

Fuzzy-Constrained Graph Pattern Matching in Medical Knowledge Graphs Postprint

Authors: Lei, Li, Xun, Du, Zan, Zhang, Zhenchao, Tao, Lei, Li

Date: 2022-11-28T00:00:00+00:00

Abstract

The research on graph pattern matching (GPM) has attracted a lot of attention. However, most of the research has focused on complex networks, and there are few researches on GPM in the medical field. Hence, with GPM this paper is to make a breast cancer-oriented diagnosis before the surgery. Technically, this paper has firstly made a new definition of GPM, aiming to explore the GPM in the medical field, especially in Medical Knowledge Graphs (MKGs). Then, in the specific matching process, this paper introduces fuzzy calculation, and proposes a multi-threaded bidirectional routing exploration (M-TBRE) algorithm based on depth first search and a two-way routing matching algorithm based on multi-threading. In addition, fuzzy constraints are introduced in the M-TBRE algorithm, which leads to the Fuzzy-M-TBRE algorithm. The experimental results on the two datasets show that compared with existing algorithms, our proposed algorithm is more efficient and effective.

Full Text

Preamble

Fuzzy-Constrained Graph Pattern Matching in Medical Knowledge Graphs

Lei Li^{1,2,3†}, Xun Du³, Zan Zhang³ & Zhenchao Tao⁴

¹Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), Hefei University of Technology, Hefei 230601, China

²Intelligent Interconnected Systems Laboratory of Anhui Province (Hefei University of Technology), Hefei 230601, China

³School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China

⁴The First Affiliated Hospital of University of Science and Technology of China, Hefei 230031, China

Keywords: Graph pattern matching; Medical Knowledge Graphs; Fuzzy constraints; Breast cancer; Diagnostic classification

Citation: Li, L. et al.: Fuzzy-Constrained Graph Pattern Matching in Medical Knowledge Graphs. *Data Intelligence* 4(3), 599-619 (2022). DOI: 10.1162/dint_a_00153}

Received: Oct. 10, 2021; **Revised:** Jan. 15, 2022; **Accepted:** Apr. 10, 2022

Abstract

Graph pattern matching (GPM) has attracted significant research attention, yet most studies have focused on complex networks, with limited research on GPM in the medical domain. This paper employs GPM to enable pre-surgical diagnosis for breast cancer. Technically, we first introduce a novel definition of GPM specifically tailored for medical applications, particularly in Medical Knowledge Graphs (MKGs). During the matching process, we incorporate fuzzy computation and propose a multi-threaded bidirectional routing exploration (M-TBRE) algorithm based on depth-first search. Additionally, we introduce fuzzy constraints into the M-TBRE algorithm, resulting in the Fuzzy-M-TBRE algorithm. Experimental results on two datasets demonstrate that our proposed algorithm achieves superior efficiency and effectiveness compared to existing methods.

Corresponding author: Lei Li (E-mail: lilei@hfut.edu.cn, ORCID: 0000-0002-5374-7293).

1. Introduction

Graphs serve as a fundamental data structure with widespread applications across numerous domains. For instance, in object anomaly detection, objects can be represented as graphs, enabling the discovery of anomalies through specialized graph algorithms [1]. Similarly, to determine user interest in specific webpages, multiple webpages can be converted into graphs and processed as a bag for classification and judgment [2]. Graph pattern matching (GPM), a prominent graph-based technology, has garnered considerable attention. Formally, GPM involves finding subgraphs within a data graph that exhibit similar or identical structure to a given pattern graph. As the application domains of GPM have evolved from initial protein isomorphism [3, 4] to community detection [5, 6], expert discovery [7], recommendation systems [8], social group identification [9–11], and criminal group detection [12], the definition of graph patterns has correspondingly adapted.

Technically, GPM was originally defined based on subgraph isomorphism. Given a data graph G_D and a pattern graph G_P as input, the method determines

whether G_D contains a subgraph with exactly the same topological structure as G_P . This approach enables property inference for unknown proteins from known ones through matching [3, 4]. However, traditional subgraph isomorphism is overly restrictive. To extend GPM's applicability, Fan et al. [12] proposed bounded simulation, which expands edge-to-edge matching to edge-to-finite-length-path matching. Nevertheless, this approach fails to leverage the rich attribute information available on vertices and edges. Consequently, Liu et al. [13] introduced the multi-constrained graph pattern matching (MC-GPM) problem to obtain more effective matching results. Subsequently, Liu et al. [14] proposed multiple fuzzy constrained graph pattern matching (MFC-GPM) based on MC-GPM, recognizing that certain attributes do not require exact matching. However, current GPM applications remain predominantly concentrated in complex networks, with minimal research exploring its use in the medical field, particularly in Medical Knowledge Graphs.

