, 1[. En cas d'absence de structure de communautés, la valeur de la modularité est négative ou nulle, alors qu'une valeur supérieure à 0,3 indique la présence d'une structure de communautés (Newman, 2006). Dans notre cas, la modularité est le critère de performance pour le modèle SPCE. Nous considérons donc l'estimation des hyperparamètres comme une tâche de recherche des valeurs de α et de β qui permettent d'obtenir la meilleure modularité.

Graphe	*Nombre de noeuds*	*Nombre de communautés*	*Nombre de liens*	*Nœuds destination*	*Nœuds source*	*orienté*
Cora	2708	7	5429	1565	2222	oui
Citeseer	2994	5	4277	1760	2099	oui
Wikipedia	5360	7	41978	5360	5360	non

TAB. 1 – *Données d'évaluation*

3. Evaluation et discussion

3.1. Données

La modularité est la mesure d'évaluation la plus répandue pour l'ISC. Néanmoins, il nous apparaît que s'agissant d'une mesure interne, il n'est pas envisageable de l'utiliser seule pour comparer des algorithmes d'ISC. Par conséquent, nous utilisons trois ensembles de données de références pour lesquels la structure de communautés est connue. Les caractéristiques de ces trois réseaux sont résumées dans le Tableau 1.

3.2. Evaluation

D'abord, nous évaluons la qualité des structures de communautés identifiées par SPCE et PHITS. Les structures de communautés identifiées par chacune des deux méthodes sont comparées avec la classification de référence. Trois mesures sont mises en œuvre pour cette évaluation : l'information mutuelle normalisée (NMI), la F-mesure, et la modularité.

Le Tableau 2 indique la moyenne de dix exécutions de PHITS et de SPCE sur les jeux de données de référence (le nombre de communautés est celui de la classification de référence). Nous avons ajouté les résultats obtenus par l'algorithme K-moyennes (KM), qui sert de base de comparaison pour les algorithmes d'identification de communautés.

Le Tableau 2 montre que les performances de SPCE sont nettement meilleures que celles de PHITS pour chaque ensemble d'évaluation. En particulier pour les réseaux de Cora et de Citeseer, elle atteint jusqu'à 100% en termes de NMI et de modularité. Les résultats de PHITS sont particulièrement faibles pour Citeseer et Cora, mais ils sont plus satisfaisants pour Wikipedia. En ce qui concerne la modularité, SPCE est toujours supérieur à KM. Néanmoins, ce dernier affiche des résultats légèrement meilleurs que SPCE en termes de NMI et de F-mesure avec les versions initiale et non-orientée de Citeseer.

3.3. Discussion

Les faibles performances que nous avons constatées pour PHITS peuvent surprendre. Lors de nos premières expérimentations, nous avons douté de notre implémentation de PHITS. Afin d'écarter cette hypothèse, nous avons effectué des essais avec un modèle équivalent à PLSA (i.e. à PHITS) : la factorisation en matrice non-négative (NMF) (Ding et al., 2008, Gaussier et Goutte, 2005, Lee et Seung, 2000). Mais là encore, NMF a donné des résultats aussi décevants que PHITS.

	NMI			F-mesure			Modularité		
Graphe	**KM**	**PHITS**	**SPCE**	**KM**	**PHITS**	**SPCE**	**KM**	**PHITS**	**SPCE**
Cora (O)	0.21	0.04	0.25	0.42	0.24	0.46	0.15	0.13	0.26
Cora (T)	0.27	0.07	0.32	0.47	0.29	0.51	0.28	0.20	0.39
Cora (N)	0.31	0.09	0.31	0.49	0.32	0.50	0.52	0.42	0.72
Citeseer (O)	0.12	0.01	0.12	0.41	0.26	0.37	0.10	0.13	0.25
Citeseer (T)	0.12	0.03	0.18	0.38	0.29	0.41	0.14	0.16	0.33
Citeseer (N)	0.18	0.02	0.16	0.43	0.28	0.41	0.38	0.34	0.71
Wikipedia (N)	0.54	0.52	0.63	0.67	0.68	0.77	0.60	0.62	0.67

TAB. 2 – *Résultats de l'évaluation (O: Original, T: Transposé, N: Non-orienté)*

Bien que les résultats obtenus avec l'algorithme des K-moyennes soient proches de ceux obtenus avec SPCE tant pour la NMI que pour la F-mesure, SPCE dépasse largement KM en ce qui concerne la modularité (c.f. Tableau 2). Il s'agit d'un constat intéressant qui a deux conséquences. Cela montre que d'une part, KM et SPCE identifient des structures de communautés différentes, et que d'autre part on ne peut se contenter de la modularité pour comparer des techniques d'identification de structures de communautés.

4. Conclusion

Dans cet article, nous avons présenté SPCE, un modèle génératif pour l'analyse de structures de communautés. L'idée de base de SPCE est d'utiliser le lissage afin d'affronter efficacement la faible densité de certains graphes. Ensuite, nous avons montré que les paramètres de lissage de notre modèle peuvent être estimés à l'aide du critère de modularité. Les résultats expérimentaux que nous avons obtenus à partir de SPCE améliorent sensiblement ceux obtenus avec PHITS.

Nous pensons que SPCE est une solution pertinente pour l'identification de structure de communautés pour différentes raisons :

- il permet d'identifier des structures de communauté cohérentes.
- il identifie des communautés qui peuvent se recouvrir.
- il permet de détecter les communautés aussi bien dans les graphes orientés que non-orientés.
- il permet enfin deux points de vue dans les graphes orientés.

Nous envisageons de poursuivre ce travail en appliquant SPCE sur des graphes que nous générerons par simulation. Le but est de comprendre le comportement du modèle avec des graphes ayant des propriétés particulières afin d'identifier les situations pour lesquelles le lissage est important. Nous pourrons aussi tester le passage à l'échelle en utilisant notre modèle avec de très grands graphes.

Une dernière perspective importante est de comparer SPCE non seulement à PHITS et l'algorithme des K-moyennes, comme nous l'avons fait dans cet article, mais de le comparer aussi avec d'autres approches pour l'identification de structures de communautés.

Références

Agresti, A. (2007). An Introduction to Categorical Data Analysis, 2nd Edition, Wiley: New York.

Bishop, C. M. (2007). Pattern Recognition and Machine Learning. Springer.

Clauset, A., M. E. J. Newman, et C. Moore (2004). Finding community structure in very large networks. Physical Review E, Vol. 70, 066111.

Cohn, D. et H. Chang (2000). Learning to probabilistically identify authoritative documents. In Proc. of the 17th ICML.

Ding, C., T. Li, et W. Peng (2008). On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. Comput. Stat. Data Anal. 52(8): 3913-3927.

Fortunato S. et C. Castellano (2008). Community Structure in Graphs. Encyclopedia of Complexity and System Science. Springer.

Gaussier, E. et C. Goutte (2005). Relation between PLSA and NMF and implications. In Proc. of the 28th annual intl. ACM SIGIR conf., Brazil, ACM.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In Proc. of the 15th UAI Conference.

Lee, D. et H. Seung (2000). Algorithms for non-negative matrix factorization. In Proc. of NIPS, pages 556–562.

Leicht, E. A. et M. E. J. Newman (2008). Community structure in directed networks. Physical Review Letter, 100:118703.

Newman, M. E. J. (2006). Modularity and community structure in networks. PNAS, USA, 103:8577.

Pons, P. et M. Latapy (2006). Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications, 10(2) :191–218.

Shi, J. et J. Malik (1997). Normalized Cuts and Image Segmentation. In Proc. of CVPR '97, IEEE Computer Society.

Summary

A large variety of techniques has been developed for community structure identification (CSI) including modularity optimization, graph partitioning, and hierarchical clustering. In this paper, we argue that generative models are a promising approach for community structure identification, although these models have received very little attention from CSI researchers. Following the work of Cohn and Chang on link analysis, we propose a new probabilistic model for community structure detection. The originality of our model is the use of smoothing in order to overcome the sparsity of network data. A method based on the modularity criterion is also proposed for the estimation of smoothing parameters.

Experiments carried out on three real datasets show that our new model SPCE (Smoothed Probabilistic Community Explorer) significantly outperforms PHITS (Probabilistic HITS).