

GNN-based Anchor Embedding for Exact Subgraph Matching

Bin Yang

Harbin Institute of Technology
Harbin, Heilongjiang, China
binyang.hit@gmail.com

Jianxiong Ye

Harbin Institute of Technology
Harbin, Heilongjiang, China
jianxiongye2023@gmail.com

Zhaonian Zou

Harbin Institute of Technology
Harbin, Heilongjiang, China
znzou@hit.edu.cn

ABSTRACT

Subgraph matching query is a classic problem in graph data management and has a variety of real-world applications, such as discovering structures in biological or chemical networks, finding communities in social network analysis, explaining neural networks in machine learning, and so on. To further solve the subgraph matching problem, several recent advanced works attempt to utilize deep-learning-based techniques to handle the subgraph matching query. However, most of these works only obtain approximate results for subgraph matching without theoretical guarantees of accuracy. In this paper, we propose a novel and effective *graph neural network (GNN)-based anchor embedding framework* (GNN-AE), which allows *exact subgraph matching*. Unlike GNN-based approximate subgraph matching approaches that only produce inexact results, in this paper, we pioneer the anchor concept (including anchor, anchor graph/path, etc) in subgraph matching and carefully devise the anchor (graph) embedding technique based on GNN models. We transform the subgraph matching problem into a search problem in the embedding space via the anchor (graph & path) embedding techniques. With the proposed anchor matching mechanism, GNN-AE can guarantee subgraph matching has no false dismissals. We design an efficient matching growth algorithm, which can retrieve the locations of all exact matches in parallel. We also propose a cost-model-based DFS query plan to enhance the parallel matching growth algorithm. Through extensive experiments on 6 real-world and 3 synthetic datasets, we confirm the effectiveness and efficiency of our GNN-AE approach for exact subgraph matching.

PVLDB Reference Format:

Bin Yang, Jianxiong Ye, and Zhaonian Zou. GNN-based Anchor Embedding for Exact Subgraph Matching. PVLDB, 14(1): XXX-XXX, 2025.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at <https://github.com/BenYoungLab/GNN-AE>.

1 INTRODUCTION

Subgraph matching [51, 70] is a fundamental problem in graph data management. Given a data graph G and a query graph Q , a subgraph matching asks for all subgraphs (also called embeddings or matches) in G that match Q in terms of isomorphism. Consider the example in Figure 1, $\{(q_1, v_1), (q_2, v_3), (q_3, v_5), (q_4, v_2), (q_5, v_7)\}$ and

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

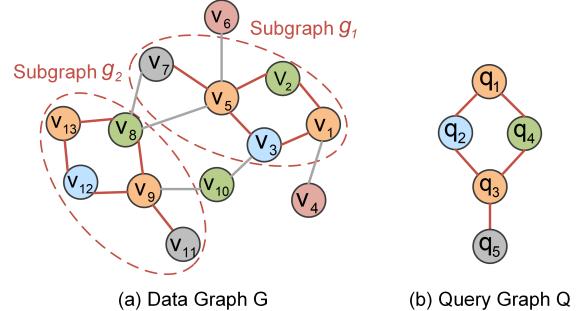


Figure 1: An example of the subgraph matching.

$\{(q_1, v_{13}), (q_2, v_{12}), (q_3, v_9), (q_4, v_8), (q_5, v_{11})\}$ are all matches of the query graph Q in the data graph G . The subgraph isomorphism problem is known to be NP-complete [14, 20, 36], so it is costly to find all exact isomorphic matches in G .

Subgraph matching has many practical applications, such as discovering structures in biological [56, 67] and chemical networks [22, 69], pattern alignment to fuse entities in knowledge graphs [28, 41, 54], finding communities in social networks [2, 39], constructing or explaining neural networks in machine learning [60, 63], and so on. Specifically, for example, in a social network, the node represents people and the edge indicates the existence of social relationships between two people. Each person is marked with a hobby label, such as 'games', 'music', and 'sports'. To recommend a new game product in a targeted manner on the social network, advertisers can extract a subgraph of a certain game community as a query and initiate subgraph matching on the entire network to obtain candidate advertisement recipients.

Prior works on the exact subgraph matching problem usually followed the exploration-based or the join-based methodology [51, 53]. The exploration-based methods adopt the backtracking search, which maps query nodes to data nodes and iteratively extends intermediate results to find all matches. The join-based methods convert the subgraph query to a relational query, and evaluate the multi-way join to obtain all results.

Since the subgraph isomorphism problem is NP-complete, exact subgraph matching has a high computation cost. A second-best alternative direction is to obtain an approximate result of subgraph matching [18, 19, 32, 73]. Previous works on approximate subgraph matching usually adopt various graph similarity measures such as edge edit distance [68] to qualify occurrences of the query Q on the data graph G . If the edit distance between Q and a subgraph g of G is no more than some threshold θ , then g is considered an approximate result of Q .

Motivation. To further solve the subgraph matching problem, several recent advanced works [15, 38, 40, 47] attempt to utilize deep-learning-based techniques such as *Graph Neural Networks* (GNNs)

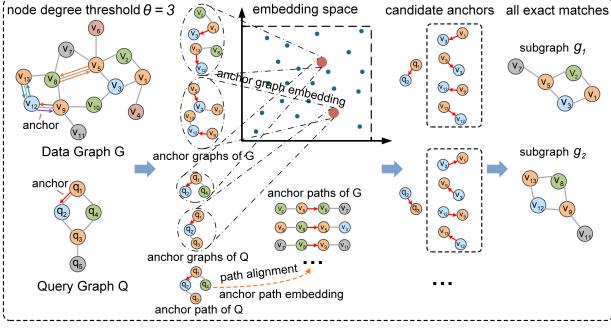


Figure 2: An example of GNN-based anchor embedding for exact subgraph matching.

to handle the subgraph matching problem. However, these works mainly focus on obtaining approximate results for subgraph matching. Specifically, GNN-based approaches transform the query graph and the data graph into vectors in an embedding space and learn the subgraph relationship between them by comparing their embedding vectors. Although GNN-based approaches are efficient, there are several drawbacks regarding these approximate approaches:

(1) They usually only approximately assert subgraph isomorphism relations between the query and the data graph, and are not suitable for retrieving locations of all matching.

(2) There are no theoretical guarantees or error bounds about the approximate accuracy.

It is known that GNNs usually provide approximate solutions, such as for classification or regression, without an accuracy guarantee. Therefore, applying GNNs to handle approximate subgraph matching is relatively straightforward. However, it is not trivial how to utilize GNNs to guarantee the accuracy of exact subgraph matching [65]. In order to obtain the exact results for subgraph matching, RL-QVO in [58] and GNN-PE in [66] provide two demonstrations of utilizing GNNs for exact subgraph matching. RL-QVO employs the *Reinforcement Learning* (RL) and GNNs to generate the high-quality matching order for subgraph matching algorithms. Although RL-QVO is effective, it relies on historical query graphs to train the model, which limits its scalability and robustness. GNN-PE trains an overfitting GNN model to learn the dominance relationships between the node (and its 1-hop neighbors) and their substructures on the data graph. These nodes with dominance relationships are concatenated into the path as the basic unit for the query to obtain candidate matches. GNN-PE requires that the GNN model be trained to overfit on the training data set. When the data graph is large, it results in a high offline computation cost.

Our Contributions. In this paper, we study the *exact subgraph matching* problem on the large-scale data graph and propose a novel GNN-AE (GNN-based Anchor Embedding) framework for exact and efficient subgraph matching. In contrast to existing GNN-based approaches for approximate subgraph matching without theoretical guarantees of accuracy, our GNN-AE can obtain exact results for subgraph matching and retrieve the locations of all matches. As shown in Figure 2, in GNN-AE, we first propose the *anchor concept* (including anchor, anchor graph/path, etc.). We design effective *anchor embedding* techniques (including *anchor graph embedding*, and *anchor path embedding*) based on the anchor concept and GNN models, which transform the subgraph matching problem into a

Table 1: Notations and Descriptions

Notations	Descriptions
Q, G and g	query graph, data graph and subgraph
g_{uv}/s_{uv}	an anchor graph (or substructure) of directed edge (u, v)
$o(g_{uv})/o(s_{uv})$	an embedding vector of anchor graph g_{uv} (or s_{uv})
p_{uv}/τ_{uv}	an anchor path (or anchor subpath) of directed edge (u, v)
C_{uv}	candidate anchors of directed edge (u, v)
ψ_{uv} and Ψ	an anchor (i.e., directed edge (u, v)) and an anchor set
g_{uv}^p/s_{uv}^p	a positive 1-hop anchor graph (or anchor substructure) of ψ_{uv}
g_{uv}^n/s_{uv}^n	a negative 1-hop anchor graph (or anchor substructure) of ψ_{uv}
p_{uv}^d/τ_{uv}^d	a dual 1-hop anchor path (or anchor subpath) of ψ_{uv}

search problem in the embedding space. If two anchor graphs/paths in the subgraph matching have an isomorphic relationship, they are mapped to the same embedding vector. Based on this principle, we can guarantee a 100% matching (i.e., candidate anchors) recall ratio for each query anchor without introducing false dismissals. GNN-AE does not rely on query graphs to train the model, and the proposed *truncated GNN model training* technique enables GNN-AE to have a low offline computation cost.

To retrieve locations of all exact matches, we propose an *anchor matching mechanism* based on anchor embeddings and design a *parallel matching growth algorithm*. At last, we also propose a *cost-model-based DFS query plan* to enhance parallel matching growth.

In summary, we make the following contributions:

- We propose a novel GNN-AE framework for exact subgraph matching via GNN-based anchor embeddings in Section 2.
- We propose the anchor concept in subgraph matching and design effective anchor (graph & path) embedding techniques for exact subgraph retrieval in Section 3.
- We develop an anchor matching mechanism based on anchor embeddings and design an efficient parallel matching growth algorithm to obtain locations of all exact matches in Section 4.
- We design a cost-model-based DFS query plan to enhance the parallel matching growth algorithm in Section 5.
- We conduct experiments on 6 real-world and 3 synthetic datasets to evaluate the effectiveness and efficiency of GNN-AE in Section 6.

Related works on subgraph matching are reviewed in Section 7 and Section 8 concludes this paper.

2 PROBLEM DEFINITION

In this section, we formally define the concepts related to the subgraph matching problem and propose the GNN-based anchor embedding framework for exact subgraph matching.

2.1 Subgraph Matching Query

In this paper, we focus on undirected vertex-labeled graphs. Let Σ be a set of labels. A vertex-labeled graph is defined as a tuple (V, E, L, Σ) , where V is a set of vertices, E is a set of edges, and $L : V \rightarrow \Sigma$ is a function that associates each vertex $v \in V$ with a label in Σ . Table 1 summarizes the frequently used notations.

DEFINITION 2.1 (GRAPH ISOMORPHISM). A graph $G = (V, E, L, \Sigma)$ is isomorphic to a graph $G' = (V', E', L', \Sigma')$, denoted as $G \cong G'$, if there exists a bijective $f : V \rightarrow V'$ such that (1) $\forall v \in V, L(v) = L'(f(v))$; (2) $\forall u, v \in V, (u, v) \in E$ if and only if $(f(u), f(v)) \in E'$.

Algorithm 1: The GNN-Based Anchor Embedding (GNN-AE) Framework for Exact Subgraph Matching

```

Input: a data graph  $G$  and a query graph  $Q$ 
Output: all subgraphs  $g$  ( $\subseteq G$ ) that are isomorphic to  $Q$ 
// Offline Pre-Computation Phase
1 initial a training data set  $D = \emptyset$  and a path set  $P = \emptyset$ 
2 define directions for edges in graph  $G$ 
3 for each directed edge  $(u, v) \in G$  do
4   if degree  $d_u \leq \theta$  or  $d_v \leq \theta$  then
5     obtain all anchor substructures  $s_{uv}$  of edge  $(u, v)$ 
6     add all substructures  $s_{uv}$  to  $D$ 
7   else
8     obtain all anchor subpaths  $\tau_{uv}$  of edge  $(u, v)$ 
9     add all subpaths  $\tau_{uv}$  to  $P$ 
/* train a GNN model for anchor graph embeddings */
10 truncated train a GNN model  $M$  over the training data set  $D$ 
11 generate embeddings  $o(s_{uv})$  for anchor substructures  $s_{uv}$  in  $D$  via  $M$ 
/* build indexes over graph  $G$  */
12 build two hash indexes  $I_s^s$  and  $I_s^d$  over embedding vectors for anchor
  substructures in  $G$ 
13 encode each anchor subpath  $\tau$  in  $P$ 
14 build a hash index  $I_\tau$  over encodings  $c(\tau)$  for anchor subpaths in  $G$ 
// Online Subgraph Matching Phase
15 compute cost-model-based DFS query plan  $\varphi$  for  $Q$ 
/* retrieve candidate anchors */
16 for each DFS edge  $(u, v) \in Q$  do
17   extract its anchor graphs  $g_{uv}$  and anchor paths  $p_{uv}$ 
18   obtain embeddings  $o(g_{uv})$  via GNNs  $M$  and encoding  $\omega(p_{uv})$  of  $p_{uv}$ 
19   find candidate anchors  $C_{uv}$  of  $(u, v)$  by retrieving  $I_s^s$ ,  $I_s^d$  and  $I_\tau$ 
/* refine and obtain all exact matches */
20 assemble candidate anchors  $C$  along the DFS edge order and refine subgraphs
   $g$  via hash join
21 return subgraphs  $g$  ( $\cong Q$ )

```

We say that a graph G' is *subgraph isomorphic* to a graph G , denoted as $G' \sqsubseteq G$, if G' is isomorphic to a subgraph g of G (i.e., $g \subseteq G$). The subgraph isomorphism problem is NP-complete [14, 20, 36].

DEFINITION 2.2 (SUBGRAPH MATCHING QUERY). *Given a data graph G and a query graph Q , a subgraph matching query find all subgraphs g in G that match Q in terms of isomorphism.*

In this paper, we consider exact subgraph matching, i.e., retrieving all subgraph locations on the data graph G that are strictly isomorphic to the query graph Q .

2.2 GNN-based Subgraph Matching Framework

To address the exact subgraph matching problem, we propose the *GNN-based anchor embedding* (GNN-AE) framework based on the GNNs technique and the anchor concept (including anchor, anchor graph/path, and anchor (graph/path) embedding, etc). Algorithm 1 shows the GNN-AE framework, which consists of offline pre-computation and online subgraph matching phases. We preprocess the data graph into anchor (graph/path) embeddings via GNNs and path encoding (lines 1-11), build indexes (lines 12-14) in the offline pre-computation phase, and then answer online subgraph matching queries over indexes (lines 15-21).

Specifically, we set anchors (i.e., some directed edges in graphs, please refer to subsection 3.1) as the basic matching units in subgraph matching, and build a matching relationship between the query anchor and the data graph anchor. In offline pre-computation,

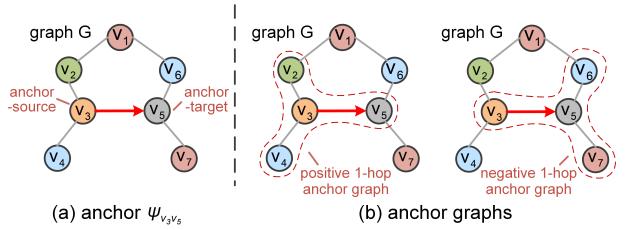


Figure 3: Illustration of the anchor and anchor graphs.

we define directions for edges in the data graph G , and divide the directed edges $(u, v) \in G$ into two categories according to the degree d_u and d_v . If $d_u \leq \theta$ or $d_v \leq \theta$, we obtain its anchor graph and all anchor substructures and transform them into anchor graph embeddings via the GNN model (refer to subsection 3.2). In subsection 3.3, we discuss how to build the subgraph matching relationships between the query graph Q and the data graph G for such directed edges (i.e., sparse-sparse anchors, sparse-dense anchors, and dense-sparse anchors) via anchor graph embeddings. If $d_u > \theta \wedge d_v > \theta$, we obtain its anchor paths and all anchor subpaths and directly encode them as anchor path embeddings. In subsection 3.4, we provide the details of building the subgraph matching relationships between Q and G for such directed edges (i.e., dense-dense anchors) via anchor path embeddings. Whether matching via anchor graph embeddings or anchor path embeddings, they all follow a basic principle: *When there is a matching relationship between query anchor $\psi_{q,j}$ and data graph anchor ψ_{uv} , the anchor graph of $\psi_{q,j}$ and the anchor graph/one of the anchor substructures of ψ_{uv} have the same embedding vector. The same rule applies to the anchor path.*

3 GNN-BASED ANCHOR EMBEDDING

In this section, we present the *anchor concept* (refer to subsection 3.1), the *anchor graph embedding* technique based on the GNN model (refer to subsection 3.2), and how to build the subgraph matching relationships between the query graph Q and the data graph G via anchor graph embedding (refer to subsection 3.3) and anchor path embedding (refer to subsection 3.4).

3.1 Anchor Concept

Here, we introduce the anchor concept in subgraph matching.

