Visualization of Matching-Graphs

An Approach with Multidimensional Scaling

Bachelor Thesis
Faculty of Science, University of Bern

submitted by

Kristjan Bardheci

from Gjakovë, Kosovo

Supervision:

PD Dr. Kaspar Riesen
Mathias Fuchs
Institute of Computer Science (INF)
University of Bern, Switzerland

Abstract

In this work, we introduce a new way to analyze and compare graphs, which are complex data structures commonly used in fields such as Machine Learning and Pattern Recognition. Graphs can be challenging to understand and compare due to their complexity and variability.

To tackle this challenge, we look into matching-graphs, a technique that captures the similarities between two different graphs. This provides a simplified, yet meaningful way to understand their structures and shared patterns. We introduce two such techniques, named Substitution Way (Subway) and Partial Edit Pat Way (Pepway).

We introduce two novel measures to assess the structural similarities between original graphs and their matching-graphs, with a specific focus on Subway. We use two-dimensional plots generated by Multidimensional Scaling to visually represent the similarities and investigate emerging patterns.

Our work analyses five datasets: COX-2, PTC(MR), NCI1, LETTER and IMDB. The results reveal that the effectiveness of pruning and selection strategies in Subway is context dependent. They show high effectiveness in simplifying complex molecular graphs, such as those in the PTC(MR), COX-2 and NCI1 dataset. However, their performance declines for smaller graphs, such as those in the LETTER dataset, and for graphs with high average node-degrees or absent node labels, such as those found in the IMDB dataset.

We also reveal the limitations of the measures when dealing with matching-graphs generated by Pepway.

As part of future work, we propose to adapt the boundary to a different model that aims to capture a broader set of matching-graphs that are in close proximity to their original graphs, but not accounted for by our current model.

Acknowledgments

I would like to express my deepest gratitude to my professor and supervisor, PD Dr. Kaspar Riesen, for his guidance, patience and invaluable advice throughout the course of this thesis. His expertise in the field has been instrumental in shaping this work and his dedication to his students is truly inspiring.

A special mention goes out to Mathias Fuchs. Despite being occupied with his own PhD studies, Mathias took on the task of guiding me through this journey and devoted a remarkable amount of time to supervising and assisting me. His patience, careful observations and tireless efforts have been of immense help and have contributed greatly to the completion of my thesis.

None of this would have been possible without the constant support from my family and friends. Their encouragement and belief in my abilities have given me the strength to overcome challenges and persevere through the most difficult times. I dedicate this work to them.

I am grateful for the journey this thesis has taken me on and for the knowledge and experience I have gained along the way. The challenges I have overcome have shaped me both as a researcher and as an individual.

Contents

Chapter 1

Introduction

Artificial Intelligence (AI) is a branch of computer science that aims to understand human intelligence and construct intelligent computers based on it [1]. For a computer to be perceived as “acting humanly”, it must successfully communicate in a human language (natural language processing), store what it knows (knowledge representation), draw conclusions (decision making), adapt to new circumstances and recognise patterns (machine learning), perceive objects (computer vision), and be able to move and manipulate objects (robotics) [2]. Rich and Knight define AI as “being concerned with how to make computers do things that humans can do better at the moment” [3]. Our ability to adapt to different environmental conditions and to change our behaviour accordingly through learning processes is an outstanding feature of human intelligence. It is precisely this distinct ability to learn, far beyond that of computers, that makes machine learning, as defined above, a core area of AI [1].

Machine learning (ML) is a sub-field of AI and is concerned with how to construct machines that improve themselves through experience. In other words, an ML system learns from data. The rapid growth and greater accessibility of online data, as well as cheaper computing capacity, together with the development of innovative learning algorithms, have contributed significantly to recent developments in machine learning. In the world of AI, many developers have found that it is often easier to train systems to learn from examples than to manually program the desired output to all possible inputs. This has led to ML becoming the preferred method in AI for many applications [4].

Pattern recognition (PR) is a subfield of AI that uses ML algorithms to try to recognise patterns in data and categorise them as accurately as possible [5]. For us humans, recognizing patterns is an intuitive process, as our brains have evolved over millions of years to ensure our survival. For example, when we perceive a green apple, we usually categorise it as “good”. A brown apple, on the other hand, we tend to categorise as “bad”. It is only natural that we also try to reproduce this ability in machines [6]. In PR, there are two approaches to represent patterns: the *statistical* and the *structural* approach.

In the statistical approach, the patterns to be classified are represented by a set of features in a vector. Each pattern represents a so-called feature vector in the multidimensional vector space. On the one hand, this form of representation enables efficient calculations between the vectors. On the other hand, the feature vectors do not allow direct connections between the patterns [7].