The incidence of breast cancer continues to rise, with increasingly younger patient populations. Breast cancer encompasses multiple types, including ductal carcinoma in situ, lobular carcinoma in situ, invasive ductal carcinoma, and invasive lobular carcinoma. Each type can be further classified according to primary tumor staging, regional lymph node staging, and distant metastasis staging. This paper aims to enable pre-surgical diagnosis through GPM technology before surgical intervention is required.

To introduce GPM into the medical domain, we formulate the GPM problem in MKGs and provide formal definitions. We propose the M-TBRE algorithm, which first decomposes the pattern graph into subgraphs, obtains matching results for each subgraph, and subsequently merges these results. M-TBRE can generate a diagnosis distribution for the pattern graph and return the top-k diagnostic classification results based on frequency. We further extend M-TBRE by introducing fuzzy constraints, yielding the Fuzzy-M-TBRE algorithm, and validate its effectiveness on two public datasets.

The remainder of this paper is organized as follows: Section 2 reviews related work on GPM. Section 3 introduces the concept of pattern matching in MKGs. Section 4 proposes the M-TBRE algorithm and Fuzzy-M-TBRE algorithm for processing GPM in MKGs. Section 5 describes the datasets and presents experimental validation of our Fuzzy-M-TBRE algorithm. Section 6 concludes the paper.

2. Related Work

Based on whether judgment relies on bijective functions or binary relationships, GPM research can be categorized into isomorphism-based GPM and simulation-based GPM.

2.1 Isomorphism-Based GPM

Isomorphism-based GPM establishes a bijective function between the pattern graph and data graph, requiring identical topological structures between matched subgraphs and pattern graphs. Ullmann [15] first proposed a depth-first search-based matching algorithm. Cordella et al. [16] improved upon Ullmann's algorithm through enhanced matching order and pruning strategies, introducing the VF2 algorithm. Tong et al. [17] proposed the G-Ray method, which employs a goodness function to measure matching degree between subgraphs and pattern graphs, enabling retrieval of optimal-k subgraphs. Cheng et al. [18] also proposed a top-k matching algorithm that ranks matched subgraphs based on spanning tree count to return optimal-k results. Cheng et al. [19] introduced the R-join algorithm based on clustering graph join indexes with subsequent optimization. Other representative algorithms include DDST [20] and IncBMatch [21]. As an NP-complete problem, isomorphism-based GPM typically employs indexing [22, 23] and parallel distributed processing [24-26] to improve matching efficiency.

Isomorphism-based GPM is primarily applied in domains with strict structural requirements, such as protein isomorphism, 3D object matching [27], and network abnormal behavior detection [28]. However, such strict matching is unsuitable for applications like social networks or knowledge graphs that do not demand exact matching precision. This limitation has motivated research into simulation-based GPM.

2.2 Simulation-Based GPM

Graph simulation, introduced by Henzinger et al. [29], relies on binary relations but remains limited to edge-to-edge matching, failing to meet many application requirements. Fan et al. [12] extended graph simulation by proposing bounded simulation, where pattern graph edges can match paths not exceeding a given constraint length k . Building on bounded simulation, Ma et al. [30] introduced strong simulation, which better preserves pattern graph topological structure. However, these existing approaches do not consider the abundant attribute information on vertices and edges in large graph data. Liu et al. [13] addressed this by extending bounded simulation to MC-GPM and proposing both a baseline exploration-based algorithm and a heuristic algorithm based on data graph compression index (HAMC). Since HAMC only considers matching path constraints without minimizing path length and lacks distributed computing support, Liu et al. [31] proposed the M-HAMC algorithm. Recognizing that vertex and edge attribute values sometimes need not match exactly, Liu et al. [14] introduced MFC-GPM and the ETOF-K algorithm, improving matching efficiency through edge matching and connection optimization. Based on topologically ordered pattern graph vertices, Liu et al. [32] proposed the NTSS algorithm, optimized through caching mechanisms and reverse edge matching. The caching mechanism prevents redundant calculations of identical candidate paths across multiple subgraphs, while reverse edge matching prunes candidate sets for edges

with zero in-degree.

3. Graph Pattern Matching

GPM aims to find all data subgraphs in a given data graph G_D that satisfy the pattern graph G_P . This section provides formal definitions for data graphs, pattern graphs, and graph pattern matching in MKGs.