DEFINITION 3.1 (ANCHOR). *Given a graph G , an anchor ψ_{uv} is a directed edge in G , where the source node u is called the anchor-source and the target node v is called the anchor-target.*

Given a graph $G = (V, E, L, \Sigma)$, we can define directions for all edges in G (line 2 of Algorithm 1), where an undirected edge produces two anchors. Therefore, we have the following Theorem 3.1.

THEOREM 3.1. *A graph $G = (V, E, L, \Sigma)$ has $2|E|$ anchors.*

Figure 3(a) shows an example of the anchor in the graph G . We assume that the directed edge (v_3, v_5) is an anchor $\psi_{v_3v_5}$ in the graph G , where v_3 is the anchor-source and v_5 is the anchor-target. In the graph G , there are 7 undirected edges and 14 anchors.

DEFINITION 3.2 (POSITIVE K-HOP ANCHOR GRAPH). *Given a graph $G = (V, E, L, \Sigma)$ and an anchor ψ_{uv} in G , the positive k -hop ($k \geq 0$) anchor graph of ψ_{uv} , denoted as $g_p^{(k)}(u, v)$, is defined as a subgraph*

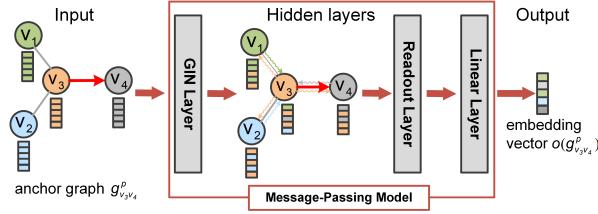


Figure 4: The GNN model for the anchor graph embedding.

induced from G by a set of nodes $\{v_i | d(u, v_i) \leq k, v_i \in V\} \cup \{v\}$, where $d(u, v_i)$ denotes the length of shortest path from u to v_i .

DEFINITION 3.3 (NEGATIVE K-HOP ANCHOR GRAPH). Given a graph $G = (V, E, L, \Sigma)$ and an anchor ψ_{uv} in G , the negative k -hop ($k \geq 0$) anchor graph of ψ_{uv} , denoted as $g_n^{(k)}(u, v)$, is defined as a subgraph induced from G by a set of nodes $\{u_i | d(v, u_i) \leq k, u_i \in V\} \cup \{u\}$, where $d(v, u_i)$ is the length of shortest path from v to u_i .

Specifically, in subgraph matching, we adopt the 1-hop anchor graph (only star shape) for anchor matching in subsequent sections. If not stated otherwise, the 1-hop anchor graph is a 1-hop anchor graph with a star shape in the rest of this paper. In Section 3.2, we will explain the reasons for using the 1-hop anchor graph for subgraph matching. Figure 3(b) illustrates an example of the positive/negative 1-hop anchor graphs.

EXAMPLE 3.1. Consider the graph G in Figure 3(b) and the anchor $\psi_{v_3v_5}$, the positive 1-hop anchor graph of $\psi_{v_3v_5}$ is the induced star subgraph formed by the node set $\{v_2, v_3, v_4, v_5\}$, and the negative 1-hop anchor graph of $\psi_{v_3v_5}$ is the induced star subgraph formed by the node set $\{v_3, v_5, v_6, v_7\}$.

In particular, for simplicity, we use g_{uv}^p to denote the positive 1-hop anchor graph of ψ_{uv} and g_{uv}^n to denote the negative 1-hop anchor graph of ψ_{uv} by default in following sections.

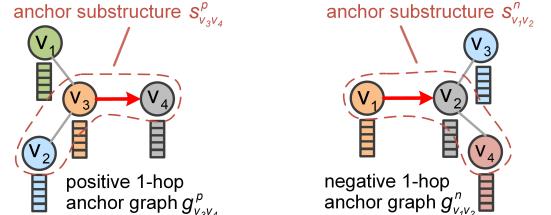
3.2 GNN Model for Anchor Graph Embedding

In this subsection, we discuss how to calculate the anchor embedding for an anchor graph/substructure via the GNNs technique (lines 1-6 and lines 10-11 of Algorithm 1).

We employ a GNN model (e.g., *Graph Isomorphism Network* (GIN) [62]) to enable the anchor graph embedding in the data graph G (lines 10-11 of Algorithm 1). Specifically, the GNN model takes a positive/negative 1-hop anchor graph g_{uv}^p/g_{uv}^n (or anchor substructure s_{uv}^p/s_{uv}^n , refer to Definition 3.4 and 3.5) as input and an embedding vector $o(g_{uv}^p)/o(g_{uv}^n)$ (or $o(s_{uv}^p)/o(s_{uv}^n)$) as output. Figure 4 illustrates an example of the GNN model.

DEFINITION 3.4 (POSITIVE ANCHOR SUBSTRUCTURE). Given an anchor ψ_{uv} and its positive k -hop anchor graph $g_p^{(k)}(u, v)$, the positive anchor substructure in $g_p^{(k)}(u, v)$ is defined as a subgraph in $g_p^{(k)}(u, v)$ that contains the edge (u, v) of ψ_{uv} , denoted as $s_p^{(k)}(u, v)$.

DEFINITION 3.5 (NEGATIVE ANCHOR SUBSTRUCTURE). Given an anchor ψ_{uv} and its negative k -hop anchor graph $g_n^{(k)}(u, v)$, the negative anchor substructure in $g_n^{(k)}(u, v)$ is defined as a subgraph in $g_n^{(k)}(u, v)$ that contains the edge (u, v) of ψ_{uv} , denoted as $s_n^{(k)}(u, v)$.



(a) substructure $s_{v_3v_4}^p \subseteq g_{v_3v_4}^p$ (b) substructure $s_{v_1v_2}^n \subseteq g_{v_1v_2}^n$

Figure 5: Illustration of the input for the GNN model.

In particular, in subgraph matching, we use s_{uv}^p to denote the positive anchor substructure of the positive 1-hop anchor graph and s_{uv}^n to denote the negative anchor substructure of the negative 1-hop anchor graph by default in following sections.

GNN Model Architecture. It is well known that the subgraph isomorphism relationship between the query and the data graph has nothing to do with the node numbering of the graph and is only constrained by the graph topology and node labels (i.e., regardless of the node IDs assignment). In GNN-AE, we set anchors as the basic unit of matching between the query Q and the data graph G and generate embedding vectors for the anchor based on anchor graphs/substructures, and GNNs. Therefore, the anchor graph embeddings obtained through the model should also be independent of the node numbering on the anchor graph. The GNN model naturally satisfies the *order-invariant* (refer to Definition 3.6) property [42].

DEFINITION 3.6 (ORDER-INVARIANT). Assume $A \in \mathbb{R}^{n \times n}$ is the adjacency matrix of a graph G and $P \in \mathbb{R}^{n \times n}$ is a permutation matrix, the function f is order-invariant if it satisfies $f(P^T AP) = f(A)$.

As shown in Figure 4, our GNN model consists of input, hidden layers, and output.

Input: As mentioned earlier, the input of the GNN model is a positive/negative 1-hop anchor graph g_{uv}^p/g_{uv}^n (or its anchor substructure s_{uv}^p/s_{uv}^n). Figure 4 shows an example of the anchor graph $g_{v_3v_4}^p$. Each vertex v_i in $g_{v_3v_4}^p$ is associated with an initial feature embedding, which is obtained via node label encoding techniques [10].

Figure 5 illustrates an example of the positive/negative 1-hop anchor graph and one of its anchor substructures.

EXAMPLE 3.2. (1) Consider the anchor $\psi_{v_3v_4}$ and its positive 1-hop anchor graph $g_{v_3v_4}^p$ in Figure 5(a), the node set $\{v_2, v_3, v_4\}$ forms one of its anchor substructures $s_{v_3v_4}^p$. (2) Consider the anchor $\psi_{v_1v_2}$ and its negative 1-hop anchor graph $g_{v_1v_2}^n$ in Figure 5(b), the node set $\{v_1, v_2, v_4\}$ forms one of its anchor substructures $s_{v_1v_2}^n$.

Hidden Layers: Most of the classic spatial-based GNN model follow the message-passing mechanism in the hidden layers. They satisfy the *order-invariant* property.

The message-passing mechanism is a paradigm that aggregates neighboring node information to update the central node information. Since messages are exchanged only between nodes connected by edges, the graph structure is incorporated into the updated embeddings. Message passing occurs one round in each layer of the network, and each round involves 2 steps [17]:

Step 1 (Aggregating): In t -th round, a node v aggregates embedding messages from all its neighbors $N(v)$. These embedding messages $\{h_u^{(t)} | u \in N(v)\}$ form a new message $m_v^{(t+1)}$ by the message function $M_t: m_v^{(t+1)} = M_t(\{h_u^{(t)} | u \in N(v)\})$. And then, $m_v^{(t+1)}$ is received by the node v .

Step 2 (Updating): The new message $m_v^{(t+1)}$ is combined with the embedding representation $h_v^{(t)}$ to form a new representation $h_v^{(t+1)}$ for $v: h_v^{(t+1)} = U_t(h_v^{(t)}, m_v^{(t+1)})$, where U_t is the update function.

Overall, the message-passing mechanism can be formulated as:

$$h_v^{(t+1)} = U_t(h_v^{(t)}, M_t(\{h_u^{(t)} | u \in N(v)\})). \quad (1)$$

The message-passing model in Figure 4 shows an example of aggregating and updating of node feature embeddings. In each convolutional layer, the nodes $\{v_1, v_2, v_4\}$ aggregate the feature embedding from node v_3 , and the node v_3 receives the feature embeddings from nodes $\{v_1, v_2, v_4\}$.

There are a lot of message-passing models in the field of graph neural networks. Specifically, we employ a *Graph Isomorphism Network* (GIN) [62]. In each GIN layer, node feature embeddings on neighbors are aggregated via a summation operator. The update function U_t is modeled by a multi-layer perceptron (MLP) rather than specified by users. In t -th round, the new hidden representation $h_v^{(t+1)}$ for v is computed as follows:

$$a_v^{(t)} = \sum_{u \in N(v)} h_u^{(t)}, \quad (2)$$

$$h_v^{(t+1)} = \text{MLP}^{(t+1)} \left((1 + \epsilon^{(t+1)}) h_v^{(t)}, a_v^{(t)} \right). \quad (3)$$

where $\epsilon^{(t+1)}$ is a learnable parameter or a fixed scalar to adjust the importance of the hidden representations $h_v^{(t)}$. Such design of GIN makes the update function closer to be injective, so GIN has stronger expressive power than traditional graph neural networks.

After the t rounds of propagation, all hidden representations $h_v^{(t)}$ at each layer of the network are aggregated to form the hidden representation of the anchor graph g_{uv}^p at the t -th layer, that is,

$$h_G^{(t)} = \text{readout}(\{h_v^{(t)} | v \in g_{uv}^p\}). \quad (4)$$

Generally, the *readout* function can be an operator that sums or means all node hidden representations in the anchor graph g_{uv}^p .

A linear layer performs a linear transformation for $h_G^{(t)} \in \mathbb{R}^{n \times 1}$ (n is the *readout* hidden dimension) and obtains the embedding vector $o(g_{uv}^p)$ with dimension m for the anchor ψ_{uv} . That is,

$$o(g_{uv}^p) = Wh_G^{(t)}, \quad (5)$$

where W is a learnable weight matrix and $W \in \mathbb{R}^{m \times n}$.

Output: The output of the GNN model is an embedding vector $o(g_{uv}^p)$ (given in Eq. 5) as the embedding of anchor ψ_{uv} on the anchor graph g_{uv}^p .

Truncated GNN Model Training. As mentioned in Theorem 3.1, a graph $G = (V, E, L, \Sigma)$ has $2|E|$ anchors. For each anchor in the data graph, we extract all anchor substructures in its positive/negative 1-hop anchor graph. In real-world graphs, node degrees usually follow the power-law distribution [3, 6, 7], and only a small fraction of nodes have high degrees [65]. For example, in DBLP graph data [51, 53], the average node degree is around 7 and the proportion of

Algorithm 2: Truncated GNN Model Training

```

Input: a data graph  $G$ , node degree threshold  $\theta$ , and the number of truncated epochs  $\rho$ 
Output: a trained GNN model  $\mathcal{M}$ 
// Generate a training data set  $D$ 
1 initial a training data set  $D = \emptyset$ 
2 define anchors in graph  $G$ 
3 for each anchor  $\psi_{uv} \in G$  do
4   if degree  $d_u \leq \theta$  then
5     obtain the positive 1-hop anchor graph  $g_{uv}^p$  of  $\psi_{uv}$ 
6     extract all positive anchor substructures  $s_{uv}^p$  in  $g_{uv}^p$ 
7     add all substructures  $s_{uv}^p$  to  $D$ 
8   else if degree  $d_v \leq \theta$  then
9     obtain the negative 1-hop anchor graph  $g_{uv}^n$  of  $\psi_{uv}$ 
10    extract all negative anchor substructures  $s_{uv}^n$  in  $g_{uv}^n$ 
11    add all substructures  $s_{uv}^n$  to  $D$ 
12 attach a unique label for each graph in  $D$ 
13 randomly shuffle graphs in  $D$ 
// Train a GNN Model Until Truncated Epochs  $\rho$ 
14 while  $\text{epochs} < \rho$  do
15   for each batch  $B \subseteq D$  do
16     obtain embedding vectors of graph in  $B$  via  $\mathcal{M}$ 
17     compute the loss function  $\mathcal{L}(D)$  of  $\mathcal{M}$  by Eq. 6
18      $\mathcal{M}.\text{update}(\mathcal{L}(D))$ 
19 return the trained GNN model  $\mathcal{M}$ 

```

nodes with a degree no more than 10 exceeds 85%. It means that the positive/negative 1-hop anchor graph of most anchors in the data graph incurs less than 512 ($= 2^9$) anchor substructures (refer to Lemma 3.1). Therefore, it is usually acceptable for these anchors with low node degrees to extract all anchor substructures in its positive/negative 1-hop anchor graph (lines 3-6 of Algorithm 1).

LEMMA 3.1. For a positive (resp., negative) 1-hop anchor graph g_{uv}^p (resp., g_{uv}^n), if its anchor-source (resp., anchor-target) has degree d , it has 2^{d-1} anchor substructures.

PROOF. Let a positive 1-hop anchor graph $g_{uv}^p = (V, E, L, \Sigma)$. According to Definition 3.4, the positive anchor substructure in a positive 1-hop anchor graph g_{uv}^p is a subgraph in g_{uv}^p that must contain the edge (u, v) of ψ_{uv} . The number of anchor substructures in g_{uv}^p is equivalent to the number of subsets of the non-anchor edge sets $\Omega = \{e | e \in E \wedge e \neq (u, v)\}$ in g_{uv}^p , where $|\Omega| = d - 1$. So g_{uv}^p has $C_{d-1}^0 + C_{d-1}^1 + \dots + C_{d-1}^{d-1} = 2^{d-1}$ anchor substructures. The same rule applies to a negative 1-hop anchor graph g_{uv}^n . \square

Algorithm 2 describes the details of the truncated GNN model training, which consists of two phases, generate a training data set D , and train a GNN model until truncated epochs ρ .

The Training Data Set: Algorithm 2 describes the building of a training data set D (lines 1-13). Specifically, we set an acceptable node degree threshold θ . For each anchor ψ_{uv} in G , if the degree of anchor-source u is not greater than θ , we obtain its positive 1-hop anchor graph g_{uv}^p and extract all anchor substructures s_{uv}^p (lines 4-7). If the degree of anchor-source u is more than θ , we check the degree of its anchor-target v . When the degree of anchor-target v is not greater than θ , we obtain its negative 1-hop anchor graph g_{uv}^n and extract all anchor substructures s_{uv}^n (lines 8-11). Each graph in D is attached to a unique m -dimensional vector as the label, which has the same dimension as the output in Eq. 5 (line 12).

For each anchor ψ_{uv} that satisfies the threshold θ in G , extracting all its positive/negative 1-hop anchor substructures s_{uv}^P/s_{uv}^N generates a large-scale training data set. In order to reduce the size of the training data set D , we provide a *node-centric view* to extract anchor substructures in G .

EXAMPLE 3.3. Consider the graph G in Figure 3(b), the anchor $\psi_{v_3v_2}$, anchor $\psi_{v_3v_4}$ and anchor $\psi_{v_3v_5}$ have the same positive 1-hop anchor graph, which is a star graph induced by the node set $\{v_2, v_3, v_4, v_5\}$. Meanwhile, the anchor $\psi_{v_2v_3}$, anchor $\psi_{v_4v_3}$ and anchor $\psi_{v_5v_3}$ have the same negative 1-hop anchor graph, which is also the star graph induced by the node set $\{v_2, v_3, v_4, v_5\}$.

In the node-centric view, we find that lots of repeated anchor substructures are extracted in Algorithm 2. Example 3.3 provides a node-centric view example. These positive/negative 1-hop anchor graphs are repeatedly enumerated and have the same center node v_3 . In other words, a 1-hop anchor graph (or substructure) corresponds to multiple anchors. Therefore, we extract 1-hop anchor graphs and substructures from the node-centric view in G to form the training data set D' and train the GNN model on D' instead of D .