This is where the structural approach comes in. A central method of structural PR deals with the search for similarities in patterns. [8]. The *graph-based pattern representation* is often used in structural PR. In these representations, patterns are mapped as graphs where nodes represent features and edges represent relationships between features. *Attributed relational graphs* are graphs in which specific attributes are assigned to both nodes and edges [9]. In the context of this work, we will use the term “label” in place of “attribute”. This form of representation provides a flexible method for modelling and analysing the structural complexity of patterns. Although graph-based pattern representation is very useful, it also brings challenges. Unlike feature vectors, which are relatively easy and efficient to analyse and compare due to their consistent dimension and uniform nature, graphs are more complex to handle. A major reason for this is that the elements of a graph, nodes and edges, are usually neither ordered nor of fixed size. A direct comparison between two graphs therefore requires the consideration of all potential node combinations, which adds considerable complexity due to the exponentially increasing number of possible combinations as the number of nodes increases [7].

An efficient approach is inexact or *error-tolerant graph matching.* One such method utilizes graph kernels, powerful tools that transform graph data into a format that machine learning algorithms, such as *support vector machines*, can efficiently process. By transforming and comparing graphs in a high-dimensional space, these kernels provide a means of assessing graph dissimilarity [10]. It’s worth noting that dissimilarity and similarity are inversely proportional in this context, a high degree of dissimilarity corresponds to a low degree of similarity (and vice versa).

The idea of *graph edit operations*, such as node addition or deletion, leads to the concept of *graph edit costs*, which quantify the *effort* needed to transform one graph into another [11]. This notion was central to the development of the first error-tolerant graph matching algorithm proposed by Tsai and Fu in 1979, which was based on tree search and graph edit costs [12]. In this context, the *graph edit distance (GED)* is introduced as a crucial tool for assessing graph dissimilarity.

It measures the minimal total cost of an *edit path*, which is a sequence of edit operations that transform a source graph into a target graph. Thus, a larger GED value implies greater dissimilarity between the two graphs [7]."

These computed dissimilarities play a critical role in graph classification tasks. For instance, the *k-nearest neighbour (k-NN)* classification algorithm can use these dissimilarities to classify graphs within a dataset [13]. However, it's worth noting that exact computation of the GED falls under the category of *quadratic assignment problems*, a class of NP-complete problems. The computational complexity of the GED, especially when based on a tree search algorithm, grows exponentially with respect to the number of nodes in the graphs, making it impractical for large graphs [7].

Riesen and Bunke have introduced an approximate graph edit distance algorithm using bipartite graph matching, which we will henceforth refer to as the *bipartite graph edit distance (BP-GED)*. Notably, the BP-GED can accomplish graph matching in cubic time, significantly enhancing efficiency [14]. It is widely used in fields such as image analysis, handwritten document analysis, biometrics, and bio- and chemoinformatics [15]. The BP-GED plays a central role in computing the GEDs between all pairs of graphs within a given class, enabling the creation of *matching-graphs* as proposed by Fuchs and Riesen [16]. This work also adopts the BP-GED as the method of choice to compute the GED for all graph pairs within a class.

The basic idea of matching-graphs is to highlight the similarities between two graphs. They are derived from the information contained within the edit path. There are two primary methods to create an matching-graph. One aims to create small, compact matching-graphs that encapsulate the core of two original graphs while the latter uses segments of the edit path to create new graph structures [16].

In this thesis, we undertake an in-depth analysis of these matching-graphs, using the statistical method of *multidimensional scaling (MDS)* for our investigation. Multidimensional scaling, a distance-based dimension reduction technique, allows us to visually examine the similarities between the original graphs and the generated matching-graphs on a two-dimensional plot, providing an intuitive understanding of their relationships. In addition, our exploration goes beyond qualitative evaluation, as we introduce a novel method to quantitatively measure the effectiveness of the matching-graphs. This combination of qualitative and quantitative analysis promises a comprehensive understanding of the potential and limitations of matching-graphs in capturing graph similarities.

This thesis is organized as follows. Chapter 2 lays the foundation by defining key terms and graph-based concepts, provides an overview of the GED, BP-GED and matching-graphs, and offers a brief discussion on MDS. In Chapter 3 we perform practical applications of these theories, introduce a novel method and present the results of experiments on various datasets. Finally, in Chapter 4 we summarize our main findings and suggest possible directions for future research.