3.1 Data Graph and Pattern Graph

3.1.1 Data Graph A data graph $G_D = (V, E, f_V^D, f_E^D)$ is a directed graph with vertex attributes and edge attributes, where V is the vertex set; E is the edge set, with $(v_i, v_j) \in E$ representing a directed edge from $v_i \in V$ to $v_j \in V$; f_V^D is a function defined on V where $\forall v \in V$, $f_V^D(v)$ represents the attribute set of v . In an MKG, each vertex v has a label rr indicating its type. Different rr values correspond to different attribute sets $f_V^D(v)$. The rr value can be DI, BI, MI, GW, OC, AL, or PD.

- **DI:** When $rr = \text{DI}$, $f_V^D(v)$ describes breast cancer diagnostic classification information, including pathological information rh , T staging rs , tumor length r_{TS} , regional lymph node N staging r_{LN} , and distant metastasis M staging r_{DM} . The value of rh can be 0, 1, 2, or 3, representing “invasive ductal carcinoma”, “invasive lobular carcinoma”, “ductal carcinoma in situ”, and “lobular carcinoma in situ”, respectively. The value of rs ranges from 0 to 4. r_{TS} is a floating-point number in centimeters. r_{LN} can be N0, N1, N2, or N3. r_{DM} can be M0 or M1.
- **BI:** When $rr = \text{BI}$, $f_V^D(v)$ describes patient basic information including r_{CN} , r_{CP} , and r_{age} . r_{CN} indicates current care needs (true/false). r_{CP} indicates current pregnancy status (true/false). r_{age} represents patient age as a positive integer.
- **MI:** When $rr = \text{MI}$, $f_V^D(v)$ describes menopausal information through r_{MS} , which can be 0 (pre-menopausal), 1 (perimenopausal), or 2 (post-menopausal).
- **GW:** When $rr = \text{GW}$, $f_V^D(v)$ describes overall well-being through r_{GWB} , ranging from 0 (“fully active, no complaints”) to 4 (“completely disabled, bed/chair-bound”).
- **OC:** When $rr = \text{OC}$, $f_V^D(v)$ describes non-breast cancers through r_{OCS} and r_{OCN} . If r_{OCS} is false, r_{OCN} is none; if true, r_{OCN} lists other cancer names.
- **AL:** When $rr = \text{AL}$, $f_V^D(v)$ describes axillary lymph nodes through r_{LN} , r_{LS} , r_{IN} , r_{SN} , and r_{CW} . r_{LS} , r_{IN} , r_{SN} , and r_{CW} are boolean values

indicating normality. r_{LN} is a positive integer indicating how many of the three (supraclavicular, subclavian, chest wall) have issues.

- **PD:** When $rr = PD$, $f_V^D(v)$ describes past diagnosis information through r_{aid} , r_{ane} , r_{aut} , r_{lun} , r_{dia} , r_{car} , r_{ost} , and r_{rep} , all boolean values indicating presence of AIDS, anemia, autoimmune disease, lung cancer, diabetes, cardiovascular disease, osteoporosis, and reproductive organ disease, respectively.

f_E^D is defined on E where $\forall e \in E$, $f_E^D(v_i, v_j)$ represents the attribute set of edge (v_i, v_j) . In an MKG, $f_E^D(v_i, v_j)$ contains only ρ , a list storing patient IDs as vertex identity information.

3.1.2 Pattern Graph A pattern graph $G_P = (V_P, E_P, f_V^P, f_E^P, f_l^P, f_m^P)$ is a directed graph with vertex and edge attributes, where V_P is the vertex set; E_P is the edge set with $(u_i, u_j) \in E_P$ representing directed edges; f_V^P is defined on V_P where $\forall v \in V_P$, $f_V^P(v)$ is the attribute set. In an MKG, the function $P(u)$ corresponding to vertex u shares the same meaning as data graph vertex attribute sets. f_E^P is defined on E_P where $\forall e \in E_P$, $f_E^P(u_i, u_j)$ is the attribute set of edge (u_i, u_j) . In an MKG, $f_E^P(u_i, u_j)$ contains only ρ , a list storing patient IDs. f_l^P is defined on E_P where $\forall (u_i, u_j) \in E_P$, $f_l^P(u_i, u_j)$ represents length constraints as positive integer k or symbol $*$, indicating the path length from v_i to v_j does not exceed k or has no limit. f_m^P is a set of membership constraint functions defined on vertex and edge attributes.