LEMMA 3.2. Let the training data set generated by Algorithm 2 is D , and the training data set obtained in the node-centric view is D' . $\forall s_{uv} \in D, \exists s'_{uv} \in D'$ that satisfies $s'_{uv} \cong s_{uv}$.

PROOF. In Algorithm 2, each substructure in D corresponds to an anchor ψ_{uv} , which satisfies either i) $d_u \leq \theta$ or ii) $d_u > \theta \wedge d_v \leq \theta$.

Case i): Algorithm 2 obtains g_{uv}^P for ψ_{uv} and extracts all anchor substructures s_{uv}^P . In the node-centric view, g_{uv}^P is a 1-hop star subgraph with the center node u , where $d_u \leq \theta$. All anchor substructures in g_{uv}^P are equivalent to all anchor substructures extracted in the 1-hop star subgraph with the center node u .

Case ii): Algorithm 2 obtains g_{uv}^N for ψ_{uv} and extracts all anchor substructures s_{uv}^N . In the node-centric view, g_{uv}^N is a 1-hop star subgraph with the center node v , where $d_v \leq \theta$. All anchor substructures in g_{uv}^N are equivalent to all anchor substructures extracted in the 1-hop star subgraph with the center node v .

In summary, for each anchor substructure s_{uv} in D , there exists an anchor substructure s'_{uv} in D' that satisfies $s'_{uv} \cong s_{uv}$. \square

According to the proof of Lemma 3.2, it is not difficult to deduce that $\forall s'_{uv} \in D', \exists s_{uv} \in D \wedge s_{uv} \cong s'_{uv}$ is also established. That is, the training data set D generated by Algorithm 2 is equivalent to the training data set D' obtained in the node-centric view, denoted as $D \equiv D'$. However, the size of the training data set D' is much smaller than that of the training data set D . We have the following Theorem 3.2 regarding the size of the training data set D' .

THEOREM 3.2. Given a data graph G and an acceptable node degree threshold θ , if the number of nodes in G with a degree not greater than θ is N , then the size of the training data set D' for the GNN model is at most $N(2^\theta - 1)$.

PROOF. According to Lemma 3.2, for each $u \in G \wedge d_u \leq \theta$, we obtain a 1-hop star subgraph g_u with the center node u . We need to extract all substructures s_u containing at least an anchor in g_u and add them to D' , where a substructure corresponds to multiple anchors. Therefore, g_u has at most $C_\theta^1 + C_\theta^2 + \dots + C_\theta^\theta = 2^\theta - 1$ substructures (similar to the proof of Lemma 3.1). The number of

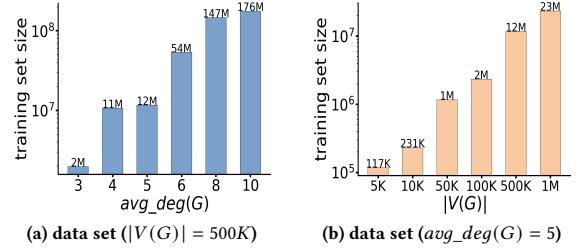


Figure 6: Illustration of the training data set.

nodes in G with a degree not greater than θ is N , so the size of the training data set D' is at most $N(2^\theta - 1)$. \square

Figure 6 illustrates the statistical number of training data in D' . We generate synthetic small-world graphs which follow the Newman-Watts-Strogatz model [59]. The number of nodes $V(G)$ varies from $5K$ to $1M$ and the average node degree $avg_deg(G)$ is adjusted from 3 to 10 . We set the default node degree threshold $\theta = 10$. In the experimental results, our GNN model can learn as many as $\geq 176M$ anchor substructures for a graph.

Train a GNN Model in a Truncated Manner: In GNN-PE [66], the GNN model is required to be trained until the loss function is equal to 0. However, GNN-AE provides an opposite way. In our GNN-AE, the GNN model aims to map isomorphic anchor substructures to the same position in the embedding space (refer to Theorem 3.3) to guarantee that we do not lose any candidate anchors for exact subgraph matching. Meanwhile, the GNN model maps non-isomorphic anchor substructures to different positions in the embedding space as much as possible to obtain a small number of candidate anchors. Therefore, it is not necessary to overfit the GNN model like GNN-PE, even the GNN model can be trained with truncated epochs.

$$\mathcal{L}(D) = \sum_{s_{uv} \in D} \|o(s_{uv}) - l(s_{uv})\|_2^2. \quad (6)$$

In each epoch, we obtain embedding vectors for graphs in each batch B , and compute the loss $\mathcal{L}(D)$ (given in Eq. 6) of the GNN model (lines 14-18 of Algorithm 2). Intuitively, from Eq. 6, each anchor substructure in the training data set tends to have a unique embedding as training proceeds.

The Quality of the Anchor Graph Embeddings. In Algorithm 2, the GNN model maps isomorphic anchor substructures to the same embedding position in the embedding space, and maps non-isomorphic anchor substructures to different embeddings as much as possible. Figure 7 illustrates an example of the generated anchor graph embeddings.

EXAMPLE 3.4. In Figure 7, we assume that the value domain of each dimension in the embedding space is $[0, 1]$. Each anchor substructure s_{uv} has a 2D embedding vector $o(s_{uv})$ via the GNN model. For example, anchor substructures $s_{v_1v_4}^P$ and $s_{v_3v_5}^P$ have embedding vectors $o(s_{v_1v_4}^P) = (0.815, 0.752)$ and $o(s_{v_3v_5}^P) = (0.298, 0.334)$ respectively. Although $s_{v_3v_5}^P$ and $s_{v_2v_6}^P$ have different assigned node IDs, they are mapped to the same embedding position $(0.298, 0.334)$ due to their isomorphic graph structures (i.e., $s_{v_3v_5}^P \cong s_{v_2v_6}^P$). This guarantees that we do not lose any candidate anchors for exact subgraph

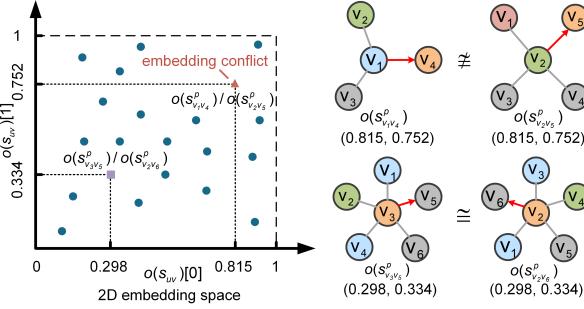


Figure 7: A visualization of anchor graph embeddings.

matching. However, $o(s_{v_1 v_4}^p)$ and $o(s_{v_2 v_5}^p)$ have non-isomorphic graph structures, but they are also mapped to the same embedding position $(0.815, 0.752)$, which incurs an embedding conflict.

An excellent GNN model should minimize occurrences of embedding conflicts. We provide a theoretical guarantee on the quality of these anchor graph embeddings generated by the GNN model.

GNN Model Expressive Power vs. WL Testing: As mentioned earlier, the GNN model naturally satisfies the *order-invariant* (refer to Definition 3.6) property. Here, we provide a rigorous theorem that a GNN model maps isomorphic anchor substructures to the same embedding position in a WL testing perspective.

LEMMA 3.3 ([62]). *Let G and G' be any two non-isomorphic graphs. If a graph neural network $M : G \rightarrow \mathbb{R}^m$ maps G and G' to different embeddings, the Weisfeiler-Lehman graph isomorphism test also decides G and G' are not isomorphic.*

The proof of Lemma 3.3 can be found in [62]. It shows that any aggregation-based GNN is at most as powerful as the WL test in distinguishing non-isomorphic graphs. Based on this, we have the following Theorem 3.3.

THEOREM 3.3. *For two graphs G and G' , if $G \cong G'$, we have $o(G) = o(G')$.*

PROOF. The proof immediately follows from the inverse negative proposition of Lemma 3.3 that if the Weisfeiler-Lehman graph isomorphism test [16, 35] decides that $G \cong G'$, a graph neural network M maps G and G' to the same embedding vector (i.e., $o(G) = o(G')$). \square

Theorem 3.3 shows that for any two isomorphic anchor substructures, the GNN model can map them to the same position in the embedding space, which is a theoretical guarantee that we do not lose any candidate anchors for exact subgraph matching.

THEOREM 3.4. *For any $t \in \mathbb{Z}^+$, if node degrees and the dimension of node features are finite, there exist parameters of t -layers GIN satisfy, for two graphs G and G' , if the 1-WL algorithm distinguishes G and G' as non-isomorphic within t rounds, the embeddings $o(G)$ and $o(G')$ obtained by the GIN are different.*

The proof of Theorem 3.4 can be found in [44, 48, 62]. It shows that the GIN is as strong as the 1-WL isomorphism test algorithm, which provides a theoretical basis for applying a GIN model in our framework for exact subgraph matching. In addition, the GIN

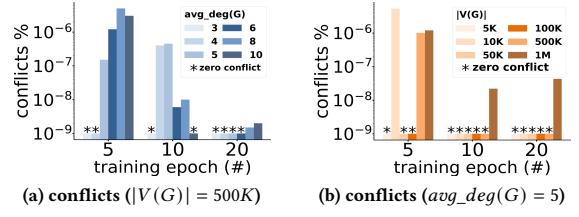


Figure 8: Illustration of anchor graph embedding quality.

is more flexible than the WL algorithm in its ability to map anchor substructures to the embedding space and is able to handle continuous node features.

The GNN Training w.r.t. The Embedding Quality: We train a GNN model (with 2 GIN layers modeled by a 1-layer MLP, SUM as the readout function, the *readout* hidden dimension $n = 10$, and the output embedding vector dimension $m = 3$) over a large training data set D' in Figure 6, where by default $|V(G)| = 500K$, $avg_deg(G) = 5$ and the node label domain size $|\Sigma| = 100$. The learning rate of the Adam optimizer $\eta = 0.001$ and the batch size is $1K \sim 2K$.

Figure 8 reports the anchor graph embedding conflict performance of our GNN-AE. We vary $avg_deg(G)$ from 3 to 10 in Figure 8(a) and adjust $|V(G)|$ from 5K to 1M in Figure 8(b). As described in Algorithm 2, a truncated epoch ρ is set to terminate the training of the GNN model early. We adjust $\rho \in \{5, 10, 20\}$. From the Figure 8, the anchor graph embedding conflict ratio via the GNN model is not greater than $10^{-5}\%$. This means that almost any two non-isomorphic anchor substructures are mapped to non-conflicting positions (refer to Figure 7) in the embedding space.

Complexity Analysis of the GNN Training. In the truncated GNN model training, we obtain the training data set D' in a node-centric view. According to Theorem 3.2, the number of anchor substructures in the training data set D' is $\sum_{v \in V(G) \wedge d_v \leq \theta} (2^{d_v} - 1)$, where θ is an acceptable node degree threshold.

For each GIN layer, let the number of layers in MLP be r . The initial feature embedding dimension of the input and the hidden dimension in MLP are both n . For a graph g_v (with a centric node v) in D' , the complexity of aggregation operations is $O(|E(g_v)|)$ and the complexity of update operations is $O(r \cdot n^2 \cdot |V(g_v)|)$. In a graph g_v , $|E(g_v)| \leq d_v$ and $|V(g_v)| = |E(g_v)| + 1$. Therefore, the time complexity of the computation on the GIN layer is $O(r \cdot n^2 \cdot |V(g_v)| + |E(g_v)|) = O(r \cdot n^2 \cdot d_v)$.

The time complexity of the *readout* function is $O(n \cdot |V(g_v)|)$ and the linear layer is $O(m \cdot n)$. We set the number of truncated epochs to be ρ . Therefore, the total time complexity of the GNN training is given by $O(\rho \cdot \sum_{v \in V(G) \wedge d_v \leq \theta} 2^{d_v} \cdot (r \cdot n^2 \cdot d_v + m \cdot n))$.

3.3 Matching via Anchor Graph Embedding

As described in Algorithm 1, we obtain all anchor substructures that satisfy the anchor-source degree $d_u \leq \theta$ or the anchor-target degree $d_v \leq \theta$ (lines 4-6). Specifically, we divide the anchors ψ_{uv} in the data graph G into 4 cases according to the anchor-source degree d_u and the anchor-target degree d_v .

- If $d_u \leq \theta \wedge d_v \leq \theta$, ψ_{uv} is a sparse-sparse anchor, i.e., ψ_{uv}^{ss} .
- If $d_u \leq \theta \wedge d_v > \theta$, ψ_{uv} is a sparse-dense anchor, i.e., ψ_{uv}^{sd} .
- If $d_u > \theta \wedge d_v \leq \theta$, ψ_{uv} is a dense-sparse anchor, i.e., ψ_{uv}^{ds} .

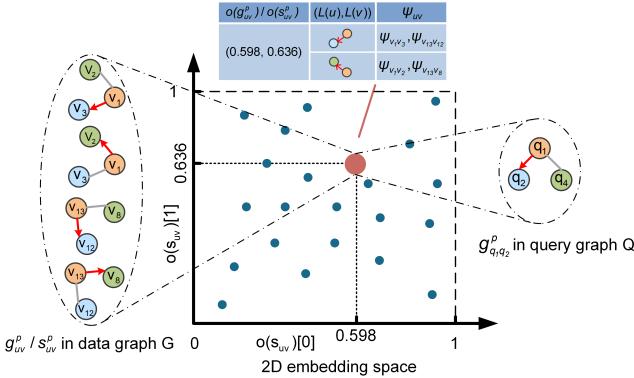


Figure 9: An example of building subgraph matching relationships via anchor graph embeddings of ψ_{uv}^{ss} and ψ_{uv}^{sd} .

- If $d_u > \theta \wedge d_v > \theta$, ψ_{uv} is a dense-dense anchor, i.e., ψ_{uv}^{dd} .

EXAMPLE 3.5. Consider the data graph G in Figure 1, and let the acceptable node degree threshold $\theta = 3$. $\psi_{v_1v_3}$ is a sparse-sparse anchor, where $d_{v_1} = 3 \wedge d_{v_3} = 3$. Similarly, $\psi_{v_2v_5}$ is a sparse-dense anchor, $\psi_{v_5v_2}$ is a dense-sparse anchor, and $\psi_{v_5v_8}$ is a dense-dense anchor, and so on. The data graph G has 16 undirected edges and 32 anchors, including 10 sparse-sparse anchors, 9 sparse-dense anchors, 9 dense-sparse anchors, and 4 dense-dense anchors.

In this subsection, we discuss how to build subgraph matching relationships between the query graph Q and the data graph G for the first 3 anchor cases (i.e., ψ_{uv}^{ss} , ψ_{uv}^{sd} and ψ_{uv}^{ds}) from the anchor graph embeddings (as discussed in subsection 3.2) (lines 12 and 16-19 of Algorithm 1). The fourth case (i.e., ψ_{uv}^{dd}) will be discussed in subsection 3.4.

Sparse-Sparse Anchor & Sparse-Dense Anchor. As described in Algorithm 2, we obtain the positive 1-hop anchor graph g_{uv}^p (and its all positive anchor substructures s_{uv}^p) for sparse-sparse anchors or sparse-dense anchors in the data graph G (lines 3-7). The truncated trained GNN model generates embedding vectors $o(g_{uv}^p)/o(s_{uv}^p)$ for these positive 1-hop anchor graphs/substructures (line 11 of Algorithm 1), which satisfy that isomorphic anchor graphs/substructures must have the same embedding vectors (refer to Theorem 3.3) and non-isomorphic anchor graphs/substructures have different embedding vectors as much as possible. Since anchor graphs/substructures with different anchors may have the same embedding vector, each embedding vector additionally maintains anchor labels (i.e., anchor-source label $L(u)$ and anchor-target label $L(v)$) to further distinguish these anchors.

Figure 9 illustrates an example of the subgraph matching relationship between data graph G and query graph Q via anchor graph embeddings of the sparse-sparse and sparse-dense anchors.

EXAMPLE 3.6. Consider the data graph G and the query graph Q in Figure 1, let the acceptable node degree threshold $\theta = 3$. $\psi_{v_1v_3}$, $\psi_{v_1v_2}$ and $\psi_{v_{13}v_{12}}$ are sparse-sparse anchors, and $\psi_{v_5v_8}$ is a sparse-dense anchor. Their positive 1-hop anchor graph/substructures have isomorphic graph structures and are mapped to the same position (0.598, 0.636) in the 2D embedding space via the GNNs. Each embedding position maintains an entry, which consists of the embedding vector, anchor labels, and anchors. For an anchor $\psi_{q_1q_2}$ in the query Q , we extract

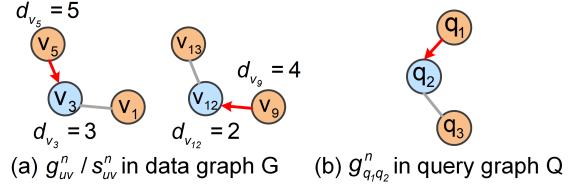


Figure 10: An example of the negative 1-hop anchor graph or substructure of g_{uv}^n/s_{uv}^n of ψ_{uv}^{ds} in the data graph G and g_{uv}^n in the query graph Q .

its positive 1-hop anchor graph $g_{q_1q_2}^p$. Then, $g_{q_1q_2}^p$ is mapped to the same embedding position in the 2D embedding space by the GNN model. After aligning the anchor labels $(L(u), L(v))$, we obtain the candidate anchors $\psi_{v_1v_3}, \psi_{v_{13}v_{12}}$ (with types ψ_{uv}^{ss} or ψ_{uv}^{sd}) in the data graph G for $\psi_{q_1q_2}$.