Chapter 2

Basic Concepts

This chapter provides an overview of the basic concepts and terminology underlying this thesis. We begin with the formal definitions of graphs, subgraphs, and graph matching, following the foundation laid by Riesen [7] in Section 2.1. The notion of GED and its more efficient variant, BP-GED, are introduced in Sections 2.2. We explore the idea of matching-graphs in Section 2.3. Finally, Section 2.4 contains a brief discussion on MDS, which is crucial for our proposed method.

2.1 Graphs

Graphs are robust tools for visualising and analysing diverse relational data, with applications ranging from representing molecular structures, to understanding handwriting patterns, to revealing connections in social networks [17].

Basic Graph Theory

A *graph* g, defined as a three-tuple $g = (V, E, \mu)$, consists of a finite set of nodes V, edges $E \subseteq V \times V$, and a node labeling function μ. We will primarily deal with graphs having either *labelled* or *unlabelled* nodes and unlabelled edges (see Figure 2.1). Edges are defined by pairs of nodes $(u, v) \in V \times V$, where nodes u and v are *adjacent* nodes. We will focus on *simple graphs*, which are undirected with at most one edge between any two nodes. The degree of node u $deg(u)$ is the number of *incident* edges to u. Empty nodes or edges are denoted by ε.

By removing nodes, their incident edges and possibly some additional edges from a graph g_1, a *subgraph* $g_2 \subseteq g_1$ is obtained. Figure 2.1 illustrates this concept, where graph (c) is a subgraph of graph (b).

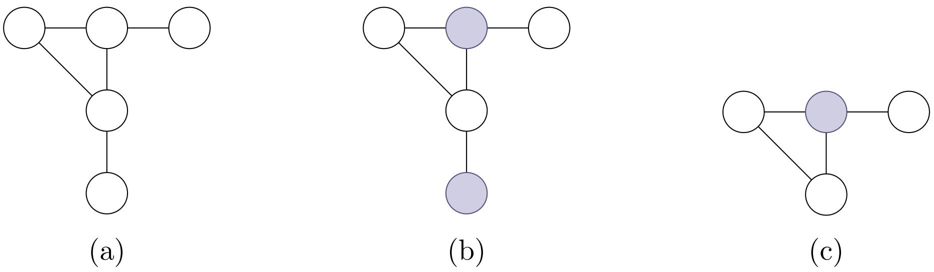


Figure 2.1: Illustration of three simple graphs: (a) has unlabelled nodes, (b) has labelled nodes, graph (c) is a subgraph of graph (b).

2.2 Graph Edit Distance

The GED is a measure that evaluates the amount of distortion required to transform the source graph g_1 into the target graph g_2 using basic edit operations such as *insertion*, *deletion* and *substitution* on nodes and edges. We denote the substitution of two nodes $u \in V_1$ and $v \in V_2$ by $(u \rightarrow v)$, the deletion of node $u \in V_1$ by $(u \rightarrow \varepsilon)$ and the insertion of node $v \in V_2$ by $(\varepsilon \rightarrow v)$. Similar notation is used for edge edit operations.

A set $\{e_1, \dots, e_k\}$ of k edit operations e_i that transform a graph g_1 completely into another graph g_2 is called an *edit path* $\lambda(g_1, g_2)$ between g_1 and g_2. The set of all such edit paths from g_1 to g_2 is denoted by $\Upsilon(g_1, g_2)$.

The cost function $c(e_i)$ measures the intensity of an edit operation e_i, indicating the amount of graph modification caused. A low-cost edit path implies minor modifications between similar graphs, while a high-cost edit path is required for more dissimilar graphs [7].

Figure 2.2 illustrates an edit path from graph g_1 to g_2. The edit path is defined as:

$$\lambda = \{(u_1 \rightarrow v_2), (u_2 \rightarrow v_3), (u_3 \rightarrow v_4), (u_4 \rightarrow v_5), (u_5 \rightarrow \varepsilon), (\varepsilon \rightarrow v_1)\} \tag{2.1}$$

and implies the following edge edit operations:

$$\begin{aligned} &\{((u_1, u_2) \rightarrow (v_2, v_3)), ((u_2, u_3) \rightarrow (v_3, v_4)), ((u_2, u_4) \rightarrow (v_3, v_5)), \\ &((u_1, u_4) \rightarrow \varepsilon), (\varepsilon \rightarrow (v_4, v_5)), (\varepsilon \rightarrow (v_1, v_2)), (\varepsilon \rightarrow (v_1, v_3))\}. \end{aligned} \tag{2.2}$$