3.1.3 Fuzzy Constraints In practice, obtaining more matching results is beneficial, as each matched subgraph corresponds to a patient with similar health information to the patient being diagnosed in the pattern graph. More matches provide better reference experience for treatment. However, subgraphs may satisfy most constraints yet be excluded due to minor attribute mismatches. Furthermore, some vertex attribute constraints need not match exactly—differences only need to fall within acceptable ranges. Therefore, we introduce fuzzy constraints to GPM in MKGs.

In MKGs, we consider membership functions to introduce fuzzy constraints for the age attribute. The membership function μ_{age} is defined as in Eq. (1), where abs is the absolute value function, ρ_{age}^P represents the age attribute constraint value of vertex v in pattern graph G_P , and ρ_{age}^D represents the age attribute value of vertex u in data graph G_D . During matching, the age attribute need only satisfy $\mu_{age} \leq \epsilon$, where ϵ is the membership constraint value set to 3.

3.2 Pattern Matching

A matched subgraph is a subgraph of data graph G_D that matches pattern graph G_P , where $G_{sub} \subset G_D$, $V_{sub} \subset V$, and $E_{sub} \subset E$. The definition of pattern matching in MKGs is as follows:

For a pattern graph $G_P = (V_P, E_P, f_V^P, f_E^P, f_I^P, f_m^P)$ and data graph $G_D = (V, E, f_V^D, f_E^D)$, G_D matches G_P , denoted $G_P \prec G_D$, if there exists a binary relationship $S \subseteq V_P \times V$ such that:

- For each pair $(u_i, v_i) \in S$, $u_i \sim v_i$, meaning v_i satisfies $f_V^P(u_i)$. If age attribute constraints are included, μ_{age} only needs to satisfy $\mu_{age} \leq \epsilon$. Except for the age attribute processed by the membership function, values of other attributes in v_i must equal corresponding attribute values in u_i .
- For each edge (u_i, u_j) in E_P , there exists a path from v_i to v_j in G_D such that $(u_j, v_j) \in S$. Since $f_I^P(u_i, u_j) = 1$, this path can be regarded as a direct edge from v_i to v_j in G_D .

Example 1: As shown in Figure 1 [Figure 1: see original paper], G_D is a data graph containing information from multiple breast cancer patients. Some vertex attributes store breast cancer diagnostic classification information. Each vertex represents patient information. For function $f_E^D(A1, B1)$, $\rho = 1375$ indicates vertex B1's information originates from patient #1375. Pattern graph G_P represents a patient's health status to be diagnosed. Vertices B, C, D, E, F, and G represent basic information, menopausal status, general well-being, other cancers, axillary lymph nodes, and past diagnoses, respectively. Vertex A contains diagnostic information that is unknown and must be obtained through GPM. Since all vertex information in G_P comes from the same patient, we must find a patient ID as edge attribute constraint information to obtain matching results.

Example 2: As shown in Figure 1, subgraph M_{sub1} can be easily found in G_D matching G_P , passing through vertices A2, B2, C1, D2, E1, F2, and G3. Vertex A2 represents the breast cancer diagnosis result for G_P . The edge attribute constraint value in M_{sub1} is 2384, indicating patient #2384 has similar health status to the patient corresponding to G_P .

After introducing fuzzy constraints, since $\mu_{age} = 0.5$ does not exceed the membership constraint value 3, vertex B3 matches vertex B. We obtain a new matched subgraph M_{sub2} passing through vertices A2, B3, C1, D2, E1, F2, and G3, with edge attribute constraint value 676.

4. Graph Pattern Matching in Medical Knowledge Graphs

This section proposes the M-TBRE algorithm to solve GPM problems in MKGs.

4.1 Algorithm Description

Multi-core CPUs enable parallel task processing and accelerate program execution. Since multi-constrained GPM is NP-complete, we adopt multi-threading to accelerate matching and quickly return results. The approach employs divide-and-conquer: pattern graph G_P is divided into several subgraphs, each matched

independently in data graph G_D , with results subsequently merged. Pattern subgraph matching can be distributed as subtasks to multiple threads for independent completion, enabling rapid result acquisition.

4.2 Algorithm Flow

In M-TBRE, since MKG pattern graphs can be regarded as paths, we segment G_P at intermediate vertices to obtain two pattern subgraphs. After matching these subgraphs, we connect their results to obtain G_P matching results.