Dense-Sparse Anchor. For a dense-sparse anchor, its positive 1-hop anchor graph has at least 2^θ anchor substructures (refer to Lemma 3.1). It is unacceptable to extract its positive 1-hop anchor graph and all anchor substructures for a dense-sparse anchor. Therefore, we obtain the negative 1-hop anchor graph g_{uv}^n (and its all negative anchor substructures s_{uv}^n) for dense-sparse anchors in the data graph G (lines 8-11 of Algorithm 2).

EXAMPLE 3.7. Consider the data graph G and the query graph Q in Figure 1, let the acceptable node degree threshold $\theta = 3$. Figure 10(a) illustrates an example of the negative 1-hop anchor graph or substructure of ψ_{uv}^{ds} in the data graph G . In Figure 10(a), the left is an anchor substructure $s_{v_5v_3}^n$ in the negative 1-hop anchor graph $g_{v_5v_3}^n$ of the dense-sparse anchor $\psi_{v_5v_3}$, and the right is a negative 1-hop anchor graph $g_{v_9v_{12}}^n$ of the dense-sparse anchor $\psi_{v_9v_{12}}$. Figure 10(b) shows the negative 1-hop anchor graph $g_{q_1q_2}^n$ of the anchor $\psi_{q_1q_2}$ in the query graph Q .

For the negative 1-hop anchor graph or substructure g_{uv}^n/s_{uv}^n and the negative 1-hop anchor graph $g_{q_1q_2}^n$ in Figure 10, we build subgraph matching relationships between the query graph Q and the data graph G similar to that in Figure 9. g_{uv}^n/s_{uv}^n and $g_{q_1q_2}^n$ in Figure 10 are mapped to the same position in the 2D embedding sapce via the GNN model. After aligning the anchor labels $(L(u), L(v))$, we obtain the candidate anchors $\psi_{v_5v_3}, \psi_{v_9v_{12}}$ (with type ψ_{uv}^{ds}) in the data graph G for $\psi_{q_1q_2}$.

Let the set of negative 1-hop anchor graphs (and their all anchor substructures) of dense-sparse anchors ψ_{uv}^{ds} in G is S_{ds} , and the set of positive 1-hop anchor graphs (and their all anchor substructures) of sparse-sparse or sparse-dense anchors $\psi_{uv}^{ss}/\psi_{uv}^{sd}$ in G is $S_{ss/sd}$. We have the following Proposition 3.1:

PROPOSITION 3.1. Let the graph structure in S_{ds} be s and the graph structure in $S_{ss/sd}$ be s' . $\forall s \in S_{ds}, \exists s' \in S_{ss/sd}$ that satisfies $s' \cong s$.

PROOF. According to Definition 3.2 and 3.3, in terms of graph structure, the positive 1-hop anchor graph g_{uv}^p (only star shape) of ψ_{uv} is equivalent to the negative 1-hop anchor graph g_{vu}^n (only star shape) of ψ_{vu} . Thus, for any positive anchor substructure s_{uv}^p in g_{uv}^p , it is also a negative anchor substructure s_{vu}^n in g_{vu}^n (refer to Definition 3.4 and 3.5).

For a dense-sparse anchor ψ_{uv}^{ds} in the data graph G , the anchor ψ_{vu} is a sparse-dense anchor, i.e., ψ_{vu}^{sd} . Therefore, any negative 1-hop

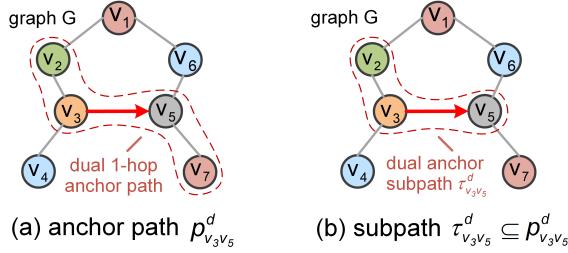


Figure 11: Illustration of the anchor path/subpath.

anchor graph/substructure of ψ_{uv}^{ds} is also a positive 1-hop anchor graph/substructure of ψ_{vu}^{sd} in the data graph G . Since $S_{ss/sd}$ contains the positive 1-hop anchor graph (and their all anchor substructures) of any ψ_{vu}^{sd} in G , for each negative 1-hop anchor graph/substructure s in S_{ds} , there exists a positive 1-hop anchor graph/substructure s' in $S_{ss/sd}$ that satisfies $s' \cong s$. \square

Proposition 3.1 indicates that S_{ds} is a subset of $S_{ss/sd}$, that is, $S_{ds} \subseteq S_{ss/sd}$. Based on this, we can map S_{ds} and $S_{ss/sd}$ to the same embedding space via a GNN model. In other words, as we do in Algorithm 2, we can use only one GNN model to obtain the embedding vectors for negative anchor graphs/substructures of dense-sparse anchors and positive anchor graphs/substructures of sparse-sparse/sparse-dense anchors in the data graph G , instead of training two GNN models for them.

Complexity Analysis. In subsection 3.3, the anchor graph embeddings are built via graphs (i.e., anchor graphs/substructures). We analyze the time complexity of building such anchor graph embeddings (lines 11-12 of Algorithm 1). We build a hash index I_s^s for ψ_{uv}^{ss} & ψ_{uv}^{sd} and a hash index I_s^d for ψ_{uv}^{ds} to store embedding vectors and their additional maintained entries.

As described in subsection 3.2, we traverse each node v that satisfies degree $d_v \leq \theta$ in the data graph G to obtain the anchor graph/substructure set D' . The time complexity of obtaining anchor graph embeddings for these anchor graph/substructure in D' is $O(\sum_{v \in V(G) \wedge d_v \leq \theta} 2^{d_v} \cdot (r \cdot n^2 \cdot d_v + m \cdot n))$ (refer to the complexity analysis of the GNN training in subsection 3.2).

In the anchor graph/substructure set D' , there are some isomorphic graphs. They obtain the same anchor graph embedding vectors via the GNN model. As shown in Figure 9, we merge the anchors of anchor graphs/substructures with the same embedding vector into the same index entry. The time complexity of the merge operation is $O(\sum_{v \in V(G) \wedge d_v \leq \theta} (2^{d_v} - 1))$. Note that due to embedding conflicts, some non-isomorphic anchor graphs/substructures in D' obtain the same embedding vector. The merge operation does not break the exact subgraph matching, but only increases the candidate anchors for the query graph.

In summary, the total time complexity of building anchor graph embeddings is $O(\sum_{v \in V(G) \wedge d_v \leq \theta} 2^{d_v} \cdot (r \cdot n^2 \cdot d_v + m \cdot n))$ (lines 11-12 of Algorithm 1).

3.4 Matching via Anchor Path Embedding

In subsection 3.3, we obtain all anchor graphs/substructures for sparse-sparse anchors, sparse-dense anchors, and dense-sparse anchors. However, for the dense-dense anchor ψ_{uv}^{dd} , it is not acceptable to obtain all anchor substructures in its anchor graph. According to

Lemma 3.1, a dense-dense anchor has at least 2^θ anchor substructures in its positive/negative 1-hop anchor graph.

In this subsection, we introduce anchor path embedding for the dense-dense anchor ψ_{uv}^{dd} in data graph G . Then, we discuss how to build subgraph matching relationships between the query Q and the data graph G via the anchor path embeddings of the dense-dense anchor (lines 8-9, 13-14, and 16-19 of Algorithm 1).

DEFINITION 3.7 (DUAL K-HOP ANCHOR PATH). Given a graph $G = (V, E, L, \Sigma)$ and an anchor ψ_{uv} in G , the dual k -hop anchor path of ψ_{uv} , denoted as $p_d^{(k)}(u, v)$, is defined as a node sequence $(u_k, u_{k-1}, \dots, u_1, u, v, v_1, v_2, \dots, v_k)$, where each node has an edge to the next node in the sequence and the adjacent edges are different.

DEFINITION 3.8 (DUAL ANCHOR SUBPATH). Given an anchor ψ_{uv} and its dual k -hop anchor path $p_d^{(k)}(u, v)$, the dual anchor subpath in $p_d^{(k)}(u, v)$ is defined as a subpath in $p_d^{(k)}(u, v)$ that contains the edge (u, v) of ψ_{uv} , denoted as $\tau_d^{(k)}(u, v)$.

Specifically, in subgraph matching, we adopt the dual 1-hop anchor path by default. For simplicity, we use p_{uv}^d to denote the dual 1-hop anchor path of ψ_{uv} and τ_{uv}^d to denote the dual anchor subpath in p_{uv}^d . Figure 11 shows an example of the dual 1-hop anchor path and its one of the dual anchor subpaths in the graph G . For an anchor in graph G , it may have multiple anchor paths. For an anchor path, it usually has multiple anchor subpaths.

EXAMPLE 3.8. In Figure 11(a), the node sequence (v_2, v_3, v_5, v_7) is a dual 1-hop anchor path of the anchor $\psi_{v_3v_5}$, denoted as $p_{v_3v_5}^d$. In Figure 11(b), the node sequence (v_2, v_3, v_5) forms one of dual anchor subpath of $p_{v_3v_5}^d$, denoted as $\tau_{v_3v_5}^d$. Note that, in Figure 11(a), the node sequences (v_2, v_3, v_5, v_6) , (v_4, v_3, v_5, v_7) and (v_4, v_3, v_5, v_6) are also dual 1-hop anchor paths of the anchor $\psi_{v_3v_5}$.

Based on Definition 3.7, we have the following lemma and proposition of the dual 1-hop anchor path p_{uv}^d .

LEMMA 3.4. For an anchor ψ_{uv} in the graph G , let the degree of anchor-source u be α and the degree of anchor-target v be β . The anchor ψ_{uv} has $\alpha\beta - \alpha - \beta + 1$ dual 1-hop anchor paths.

PROOF. According to Definition 3.7, the dual 1-hop anchor path of ψ_{uv} is a node sequence (u_1, u, v, v_1) . For ψ_{uv} , $u_1 \in \{u_i | u_i \in N(u) \wedge u_i \neq v\}$ and $v_1 \in \{v_j | v_j \in N(v) \wedge v_j \neq u\}$. Thus, the number of dual 1-hop anchor paths is equivalent to the number of combinations of u_i and v_i . So ψ_{uv} has $(\alpha - 1) \cdot (\beta - 1) = \alpha\beta - \alpha - \beta + 1$ dual 1-hop anchor paths. \square

PROPOSITION 3.2. Given a data graph G and an acceptable node degree threshold θ , if the number of dense-dense anchors is M and the max node degree is ξ , then the number of dual 1-hop anchor path $|\mathcal{P}|$ in G is $M \cdot \theta^2 \leq |\mathcal{P}| \leq M \cdot (\xi - 1)^2$.

PROOF. In subsection 3.3, the node degrees on a dense-dense anchor is at least $\theta + 1$. According to the proof of Lemma 3.4, a dense-dense anchor has at least θ^2 dual 1-hop anchor paths. Thus, the lower bound on the number of dual 1-hop paths in G is $M \cdot \theta^2$. Similarly, a dense-dense anchor has at most $(\xi - 1)^2$ dual 1-hop anchor paths. Therefore, the upper bound on the number of dual 1-hop paths in G is $M \cdot (\xi - 1)^2$. \square

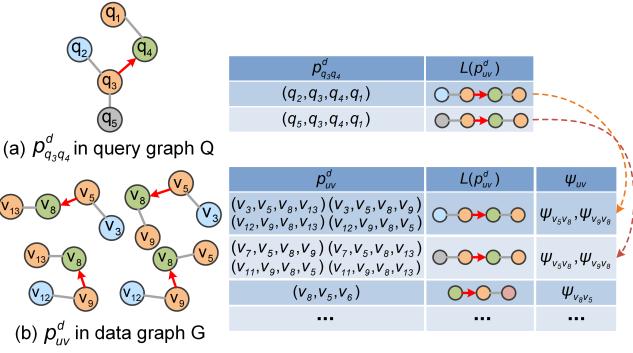


Figure 12: An example of building subgraph matching relationships via anchor path/subpath alignment.

Proposition 3.2 provides the theoretical upper and lower bounds on the number of dual 1-hop paths in the data graph G . For a dense-dense anchor in the data graph G , we obtain its all dual 1-hop anchor paths and extract all dual anchor subpaths for each anchor path (lines 8-9 of Algorithm 1).

Path Alignment & Encoding. Unlike the diverse anchor graph/substructure, the anchor path/subpath has a relatively simple graph structure. It is not cost-effective to obtain embedding vectors for anchor paths/subpaths via a GNN model. Therefore, we provide a *path alignment* to directly build subgraph matching relationships between the query graph Q and the data graph G for the dense-dense anchor ψ_{uv}^{dd} . Figure 12 illustrates an example of the subgraph matching relationship between data graph G and query graph Q .

EXAMPLE 3.9. Consider the data graph G and the query Q in Figure 1, and let the acceptable node degree threshold $\theta = 3$. $\psi_{v_5v_8}$ and $\psi_{v_9v_8}$ are dense-dense anchors. We obtain all their dual 1-hop anchor paths and dual anchor subpaths of each anchor path. In Figure 12, the anchor paths $(v_3, v_5, v_8, v_{13}), (v_3, v_5, v_8, v_9), (v_{12}, v_9, v_8, v_{13})$ and (v_{12}, v_9, v_8, v_5) in G have the isomorphic graph structures. Thus, they are merged into the same entry. For an anchor $\psi_{q_3q_4}$ in Q , there exists a dual 1-hop anchor path (q_2, q_3, q_4, q_1) that can be aligned with anchor paths in the above entry in terms of path labels, so we obtain candidate anchors $\{\psi_{v_5v_8}, \psi_{v_9v_8}\}$ of $\psi_{q_3q_4}$. Similarly, we can align path labels of (q_5, q_3, q_4, q_1) with path labels of anchor paths $(v_7, v_5, v_8, v_9), (v_7, v_5, v_8, v_{13}), \dots$, and obtain candidate anchors $\{\psi_{v_5v_8}, \psi_{v_9v_8}\}$. The intersection of two candidate anchor sets $\{\psi_{v_5v_8}, \psi_{v_9v_8}\}$ is taken as final candidate anchors of $\psi_{q_3q_4}$ in dense-dense anchors.

Specifically, we provide an anchor path encoding for efficient anchor path/subpath alignment. For an anchor ψ_{uv} , there are 5 possible dual 1-hop anchor path patterns. Figure 13 shows the 5 patterns. We encode a dual 1-hop anchor path/subpath as a 4-tuple consisting of the node labels $(L(u_1), L(u), L(v), L(v_1))$ on the path. If a node on the anchor path is missing, we replace it with the value -1. For example, in Figure 13(b), the anchor path is (u, v, v_1) , which is encoded as $(-1, L(u), L(v), L(v_1))$. In particular, for the path pattern (u_1, u, v, v_1) in Figure 13(e), where $u_1 = v_1$, we encode it as $(-2, L(u), L(v), L(v_1))$. These encodings can be viewed as embeddings of the anchor paths/subpaths. If an anchor path from the query Q and the anchor paths/subpaths in the data graph G have the same encoding, they can be aligned.

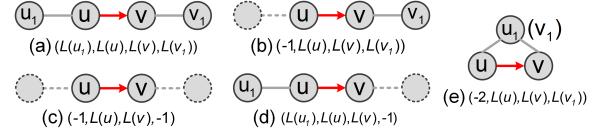


Figure 13: Illustration of dual 1-hop anchor path encoding.

PROPOSITION 3.3. Given a graph $G = (V, E, L, \Sigma)$, the upper bound on the number of dual 1-hop anchor path/subpath embeddings is $|\Sigma|^4 + 3|\Sigma|^3 + |\Sigma|^2$.

PROOF. According to Figure 13, the proof is intuitive. In case (a), the number of possible embeddings is equivalent to the number of combinations of node labels on (u_1, u, v, v_1) . Therefore, the path pattern (a) has $|\Sigma|^4$ embeddings. The same rule applies to case (b)-(e) in Figure 13. Thus, the proposition holds. \square

Complexity Analysis. As described in Algorithm 1, we obtain all dual 1-hop anchor paths p_{uv}^d and their anchor subpaths r_{uv}^d for each dense-dense anchor in the data graph G (lines 8-9). Each anchor path/subpath is encoded and stored in a hash index I_T (lines 13-14 of Algorithm 1). We analyze the time complexity of building such anchor path/subpath embeddings.

