L3 Research Internship Topic in Computer Science

Hierarchical Query-Aware Graph Summarization

Keywords: Hybrid Artificial Intelligence, Semantic Web, Big Data, Query Processing

Supervision:

- **Prof. Angela Bonifati**, angela.bonifati@univ-lyon1.fr
  Équipe BD, Laboratoire Liris, Lyon 1 University & INRIA Grenoble Rhône-Alpes, Lyon.
- **Dr. Stefania Dumbrava**, stefania.dumbrava@ensiie.fr

Setting: The explosion in the amount of the data on the Web has led to the need to explore, query and understand the available data sources. The task is challenging due to the complex and heterogeneous structure of those graphs which, unlike relational databases, do not come with a structure-dictating schema. Summarization has been applied to large data graphs to facilitate these tasks. Its purpose is to extract concise and meaningful information from knowledge bases, representing their content as faithfully as possible. There is no single concept of a graph summary, as their construction is application dependent and relies on various techniques and computational tools in data mining and data management [1,2,3,5].

Goal: In this internship, we will focus on designing multi-level summaries, optimized for approximate query answering. As such, we will extract and analyze the hierarchical relations between the represented concepts, layering the data into different levels of granularity, from the most abstract to the most specific. At each level, we will apply compactification heuristics to facilitate the efficient, albeit approximate [4], evaluation of complex queries. Implementing and validating such a methodology 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 large scale graphs. For example, when dealing with social graphs, with billions of nodes and trillions of edges, neighborhood queries are often bottlenecks, whose execution is prone to timeouts. The ability to provide an answer to these, even an inexact one, is extremely important for use cases requiring real-time analytics, e.g., in retail, finance, forensics, cybersecurity and fraud detection etc. Depending on the candidate’s abilities and interests, we can further explore different applications and techniques.

Opportunities:

The selected students (the topic can be pursued by a team of students) will have the opportunity (to be involved in the writing of a research article documenting the obtained results.)
Requirements:

- Very good programming skills (Java);
- Familiarity with Semantic Web technologies (RDF, SPARQL) and with data management and exploration methods are appreciated, but not necessary.

Bibliography:


