Analyse topologique et requêtes interactives dans des grands graphes sémantiques

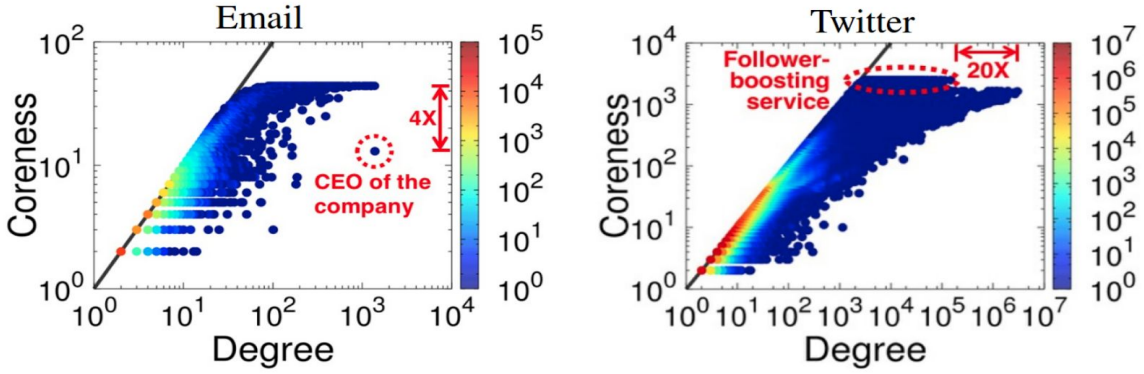
Equipes participantes

- Complex Networks (CN) : Maximilien Danisch
- Bases de Données (BD) : Amine Baazizi, Hubert Naacke

Contexte et Problème

Le projet étudie l'interrogation (ou requêtage) des données du Linked Open Data (LOD) un graphe sémantique de très grande taille contenant plusieurs milliards d'arcs. Le LOD est appelé **graphe sémantique** car ses arcs ont des labels (ou étiquettes) sémantiques.
Le LOD réunit des données de domaines divers. Par exemple le dataset YAGO[1] [4], inclus dans le LOD, contient les connaissance issues de Wikipedia et Geonames. Sa taille dépasse les 10 milliards d'arcs avec plus d'un million de labels différents.
L'interrogation d'un graphe consiste à poser une requête, appelée *Basic Graph Pattern*. Une requête est un motif constitué d'arcs (chaque arc possédant un label) connectés entre eux. Calculer le résultat d'une requête consiste à retrouver tous les sous-graphes qui correspondent au motif de la requête. Dans ce contexte, interroger le LOD pose deux problèmes que nous détaillons ci-dessous :

Difficulté d'exprimer les requêtes à cause de la diversité sémantique du LOD.
Exprimer des requêtes sensées est difficile à cause de l'abondance d'étiquettes (qui se comptent en centaines de milliers) et à cause de la forte irrégularité du graphe (toutes les étiquettes ne sont pas équi-fréquentes). Afin de palier à ce manque, certaines approches proposent de construire un synopsis au préalable dans le but de révéler les informations structurelles intéressantes et qui pourraient donc potentiellement faciliter la formulation des requêtes. La construction et la maintenance de ce type de structure engendrent un important coût qu'il faudra supporter. Il s'avère que les graphes réels (appelés par la suite graphes de terrain) jouissent de certaines propriétés topologiques qui sont relativement faciles à calculer et qui, surtout, peuvent révéler certaines informations suffisantes pour explorer les données au moyen des requêtes de graphe. Par exemple, la valeur de core d'un noeud (qui indique le nombre de voisins importants) peut être combinée avec le degré pour identifier des noeuds intéressants par rapport aux autres noeuds du graphe. Les figures suivantes



extraites de [3], montrent (à gauche) un nœud avec un degré élevé par rapport à sa valeur de core. Ce noeud qui ressort d'un réseau d'échange de mail au sein d'une société, est en fait son PDG et pourrait servir comme point de départ pour une requête qui voudraient retourner la couche des décideurs d'une entreprise. Une telle approche éviterait à l'utilisateur d'effectuer différentes itérations en terme de requêtes avant de localiser le point central qui est le PDG. Dans un autre exemple (cf figure de droite), un groupe d'utilisateurs avec une valeur de core élevée par rapport à leur degré sur Twitter est en fait un groupe d'utilisateurs robots que l'on peut payer pour augmenter son réseau de followers. Là encore, un utilisateur désireux de connaître une information pourrait utiliser la sémantique véhiculée par le graphe de terrain pour entamer la formulation de

[1] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

sa requête. Des travaux menés au sein de l'équipe Complexe Networks consistent à affiner cette notion de valeur de core [1].
Pour conclure, faciliter l'interrogation (ou le requêtage) du LOD nécessite une étape d'**analyse topologique** pour extraire certains motifs caractéristiques au graphe.