In Algorithm 1, we traverse each directed edge (u, v) (i.e., anchor ψ_{uv}) that satisfies degree $d_u > \theta \wedge d_v > \theta$ to obtain the anchor path/subpath set P . For each dense-dense anchor ψ_{uv}^{dd} , the time complexity of extracting its all anchor paths and subpaths is $O(d_u \cdot d_v)$. Therefore, the time complexity of obtaining all anchor paths/subpaths in P is $O(\sum_{\psi_{uv}^{dd} \in \Psi(G)} (d_u \cdot d_v))$, where $\Psi(G)$ is the anchor set of data graph G and $|\Psi(G)| = 2|E(G)|$ (refer to Theorem 3.1).

For each anchor path/subpath in P , we scan node labels to encode its embedding. According to Figure 12, the anchor paths/subpaths with the same embedding are merged into the same index entry. The time complexity of encoding and merging operation is $O(\sum_{\psi_{uv}^{dd} \in \Psi(G)} (d_u \cdot d_v))$.

In summary, the overall time complexity of building anchor path/subpath embeddings is $O(\sum_{\psi_{uv}^{dd} \in \Psi(G)} (d_u \cdot d_v))$, where $\Psi(G)$ is the anchor set of data graph G .

Hybrid Positive & Negative Anchor Path. Generally, the dual 1-hop paths can be adapted to most data graphs, such as WordNet and Human graph data [34, 46]. For the larger data graphs, such as Youtube and US Patents [51], we provide hybrid positive & negative anchor paths with a lower offline build time cost.

DEFINITION 3.9 (POSITIVE K-HOP ANCHOR PATH). Given a graph $G = (V, E, L, \Sigma)$ and an anchor ψ_{uv} in G , the positive k -hop anchor path of ψ_{uv} , denoted as $p_p^{(k)}(u, v)$, is defined as a node sequence $(u_k, u_{k-1}, \dots, u_1, u, v)$, where each node has an edge to the next node in the sequence and the adjacent edges are different.

DEFINITION 3.10 (NEGATIVE K-HOP ANCHOR PATH). Given a graph $G = (V, E, L, \Sigma)$ and an anchor ψ_{uv} in G , the negative k -hop anchor path of ψ_{uv} , denoted as $p_n^{(k)}(u, v)$, is defined as a node sequence $(u, v, v_1, v_2, \dots, v_k)$, where each node has an edge to the next node in the sequence and the adjacent edges are different.

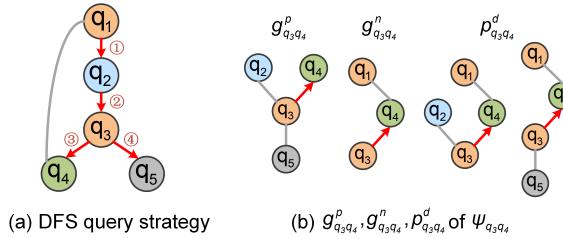


Figure 14: DFS query strategy for query graph Q and anchor graphs/path for $\psi_{q_3q_4}$ in Q .

Similar to Definition 3.8, the positive (resp., negative) anchor subpath is a subpath in $p_p^{(k)}(u, v)$ (resp., $p_n^{(k)}(u, v)$) that contains the edge (u, v) of ψ_{uv} .

Specifically, in subgraph matching, we use positive & negative 1-hop anchor paths. For example, in Figure 11, the node sequence (v_2, v_3, v_5) is a positive 1-hop anchor path of $\psi_{v_3v_5}$ and (v_3, v_5, v_7) is a negative 1-hop anchor path of $\psi_{v_3v_5}$. For a dense-dense anchor in the data graph G , we obtain its all positive & negative 1-hop anchor paths and extract all anchor subpaths for each anchor path. The building principle of the subgraph matching relationship is similar to that in Figure 12. For the anchor $\psi_{q_3q_4}$ in query Q , we extract all positive & negative 1-hop anchor paths and align them with the positive & negative 1-hop anchor paths/subpaths in data graph G to obtain the candidate anchor. However, using hybrid positive & negative 1-hop anchor paths has a lower time complexity for building anchor path/subpath embeddings.

LEMMA 3.5. *For an anchor ψ_{uv} in the graph G , let the degree of anchor-source u be α and the degree of anchor-target v be β . The anchor ψ_{uv} has $\alpha - 1$ positive 1-hop anchor paths and $\beta - 1$ negative 1-hop anchor paths.*

It is obvious Lemma 3.5 holds (refer to the proof of Lemma 3.4). According to Lemma 3.5, for a dense-dense anchor ψ_{uv}^{dd} , the time complexity of extracting its all positive & negative anchor paths and subpaths is $O(d_u + d_v)$. Thus, the time complexity of building hybrid positive & negative anchor path/subpath embeddings is $O(\sum_{\psi_{uv}^{dd} \in \Psi(G)} (d_u + d_v))$ (refer to the complexity analysis in subsection 3.4), where $\Psi(G)$ is the anchor set of data graph G .

4 SUBGRAPH MATCHING WITH GNN-BASED ANCHOR EMBEDDING

In this section, we introduce the details of subgraph matching with anchor (graph & path) embeddings and how to obtain the locations of all matches in parallel (lines 15-21 of Algorithm 1).

4.1 DFS Query Strategy

In this subsection, we present a *depth-first search (DFS) query strategy* to obtain the anchors in query graph Q , which divides the query edges into two categories: anchors and non-anchor edges.

Specifically, starting from a node in the query graph Q , we perform a depth-first search and obtain anchors in the search order. Figure 14(a) provides an example of the DFS query strategy. Consider the query graph Q in Figure 1, we generate a DFS search tree of Q . According to the DFS search process, we obtain the anchors in query graph Q . In Figure 14(a), the query graph Q has four anchors

Algorithm 3: Exact Subgraph Matching with GNN-based Anchor Embedding

```

Input: a query graph  $Q$ , a trained GNN model  $\mathcal{M}$ , and hash indexes  $I_s^s, I_s^d$  and  $I_\tau$  over data graph  $G$ 
Output: a set  $S$  of matching subgraphs
1 obtain query anchor set  $\Psi(Q)$  and query node permutation  $\pi$  from a cost-model-based DFS query plan  $\varphi$ 
2 for each query anchor  $\psi_{q_iq_j} \in \Psi(Q)$  do
3   extract positive/negative 1-hop anchor graphs  $g_{q_iq_j}^p/g_{q_iq_j}^n$ 
4   obtain  $o(g_{q_iq_j}^p)$  and  $o(g_{q_iq_j}^n)$  via GNN model  $\mathcal{M}$ 
5   encode each dual 1-hop anchor path  $p_{q_iq_j}^d$ 
6   obtain the encoding set  $\Omega$  of all dual 1-hop anchor paths
7    $C_{q_iq_j} \leftarrow \text{GetCandidate}(o(g_{q_iq_j}^p), o(g_{q_iq_j}^n), \Omega, I_s^s, I_s^d, I_\tau)$ 
8   add  $\{\psi_{q_iq_j} : C_{q_iq_j}\}$  to  $C$ 
9 return  $\text{MatchGrowth}(\pi, \varphi, C)$ 
10 Function  $\text{GetCandidate}(o(g_{q_iq_j}^p), o(g_{q_iq_j}^n), \Omega, I_s^s, I_s^d, I_\tau)$ 
11   obtain candidate anchor set  $C_{ss/sd}$  by retrieving  $I_s^s$  via key  $o(g_{q_iq_j}^p)$ 
12   obtain candidate anchor set  $C_{ds}$  by retrieving  $I_s^d$  via key  $o(g_{q_iq_j}^n)$ 
13   initial path candidate anchor set  $C_{dd} = \Psi(G)$ 
14   for each path encoding  $\omega \in \Omega$  do
15     obtain candidate anchor set  $C_p$  by retrieving  $I_\tau$  via key  $\omega$ 
16      $C_{dd} \leftarrow C_{dd} \cap C_p$ 
17   return  $C_{ss/sd} \cup C_{ds} \cup C_{dd}$ 
18 Procedure  $\text{MatchGrowth}(\pi, \varphi, C)$ 
19   build the candidate anchor table  $T$  based on  $\varphi$  and  $C$ 
20   initial tree seeds are candidate anchors of first anchor  $\psi_{qq'}$  in  $\varphi$ 
21   for each candidate anchor  $\psi_{uv} \in C_{qq'}$  do // In parallel
22     initial a match tree  $S_t = \psi_{uv}$ 
23     while the depth of  $S_t$ .depth <  $|\pi|$  do // In parallel
24       assemble candidate anchors in  $T$  via hash join
25       if no permutation conflicts & non-anchor match exists then
26         insert candidate anchor nodes to  $S_t$ 
27       else
28         Stop the growth of this branch in  $S_t$ 
29   add all branches  $g$  with depth  $|\pi|$  to  $S$ 
30 return  $S$ 

```

$\{\psi_{q_1q_2}, \psi_{q_2q_3}, \psi_{q_3q_4}, \psi_{q_3q_5}\}$. The remaining undirected edge (q_1, q_4) is a non-anchor edge. For each anchor in query graph Q , we extract its positive 1-hop anchor graph, negative 1-hop anchor graph, and all dual 1-hop anchor paths. Figure 14(b) show an example.

EXAMPLE 4.1. Consider the anchor $\psi_{q_3q_4}$ in query graph Q , its positive 1-hop anchor graph $g_{q_3q_4}^p$ is a induced star subgraph formed by the node set $\{q_2, q_3, q_4, q_5\}$. The negative 1-hop anchor graph $g_{q_3q_4}^n$ is a induced star subgraph formed by the node set $\{q_1, q_3, q_4\}$. The anchor $\psi_{q_3q_4}$ has 2 dual 1-hop paths (q_2, q_3, q_4, q_1) and (q_5, q_3, q_4, q_1) .

4.2 Anchor Matching Mechanism

In this subsection, we discuss how to obtain candidate anchors over data graph G for the anchor in query graph Q . The principles have been described in subsection 3.3 - 3.4. Algorithm 3 shows the whole process of getting candidate anchors (lines 2-7 and 10-17).

Specifically, for each query anchor $\psi_{q_iq_j}$ in query graph Q , we extract its positive (and negative) 1-hop anchor graphs $g_{q_iq_j}^p$ (and $g_{q_iq_j}^n$) and obtain embedding vectors $o(g_{q_iq_j}^p)$ (and $o(g_{q_iq_j}^n)$) via the trained GNN model (lines 3-4 of Algorithm 3). Then, all dual 1-hop anchor paths $p_{q_iq_j}^d$ of $\psi_{q_iq_j}$ are extracted and encoded (lines 5-6 of

Algorithm 3). The hash indexes I_s^s , I_s^d and I_t are built in the offline pre-computation phase (refer to lines 1-14 of Algorithm 1). They maintain the entries in Figure 9 or Figure 12. In GetCandidate, we take $o(g_{q_i q_j}^p)$ as the key to retrieve I_s^s to obtain sparse-sparse & sparse-dense candidate anchors. In a similar way, we use $o(g_{q_i q_j}^n)$ to retrieve I_s^d to obtain dense-sparse candidate anchors (lines 11-12 of Algorithm 3). To obtain dense-dense candidate anchors, we retrieve I_t with the encoding of each dual 1-hop anchor path as the key, and calculate the intersection of candidate anchors obtained by all anchor paths (lines 13-16 of Algorithm 3).

According to the introduction in subsection 3.3, the intersection of the 4 anchor types (i.e., ψ_{uv}^{ss} , ψ_{uv}^{sd} , ψ_{uv}^{ds} and ψ_{uv}^{dd}) in data graph G is \emptyset and their union is the anchor set $\Psi(G)$ of G . Thus, we have the proposition of candidate anchors below.

PROPOSITION 4.1. *Given the data graph G and an anchor $\psi_{q_i q_j}$ in query graph Q , the candidate anchor set of $\psi_{q_i q_j}$ on G is $C_{q_i q_j} = C_{ss/sd} \cup C_{ds} \cup C_{dd}$, where $C_{ss/sd} \cap C_{ds} \cap C_{dd} = \emptyset$.*

Graph-based Matching. In subsection 3.3, we illustrate the principle of building subgraph matching relationships via anchor graph embeddings based on anchor graphs/substructures. We can derive a necessary condition of an anchor ψ_{uv} in data graph G to be a match of an anchor $\psi_{q_i q_j}$ in query graph Q . Therefore, we have the following Lemma 4.1 and Lemma 4.2.

LEMMA 4.1 (POSITIVE GRAPH CANDIDATE ANCHORS). *Given an anchor ψ_{uv} in data graph G and an anchor $\psi_{q_i q_j}$ in query graph Q , the anchor ψ_{uv} can be a candidate anchor of $\psi_{q_i q_j}$, if ψ_{uv} has the positive 1-hop anchor graph g_{uv}^p (or an anchor substructure s_{uv}^p) such that $o(g_{uv}^p)$ (or $o(s_{uv}^p)$) = $o(g_{q_i q_j}^p)$ and $L(u) = L(q_i) \wedge L(v) = L(q_j)$.*

LEMMA 4.2 (NEGATIVE GRAPH CANDIDATE ANCHORS). *Given an anchor ψ_{uv} in data graph G and an anchor $\psi_{q_i q_j}$ in query graph Q , the anchor ψ_{uv} can be a candidate anchor of $\psi_{q_i q_j}$, if ψ_{uv} has the negative 1-hop anchor graph g_{uv}^n (or an anchor substructure s_{uv}^n) such that $o(g_{uv}^n)$ (or $o(s_{uv}^n)$) = $o(g_{q_i q_j}^n)$ and $L(u) = L(q_i) \wedge L(v) = L(q_j)$.*

Path-based Matching. In subsection 3.4, we illustrate the principle of building subgraph matching relationships via anchor path embeddings based on anchor paths/subpaths. Similarly, we can derive another necessary condition of an anchor ψ_{uv} in data graph G to be a match of an anchor $\psi_{q_i q_j}$ in query graph Q . Thus, we have the Lemma 4.3 of path-based candidate anchors below.

LEMMA 4.3 (DUAL PATH CANDIDATE ANCHORS). *Given an anchor ψ_{uv} in data graph G and an anchor $\psi_{q_i q_j}$ in query graph Q , the anchor ψ_{uv} can be a candidate anchor of $\psi_{q_i q_j}$, if for any dual 1-hop anchor path $p_{q_i q_j}^d$ of $\psi_{q_i q_j}$, ψ_{uv} has a dual 1-hop anchor path p_{uv}^d aligned with $p_{q_i q_j}^d$.*

In addition to the dual 1-hop anchor path, the hybrid positive & negative anchor paths also have similar corollaries as Lemma 4.3. Here we do not repeat them.

4.3 Matching Growth Algorithm

In this subsection, we present the effective matching growth method to obtain all exact matches for the query graph Q (lines 18-30 of Algorithm 3). In subsection 4.2, Lemmas 4.1- 4.3 ensure that there

$\psi(Q)$	$\psi_{q_1 q_2}$	$\psi_{q_2 q_3}$	$\psi_{q_3 q_4}$	$\psi_{q_4 q_5}$
$C_{q_i q_j}$	(V_1, V_3)	(V_3, V_1)	(V_5, V_2)	(V_5, V_7)
	(V_5, V_3)	(V_3, V_5)	(V_5, V_8)	(V_9, V_{11})
	(V_9, V_{12})	(V_{12}, V_9)	(V_9, V_8)	
	(V_{13}, V_{12})	(V_{12}, V_{13})		

Figure 15: An example of the table of candidate anchors.

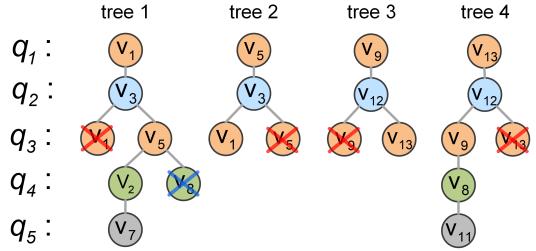


Figure 16: An example of the match tree growth.

are no false dismissals in candidate anchors for each anchor in the query Q . The matching growth method assembles these candidate anchors to output all exact matches.

Consider the data graph G and query graph Q in Figure 1. According to the anchor matching mechanism in subsection 4.2, we obtain candidate anchors for each anchor in query graph Q . Figure 15 illustrates the candidate anchors. We organize them into a table based on the order of query anchors obtained by the DFS query strategy (line 19 of Algorithm 3).

Hash Join & Refinement Pruning. We take candidate anchors $C_{qq'}$ of first anchor $\psi_{qq'}$ in the DFS query plan ϕ as initial tree seeds (line 20 of Algorithm 3). For example, in Figure 15, the candidate anchors $\{(v_1, v_3), (v_5, v_3), (v_9, v_{12}), (v_{13}, v_{12})\}$ of $\psi_{q_1 q_2}$ are selected as initial seeds of the match tree. For each tree seed, we perform the *hash join* on the table of candidate anchors to grow the match tree. Figure 16 shows an example of the match tree growth.