The GED is defined as follows. Let g_1 be the source and g_2 be the target graph. The *GED* between g_1 and g_2 is defined as:

$$d_{\lambda_{min}}(g_1, g_2) = \min_{\lambda \in \Upsilon(g_1, g_2)} \sum_{e_i \in \lambda} c(e_i) \tag{2.3}$$

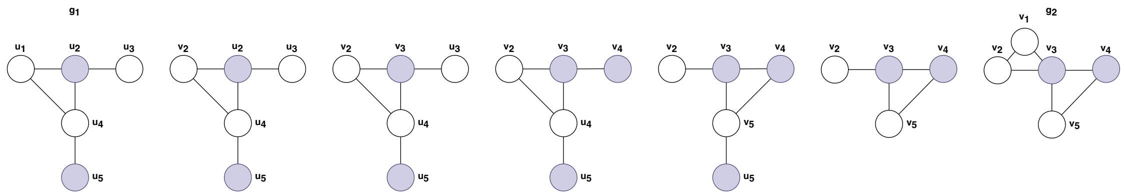


Figure 2.2: Transformation from graph g_1 into g_2 using the edit path $\lambda = \{(u_1 \to v_2), (u_2 \to v_3), (u_3 \to v_4), (u_4 \to v_5), (u_5 \to \varepsilon), (\varepsilon \to v_1)\}$.

where $d_{\lambda_{min}}(g_1, g_2)$ is not necessarily unique.

Calculating the optimal path is inherently complex due to the exponentially large number of possible edit paths. This complexity places GED within the realm of NP-complete problems, under the family of *quadratic assignment problems (QAPs)* [18].

Bipartite Graph Edit Distance

Bipartite graph edit distance simplifies the GED problem by transforming it into a *Linear Sum Assign Problem (LSAP)*, which shifts the complexity from exponential to cubic. In addition to this transformation, deriving upper and lower bounds as an approximation for the GED can guide the search for optimal solutions and help to eliminate non-optimal paths early in the process.

The shift also allows for the use of established LSAP-solving algorithms, such as Munkren's. As a result, BP-GED provides a more practical approach to graph matching [7]. This work employs the BP-GED algorithm.

2.3 Matching-Graphs

The concept of matching graphs is introduced as an innovative data structure to embody information about the matches between graph pairs. Proposed by Fuchs and Riesen [16], it extracts and encapsulates data about matching components from two graphs, creating a new structure that encapsulates matching nodes and edges.

Two methods for forming matching-graphs, *Substitution Way (Subway)* [19] and *Partial Edit Path Way (Pepway)* [20], are presented. Both methods depend on the GED computation, which results in an edit path for each pair of graphs from the same class. This edit path forms the basis for matching-graph construction.

A matching-graph created by a source graph g_1 and a target graph g_2 is denoted as $m_{g_1 \times g_2}$. For each edit path $\lambda(g_1, g_2)$, two matching-graphs $m_{g_1 \times g_2}$ and $m_{g_2 \times g_1}$ are built for the source and the target graph g_1 and g_2, respectively.

For the sake of simplicity, a unit cost of 1.0 for deletions and insertions of both nodes and edges were employed and for the BP-GED algorithm, that approximates the GED, a weighting parameter $\alpha \in [0, 1]$ that is used to trade-off the relative importance of node and edge edit costs was optimized (on all datasets) [16].

2.3.1 Subway

The Subway approach creates compact matching graphs that represent *core* or *diversified* parts of the original graphs. These matching-graphs are always subgraphs of the original graphs, implying that $m_{g_1 \times g_2}$ and $m_{g_2 \times g_1}$ are subgraphs of g_1 and g_2, respectively. They also keep the node labels from the original graphs.

All nodes of g_1 and g_2 that are actually substituted in the edit path $\lambda(g_1, g_2)$ are added to $m_{g_1 \times g_2}$ and $m_{g_2 \times g_1}$, respectively. Conversely all nodes that are deleted in g_1 or inserted in g_2 are not considered in the matching-graphs. Isolated nodes, i.e. nodes without any adjacent nodes are removed in the process [19].