Algorithm 1 details the M-TBRE steps. First, we obtain intermediate vertex mid of G_P and its candidate vertex set $cand_{mid}$ (lines 1-2). Line 3 initializes $pool$ (thread pool) and $tempInfo$ (temporary matching results), with thread count configurable based on requirements. G_P is then divided into two sub-pattern graphs G_P^{sub1} and G_P^{sub2} using mid as the dividing point. We traverse $cand_{mid}$ to complete subgraph matching (lines 4-27). For each $cand_{mid}[i]$, we use edge attribute constraint ρ to obtain $rpids$, $De_{forward}$, and $De_{successor}$ (lines 6-25). For each patient ID $rpids$ in $rpids$, we use $rpids$ as the edge attribute constraint to complete matching, storing partial results in $tempInfo$ (lines 15-16). The thread pool submits subtasks MC-SEM and MC-FEM to match G_P^{sub1} and G_P^{sub2} (lines 19-20). Algorithm RM merges the matched results (line 28).

The MC-SEM algorithm completes matching of pattern subgraph G_P^{sub1} . Parameters ver , $cand$, $rpids$, and $tempInfo$ represent the pattern vertex to match, candidate vertex, edge attribute constraint value (patient ID), and temporary results. If ver has no successor edges (out-degree 0), matching completes and results are saved to $tempInfo$ (Algorithm 2, lines 2-7). If ver has successor edges, we traverse successor edges where attribute constraint ρ includes $rpids$, recursively completing matching (lines 8-16).

The MC-FEM algorithm matches pattern subgraph G_P^{sub2} using reverse depth-first search, with processing similar to MC-SEM.

The RM algorithm connects matching results of G_P^{sub1} and G_P^{sub2} . When a given $rpids$ serves as attribute constraint on all edges and both flag bits are 1, combining results from G_P^{sub1} and G_P^{sub2} yields a matching result for G_P (Algorithm 4, lines 4-6).

Example 3: Using G_P and G_D from Figure 1, we first obtain intermediate vertex D and candidate set $cand_{mid} = \{D2\}$. G_P is divided into G_P^{sub1} (vertices A, B, C) and G_P^{sub2} (vertices E, F, G). Forward edge $(C1, D2)$ and successor edge $(D2, E1)$ share attribute constraint $rpids = \{676, 2384\}$. Matching edge $(C1, D2)$ with $rpids = 2384$ while C1 matches C yields $(C1, D2, G_D) \sim (C, D, G_P)$. Similarly, we obtain $(A2, B2, G_D) \sim (A, B, G_P)$. Matching G_P^{sub1} with $rpids = 2384$ yields diagnostic classification information from vertex A2. Likewise, G_P^{sub2} matching produces $(D2, E1, G_D) \sim (D, E, G_P)$, $(E1, F2, G_D) \sim (E, F, G_P)$, and $(F2, G3, G_D) \sim (F, G, G_P)$. Since both subgraphs have matching results, G_P 's diagnostic classification comes from G_P^{sub1} .

with patient ID 2384. M-TBRE obtains two matching subgraphs: M_{sub1} with vertices $\{A2, B2, C1, D2, E1, F2, G3\}$ and edge attribute 2384, and M_{sub2} with vertices $\{A2, B3, C1, D2, E1, F2, G3\}$ and edge attribute 676. Patients #676 and #2384 can serve as references for treating the patient corresponding to G_P .

5. Experiments

We conduct experiments on two public MKGs, with dataset details shown in Table 1. We implement M-TBRE for MKG pattern matching. Since M-TBRE divides pattern graphs into two subgraphs, matching for different edge attribute constraint values rp_{id} is submitted to the thread pool as subtasks. Thread pool size affects algorithm performance, so we vary thread count to measure efficiency dynamics. To obtain more matches, we introduce fuzzy constraints via a membership function for vertex age attributes, allowing age values to differ within calculated membership constraint bounds. This yields the Fuzzy-M-TBRE algorithm, which we compare against M-TBRE to demonstrate fuzzy constraint effectiveness.

Table 1. Detail information of two datasets.

Dataset	Vertices	Edges	Description
Female-breast-cancer-2013a			A graph about breast cancer patients
Breastcancer-femalepatient-2016A			A graph about breast cancer patients

5.1 Experimental Settings and Implementation

The experimental MKG focuses on breast cancer. Dataset-1 and Dataset-2 represent Female-breast-cancer-2013a and Breastcancer-femalepatient-2016A, respectively. Dataset-1 contains physical condition information from 10,000 patients; Dataset-2 contains 100,000 patients. We use several pattern graphs similar to Figure 1. Our membership function applies only to vertex age attributes with constraint value 3. Both algorithms are implemented in Java on a PC with Intel(R) Core(TM) i9-10900F CPU @ 2.81 GHz, 32 GB RAM, and Windows 10.

