Implementing and Analyzing SAPPER for Subgraph Indexing and Approximate Matching in Large Graph

Wedad Alahamade  Woa2326@rit.edu
Supervised by:  Prof. Carlos Rivero
Rochester Institute of Technology

BACKGROUND

• Graphical data has become very important to manage data in number of fields such as bioinformatics, biology, and social networking.
• Noisy or inconsistent data are common with large data.
• SAPPER studies the subgraph query in the context of noise to retrieve all occurrences of the subgraphs that are approximately isomorphic to the query graph.
• SAPPER takes advantage of pre-generated random spanning trees and a graph enumeration order to prune some of the subgraphs as early as possible before matching with the database.
• We implemented and analyzed SAPPER to identify the cases where SAPPER can and cannot work correctly to be used in future work.

SAPPER PROCESSING

Input: Database Graph G, Query Graph q, and number of missing edges θ. Output: All occurrences of the approximate subgraph matches of q in G.

Process 1: Construct the Hybrid Neighborhood Unit (HNU) index to store the local information for each vertex in G to speed up the query processing.

Process 2: Identify the approximate matches of the query graph in the database graph using:
• Vertex Matching to find all candidates matches for each vertex in the query graph in the database graph.
• Query Graph Enumeration to enumerate the query graph’s edges lexicographically based on their occurrences in the database.
• Random Spanning Tree Generation to generate all possible subgraphs of query graph that are D_q(G, q) ≤ θ.
• Depth-First Enumeration Order to enumerate all candidate graphs of query graph that has at least one exact match in the database.

For each subgraph query:
If the longest prefix is not matched yet find the exact matches in G.
else find and output the exact matches according to the prefix’s matches.

EXPERIMENT

• A biological dataset for aids is used to evaluate the performance of SAPPER in the query time and the number of exact and approximate matches.
• The dataset contains 240,000 query graphs with different size.
• We tested SAPPER’s ability to retrieve all exact and approximate matches of the query graph and compare it with GraphQL.

FUTURE WORK

• Improve the execution time with large query.
• Improve the scalability of SAPPER to query larger graphs.