EXAMPLE 4.2. Consider the data graph G and the query graph Q in Figure 1, we obtain query node permutation $\pi = (q_1, q_2, q_3, q_4, q_5)$ according to the search order of the DFS query strategy. We grow the match tree along the query node permutation π . In Figure 15, the anchor $\psi_{q_1 q_2}$ has 4 candidate anchors, so there are 4 match trees required to grow. All exact matches are produced in these match trees.

In Figure 16, there are 3 cases in which branches stop growing in the match tree.

Case 1. (Permutation Conflict Pruning) If the matches of two query nodes along the query node permutation π conflict, the branch stops growing (or is pruned).

EXAMPLE 4.3. The match $\{(q_1, v_1), (q_2, v_3), (q_3, v_1)\}$ in tree 1 has a conflict and this branch stops growing.

Case 2. (Non-anchor Edges Pruning) If a non-anchor edge (q_i, q_j) exists between the query node q_j and its forward query node q_i in permutation π , we check whether there is an edge (v_i, v_j) between their corresponding matches in the data graph G . If (v_i, v_j) does not exist, the branch stops growing (or is pruned).

EXAMPLE 4.4. For the match $\{(q_1, v_1), (q_2, v_3), (q_3, v_5), (q_4, v_8)\}$ in tree 1, there is a non-anchor edge (q_1, q_4) between q_4 and q_1 , but no edge between v_1 and v_8 in G . Thus, the branch (v_1, v_3, v_5, v_8) in tree 1 stops growing.

Case 3. (Hash Join Failure Pruning) If the anchor-target v of a candidate anchor ψ_{uv} cannot find a joined anchor-source in the hash join operation, the branch stops growing (or is pruned).

EXAMPLE 4.5. For the match $\{(q_1, v_5), (q_2, v_3), (q_3, v_1)\}$ in tree 2, the match v_1 cannot find a joined anchor-source in candidate anchors of $\psi_{q_3q_4}$, so this branch stops growing.

Note that since the anchor-source of $\psi_{q_3q_5}$ is q_3 , its candidate anchors should perform a *hash join* with the candidate anchors of the anchor whose anchor-target is q_3 . That is, we should perform a *hash join* between the candidate anchors of $\psi_{q_2q_3}$ and $\psi_{q_3q_5}$ to obtain the match for q_5 . Finally, all branches g with depth $|V(Q)|$ are all exact matches (line 29 of Algorithm 3).

Parallel Growth. For the query graph Q , its candidate anchors for each anchor have been obtained via the anchor (graph & path) embedding techniques. In the match tree growth, the match of each query node q_i only relies on the matches of its forward query nodes in the permutation π . Thus, the growth of each match tree is independent of each other, and the growth of each branch in the tree is also independent of each other. We have the Lemma 4.4.

LEMMA 4.4 (GROWTH ISOLATION). All match trees or branches in the tree grow isolated from each other.

As described in Algorithm 3, we can arrange a thread for each match tree to grow them in a parallel manner (line 21). In addition, we can also perform the growth of branches in a tree in parallel to further improve efficiency (line 23 of Algorithm 3).

Complexity Analysis. In Algorithm 3, the time complexity of obtaining query anchor set $\Psi(Q)$ and query node permutation π (line 1) is given by $O(|V(Q)| + |E(Q)|)$. For each query anchor $\psi_{q_iq_j} \in \Psi(Q)$, we extract its positive/negative 1-hop anchor graphs. Since the GNN computation cost is $O(|E(g_{q_iq_j})|)$ [61], the time complexity of obtaining embedding vectors of anchor graphs $g_{q_iq_j}^p/g_{q_iq_j}^n$ in Q via the GNNs (lines 2-4) is given by $O(\sum_{\psi_{q_iq_j} \in \Psi(Q)} (d_{q_i} + d_{q_j}))$, where $\Psi(Q)$ is the anchor set of Q . According to Lemma 3.4, a query anchor $\psi_{q_iq_j}$ has $(d_{q_i} - 1) \cdot (d_{q_j} - 1)$ dual 1-hop anchor paths. Thus, the time complexity of encoding all dual 1-hop anchor paths for Q is given by $O(\sum_{\psi_{q_iq_j} \in \Psi(Q)} (d_{q_i} \cdot d_{q_j}))$.

For the GetCandidate function (lines 10-17), the time cost of retrieving hash indexes I_s^s, I_s^d and I_t is $O(1)$. We assume that the encoding set of all dual 1-hop anchor paths for $\psi_{q_iq_j}$ is $\Omega_{q_iq_j}$. The time complexity of obtaining path candidate anchor set C_{dd} (lines 14-16) for $\psi_{q_iq_j}$ is given by $O(\sum_{\omega \in \Omega_{q_iq_j}} |C_p^\omega|)$, where $|C_p^\omega|$ is the number of candidate anchors for the path encoding $\omega \in \Omega_{q_iq_j}$. Therefore, the time complexity of the GetCandidate function is given by $O(\sum_{\psi_{q_iq_j} \in \Psi(Q)} (\sum_{\omega \in \Omega_{q_iq_j}} |C_p^\omega| + |C_{ss/sd}(\psi_{q_iq_j})| + |C_{ds}(\psi_{q_iq_j})|))$, where $|C_{ss/sd}(\psi_{q_iq_j})|$ is the number of sparse-sparse & sparse-dense candidate anchors for $\psi_{q_iq_j}$ and $|C_{ds}(\psi_{q_iq_j})|$ is the number of dense-sparse candidate anchors for $\psi_{q_iq_j}$.

Finally, we use hash join to assemble candidate anchors and perform refinement pruning in the MatchGrowth procedure (lines 18-30) to obtain all exact matches. The time complexity is given by

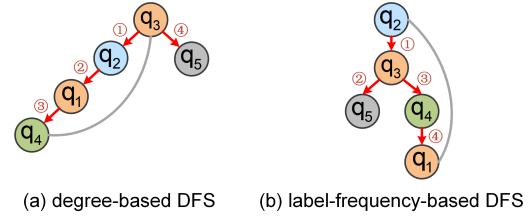


Figure 17: An example of the degree-based and label-frequency-based DFS query plan.

Algorithm 4: Cost-Model-Based DFS Query Plan

```

Input: a query graph  $Q$ 
Output: a query anchor set  $\Psi(Q)$  in the query plan  $\varphi$ 
1 initial  $\Psi(Q) = \emptyset$  and  $Cost(\varphi) = +\infty$ 
// apply MaxDeg, MinLF, or Rand strategy in subsection 5.2
2 obtain a set  $M_{st}$  of DFS start nodes
3 for each start node  $\mu \in M_{st}$  do
4   initial  $\Psi_{tmp}(Q) = \emptyset$  and  $Cost_{tmp}(\varphi) = 0$ 
5   perform DFS to select a node  $q_i$  with minimum query cost  $f(\psi_{\mu q_i})$ 
6   add  $\psi_{\mu q_i}$  to  $\Psi_{tmp}(Q)$ 
7   while at least one node in  $V(Q)$  is not visited do
8     continue DFS to select a node with minimum query cost  $f(\psi_{q_i q_j})$ 
9     add  $\psi_{q_i q_j}$  to  $\Psi_{tmp}(Q)$ 
10     $Cost_{tmp}(\varphi) \leftarrow Cost_{tmp}(\varphi) + f(\psi_{q_i q_j})$ 
11   if  $Cost_{tmp}(\varphi) < Cost(\varphi)$  then
12      $\Psi(Q) \leftarrow \Psi_{tmp}(Q)$ 
13      $Cost(\varphi) \leftarrow Cost_{tmp}(\varphi)$ 
14 return  $\Psi(Q)$ 

```

$O(\sum_{\psi_{q_iq_j} \in \Psi(Q)} (|C_{ss/sd}(\psi_{q_iq_j})| + |C_{ds}(\psi_{q_iq_j})| + |C_{dd}(\psi_{q_iq_j})|))$, in which $|C_{dd}(\psi_{q_iq_j})|$ is the number of dense-dense candidate anchors.

Therefore, the overall time complexity of Algorithm 3 is given by $O(\sum_{\psi_{q_iq_j} \in \Psi(Q)} (\sum_{\omega \in \Omega_{q_iq_j}} |C_p^\omega| + |C_{q_iq_j}| + d_{q_i} d_{q_j}))$, where $|C_{q_iq_j}|$ is the number of candidate anchors for query anchor $\psi_{q_iq_j}$.

5 COST-MODEL-BASED DFS QUERY PLAN

In this section, we provide a cost-model-based DFS query plan (line 15 of Algorithm 1) to enhance the parallel matching growth algorithm in subsection 4.3.

5.1 Cost Model

In this subsection, we design a formal cost model to estimate the query cost of the query plan φ . According to the complexity analysis in subsection 4.3, the query cost of subgraph matching is mainly limited by the number of candidate anchors $|C_p^\omega|$ for the anchor path and the number of candidate anchors $|C_{q_iq_j}|$ for each anchor $\psi_{q_iq_j}$ in the query graph Q . Thus, intuitively, fewer candidate anchors for an anchor would lower query cost.

Therefore, we define the query cost $Cost(\varphi)$ for the query plan φ as follows:

$$Cost(\varphi) = \sum_{\psi_{q_iq_j} \in \Psi(Q)} f(\psi_{q_iq_j}), \quad (7)$$

where $\Psi(Q)$ is the query anchor set and $f(\psi_{q_i q_j})$ is the query cost of a query anchor $\psi_{q_i q_j}$. We aim to find a good DFS query plan φ to minimize the total query cost $Cost(\varphi)$ given in Eq. 7.

Anchor Query Cost. We discuss how to compute the query cost $f(\psi_{q_i q_j})$ for a query anchor $\psi_{q_i q_j}$ in Eq. 7. We provide two intuitive and lightweight methods as examples.

Intuitively, for an anchor $\psi_{q_i q_j}$, if the degrees of the anchor-source q_i and the anchor-target q_j are high, the number of candidate anchors of $\psi_{q_i q_j}$ is expected to be small, which incurs a low query cost. Thus, we can set $f(\psi_{q_i q_j}) = -(d_{q_i} + d_{q_j})$, where d_{q_i} is the degree of node q_i and d_{q_j} is the degree of node q_j .

Alternatively, we can adopt other query cost metrics, such as node label frequency. We count and record the node label frequency in the data graph G offline. In the online subgraph matching phase, for an anchor $\psi_{q_i q_j}$ in the query Q , we can compute the minimum label frequency in the 1-hop neighborhood of the anchor-source q_i and the minimum label frequency in the 1-hop neighborhood of the anchor-target q_j . We can thus set $f(\psi_{q_i q_j}) = \min(\Gamma(q_i)) + \min(\Gamma(q_j))$, where $\Gamma(q_i)$ is the node label frequency set in the 1-hop neighborhood of q_i and $\Gamma(q_j)$ is the node label frequency set in the 1-hop neighborhood of q_j .

Figure 17 shows an example of the degree-based DFS query plan and the label-frequency-based DFS query plan. Consider the data graph G and the query graph Q in Figure 1. In Figure 17(a), we assume that DFS starts from the node q_3 with the maximum degree in query Q . In the DFS procedure, we select the node with the minimum anchor query cost $f(\psi_{q_i q_j}) = -(d_{q_i} + d_{q_j})$ from the neighbors of the current node as the next node to traverse. We thus obtain a degree-based DFS query plan and the anchors $\{\psi_{q_3 q_2}, \psi_{q_2 q_1}, \psi_{q_1 q_4}, \psi_{q_3 q_5}\}$. Similarly, in Figure 17(b), we assume that DFS starts from the node q_2 (or q_5) with the minimum node label frequency. We can obtain a label-frequency-based DFS query plan and the anchors $\{\psi_{q_2 q_3}, \psi_{q_3 q_5}, \psi_{q_3 q_4}, \psi_{q_4 q_1}\}$ by using the anchor query cost function $f(\psi_{q_i q_j}) = \min(\Gamma(q_i)) + \min(\Gamma(q_j))$.

5.2 DFS Query Plan

Algorithm 4 illustrates the cost-model-based (given in Eq. 7) DFS query plan selection. For a query graph Q , Algorithm 4 returns a query anchor set $\Psi(Q)$. Specifically, we initialize $\Psi(Q) = \emptyset$ and query cost $Cost(\varphi)$ for the query plan φ (line 1). Then, we select a node set M_{st} in the query Q , each of which can be used as a start node of DFS (line 2). For a start node $\mu \in M_{st}$, we perform DFS and minimize the query cost $f(\psi_{q_i q_j})$ at each step expansion (lines 4-10). For different start nodes in M_{st} , we always keep the best-so-far query anchor set in $\Psi(Q)$ and the minimum query cost in $Cost(\varphi)$ (lines 11-13). Finally, we obtain a heuristic query anchor set $\Psi(Q)$ with the lowest query cost (line 14).

DFS Start Node Selection Strategy. We provide three strategies to select the set M_{st} of DFS start nodes (line 2 of Algorithm 4).

- **Top-K Maximum Degree Nodes (MaxDeg):** select the top- k nodes with maximum degree in the query graph Q .
- **Top-K Minimum Label Frequency Nodes (MinLF):** select the top- k nodes with the minimum label frequency (in the data graph G) in the query graph Q .
- **K Random Nodes (Rand):** randomly select k nodes in the query graph Q to form M_{st} .

Table 2: Properties of real-world graph datasets

Category	Dataset	$ V(G) $	$ E(G) $	$ \Sigma $	$avg_deg(G)$
Biology	Yeast (ye)	3,112	12,519	71	8.0
	HPRD (hp)	9,460	34,998	307	7.4
Lexical	WordNet (wn)	76,853	120,399	5	3.1
Social	DBLP (db)	317,080	1,049,866	15	6.6
	Youtube (yt)	1,134,890	2,987,624	25	5.3
Citation	US Patents (up)	3,774,768	16,518,947	20	8.8

Table 3: Parameter settings

Parameter	Values
acceptable node degree threshold θ	6, 8, 10 , 12, 14, 16
the size, $ V(Q) $, of the query graph Q	4, 6, 8 , 10, 12, 16, 24, 32
$avg_deg(G)$ of the synthetic data graph G	3, 4, 5, 6, 8
the number, $ \Sigma $, of the synthetic data graph G	20, 50, 100 , 200, 300, 500
the size, $ V(G) $, of the synthetic data graph G	10K, 50K, 80K , 100K, 500K, 1M

6 EXPERIMENTAL EVALUATION

In this section, we evaluate the effectiveness and efficiency of our proposed GNN-AE approach.

6.1 Experimental Setup

To evaluate our GNN-AE, we conduct subgraph matching experiments on a Linux machine equipped with an Intel Xeon E5-2620 v4 2.1GHz CPU and 220GB memory. The GNN model of GNN-AE is implemented by PyTorch, where embedding vectors are offline computed on a Linux machine equipped with an Nvidia GeForce GTX 2080 Ti. The online subgraph matching is implemented in C++, which is consistent with previous work GNN-PE [66].

For the GNN model (as mentioned in subsection 3.2), we set 2 GIN layers modeled by a 1-layer MLP, the hidden dimension of the *readout* function $n = 10$, and the output embedding vector dimension $m = 3$. During the training process, the truncated epochs $\rho = 20$ by default. We use the Adam optimizer to update parameters and set the learning rate $\eta = 0.001$. Due to different data sizes, we adjust batch sizes for different datasets from 1,024 to 2,048. In the matching growth algorithm, we adopt a parallel mode with 8 threads. For the query plan, we adopt the degree-based DFS query plan and the DFS start node selection strategy with MaxDeg.

Baseline Methods. We compare the performance of our GNN-AE with the following 7 representative subgraph matching baseline methods: GraphQL (GQL) [26], QuickSI (QSI) [49], RI [11], Hybrid (a combination of candidate filtering, vertex ordering and enumeration of CFL [9], RI [11] and GQL respectively) [51], RapidMatch (RM) [53], BICE [13], and GNN-PE [66]. Since the RL-QVO [58] relies on historical query graphs to generate the matching order for subgraph matching, we do not take it as one of the baselines.

Real/Synthetic Datasets. We conduct experiments on 6 standard real-world data graphs from various domains and 3 synthetic data graphs to evaluate our GNN-AE approach and baselines.

Real-world Graphs: We selected six real-world datasets from four categories used by previous works [9, 23, 24, 26, 49, 51, 55, 72]. Table 2 lists the properties of the real-world graph datasets. For the

anchor path embedding, we use dual 1-hop anchor paths for small data graphs (e.g., Yeast, HPRD, WordNet, and DBLP) by default. For large data graphs (e.g., Youtube, and US Patents), we use the hybrid positive & negative anchor paths (as mentioned in subsection 3.4).

Synthetic Graphs: The synthetic graphs were generated by the NetworkX [21]. We produce three synthetic graphs: a random regular graph (Syn-RG), a small-world graph following the Newman-Watts-Strogatz model (Syn-WS) [59], and a scale-free graph following the Barabasi-Albert model (Syn-BA) [7]. To evaluate the GNN-AE efficiency, we vary the characteristics of synthetic graphs. Table 3 depicts the parameters of synthetic graphs, where default parameter values are in bold. In the evaluation, we vary one parameter value while keeping the other parameters at default values.

