Anomaly detection in heterogeneous graphs

In recent years, a large number of companies have begun to collect and store all the data generated through their activities in form of graphs, due to the great potential of this type of structures. Graphs are able to expand a business data model in an efficient way. The topology of a graph with the inherent knowledge of nodes and relationships contains valuable information that is needed to process in order to predict new insights from data. This analysis allows companies to compute personalized recommendations, to efficiently manage authentication on Internet of Things environments, or to analyze social behaviours in large networks, among other tasks.

Our main goal is the detection of anomalies or changes in time-varying or dynamic large graph data, based on distances and connectivity structure, focusing our attention on heterogeneous and attributed graphs. A heterogeneous graph is a type of graph that is made of various types of nodes and edges, where they represent various types of entities and relationships between two entities. Each edge has a corresponding weight which describes the strength of relationship between nodes. Notice that heterogeneous graph model is a promising data model for dealing with heterogeneous information due to its generality and expressiveness.

On the other hand, an attributed graph is a structure where nodes and/or edges have features associated with them. For example, in a social network, users may have various interests, work/live at different locations, be of diverse education levels, etc. while the relational links may have various strengths, types, frequency, etc.

We performed several community-based anomaly detection algorithms for attributed and heterogeneous graphs. They were included in a web application. We tested different database engines and query languages and we selected the Neo4J and Cypher for our proposal. However, we expect to obtain a solution which is not dependant of the selected graph database and query languages. Given a stream of heterogeneous graphs, i.e. containing different types of nodes and edges, we spot anomalous nodes and relationships in real-time while consuming bounded memory. The main idea of the community or clustering-based approaches is to monitor graph communities or clusters over time and report an event when there is structural or contextual change in any of them instead of monitoring the changes in the whole network.

The main problems we encountered in this approach were:
The non-existence of a standard query language that fulfils all possible databases.
The way in which databases for graphs deal with heterogeneous graphs.
The way in which databases for graphs deal with attributed graphs.