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.