Évaluer la prévision de liens dans les graphes avec les distances de centralité

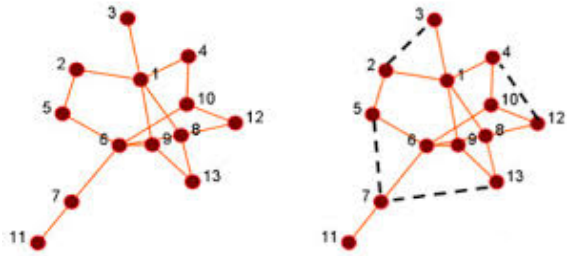
Lionel Tabourier (LIP6 – CNRS / UPMC – Paris)

Gilles Tredan (LAAS – CNRS – Toulouse)

stages@complexnetworks.fr, http://complexnetworks.fr

Qu'il s'agisse de relations sociales, de transactions commerciales ou encore de machines échangeant des paquets d'information, la structure des réseaux complexes auto-organisés évolue au cours du temps. Les mécanismes microscopiques qui font que des liens se créent et que d'autres disparaissent au cours du temps sont souvent mal identifiés. Par conséquent, comprendre ce qui fait que telle interaction a plus de chance de se produire que telle autre est un enjeu majeur pour comprendre la dynamique de tels systèmes.

Une formulation classique du **problème de prévision de liens** est la suivante : étant donné un réseau à un instant t donné, comment prévoir quels liens vont apparaître au pas de temps suivant [LNK07] ? En apprentissage automatique, ce problème est souvent interprêté comme une question de classification à deux classes. En effet, une paire de nœuds appartient soit à l'ensemble des paires qui vont être liées à $t+1$, soit non. On recherche alors dans la structure du graphe à t des informations corrélées à la probabilité d'apparition d'un lien à l'instant $t+1$ (e.g. [AHCSZ06]). Par exemple deux nœuds qui ont un grand nombre de voisins en commun ont une probabilité plus élevée de se lier l'un à l'autre.



Exemple de prévision de liens sur un graphe.

Dans la littérature, la qualité de la prévision est évaluée en général selon le résultat de la tâche de classification. On considère en effet qu'une prévision est correcte si le lien prévu apparaît (vrai positif), mauvaise s'il n'apparaît pas (faux positif). On quantifie également les liens qui apparaissent mais ne sont pas prévus (faux négatifs), et les liens non-prévus qui n'apparaissent pas (vrais négatifs).

Une manière simple et naturelle de résumer par une valeur numérique la qualité de la prévision est de mesurer la distance d'édition du graphe (*graph edit distance*, ou GED) : on

compte le nombre de liens qui diffèrent entre le graphe attendu et le graphe observé. En pratique, cela revient à faire la somme des faux positifs et des faux négatifs de la prévision. Cette mesure semble judicieuse dans de nombreux contextes, cependant elle souffre de plusieurs défauts. D'abord, elle ne fait pas de distinction entre les catégories de mauvaises prévisions. Faux positifs et faux négatifs sont mis au même niveau alors que selon les applications, un type d'erreur peut être beaucoup plus critique que l'autre. De plus, cette mesure ne fait pas de différence entre les liens prévus alors que connecter deux nœuds très éloignés dans le graphe a beaucoup plus d'impact sur son fonctionnement que connecter deux nœuds qui font déjà partie de la même communauté.

Dans un travail récent, Pignolet et al. [PRST16] proposent de mesurer l'évolution dynamique du graphe à l'aide d'autres distances que la GED. Ces nouvelles distances sont basées sur la notion de **centralité** dans les graphes. Il existe plusieurs définitions au concept de centralité, mais toutes visent à évaluer "l'importance" d'un nœud dans la structure : son nombre de voisin, sa proximité avec les autres nœuds du réseau etc. En mesurant la distance de centralité entre le graphe obtenu et le graphe prévu, on donne une autre mesure de la réussite de la prévision, complémentaire de la GED.

Au cours de ce stage, nous proposons de revisiter l'évaluation de la qualité des prévisions à l'aide des distances de centralité. Il s'agit en particulier de répondre à la question : comment peut-on modifier les techniques d'apprentissage de manière à minimiser les distances de centralité pertinentes, tout en conservant une GED faible ? Le stage s'adresse à des étudiants issus de formations variées (réseaux complexes, fouille de données, apprentissage ...). Les principales compétences recherchées sont une certaine connaissance de l'algorithmique de graphes, la capacité d'adaptation à des méthodes nouvelles et de l'intérêt pour les thématiques interdisciplinaires. Du point de vue des langages et outils utilisés, le/la stagiaire bénéficiera d'une grande liberté. Le stage pourra être effectué au LIP6 (Paris) ou au LAAS (Toulouse).

Références

[AHCSZ06] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *SDM06 : workshop on link analysis, counter-terrorism and security*, 2006.

[LNK07] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7) :1019–1031, 2007.

[PRST16] Yvonne Anne Pignolet, Matthieu Roy, Stefan Schmid, and Gilles Tredan. The many faces of graph dynamics. *arXiv preprint arXiv :1608.01911*, 2016.