ABSTRACT

Name of student: **Shubhangi Agarwal** Roll no: **14111268**

Degree for which submitted: **Doctor of Philosophy**

Department: **Computer Science and Engineering**

Thesis title: **Subgraph Matching and Mining in Large Graphs**

Name of Thesis Supervisor: **Prof. Arnab Bhattacharya**

Month and year of thesis submission: **April, 2023**

The rapid progress in internet connectivity and the continual evolution of data storage and sharing techniques have substantially widened access to vast amounts of information. This surge in data availability is especially evident in the proliferation of linked data across diverse domains such as social networks, chemoinformatics, bioinformatics, and road networks. This expanding landscape underscores the imperative of harnessing the power of *graphs* as an instrumental means towards enhanced comprehension and deeper insights. Graphs provide a strong foundation for analyzing interconnected networks and can be automatically created by extracting entities and relationships. A labeled graph is a way of representing data where entities are modeled by nodes with associated labels, and their relationships are depicted by edges or links between them.

Graph mining is an important field of research for efficiently analyzing large real-world graphs. To extract useful information from graphs or networks various data mining techniques are applied to graphs, like identifying clusters and mining frequent patterns. A popular subfield of graph mining, *Subgraph querying*, focuses on identifying all occurrences of a specific subgraph pattern in a larger graph or a set of graphs. The goal is to identify specific subgraphs of interest, which can help

in answering specific questions about the graph or verifying the presence of a known structure. For instance, the presence of a substructure in a chemical compound can help us understand its properties and behavior.

The applications of subgraph querying across a wide range of domains and over large amounts of data encourage the development of efficient approaches for label and structural pattern matching on large graphs. As subgraph isomorphism is NP-complete, efficient heuristics have been proposed. However, in certain cases the graph may have missing labels or edges, making it unsuitable to apply exact graph querying paradigms. Technological advancements have made possible the automatic construction of graphs from raw data. However, such graph constructing models assign a *probability value* or confidence score to the extracted entities or relationships or their attributes. In such a scenario, an exact subgraph match may not exist and an *approximate subgraph match* may be more suitable.

Approximate subgraph matching (ASM) involves finding subgraph patterns in a larger input graph that are *similar* to a given subgraph pattern, generally referred to as a query, but not necessarily identical. This provides more flexibility and robustness in graph analysis and enables the detection of subgraphs that may have different labels, sizes, or shapes. Applications of ASM range from recommendation systems to medical diagnostics based on symptom-disease association.

Various similarity measures can be used to compute the degree of similarity between the query graph pattern and subgraphs in the input graph, such as graph edit distance, maximum common subgraph or graph based statistical measures like degree distribution. However, methods based on such similarity measures commonly define application dependent thresholds. Moreover, if the input graph is probabilistic, it may be necessary to establish an acceptable probability threshold for existence. Determining an appropriate threshold is necessary to capture important patterns, however, this process can be challenging, especially in cases where the underlying distribution of data is unknown or the data is noisy.

In this work, we address this issue and propose *statistical significance* based sub-

graph querying methods for both deterministic and probabilistic graphs. *Statiscal significance* measures like the *Pearson's chi-squared statistic* allows to integrate the label and the topological similarity of a subgraph as well as the associated probabilities. Another widespread way is the use of neural network-based models to map the input subgraphs and the query to a feature space and compare the embedded vectors to search for answer retrieval. As the embeddings of the graphs are often obtained by aggregating the embeddings of the constituent nodes, we present a *graph neural network* for generating node embedding vectors that capture *global position* of the nodes in the context of the graph with respect to a chosen set of nodes, called *anchors*.

We propose two novel chi-squared based approaches, VerSaChI and ChiSeL, to search for approximate subgraph patterns in deterministic and probabilistic input graphs, respectively. VerSaChI computes a similarity of the neighborhood until two-hops and uses Chebyshev's inequality along with χ^2 measure. On the other hand, ChiSeL takes into account the possible world semantics while avoiding enumeration of all possible instantiations of the (sub)graph. We also conduct a study using VeNoM, a variant of an existing ASM approach, by parametrizing the depth and breadth of the neighborhood considered. We also discuss GraphReach, a random walk based *position-aware graph neural network*. It uses a diversified anchor selection algorithm for a more meaningful node embedding. Extensive experiments demonstrate the robustness of various methods and showcase their efficacy on different datasets.