5.2.1 Experiments on Execution Time

This experiment examines execution time variation with different thread pool sizes in M-TBRE. Results represent arithmetic means across 10 runs to prevent error interference.

As shown in Figure 2 [Figure 2: see original paper] and Figure 3 [Figure 3: see original paper], the x-axis represents different pattern graphs; the y-axis represents matching time. M-TBRE-1 indicates a single thread. Since MKG pattern graphs are paths, the ETOF-K algorithm's edge join strategy is ineffective, making its performance nearly identical to M-TBRE-1 on both datasets. NTSS's reverse matching strategy is also ineffective, but its caching mechanism avoids duplicate path calculations, making it superior to ETOF-K and M-TBRE-1 on both datasets.

Figure 2. Matching time of different pattern graphs on Female-breast-cancer-2013a.

Figure 3. Matching time of different pattern graphs on Breastcancer-femalepatient-2016A.

However, M-TBRE-1 extends to multi-threaded variants: M-TBRE-2, M-TBRE-4, and M-TBRE-8 (2, 4, and 8 threads). Figures 2 and 3 show M-TBRE-2 already outperforms NTSS, validating M-TBRE's effectiveness. Tables 2 and 3 provide detailed execution times in seconds. Table 4 compares average execution times across both datasets, showing decreasing execution time as thread count increases.

Table 2. Execution time on the Female-breast-cancer-2013a dataset.

Algorithm	Time (s)
ETOF-K	
M-TBRE-1	
M-TBRE-2	
M-TBRE-4	
M-TBRE-8	

Table 3. Execution time on the Breastcancer-femalepatient-2016A dataset.

Algorithm	Time (s)
ETOF-K	
M-TBRE-1	
M-TBRE-2	
M-TBRE-4	
M-TBRE-8	

Table 4. Comparison of execution time on two datasets.

Dataset	M-TBRE-1	M-TBRE-2	M-TBRE-4	M-TBRE-8	Percentage
Female-breast-cancer-2013a					43.11%

Dataset	M-TBRE-1	M-TBRE-2	M-TBRE-4	M-TBRE-8	Percentage
Female-breast-cancer-2013a					39.56%
Female-breast-cancer-2013a					27.81%
Breastcancer-femalepatient-2016A					40.45%
Breastcancer-femalepatient-2016A					42.44%
Breastcancer-femalepatient-2016A					36.06%

For Dataset-1, M-TBRE-2 improves execution time by 43.11% over M-TBRE-1; M-TBRE-4 improves by 39.56% over M-TBRE-2; but M-TBRE-8 only improves by 27.81% over M-TBRE-4. This occurs because Dataset-1 is relatively small, making thread context switching and system state transition overhead significant. For the larger Dataset-2, these overheads constitute a smaller proportion, allowing M-TBRE-8 to achieve 36.06% improvement over M-TBRE-4.

As thread count increases, M-TBRE execution speed accelerates, but improvements diminish beyond a certain point, as shown by M-TBRE-8 in Figure 2. For larger datasets or more subtasks, this slowdown becomes less pronounced.

5.2.2 Experiments on Fuzzy Constraints

This experiment examines changes in matching subgraph count when introducing fuzzy constraints. Fuzzy-M-TBRE represents M-TBRE with fuzzy constraints. Since Fuzzy-M-TBRE can retrieve all pattern graph matches, we compare total match counts before and after fuzzy constraint introduction.

As shown in Figure 4 [Figure 4: see original paper] and Figure 5 [Figure 5: see original paper], the x-axis represents different pattern graphs; the y-axis represents matched subgraph count. On both datasets, Fuzzy-M-TBRE returns more matches than M-TBRE for the same pattern graph. Each matched subgraph corresponds to a breast cancer patient whose treatment plan can inform the pattern graph patient's treatment. Fuzzy constraints thus provide more treatment options, demonstrating their necessity.

Figure 4. Number of matched subgraphs of different pattern graphs on Female-breast-cancer-2013a.

Figure 5. Number of matched subgraphs of different pattern graphs on Breastcancer-femalepatient-2016A.