Interrogation et indexation efficace du LOD
Calculer le résultat d'une requête est d'autant plus complexe dans le LOD que les requêtes sont diverses. Les motifs potentiellement pertinents à retrouver ont des formes très variées et sont très nombreux. Or l'analyse topologique du LOD produit l'ensemble des motifs identifiés comme étant les plus caractéristiques, mais cet ensemble est potentiellement très grand (très supérieur à la taille même du LOD). Cela soulève de problème d'exploiter le résultat de l'analyse topologique pour évaluer une requête. Cela nécessite d'indexer les motifs pour les retrouver efficacement et de déterminer quels sont les motifs utiles pour une requête.

Objectifs

- Le premier objectif de ce projet est d'identifier les propriétés topologiques particulières existant dans le LOD puis d'utiliser ces propriétés pour concevoir un algorithme capable d'énumérer et quantifier certains motifs dont l'occurrence dépend fortement des propriétés topologiques du graphe. Le défi consiste ici à tenir compte des *labels* dans la caractérisation des particularités et dans l'énumération des motifs. Pour cela, nous nous appuyons sur les travaux menés au sein de l'équipe CN, actuellement en cours de révision, montrant qu'il est possible de lister efficacement les k-cliques et les k-motifs dans de tels graphes appelés graphes de terrain en exploitant une propriétée appelée valeur de core. Cette valeur de core s'avère très faible dans la plupart des graphes de terrains, bien qu'en général la valeur de core soit très élevée. En effet, il a été démontré que si la valeur de core d'un graphe est faible alors il est possible de lister efficacement toutes les cliques maximales d'un graphe de terrain ayant plusieurs milliards de liens même si ce problème est NP-hard [2] dans le cas général;

- Le deuxième objectif est de proposer une méthode pour exploiter le résultat de l'analyse topologique afin de faciliter la formulation de requêtes de motifs. Le défi consiste à construire à partir du résultat de l'analyse topologique une vue agrégée du graphe décrivant de manière quantitative les particularités topologiques. Ceci doit apporter à l'utilisateur une information plus riche que les statistiques habituelles (ex fréquence des labels ou des paires de labels). Ceci doit guider l'utilisateur dans le choix des labels (parfois rares) qui composeront ses requêtes.

- Le troisième objectif est d'indexer plus efficacement les données en bénéficiant de la connaissance acquise sur la structure du graphe sous jacent. En rupture avec les approches récentes [8] communément employées qui indexent les structures simples du graphe telles que les étoiles ou les chemin de longueur 2, nous proposons d'indexer les structures topologiques particulières et complexes produites par la phase d'analyse topologique (cf objectif 1). Le défi est de faire face au volume souvent très important des structures à indexer. Par exemple indexer les cliques n'est pas trivial car leur nombre est souvent beaucoup plus grand que le nombre d'arcs. En nous appuyant sur les travaux de dénombrement de structure semi-régulières [7], nous proposerons une organisation hiérarchique des structures particulières. Nous étudierons le gain en performance que cela apporte pour le traitement des requêtes des utilisateurs.

Résultats attendus

D'une part, dans le domaine de l'analyse de graphes : caractérisation de nouvelles propriétés des graphes de terrains étiquetés, et nouveaux algorithmes de fouille de graphes de terrains basés sur les étiquettes. D'autre part, dans le domaine de la gestion et l'interrogation de données sémantiques : conception de nouveaux algorithmes pour indexer efficacement des bases de données graphe; nouvelles méthodes d'interrogation interactive basée sur l'extraction de motifs complexes.

Implémentation d'algorithmes et réalisation d'un nouvel outil pour interroger YAGO plus efficacement et plus ergonomiquement (guider l'utilisateur à l'aide des motifs extraits et des statistiques associées).

Références

1. Maximilien Danisch, T.-H. Hubert Chan, Mauro Sozio: Large Scale Density-friendly Graph Decomposition via Convex Programming. WWW 2017.
2. Eppstein, D., Strash, D.: Listing all maximal cliques in large sparse real-world graphs. Experimental Algorithms. 2011.
3. Shin, K., Eliassi-Rad, T., Faloutsos, C.: CoreScope: Graph Mining Using k-Core Analysis—Patterns, Anomalies and Algorithms. ICDM 2016.
4. F. Mahdisoltani, J. Biega, F. M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. *Innovative Data Systems Research (CIDR)*. 2015.
5. O. Curé, H. Naacke, M.-A. Baazizi, B. Amann. On the Evaluation of RDF Distribution Algorithms Implemented over Apache Spark. I*nternational Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS)*, pp.16-31, 2015.
6. H. Naacke, B. Amann, O. Curé : SPARQL Graph Pattern Processing with Apache Spark. *Graph Data-management Experiences & Systems, Workshop, SIGMOD,* 2017.
7. M.-A. Baazizi, H. Ben Lahmar, D. Colazzo, G. Ghelli, C. Sartiani : Schema Inference for Massive JSON Datasets. *Extending Database Technology (EDBT)*, 2017.
8. A. Khandelwal, Z. Yang, Evan Ye, R. Agarwal, Ion Stoica. ZipG: A Memory-efficient Graph Store for Interactive Queries. SIGMOD. pp. 1149-1164. 2017