SUJET DE THÈSE

Contexte. L'abondance de données générées par les diverses activités numérisées constitue un potentiel de connaissances largement sous utilisé. Par exemple, au niveau des entreprises, le suivi informatique des expéditions, ventes, fournisseurs, clients, l'observation des réseaux sociaux, peut être exploité pour concevoir de nouveaux produits et services, ou encore prendre des décisions. De nombreuses données sont maintenant largement disponibles dans les entreprises, mais aussi via des portails « open data » (`https://www.data.gouv.fr/fr/`) et les médias sociaux (Twitter, Instagram, Foursquare, Googleplace, etc.). Ces différentes données permettent d'accéder en temps-réel aux comportements d'un grand nombre d'individus. La collecte et l'analyse de leurs contenus peuvent alors offrir une vision incomparable de comportements à grande échelle et de leurs dynamiques.

Ces données ont la particularité d'être à la fois hétérogènes et structurées pouvant être représentées sous la forme d'un réseau d'information, c'est-à-dire un multi-graphe attribué. Peu d'outils d'analyse de ce type de données existent à ce jour, alors que les besoins sont très nombreux. L'objectif de cette thèse est de définir des outils génériques pour analyser ce type de données, par la fouille de sous-graphes et la découverte de communautés. L'équipe DM2L du LIRIS a produit de nombreux résultats sur l'analyse de graphes dynamiques attribués ces dernières années [Rob09, DPRB13, PPRB13, OLPB15, BPR16, KPZ$^+$17]. Ces développements se traduisent par une très bonne connaissance de l'état de l'art mais aussi par la maintenance de prototypes combinant de nombreux algorithmes [HZK$^+$15, KPPR14, RSPF13].

Points durs et verrous à lever. Ce sujet de thèse propose à la fois des travaux fondamentaux et des applications concrètes sur des données réelles de réseaux sociaux et de réseaux d'entreprises.

Les travaux de recherche fondamentaux porteront sur l'élaboration de nouveaux algorithmes de détection de communautés dans des réseaux d'information hétérogènes (en continuation de [CAH10, CTH15] et en explorant des méthodes de marches aléatoires sur des graphes), couplés à des méthodes d'extraction de sous-graphes exceptionnels. De plus, la fouille de données structurées doit être pensée dans un environnement dynamique (évolution temporelle du réseau) et pleinement interactif afin que l'analyste puisse exprimer ses préférences, ses connaissances expertes, et enrichir dynamiquement les modèles construits. Cela passe par la définition de nouveaux modes d'interaction entre l'analyste et les algorithmes d'exploration de données, la prise en compte des connaissances du domaine, et l'apprentissage de préférences implicites des utilisateurs, qui sont autant de verrous actuellement.

Les travaux de recherche appliqués porteront sur les applications à des données réelles de réseaux sociaux et de réseaux d'entreprises à des fins de recommandations de nouvelles

relations ou de chemins dans le réseau, à l'identification des acteurs les plus influençant ou précurseur du réseau, à la diffusion d'information sur le réseau, ou encore l'analyse de son évolution temporelle.

Collaboration scientifique. La thèse se déroulera en collaboration entre l'équipe Data Mining & Machine Learning (DM2L) du LIRIS et le Data R&D Institute de EMLyon Business School. Le candidat partagera son temps d'étude et de recherche entre ces deux institutions. Il est attendu du candidat d'avoir de solides connaissances en informatiques et une appétence pour les applications business. Les publications visées seront à la fois appliquées : dans des journaux de sciences de gestion (marketing quantitatif par exemple), et fondamentales : dans des journaux et proceedings de conférence en informatique (machine learning, data mining, network science).

Contact.

— Céline Robardet, professeur à l'INSA Lyon, responsable de l'équipe Data Mining & Machine Learning au LIRIS (UMR CNRS 5205), celine.robardet@insa-lyon.fr
— Jean Savinien, professeur associé à EMLyon et maître de conférence à l'IECL (UMR CNRS 7502), responsable recherche du Data R&D Institute, savinien@em-lyon.com

Références

[BPR16] Ahmed Anes Bendimerad, Marc Plantevit, and Céline Robardet. Unsupervised exceptional attributed sub-graph mining in urban data. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 21–30, 2016.

[CAH10] Rémy Cazabet, Frédéric Amblard, and Chihab Hanachi. Detection of overlapping communities in dynamical social networks. In Ahmed K. Elmagarmid and Divyakant Agrawal, editors, *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SocialCom / IEEE International Conference on Privacy, Security, Risk and Trust, PASSAT 2010, Minneapolis, Minnesota, USA, August 20-22, 2010*, pages 309–314. IEEE Computer Society, 2010.

[CTH15] Rémy Cazabet, Hideaki Takeda, and Masahiro Hamasaki. Characterizing the nature of interactions for cooperative creation in online social networks. *Social Netw. Analys. Mining*, 5(1) :43 :1–43 :17, 2015.

[DPRB13] Elise Desmier, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Trend mining in dynamic attributed graphs. In *Machine Learning and Knowledge Discovery in Databases*, pages 654–669. Springer, 2013.

[HZK+15] Pierre Houdyer, Albrecht Zimmerman, Mehdi Kaytoue, Marc Plantevit, Joseph Mitchell, and Céline Robardet. Gazouille : Detecting and illustrating local events from geolocalized social media streams. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III*, pages 276–280, 2015.

[KPPR14] Mehdi Kaytoue, Yoann Pitarch, Marc Plantevit, and Céline Robardet. Triggering patterns of topology changes in dynamic graphs. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, Beijing, China, August 17-20, 2014*, pages 158–165, 2014.

[KPZ+17] Mehdi Kaytoue, Marc Plantevit, Albrecht Zimmermann, Anes Bendimerad, and Céline Robardet. Exceptional contextual subgraph mining. *Machine Learning*, pages 1–41, 2017.

[OLPB15] Günce Keziban Orman, Vincent Labatut, Marc Plantevit, and Jean-François Boulicaut. Interpreting communities based on the evolution of a dynamic attributed network. *Social Netw. Analys. Mining*, 5(1) :20 :1–20 :22, 2015.

[PPRB13] Adriana Prado, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Mining graph topological patterns : Finding co-variations among vertex descriptors. *IEEE Trans. Knowl. Data Eng.*, 25(9) :2090–2104, 2013.

[Rob09] Céline Robardet. Constraint-based pattern mining in dynamic graphs. In *ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009*, pages 950–955, 2009.

[RSPF13] Céline Robardet, Vasile-Marian Scuturici, Marc Plantevit, and Antoine Fraboulet. When TEDDY meets grizzly : temporal dependency discovery for triggering road deicing operations. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 1490–1493, 2013.