Master Internship position

Title: Schema-based Graph Data Integration

Supervisors: Angela BONIFATI (LIRIS/Lyon,angela.bonifati@liris.cnrs.fr)

Rachid ECHAHED (LIG/Grenoble, rachid.echahed@imag.fr)

Location: LIRIS (Lyon) and/or LIG (Grenoble)

URL : https://perso.liris.cnrs.fr/angela.bonifati/

Keywords: Graph databases, Property Graph Data Integration, Property Graph Schemas, Tool Development

Short Description of the Master (M2) Thesis Project:

Graphs are a flexible and agile data model for representing complex network-structured data used in a wide range of application domains, including social networks, biological networks, bioinformatics, medical data, quantum calculi and knowledge management [2].

Graph database systems are becoming increasingly popular due to their high flexibility. Various graph query languages are being proposed such as Cypher, PGQL, GSQL, and G-CORE, leading to an effort to standardize a graph query language, resulting in two separate ISO/IEC standards: GQL and SQL/PGQ.

The aforementioned languages are based on a particular definition of graphs called "property-graphs". These graphs feature nodes and edges, which can be enriched by means of labels and can also be endowed with finite records. Such graphs provide great flexibility either in data representation as well as in query formulation.

The advent of this new generation of graph-oriented databases, in addition to the newly released standard query languages, give rise to a bunch of new research problems to be solved both in industry and in academia. In this project, we are concerned with graph data integration issues in the context of the ISO standard language GQL and particularly by using typing information based on the recent PG-Schema, a schema language for property graphs from academia and industry [3]. Actually, organizations cannot operate, in today's digital context, without collecting data from various sources. Data integration plays an important role when various data are gathered from different places and transformed into a coherent set of data structures. Hence, property graphs constantly need to be transformed in order to be updated with new information or to be transferred between applications [2]. It is thus desirable to use tools to ensure correctness of such transformations, such as adherence to the typing information or being compatible with the requirements of an application that uses the output of a graph query etc. Such tools and their underlying theory are still lacking.

In this project, we consider the investigation of property graph transformations guided by schema conformance. The targeted framework will feature different kinds of property graph transformations which may occur, for instance, when performing data graph integration processes. In the case where such transformations are not well specified, they may lead to inconsistent graph databases. To avoid such inconveniences, it is well known that verification techniques or type-checking methods can be

used successfully to help ensure the correctness of the considered transformations. We will follow these lines of work in this project and extend schema-less transformation techniques, such as [1], to support typing schemas [3].

The successful candidate should have good programming skills. The 6 months Master project (approx. 600 Euros per month) can be easily adapted to the skills and motivations of the candidate, from theory to practice, and can be extended to a PhD thesis. The candidate will work with top researchers in the area of data management and graph rewriting in the context of a larger grant funded by the French ANR agency.

References

[1] A. Bonifati, F. Murlak, Y. Ramusat: Transforming Property Graphs. Proc. VLDB Endow. 17(11): 2906-2918 (2024).

[2] S. Sakr et al. The future is big graphs: a community view on graph processing systems. Com. ACM 64(9): 62-71 (2021).

[3] R. Angles et al.: PG-Schema: Schemas for Property Graphs. Proc. ACM Manag. Data 1(2): 198:1-198:25 (2023).