Two different strategies for *edge handling* were proposed [16]:

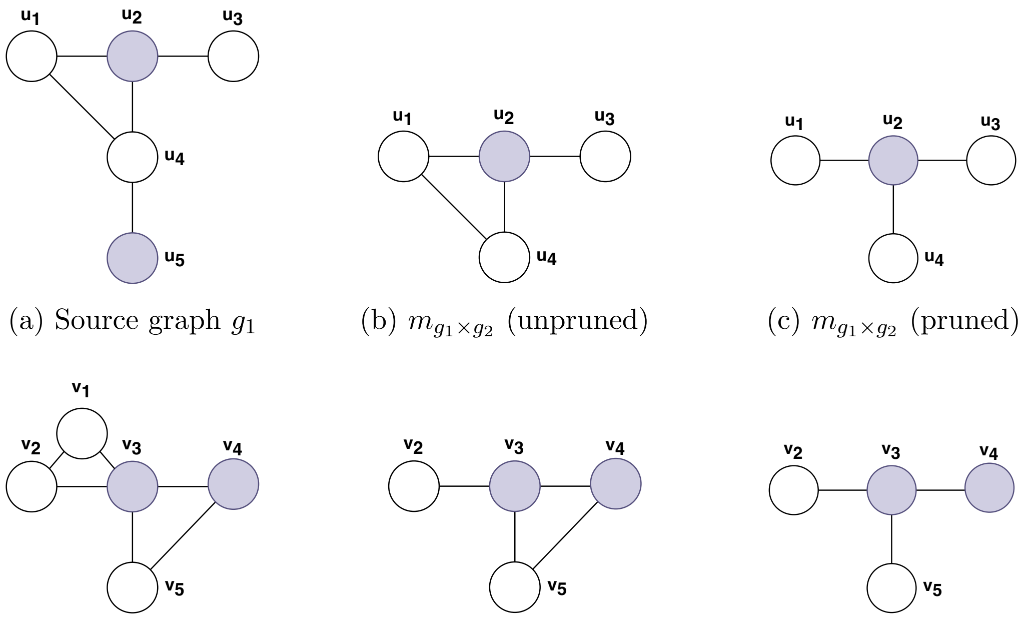
- *No Pruning:* If two nodes $u_1, u_2 \in V_1$ of a source graph g_1 are substituted with nodes $v_1, v_2 \in V_2$ in a target graph g_2 and there exists an edge $(u_1, u_2) \in E_1$, (u_1, u_2) is included in the matching-graph $m_{g_1 \times g_2}$ regardless whether or not edge (v_1, v_2) is available in E_2. The edge is *unpruned.*

- *Pruning:* We assume the same scenario as before. Edge (u_1, u_2) is only included in the matching-graph $m_{g_1 \times g_2}$ if and only if there exists an edge (v_1, v_2) in E_2 . So in cases where there is no corresponding edge in the other graph, the edge is *pruned.*

Figure 2.3 illustrates the Subway procedure on two graphs: (a) is the source graph g_1 and (d) is the target graph g_2. The edit path for this example is referred to in Equation 2.1. The two graphs that are generated without pruning are graphs (b) and (e). With edge pruning applied, matching-graphs (c) and (f) are generated. In the pruned matching-graph (c), we can see that the edge $(u_1, u_4) \in E_1$ has no counterpart in the target graph ($(v_2, v_5) \notin E_2$) and is therefore not included in the matching-graph. Matching-graph (f) follows the same principle.

Consider a set of training graphs G_{ω_l} containing n graphs that belong to the same class ω_l. The generation of all possible matching-graphs for all pairs of graphs (g_i, g_j) for $i, j = 1, \ldots n$ results in $n \times (n-1)$ matching-graphs. This can lead to a significantly large set of matching-graphs M_{ω_l}, especially when n is large. To maintain a manageable size, two selection methods were proposed with the aid of the *set median graph* [21]. The set median graph $g_{mdn} \in S$ is defined as:

$$g_{mdn} = \underset{g_i \in S}{\operatorname{argmin}} \sum_{g_j \in S} d(g_i, g_j) \tag{2.4}$$



(a) Source graph g_1 (b) $m_{g_1 \times g_2}$ (unpruned) (c) $m_{g_1 \times g_2}$ (pruned)

(d) Target graph g_2 (e) $m_{g_2 \times g_1}$ (unpruned) (f) $m_{g_2 \times g_1}$ (pruned)

Figure 2.3: Matching-graphs with unpruned and pruned edges derived from source graph g_1 and target graph g_2.

where S is an arbitrary set of graphs. The set median graph is the graph of S whose sum of distances to all other graphs in S is minimal. The selection process is iterative and the two methods are described as follows [21]:

- *Center selection*: This involves selecting matching-graphs that are considered *similar* as they are positioned near the center of the total set of matching-graphs. 80 of these graphs are chosen for each class, characterized by having the minimum total graph edit distances when compared with all other matching-graphs in the same class. They symbolize the *core* or *central* features of the original graphs.

- *Spanning selection*: This involves choosing *diverse* matching-graphs, which are located farthest from the center of the set. Like the center selection, 80 graphs per class are selected, but these have the maximum total graph edit distances when compared to all other matching-graphs. They represent the *diversified* or *varied* aspects of the original graphs.