Query Graphs. We follow previous works [9, 13, 23, 50, 51, 53, 66] to generate query graphs for each data graph G . Specifically, for each dataset, we perform a random walk on G until getting the specified number of nodes and extract the induced subgraph. Table 3 lists the size $|V(Q)|$ of the query graph Q . Except for query graphs with $|V(Q)| = 4$, each query set on real-world graphs contains two categories: dense query graph (i.e., $\text{avg_deg}(Q) > 3$) and sparse query graph (i.e., $\text{avg_deg}(Q) \leq 3$). For each query size and category, we generate 100 query graphs. Since extracting dense query graphs from synthetic graphs is difficult, we extract queries without distinguishing between dense and sparse on synthetic graphs.

Evaluation Metrics. In our experiments, we report the efficiency of our GNN-AE approach and baselines. Due to the complexity of some query graphs, it cannot be found all matches within a limited time. We follow previous works [5, 9, 23, 30, 31, 51–53] to terminate the search for a query graph when 10^5 matches are discovered. We report the results of the query graph which can be processed within five minutes so that our evaluation can be finished in a reasonable time. We also evaluate the *filtering power* of our anchor embeddings (including both anchor graph embeddings and anchor path embeddings), which is the percentage of candidate anchors that can be filtered out by our anchor embeddings. Specifically, given a data graph $G = (V, E, L, \Sigma)$, let the number of candidate anchors for a query anchor $\psi_{q_i q_j}$ be $|C_{q_i q_j}|$, which obtained via anchor embeddings. For the query anchor $\psi_{q_i q_j}$, the number of matched anchors obtained from all its exact matching results is $|M_{q_i q_j}|$. Thus, the filtering power, $FP_{q_i q_j}$, of anchor embeddings for the query anchor $\psi_{q_i q_j}$ is

$$FP_{q_i q_j} = \frac{2|E| - |C_{q_i q_j}|}{2|E| - |M_{q_i q_j}|}. \quad (8)$$

where $2|E|$ is the number of anchors in G (refer to Theorem 3.1). We also report offline pre-computation costs of our GNN-AE, including the time to obtain anchor structures, the GNN training time, anchor (graph & path) embedding time, and index construction time.

6.2 Parameter Tuning

In this subsection, we tune parameters for our GNN-AE approach. **The GNN-AE Efficiency w.r.t Node Degree Threshold θ .** In Section 3, we set an acceptable node degree threshold θ on the data graph. For the dense-dense anchor ψ_{uv} with $d_u > \theta \wedge d_v > \theta$, we obtain its anchor path embeddings, and anchor graph embeddings for other anchor types (i.e., sparse-sparse, sparse-dense, and

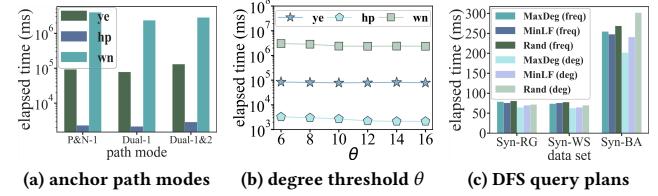


Figure 18: GNN-AE efficiency w.r.t. varying parameters θ , anchor path patterns, and DFS query plan strategies.

dense-sparse anchors). Figure 18(b) illustrates the GNN-AE efficiency on three real-world graphs, by varying the acceptable node degree threshold θ from 6 to 16. When θ increases, more anchors on the data graph obtain anchor embeddings via the graph rather than the path, which incurs higher filtering power for the anchor embeddings. The elapsed time of our GNN-AE approach shows a downward trend. However, when θ exceeds a certain threshold (e.g., 12), the performance of our GNN-AE approach is not sensitive to θ . This is because those anchors with high node degrees often have a low probability of being filtered out. The anchor embeddings via the graph or the path have similar filtering power for them.

The GNN-AE Efficiency w.r.t Anchor Path Modes. Figure 18(a) illustrates the efficiency of our GNN-AE approach with different anchor path modes on three real-world graphs. In addition to the dual 1-hop anchor path (i.e., Dual-1) and the hybrid positive & negative 1-hop anchor paths (i.e., P&N-1) modes in subsection 3.4, we also evaluate the hybrid dual 1 & 2-hop anchor paths mode (i.e., Dual-1&2). That is, given an anchor ψ_{uv} of the data graph, if $\theta < d_u \leq \zeta \vee \theta < d_v \leq \zeta$, then we adopt dual 2-hop anchor paths; if $d_u \geq \zeta \wedge d_v \geq \zeta$, then we adopt dual 1-hop anchor paths. we set $\zeta = 40$ by default. It is obvious that the dual 1-hop anchor path has a stronger filtering power than the hybrid positive & negative 1-hop anchor paths, which incurs higher performance of our GNN-AE approach. Although the hybrid dual 1 & 2-hop anchor paths have a stronger filtering power than the dual 1-hop anchor path in principle, our GNN-AE has to operate dual 1-hop and 2-hop anchor paths for each anchor in the query graph in the hybrid dual 1 & 2-hop anchor paths mode. Thus, the hybrid dual 1 & 2-hop anchor paths mode does not improve our GNN-AE performance.

The GNN-AE Efficiency w.r.t DFS Query Plan Strategies. In Section 5, we provide two methods (i.e., degree-based DFS query plan, and label-frequency-based DFS query plan) to compute the query cost $f(\psi_{q_i q_j})$ in Eq. 7 and three DFS start node selection strategies (i.e., MaxDeg, MinLF, and Rand). Figure 18(c) shows the performance of our GNN-AE approach with different DFS query plan strategies on three synthetic graphs. Generally, degree-based query cost methods perform better than label-frequency-based ones on three synthetic graphs. This implies that the degree-based query cost method may obtain fewer candidate anchors for each anchor in the query graph. In addition, the combination MaxDeg (deg) obtains the optimal performance. Thus, we adopt MaxDeg (deg) as our default DFS query plan strategy in the following experiments.

6.3 Efficiency and Effectiveness Evaluation

In this subsection, we evaluate the efficiency and effectiveness of our GNN-AE approach.

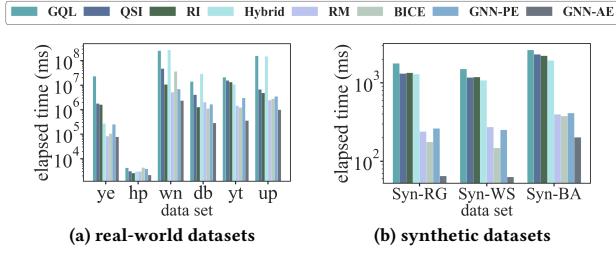


Figure 19: GNN-AE efficiency on real/synthetic datasets.

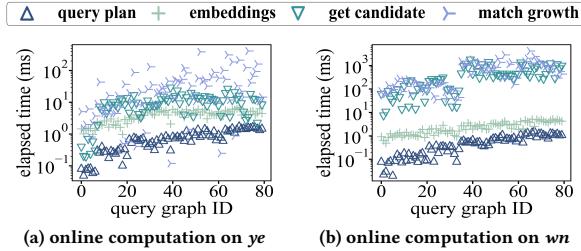


Figure 20: The GNN-AE online computation cost.

The GNN-AE Efficiency on Real/Synthetic Datasets. In Figure 19 and Figure 20, we report the efficiency of our GNN-AE approach on online subgraph matching.

Compare with Baselines: Figure 19 illustrates the efficiency comparisons between our GNN-AE and 7 baseline methods over real-world and synthetic datasets. Overall, our GNN-AE consistently outperforms baseline methods. Specifically, for some real-world (e.g., *db*, *yt*, and *up*) and synthetic (e.g., *Syn-RG*, and *Syn-WS*) datasets, our GNN-AE can perform better than baselines by up to 1-2 orders of magnitude. In particular, our GNN-AE significantly outperforms traditional exploration-based backtracking search methods (e.g., GQL, QSI, and Hybrid).

In subsequent experiments, we also evaluate the efficiency of our GNN-AE under varying query and data graph characteristics (e.g., query graph patterns, $|V(Q)|$, $|V(G)|$, $\text{avg_deg}(G)$, and $|\Sigma|$ of G). To better show the curve trends, we omit the baseline results.

Online Computation Cost: Figure 20 reports the computation cost of four steps (including DFS query plan, obtain anchor embeddings, get candidate anchors, and match growth) in the online subgraph matching phase (i.e., Algorithm 3). We select 80 query graphs of the real-world graphs *ye* and *wn* respectively, and record the elapsed time of four steps for each query graph. From each subfigure, we find that the get candidate anchors and match growth steps take up most of the elapsed time for subgraph matching, which is consistent with the complexity analysis in the subsection 4.3. For the DFS query plan and obtaining anchor embeddings, the time cost of our GNN-AE remains low (i.e., $< 1.7\text{ ms}$, and $< 6.5\text{ ms}$, respectively).

The GNN-AE Efficiency w.r.t Query Graph Patterns. Figure 21 compares the efficiency of our GNN-AE approach with that of 7 baseline methods on different query graph patterns, including dense (i.e., $\text{avg_deg}(Q) > 3$, denoted as *D*) and sparse (i.e., $\text{avg_deg}(Q) \leq 3$, denoted as *S*) query graphs with varying sizes. We take close observation of seven query graph patterns ranging from 4S to 24S on the real-world graph *wn*. The figure shows that our GNN-AE can

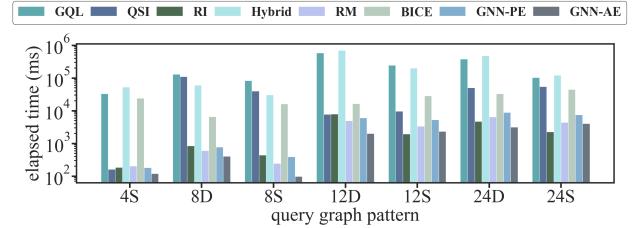


Figure 21: GNN-AE efficiency w.r.t query graph patterns.

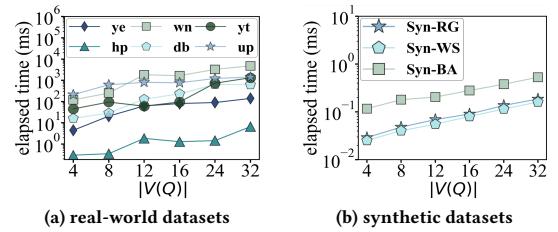


Figure 22: GNN-AE efficiency w.r.t query graph size $|V(Q)|$.

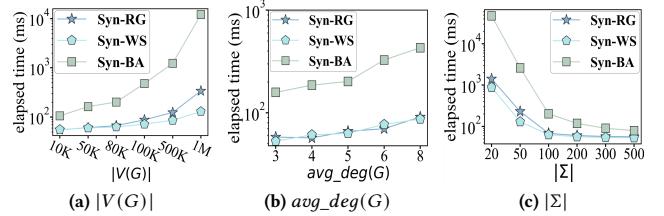


Figure 23: The GNN-AE scalability w.r.t varying data graph.

perform better than baselines on most query graph patterns. Especially, for most query graph patterns, GNN-AE can perform better than traditional exploration-based backtracking search methods (e.g., GQL, QSI, and Hybrid) and BICE by 1-2 orders of magnitude.

The GNN-AE Efficiency w.r.t Query Graph Size. Figure 22 illustrates the efficiency of our GNN-AE approach with varying the query graph size $|V(Q)|$ from 4 to 32. When $|V(Q)|$ increases, more anchors in the query graph are processed. According to the time complexity analysis in the subsection 4.3, the elapsed time of online subgraph matching tends to increase. Therefore, larger query graph size $|V(Q)|$ incurs higher elapsed time. For varying query graph sizes, our GNN-AE approach can achieve a low average time cost.

The GNN-AE Scalability w.r.t Data Graph. Figure 23 evaluates the scalability of our GNN-AE approach with different data graph characteristics, including the data graph size $|V(G)|$, $\text{avg_deg}(G)$, and $|\Sigma|$ of G . We vary the synthetic graph size $|V(G)|$ ranging from $10K$ to $1M$, $\text{avg_deg}(G)$ ranging from 3 to 8, and $|\Sigma|$ ranging from 20 to 500. Intuitively, larger size $|V(G)|$ and average degree $\text{avg_deg}(G)$ in the data graph G , more candidate anchors for the query graph, which incurs an increased time cost. When $|\Sigma|$ of the data graph G increases, the anchor embeddings in our GNN-AE approach have a stronger filtering power, which produces fewer candidate anchors for the query graph. Thus, the elapsed time of our GNN-AE shows a downward trend. Nevertheless, for varying data graph characteristics, GNN-AE still remains a low time cost (i.e., $< 47\text{ sec}$), which confirms the scalability of our GNN-AE approach.

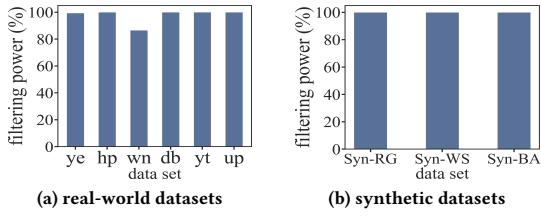


Figure 24: The anchor embedding filtering power.

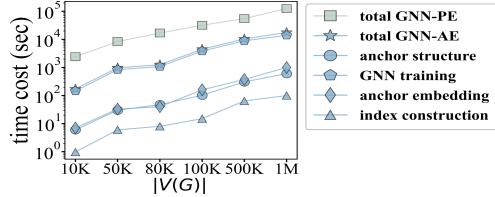


Figure 25: The GNN-AE offline pre-computation cost.

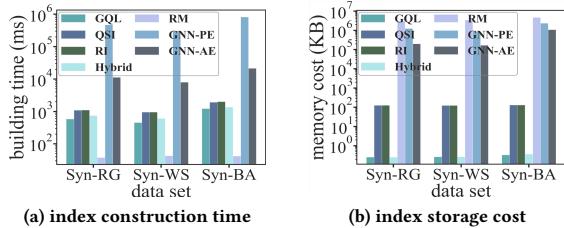


Figure 26: The GNN-AE index construction time/space costs.

The Anchor Embedding Filtering Power. Figure 24 presents the filtering power of our proposed anchor (graph & path) embeddings in Section 3 over real-world and synthetic datasets. According to Eq. 8, we compute the average filtering power of anchor embeddings from query graphs for which our GNN-AE can obtain all exact matches. From subfigures, we can see that for all real-world (except *wn*) and synthetic datasets, the filtering power of our anchor embeddings can reach 99.37% ~ 99.99%, which confirms the effectiveness of the proposed anchor embeddings in our GNN-AE approach. For the data graph *wn*, most vertices in *wn* have the same label ($|\Sigma|$ of *wn* is only 5), which makes it challenging. Therefore, anchor embeddings have a relatively low filtering power on *wn*.

6.4 Offline Pre-Computation Evaluation

In this subsection, we evaluate the cost in the offline pre-computation phase for our GNN-AE approach.

The GNN-AE Offline Pre-Computation Cost. The GNN-AE offline pre-computation consists of obtaining anchor structures (graphs & paths), GNN training, generating anchor embeddings, and index construction. Figure 25 evaluates the offline pre-computation cost of our GNN-AE approach with different data graph size $|V(G)|$ on the synthetic graph *Syn-WS*. In the subfigure, we find that GNN training takes up most of the offline pre-computation cost. However, it should be noted that our GNN model can be trained in a truncated manner (refer to Algorithm 2). We can set a lower truncated epoch value to further reduce the offline per-computation cost of our GNN-AE approach. Since GNN-PE requires the GNN model to be

trained to overfit the training data set, GNN-PE has a high offline pre-computation cost. For varying $|V(G)|$, GNN-AE can achieve a lower per-computation cost than GNN-PE by 1 order of magnitude.

The GNN-AE Index Construction Time & Space Costs. Figure 26 compares the index construction time and storage cost of GNN-AE with 6 baseline methods that require indexes or relations. Although GNN-AE requires higher index construction time than exploration-based backtracking search methods (e.g., GQL, QSI, RI, and Hybrid) and RM, our GNN-AE has a lower index construction cost than the advanced GNN-PE approach. For the index storage cost, we calculate the average index memory cost for query graphs. The memory cost of RM on some query graphs exceeds 500MB. GNN-AE only stores the anchor (graph & path) embeddings and their corresponding data graph anchors in the indexes, rather than storing a larger number of paths and their embeddings like GNN-PE. Thus, our GNN-AE has a lower storage cost. Note that, the baseline methods (except GNN-PE) construct an index for each query graph in subgraph matching. Our GNN-AE is similar to GNN-PE in that both their index construction is *offline* and *one-time only*. Thus, our GNN-AE can also process numerous subgraph matching requests simultaneously with high throughput like the GNN-PE approach.

7 RELATED WORK

In this section, we introduce related work on subgraph matching. **Exact Subgraph Matching.** Existing work on the exact subgraph matching problem usually follows the *exploration-based* or *join-based* paradigms [51, 53]. The exploration-based methods adopt the backtracking search, which recursively extend intermediate results by mapping query graph nodes to data graph nodes along a matching order [8, 12, 13, 23, 24, 30, 55, 67, 72]. The join-based methods model the subgraph matching as a relational query problem, and perform a multi-way join to obtain all results [1, 4, 33, 43, 45, 53].