6. Conclusion

This paper introduces the GPM problem in MKGs with formal definitions. To address this problem, we propose the M-TBRE algorithm, which decomposes pattern graphs into subgraphs, employs multi-threaded bidirectional routing for

subgraph matching, and merges results. Fuzzy constraints are introduced to obtain additional matching subgraphs, where each match corresponds to a past patient with similar physical condition to the pattern graph patient. Treatment plans from matched patients can inform the pattern graph patient's treatment, enabling better and more effective treatment development. Experimental validation on two public MKG datasets demonstrates M-TBRE's superior performance. The necessity of fuzzy constraints is also proven, as Fuzzy-M-TBRE outperforms M-TBRE. Future work will further improve M-TBRE and investigate dynamic graph pattern matching in MKGs with evolving pattern graph content.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 62076087 & 61906059, and the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education of China under grant IRT17R32.

The first author thanks his wife Jun Zhang, parents, and friends during his fight with lung adenocarcinoma: "I leave no trace of wings in the air, but I am glad I have had my flight."

Author Contributions: All authors contributed to writing the paper. L. Li (lilei@hfut.edu.cn) designed algorithms and experiments and provided funding; X. Du (dlx4339@163.com) designed and conducted experiments and analyzed data; Z. Tao (zctao@ustc.edu.cn) analyzed data.

References

- [1] Ma, X., Wu, J., Xue, S., et al.: A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering* (2021)
- [2] Wu, J., Zhu, X., Zhang, C., et al.: Bag constrained structure pattern mining for multi-graph classification. *IEEE Transactions on Knowledge and Data Engineering* 26(10), 2382-2396 (2014)
- [3] Hu, J., Ferguson, A.: Global graph matching using diffusion maps. *Intelligent Data Analysis* 20(3), 637-654 (2016)
- [4] Tian, Y., Patel, J.: TALE: A tool for approximate large graph matching. In: *Proceedings of the IEEE 24th International Conference on Data Engineering*, pp. 963-972 (2018)
- [5] Liu, F., Xue, S., Wu, J., et al.: Deep learning for community detection: progress, challenges and opportunities. In: *Proceedings IJCAI*, pp. 4981-4987

(2020)

- [6] Su, X., Xue, S., Liu, F., et al.: A comprehensive survey on community detection with deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21 (2021)
- [7] Fan, W., Wang, X., Wu, Y.: Finding experts by graph pattern matching. In: *Proceedings of the IEEE 29th International Conference on Data Engineering*, pp. 1316-1319 (2008)
- [8] Fan, W., Wang, X., Wu, Y.: Incremental graph pattern matching. *ACM Transactions on Database Systems* 38(3), 1-47 (2013)
- [9] Khan, A., Golab, L., et al.: Compact group discovery in attributed graphs and social networks. *Information Processing & Management* 57(2), 102054 (2020)
- [10] Ryota, S., Hitoshi, H., et al.: Social Group Discovery Extracting Useful Features using Multiple Instance Learning. *Journal of Japan Society for Fuzzy Theory & Intelligent Informatics* 28(6), 920-931 (2016)
- [11] Chikhaoui, B., Shimura, J., Wang, S.: Community Mining and Cross-Community Discovery in Online Social Networks. In: *Proceedings of the International Conference on Network-Based Information Systems*, pp. 176-187 (2020)
- [12] Fan, W., Li, J., et al.: Graph pattern matching: from intractable to polynomial time. *Proceedings of the VLDB Endowment* 3(1-2), 264-275 (2010)
- [13] Liu, G., Zheng, K., et al.: Multi-constrained graph pattern matching in large-scale contextual social graphs. In: *Proceedings of the IEEE 31st International Conference on Data Engineering*, pp. 351-362 (2015)
- [14] Liu, G., Li, L., Wu, X.: Multi-fuzzy-constrained graph pattern matching with big graph data. *Intelligent Data Analysis* 24(4), 941-958 (2020)
- [15] Ullmann, J.: An Algorithm for Subgraph Isomorphism. *Journal of the ACM* 23(1), 31-42 (1976)
- [16] Cordella, L., Foggia, et al.: A (Sub) Graph Isomorphism Algorithm for Matching Large Graphs. *IEEE transactions on pattern analysis and machine intelligence* 26(10), 1367-1372 (2004)
- [17] Tong, H., Faloutsos, C., et al.: Fast best-effort pattern matching in large attributed graphs. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 737-746 (2007)
- [18] Cheng, J., Zeng, X., Yu, J.: Top-k graph pattern matching over large graphs. In: *Proceedings of the IEEE 29th International Conference on Data Engineering*, pp. 1033-1044 (2013)
- [19] Cheng, J., Yu, J., et al.: Fast Graph Pattern Matching. In: *Proceedings of the IEEE 24th International Conference on Data Engineering*, pp. 913-922