2.3.2 Pepway

This method revolves around the idea of randomly choosing a certain proportion of all available edit operations in the edit path, resulting in a partial edit path with a subset of operations. These operations are then implemented on pairs of graphs according to specific rules, generating two new graphical representations.

Suppose we have a set of training graphs G_ω, and for every pair of graphs, (g_i, g_j) within this set, we derive an edit path $\lambda(g_i, g_j) = \{e_1, \ldots, e_s\}$. Matching-graphs are created for each edit path $\lambda(g_i, g_j)$.

In the context of this work, matching-graphs are created according to the following procedure. We randomly select a percentage $p \in \{0.25, 0.5, 0.75, 1.0\}$ of all s available edit operations in $\lambda(g_i, g_j)$. This provides us with a partial edit path $\tau(g_i, g_j) = \{e_1, \ldots, e_t\} \in \lambda(g_i, g_j)$ with $t = \lfloor p \cdot s \rfloor$ edit operations. Each operation $e_i \in \tau(g_i, g_j)$ is applied on both graphs g_i and g_j according to following rules:

- If e_i indicates a deletion, e_i is applied only to $m_{g_i \times g_j}$.
- If e_i indicates an insertion, the node that would be inserted in $m_{g_i \times g_j}$ is instead deleted in $m_{g_j \times g_i}$.
- If e_i refers to a substitution, e_i is applied on both g_i and g_j, swapping the labels of the matching nodes in $m_{g_i \times g_j}$ and $m_{g_j \times g_i}$.

We would like to use the example of a transformation from a source graph g_1 to a target g_1 that we used in the section before (see Figure 2.2). For clarity, we will display the edit graph here of that transformation again:

$$\lambda = \{(u_1 \to v_2), (u_2 \to v_3), (u_3 \to v_4), (u_4 \to v_5), (u_5 \to \varepsilon), (\varepsilon \to v_1)\}.$$

Suppose $p = 0.5$ and $s = 6$, this leads to $t = \lfloor 0.5 \cdot 6 \rfloor = 3$, which means that three edit operations from the edit path will be performed. With this information, we choose the partial edit path τ randomly as follows:

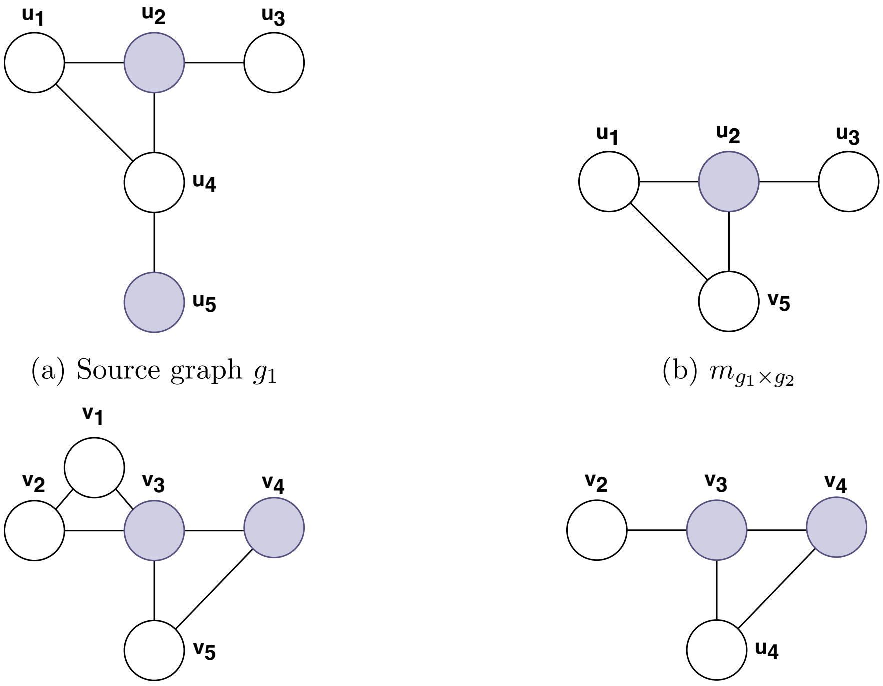
$$\lambda = \{(u_1 \to v_2), (u_2 \to v_3), (u_3 \to v_4), \overbrace{(u_4 \to v_5), (u_5 \to \varepsilon), (\varepsilon \to v_1)}^{\tau}\}.$$

This leads to following partial edit path τ:

$$\tau = \{(u_4 \to v_5), (u_5 \to \varepsilon), (\varepsilon \to v_1)\} \tag{2.5}$$

We apply the edit operations on g_1 and g_2 according to the rules mentioned to receive the matching-graphs $m_{g_1 \times g_2}$ and $m_{g_2 \times g_1}$ as shown in Figure 2.4.

