Impact de l'échantillonnage sur l'inférence de structure dans les réseaux

Timothée Tabouy [1]

[1] : Mathématiques et Informatique Appliquées (MIA) - Site web
AgroParisTech, Institut national de la recherche agronomique (INRA) : UMR0518
INRA Domaine de Vilvert F-78352 Jouy-en-Josas Cedex - France

Le modèle probabiliste le plus simple sur les graphes est le modèle d'Erd\H{o}s-Reyni (\cite{Erdos1959}). Afin d'aller plus loin et de modéliser une hétérogénéité de la topologie dans le réseau, des modèles probabilistes à espace d'états latents ont été proposés. Parmi la collection de modèles proposés, les Stochastic Block Models (SBM) sont sans doute les modèles les plus utilisés (\cite{Nowicki2001}). L'inférence de ces modèles est complexe et des algorithmes type EM (Expectation Maximization) ont été proposés pour l'estimation des paramètres et le recouvrement de la structure en blocs latents (\cite{Robin2008}). L'obtention de données de type réseau se fait souvent à partir d'un échantillonnage centré sur les noeuds. De plus, l'échantillonnage est souvent partiel et à un impact sur l'inférence des réseaux. En général, on néglige ce fait et on considère le réseau parfaitement observé. \\

Le but de ce travail est de prendre en compte dans l'inférence la stratégie d'échantillonnage utilisée pour échantillonner un graphe afin de corriger les biais d'estimation. La théorie des données manquantes développée par D. Rubin dans \cite{Rubin1976} nous a permis de classer certaines stratégies d'échantillonnages adaptées aux réseaux dans plusieurs catégories dans lesquelles la prise en compte de la stratégie dans l'inférence varie. Les stratégies se regroupent essentiellement en 2 grandes classes, celles Missing At Random (MAR), quand la probabilité d'être échantillonné est indépendante de la valeur des données manquantes, et celles qui ne le sont pas. Pour les stratégies MAR, la stratégie d'échantillonnage ne perturbe pas l'inférence, en particulier l'inférence est conduite uniquement sur la partie observée du graphe. Les stratégies non MAR quant à elles nécessitent la prise en compte dans l'inférence de la stratégie d'échantillonnage employée pour récolter les données.\\
Nous avons adapté les algorithmes EM dans leur forme variationnelle pour l'inférence des paramètres du SBM binaire de toutes les stratégies MAR, ainsi que pour la stratégie \og 2 poids 2 mesures \fg, c'est-à-dire quand la probabilité pour une arête d'être échantillonnée dépend de sa valeur.