Research Internship

Title:Graph Constraints from Theory to Practice

Name and mail of the advisors:
Angela Bonifati, Lyon 1 University & INRIA, France
Wim Martens, Univ. of Bayreuth, Germany

General presentation of the topic: Graph-structured data is highly popular in a wide range of domains, such as social networks, knowledge bases, online retailer customer analysis, and fraud detection. Graphs are applicable to all settings in which interconnected data is used to represent, explore and predict digital-world and real-world phenomena. *Understanding the connections in their data* can be crucial for companies: we have seen over and over again how companies that understood how to leverage the connectivity of their data were successful (e.g. Google with Pagerank, Facebook with their social network graph). Graph-structured data is designed with precisely this goal in mind.
Oftentimes, graph data comes without a structure and in a constraint-less fashion, thus leading to inconsistency and poor quality. According to a recent survey [1], the most recurrent task for users of real-world graphs is data integration. As a matter of fact, graphs naturally lend themselves to information reconciliation and integration. However, the result of integrating large-scale graphs, e.g. in knowledge bases such as DBPedia, Wikipedia or the more recent CovidGraph[1] might turn out to be incorrect and error-prone if not followed by a posteriori verification by means of graph constraints. Constraints are also important pillars for query optimization and metadata management, the latter being unexplored topics in graph data management.
There are several models for representing graphs. Among the most popular is the property graph data model, which is a multigraph with both labelled nodes and edges, along with property value pairs associated to both. It has gained adoption with systems such as AgensGraph, Amazon Neptune, ArangoDB, Blazegraph, CosmosDB, DataStax Enterprise Graph, HANA Graph, JanusGraph], Neo4j, Oracle PGX, OrientDB, Sparksee, Stardog, TigerGraph, Titan, etc. All these systems possess disparate storage models and APIs and some of them support limited forms of graph query languages.
A lingua franca for querying graphs is currently missing and standardization efforts are undergoing to that purpose (such as the ISO/IEC GQL initiative, supported by the community technical working groups of the Linked Data Benchmark Council (LDBC)).

Objective of the internship:
In this internship, we will focus on cardinality constraints for property graphs that are able to encode a wide spectrum of constraints spanning from relational primary keys to cardinality constraints for associations in conceptual modeling. We will study the problem of cardinality constraint validation, by translating the constraints into conjunctive queries with aggregates and into threshold queries, respectively. Implementing and validating such algorithms has tangible applications as most commercial graph databases, such as Neo4j, Oracle PGX, and Microsoft Azure Cosmos DB, are faced with the challenge of processing schema-less and constraint-less graphs.

1 https://covidgraph.org/

We expect that different solutions can be thought of depending on the chosen backend, i.e. relational or graph-based as in [2], for the implementation.
We will design an experimental study devoted to comparing the validation algorithms and to work with both benchmark and real-life graph data. The study will include the consideration of real-life constraints built upon real-life queries found in concrete query logs [3,4].

International opportunities: The selected candidate will have the chance to work with world-class researchers in the fields of graph data management, data systems and database theory.

Expected ability of the student: Very good problem solving and programming abilities, familiarity with data management techniques, good understanding of database theory and algorithms.

Bibliographic references

[1] Siddhartha Sahu, Amine Mhedhbi, Semih Salihoglu, Jimmy Lin, M. Tamer Özsu: **The ubiquity of large graphs and surprising challenges of graph processing: extended survey**. VLDB J. 29(2-3): 595-618 (2020)
[2] Guillaume Bagan, Angela Bonifati, Radu Ciucanu, George H. L. Fletcher, Aurélien Lemay, Nicky Advokaat: **gMark: Schema-Driven Generation of Graphs and Queries**. IEEE Trans. Knowl. Data Eng. 29(4): 856-869 (2017)
[3] Angela Bonifati, Wim Martens, Thomas Timm: **An Analytical Study of Large SPARQL Query Logs**. Proc. VLDB Endow. 11(2): 149-161 (2017)
[4] Angela Bonifati, Wim Martens, Thomas Timm: **Navigating the Maze of Wikidata Query Logs**. WWW 2019: 127-138