In this work, we randomly select 10 graphs from each class. Subsequently, we create all possible matching-graph combinations. This process results in a total of 90 unique matching-graphs.



(a) Source graph g_1

(b) $m_{g_1 \times g_2}$

(c) Target graph g_2

(d) $m_{g_2 \times g_1}$

Figure 2.4: Matching-graphs derived from source graph g_1 and target graph g_2 using Pepway.

2.4 Multidimensional Scaling

There exist several techniques for dimensionality reduction, each with its own strengths and unique approach. *Multidimensional Scaling (MDS)*, *Principal Component Analysis (PCA)*, *t-Distributed Stochastic Neighbor Embedding (t-SNE)*, and *Linear Discriminant Analysis (LDA)* are notable examples. All of the methods mentioned are multivariate techniques used to analyze multiple variables simultaneously. In PCA, original variables are transformed into a new set of uncorrelated variables (principal components), ordered by the amount of variance they explain [22]. LDA aims to find a linear combination of features that characterizes or separates two or more classes of objects [23]. The t-SNE technique is a method, which uses probability distributions to represent similarities between points [24]. Multidimensional scaling attempts to represent the data in a lower dimensional space, often two or three dimensions, such that the distances (or dissimilarities) between points are preserved as much as possible [25].

For our specific application, we have chosen to use MDS, more precisely, the *metric* MDS variant. The reason for this is because we will be dealing with a *dissimilarity matrix* encompassing all pairs of graphs and matching-graphs within a class, and metric MDS is a distance-based method, making it an appropriate

	$\mathbf{g_1}$	$\mathbf{g_2}$	$\mathbf{g_3}$	$\mathbf{g_4}$
$\mathbf{g_1}$	0	1.2	3.6	2.4
$\mathbf{g_2}$	1.2	0	2.4	1.2
$\mathbf{g_3}$	3.6	2.4	0	1.2
$\mathbf{g_4}$	2.4	1.2	1.2	0

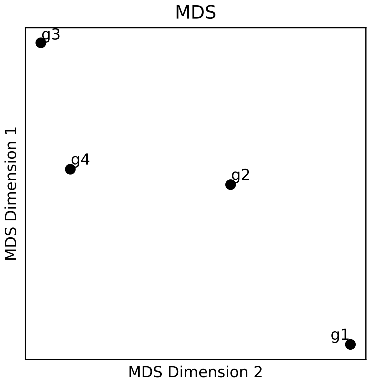


Figure 2.5: Symmetric dissimilarity matrix (left) and the corresponding MDS plot (right) for four graphs. The entries in the matrix are the pairwise GED values. A high dissimilarity in the matrix corresponds to a high distance in the plot.

choice for our data.

Assume you have a dissimilarity matrix $\mathbf{D} = [d_{ij}]$ for $i, j = 1, \ldots, n$, where d_{ij} represents the dissimilarity between the i^{th} and j^{th} objects in your dataset. MDS seeks a set of points in a lower-dimensional space that preserves these dissimilarities as much as possible.

We operate under the assumption of a symmetric dissimilarity matrix, despite the approximated nature of the GED using the BP-GED algorithm. This symmetry ensures equal distances $d(g_i, g_j) = d(g_j, g_i)$ for all graph pairs (g_i, g_j).

Figure 2.5 illustrates this process for a simplified example of four graphs. The dissimilarity matrix (left) maps the graph edit distances between the graphs, and the MDS plot (right) visually represents these distances in two dimensions. Note that the points in the MDS plot have been positioned such that their Euclidean distances approximate the original graph edit distances from the dissimilarity matrix as closely as possible.

The objective cost function that MDS tries to minimize (stress) is defined as follows:

$$\text{stress}(X) = \frac{\sum_{i<j}(d_{ij} - ||x_i - x_j||)^2}{\sum_{i<j} d_{ij}^2} \tag{2.6}$$

Here, $X = \{x_1, x_2, ..., x_n\}$ are the points in the low-dimensional space, d_{ij} is the dissimilarity between objects i and j, and $||x_i - x_j||$ is the Euclidean distance between points x_i and x_j in the low-dimensional space [25].