(2008)

- [20] Song, C., Ge, T., Chen, C., Wang, J.: Event pattern matching over graph streams. *Proceedings of the VLDB Endowment* 8(4), 413–424 (2014)
- [21] Fan, W., Li, J., Luo, J., Tan, Z., Wang, X., Wu, X.: Incremental graph pattern matching. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 925–936 (2011)
- [22] Yan, X., Yu, P., Han, J.: Graph indexing: a frequent structure-based approach. In: *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pp. 335–346 (2004)
- [23] Shasha, D., Wang, J., Giugno, R.: Algorithmics and applications of tree and graph searching. In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 39–52 (2002)
- [24] Afrati, F., Fotakis, D., Ullman, J.: Enumerating subgraph instances using map-reduce. In: *Proceedings of the IEEE 29th International Conference on Data Engineering*, pp. 62–73 (2013)
- [25] Shao, Y., Cui, B., Chen, L., Ma, L., Yao, J., Xu, N.: Parallel subgraph listing in a large-scale graph. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 625–636 (2014)
- [26] Huang, J., Venkatraman, K., Abadi, D.: Query optimization of distributed pattern matching. In: *Proceedings of the 2014 IEEE 30th International Conference on Data Engineering*, pp. 64–75 (2014)
- [27] Demirci, M.: Graph-based shape indexing. *Machine Vision and Applications* 23(3), 541–555 (2012)
- [28] Choudhury, S., Holder, L., Chin, G., Ray, A., Beus, S., Feo, J.: StreamWorks: a system for dynamic graph search. In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 1101–1104 (2013)
- [29] Henzinger, M., Henzinger, T., Kopke, P.: Computing simulations on finite and infinite graphs. In: *Proceedings of the IEEE 36th Annual Foundations of Computer Science*, pp. 453–462 (1995)
- [30] Ma, S., Cao, Y., Fan, W., Huai, J., Wo, T.: Capturing topology in graph pattern matching. *ACM Transactions on Database Systems* 39(1), 4:1–4:46 (2014)
- [31] Liu, G., Liu, Y., Zheng, K., Liu, A., Li, Z., Wang, Y., Zhou, X.: MCS-GPM: Multi-Constrained Simulation Based Graph Pattern Matching in Contextual Social Graphs. *IEEE Transactions on Knowledge and Data Engineering* 30(6), 1050–1064 (2018)
- [32] Liu, G., Li, L., Liu, G., Wu, X.: Social Group Query Based on Multi-Fuzzy-Constrained Strong Simulation. *Transactions on Knowledge Discovery*

from Data 16(3), 1-27 (2021)

Author Biography

Lei Li received his Bachelor' s degree from Jilin University, Changchun, China, in 2004, his Master' s degree from Memorial University of Newfoundland, St. John' s, Canada, in 2006, and his Ph.D. from Macquarie University, Sydney, Australia, in 2012. He is currently an Associate Professor at the Key Laboratory of Knowledge Engineering with Big Data (Ministry of Education of China), Intelligent Interconnected Systems Laboratory of Anhui Province (Hefei University of Technology), and School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. His research interests include data mining, social computing, and graph computing. He is a senior member of IEEE. ORCID: 0000-0002-5374-7293

Xun Du received his Bachelor' s degree in 2019 and Master' s degree in 2022 from Hefei University of Technology, China. His research interests are in graph computing and social computing. He currently works as a software development engineer at Honor Terminals LTD. ORCID: 0000-0002-7020-5883

Zan Zhang received his Ph.D. in Computer Science from Hefei University of Technology, China, in 2018. He is currently a lecturer at Hefei University of Technology. His research interests include data mining and knowledge engineering. ORCID: 0000-0002-6383-1683

Zhenchao Tao was born in Wuhu, Anhui, China, in 1985. He received his M.M. degree in oncology from Anhui Medical University in 2012. From 2012 to 2015, he was a Resident, and from 2015 to 2017, an Attending Physician at the Department of Radiotherapy, Anhui Cancer Hospital. Since 2017, he has been an Attending Physician with the Department of Radiotherapy, The First Affiliated Hospital of University of Science and Technology of China. He has published five articles as first author and participated in three provincial and national projects. ORCID: 0000-0001-8142-9164

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv –Machine translation. Verify with original.