In recent years, several new works [58, 66] attempted to utilize deep-learning-based techniques to deal with the exact subgraph matching, but they both have obvious drawbacks. RL-QVO in [58] employs the *Reinforcement Learning* (RL) and GNNs to generate the high-quality matching order for subgraph matching algorithms, but it relies on historical query graphs to train the model, which limits its scalability and robustness. GNN-PE in [66] adopts a GNN model to learn the dominance relationships between the node (and its 1-hop neighbors) and their substructures on the data graph, and then uses the path with dominance relationships to obtain matches. GNN-PE requires that the GNN model be trained to overfit on the training data set, which results in a high offline computation cost. **Approximate Subgraph Matching.** An alternative methodology is to obtain top- k approximate results for subgraph matching, which returns subgraphs in the data graph G that are similar to the given query Q [18, 19, 32, 73]. Approximate subgraph matching usually adopt various graph similarity measures such as edge edit distance [68] to qualify occurrences of the query Q on the data graph G . If the edit distance between Q and a subgraph g of G is no more than some threshold θ , then g is an approximate result of Q .

Recently, several advanced works [15, 38, 40, 47] attempt to utilize deep-learning-based techniques to handle the approximate subgraph matching. However, these approaches usually only approximately assert subgraph isomorphism relations between the

query and the data graph, and can not retrieve locations of all matching. Beyond that, they have no theoretical guarantees or error bounds about the approximate accuracy.

Other Learning-based Graph Computation. In recent years, learning-based methods have emerged in a few related graph computation work, such as subgraph counting [27, 37, 57, 71], shortest path distance [29], graph edit distance computation [64] and graph keyword search [25]. However, these graph computation problems (e.g., count or graph edit distance) are different from the subgraph matching problem, and the learning techniques in these works cannot be applied to solve our subgraph matching problem.

8 CONCLUSION

In this paper, we propose a novel *GNN-based anchor embedding* (GNN-AE) framework for exact and efficient subgraph matching. We first propose the anchor concept in subgraph matching. Based on the anchor concept and GNN models, we carefully design effective anchor embedding techniques (including anchor graph embedding, and anchor path embedding), which transform the subgraph matching problem into a search problem in the embedding space. In addition, a parallel matching growth algorithm is developed to obtain the locations of all exact matches. Extensive experiments have been conducted on 6 real-world and 3 synthetic datasets to show the effectiveness and efficiency of our GNN-AE approach.

REFERENCES

- [1] Christopher R Aberger, Andrew Lamb, Susan Tu, Andres Nötzli, Kunle Olukotun, and Christopher Ré. 2017. Emptyheaded: A relational engine for graph processing. *ACM Transactions on Database Systems* 42, 4 (2017), 1–44.
- [2] Anagha Ajoykumar and M Venkatesan. 2023. Study of anomalous subgraph detection in social networks. *International Research Journal on Advanced Science Hub* 5, 05 (2023), 287–300.
- [3] Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 1 (2002), 47.
- [4] Khaled Ammar, Franck McSherry, Semih Salihoglu, and Manas Joglekar. 2018. Distributed evaluation of subgraph queries using worst-case optimal and low-memory dataflows. *Proceedings of the VLDB Endowment* 11 (2018), 691–704.
- [5] Junya Arai, Yasuhiro Fujiwara, and Makoto Onizuka. 2023. Gup: Fast subgraph matching by guard-based pruning. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 1–26.
- [6] Albert-László Barabási. 2002. Emergence of scaling in complex networks. *Handbook of Graphs and Networks: From the Genome to the Internet* (2002), 69–84.
- [7] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.
- [8] Bibek Bhattacharai, Hang Liu, and H Howie Huang. 2019. Ceci: Compact embedding cluster index for scalable subgraph matching. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 1447–1462.
- [9] Fei Bi, Lijun Chang, Xuemin Lin, Lu Qin, and Wenjie Zhang. 2016. Efficient subgraph matching by postponing cartesian products. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 1199–1214.
- [10] Ekaba Bisong and Ekaba Bisong. 2019. Introduction to Scikit-learn. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (2019), 215–229.
- [11] Vincenzo Bonnici, Rosalba Giugno, Alfredo Pulvirenti, Dennis Shasha, and Alfredo Ferro. 2013. A subgraph isomorphism algorithm and its application to biochemical data. *BMC Bioinformatics* 14, 7 (2013), 1–13.
- [12] Vincenzo Carletti, Pasquale Foggia, Alessio Saggesse, and Mario Vento. 2017. Challenging the time complexity of exact subgraph isomorphism for huge and dense graphs with VF3. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2017), 804–818.
- [13] Yunyoung Choi, Kunsoo Park, and Hyunjoon Kim. 2023. BICE: Exploring compact search space by using bipartite matching and cell-wide verification. *Proceedings of the VLDB Endowment* 16, 9 (2023), 2186–2198.
- [14] Stephen A Cook. 2023. The complexity of theorem-proving procedures. In *Logic, Automata, and Computational Complexity: The Works of Stephen A. Cook*. 143–152.
- [15] Khoa D Doan, Saurav Manchanda, Suchismit Mahapatra, and Chandan K Reddy. 2021. Interpretable graph similarity computation via differentiable optimal alignment of node embeddings. In *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 665–674.
- [16] Brendan L Douglas. 2011. The weisfeiler-lehman method and graph isomorphism testing. *arXiv preprint arXiv:1101.5211* (2011), 1–43.
- [17] Chi Thang Duong, Hongzhi Yin, Dung Hoang, Minn Hung Nguyen, Matthias Weidlich, Quoc Viet Hung Nguyen, and Karl Aberer. 2020. Graph embeddings for one-pass processing of heterogeneous queries. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 1994–1997.
- [18] Sourav Dutta, Pratik Nayek, and Arnab Bhattacharya. 2017. Neighbor-aware search for approximate labeled graph matching using the chi-square statistics. In *Proceedings of the International Conference on World Wide Web (WWW)*. 1281–1290.
- [19] Marco Gori, Marco Maggini, and Lorenzo Sarti. 2005. Exact and approximate graph matching using random walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 7 (2005), 1100–1111.
- [20] Martin Grohe and Pascal Schweitzer. 2020. The graph isomorphism problem. *Commun. ACM* 63, 11 (2020), 128–134.
- [21] Aric Hagberg and Drew Conway. 2020. Networkx: Network analysis with python. *URL: https://networkx.github.io* (2020).
- [22] Mai Hamdalla, David Grant, Ion Mandoiu, Dennis Hill, Sanguthevar Rajasekaran, and Reda Ammar. 2012. The use of graph matching algorithms to identify biochemical substructures in synthetic chemical compounds: Application to metabolomics. In *Proceedings of the International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*. 1–6.
- [23] Myoungji Han, Hyunjoon Kim, Geommo Gu, Kunsoo Park, and Wook-Shin Han. 2019. Efficient subgraph matching: Harmonizing dynamic programming, adaptive matching order, and failing set together. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 1429–1446.
- [24] Wook-Shin Han, Jinsoo Lee, and Jeong-Hoon Lee. 2013. Turboiso: towards ultrafast and robust subgraph isomorphism search in large graph databases. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 337–348.
- [25] Yu Hao, Xin Cao, Yufan Sheng, Yixiang Fang, and Wei Wang. 2021. KS-GNN: keywords search over incomplete graphs via graph neural network. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*. 1700–1712.
- [26] Huahai He and Ambuj K Singh. 2008. Graphs-at-a-time: query language and access methods for graph databases. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 405–418.
- [27] Wenzhe Hou, Xiang Zhao, and Bo Tang. 2024. LearnSC: An efficient and unified learning-based framework for subgraph counting problem. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 2625–2638.
- [28] Sen Hu, Lei Zou, Jeffrey Xu Yu, Haixun Wang, and Dongyan Zhao. 2017. Answering natural language questions by subgraph matching over knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering* 30, 5 (2017), 824–837.
- [29] Shuai Huang, Yong Wang, Tianyu Zhao, and Guoliang Li. 2021. A learning-based method for computing shortest path distances on road networks. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 360–371.
- [30] Zite Jiang, Shuai Zhang, Xingzhong Hou, Mengting Yuan, and Haihang You. 2024. IVE: Accelerating enumeration-based subgraph matching via exploring isolated vertices. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 4208–4221.
- [31] Hyunjoon Kim, Yunyoung Choi, Kunsoo Park, Xuemin Lin, Seok-Hee Hong, and Wook-Shin Han. 2023. Fast subgraph query processing and subgraph matching via static and dynamic equivalences. *The VLDB Journal* 32, 2 (2023), 343–368.
- [32] Segla Kpodjedo, Philippe Galinier, and Giulio Antoniol. 2014. Using local similarity measures to efficiently address approximate graph matching. *Discrete Applied Mathematics* 164 (2014), 161–177.
- [33] Longbin Lai, Zhu Qing, Zhengyi Yang, Xin Jin, Zhengmin Lai, Ran Wang, Kongzhang Hao, Xuemin Lin, Lu Qin, Wenjie Zhang, et al. 2019. Distributed subgraph matching on timely dataflow. *Proceedings of the VLDB Endowment* 12, 10 (2019), 1099–1112.
- [34] Jinsoo Lee, Wook-Shin Han, Romans Kasperovics, and Jeong-Hoon Lee. 2012. An in-depth comparison of subgraph isomorphism algorithms in graph databases. *Proceedings of the VLDB Endowment* 6, 2 (2012), 133–144.
- [35] Andrei Leman and Boris Weisfeiler. 1968. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya* 2, 9 (1968), 12–16.
- [36] Harry R Lewis. 1983. Michael R. PiGarey and David S. Johnson. Computers and intractability. A guide to the theory of NP-completeness. *The Journal of Symbolic Logic* 48, 2 (1983), 498–500.
- [37] Xin Liu, Haojie Pan, Mutian He, Yangqiu Song, Xin Jiang, and Lifeng Shang. 2020. Neural subgraph isomorphism counting. In *Proceedings of the International Conference on Knowledge Discovery & Data Mining (SIGKDD)*. 1959–1969.
- [38] Xuanzhou Liu, Lin Zhang, Jiaqi Sun, Yujiu Yang, and Haiqin Yang. 2023. D2Match: leveraging deep learning and degeneracy for subgraph matching. In *Proceedings of the International Conference on Machine Learning (ICML)*. 22454–22472.

- [39] Yunkai Lou and Chaokun Wang. 2023. A generalized community-structure-aware optimization framework for efficient subgraph matching in social network analysis. *IEEE Transactions on Computational Social Systems* 11 (2023), 2545–2557.
- [40] Zhaoyu Lou, Jiaxuan You, Chengtao Wen, Arquimedes Canedo, Jure Leskovec, et al. 2020. Neural subgraph matching. *arXiv preprint arXiv:2007.03092* (2020), 1–16.
- [41] Cai Lu, Xinran Xu, and Bingbin Zhang. 2024. Modification and completion of geological structure knowledge graph based on pattern matching. *Scientific Reports* 14, 1 (2024), 9825.
- [42] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. 2019. Invariant and equivariant graph networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 1–14.
- [43] Amine Mhedhibi and Semih Salihoglu. 2019. Optimizing subgraph queries by combining binary and worst-case optimal joins. *Proceedings of the VLDB Endowment* 12, 11 (2019), 1692–1704.
- [44] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, Vol. 33. 4602–4609.
- [45] Dung Nguyen, Molham Aref, Martin Bravenboer, George Kollias, Hung Q Ngo, Christopher Ré, and Atri Rudra. 2015. Join processing for graph patterns: An old dog with new tricks. In *Proceedings of the GRADES’15*. 1–8.
- [46] Yeonsu Park, Seongyun Ko, Sourav S Bhowmick, Kyoungmin Kim, Kijae Hong, and Wook-Shin Han. 2020. G-CARE: A framework for performance benchmarking of cardinality estimation techniques for subgraph matching. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 1099–1114.
- [47] Indradumna Roy, Venkata Sai Baba Reddy Velugoti, Soumen Chakrabarti, and Abir De. 2022. Interpretable neural subgraph matching for graph retrieval. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*. 8115–8123.
- [48] Ryoma Sato. 2020. A survey on the expressive power of graph neural networks. *arXiv preprint arXiv:2003.04078* (2020), 1–42.
- [49] Haichuan Shang, Ying Zhang, Xuemin Lin, and Jeffrey Xu Yu. 2008. Taming verification hardness: an efficient algorithm for testing subgraph isomorphism. *Proceedings of the VLDB Endowment* 1, 1 (2008), 364–375.
- [50] Shixuan Sun and Qiong Luo. 2019. Scaling up subgraph query processing with efficient subgraph matching. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 220–231.
- [51] Shixuan Sun and Qiong Luo. 2020. In-memory subgraph matching: An in-depth study. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 1083–1098.
- [52] Shixuan Sun and Qiong Luo. 2020. Subgraph matching with effective matching order and indexing. *IEEE Transactions on Knowledge and Data Engineering* 34, 1 (2020), 491–505.
- [53] Shixuan Sun, Xibo Sun, Yulin Che, Qiong Luo, and Bingsheng He. 2021. Rapid-match: A holistic approach to subgraph query processing. *Proceedings of the VLDB Endowment* 14, 2 (2021), 176–188.
- [54] Yunhao Sun, Guanyu Li, Jingjing Du, Bo Ning, and Heng Chen. 2022. A subgraph matching algorithm based on subgraph index for knowledge graph. *Frontiers of Computer Science* 16 (2022), 1–18.
- [55] Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. 2012. Efficient subgraph matching on billion node graphs. *Proceedings of the VLDB Endowment* 5, 9 (2012), 788–799.
- [56] Yuanyuan Tian, Richard C McEachin, Carlos Santos, David J States, and Jignesh M Patel. 2007. SAGA: A subgraph matching tool for biological graphs. *Bioinformatics* 23, 2 (2007), 232–239.
- [57] Hanchen Wang, Rong Hu, Ying Zhang, Lu Qin, Wei Wang, and Wenjie Zhang. 2022. Neural subgraph counting with Wasserstein estimator. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 160–175.
- [58] Hanchen Wang, Ying Zhang, Lu Qin, Wei Wang, Wenjie Zhang, and Xuemin Lin. 2022. Reinforcement learning based query vertex ordering model for subgraph matching. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 245–258.
- [59] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393, 6684 (1998), 440–442.
- [60] Fang Wu, Siyuan Li, Xurui Jin, Yinghui Jiang, Dragomir Radev, Zhangming Niu, and Stan Z Li. 2023. Rethinking explaining graph neural networks via non-parametric subgraph matching. In *Proceedings of the International Conference on Machine Learning (ICML)*. 3751–3753.
- [61] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2020), 4–24.
- [62] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? In *Proceedings of the International Conference on Learning Representations (ICLR)*. 1–17.
- [63] Dominic Tianli Yang. 2023. *Exploiting symmetry in subgraph isomorphism and formulating neural network constrained optimization problems*. University of California, Los Angeles.
- [64] Lei Yang and Lei Zou. 2021. Noah: Neural-optimized A* search algorithm for graph edit distance computation. In *Proceedings of the International Conference on Data Engineering (ICDE)*. 576–587.
- [65] Yutong Ye, Xiang Lian, and Mingsong Chen. 2023. Efficient Exact subgraph matching via GNN-based path dominance embedding (technical report). *arXiv preprint arXiv:2309.15641* (2023), 1–20.
- [66] Yutong Ye, Xiang Lian, and Mingsong Chen. 2024. Efficient exact subgraph matching via GNN-based path dominance embedding. *Proceedings of the VLDB Endowment* 17, 7 (2024), 1628–1641.
- [67] Shijie Zhang, Shiron Li, and Jiong Yang. 2009. GADDI: distance index based subgraph matching in biological networks. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*. 192–203.
- [68] Shijie Zhang, Jiong Yang, and Wei Jin. 2010. Sapper: Subgraph indexing and approximate matching in large graphs. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 1185–1194.
- [69] Wentao Zhang, Mouyi Weng, Mingzheng Zhang, Zhefeng Chen, Bingxu Wang, Shunning Li, and Feng Pan. 2024. Rapid mining of fast ion conductors via subgraph isomorphism matching. *Journal of the American Chemical Society* 146, 27 (2024), 18535–18543.
- [70] Zhijie Zhang, Yujie Lu, Weiguo Zheng, and Xuemin Lin. 2024. A comprehensive survey and experimental study of subgraph matching: trends, unbiasedness, and interaction. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 1–29.
- [71] Kangfei Zhao, Jeffrey Xu Yu, Hao Zhang, Qiyan Li, and Yu Rong. 2021. A learned sketch for subgraph counting. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 2142–2155.
- [72] Peixiang Zhao and Jiawei Han. 2010. On graph query optimization in large networks. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 340–351.
- [73] Linhong Zhu, Wee Keong Ng, and James Cheng. 2011. Structure and attribute index for approximate graph matching in large graphs. *Information Systems* 36, 6 (2011), 958–972.