The *SMACOF (Scaling by MAjorizing a COmplicated Function)* algorithm updates the positions of points x_i iteratively to minimize the stress. In each iteration, the algorithm makes use of a B-matrix, whose elements $b_{ij}^{(t)}$ are calculated as $b_{ij}^{(t)} = -\frac{d_{ij}}{||x_i^{(t)} - x_j^{(t)}||}$. The B-matrix contains scaled dissimilarities between the points [25].

The update in each iteration is given as:

$$x_i^{(t+1)} = \frac{\sum_{j \neq i} w_{ij}^{(t)} b_{ij}^{(t)} x_j^{(t)}}{\sum_{j \neq i} w_{ij}^{(t)} b_{ij}^{(t)}} \tag{2.7}$$

Here, t is the iteration number, and $w_{ij}^{(t)} = \frac{d_{ij}}{||x_i^{(t)} - x_j^{(t)}||}$ are the weights that help in the minimization of the stress. The algorithm continues until the change in stress is smaller than a given threshold, or a maximum number of iterations is reached.

Chapter 3

Experiment

In this chapter, we engage in a comprehensive analysis of matching-graphs derived from a single class. We use MDS to identify patterns within these graphs and introduce several methods and measures for quantitative and qualitative analysis of matching-graphs. We then present the datasets used in our experiments. The chapter concludes with a detailed analysis of our experimental results, highlighting examples that clearly show unique patterns in the matching-graphs.

For the purpose of our analysis, we use the best matching graphs for each class, as determined by their highest classification accuracy [16].

3.1 Methods and Measures

In this section, we present the methods and key measures that form the basis of our comprehensive analysis. This includes simple MDS plots with all graphs present, the innovative concept of *Circle-Bound*, as well as quantitative measures such as the average node degree and the average node/edge size of original and matching-graphs.

Circle-Bound

Circle-Bound is a key element of our analytical framework. The concept works with a set of matching-graphs M_{ω_l}, where ω_l denotes a specific class. As seen in Figure 3.1, the circle is defined by selecting its center to be the midpoint between the two original graphs g_1 and g_2. The radius of this circle is then the distance from this midpoint to either of the original graphs. The circle acts as a boundary. Central to this idea is our hypothesis that matching-graphs, that encapsulate the core parts of two original graphs, should ideally be located within the circle. If a matching-graph is located near the center of the circle, it suggests that it represents

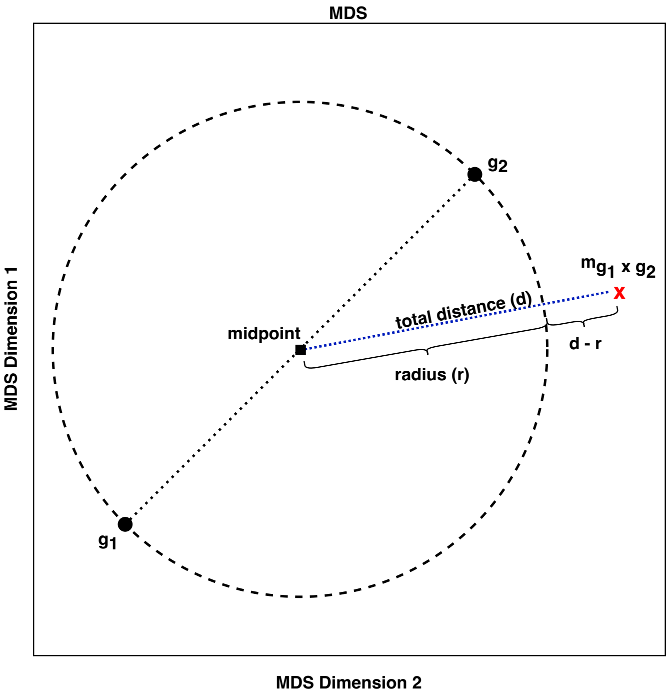


Figure 3.1: Illustration of the Circle-Bound concept on an MDS plot. The black dots g_1 and g_2 refer to source and target graphs, respectively. The red x refers to the matching-graph $m_{g_1 \times g_2}$ that was created from g_1 and g_2.

both original graphs equally well.

If a matching-graph lies outside the circle, it means that the matching-graph has properties that go beyond a balanced representation of the original graphs' properties. In our analysis, we aim to quantify the similarities between graphs. We do this by calculating the percentage of matching-graphs within a given category that fall within a defined circle. The equation we use to determine this is as follows:

$$\text{In-Circle} = \frac{|M_{\omega_l}^{in}|}{|M_{\omega_l}|} \times 100 \tag{3.1}$$

In this equation, $\pi_{\text{in}}(M_{\omega_l})$ denotes the In-Circle percentage for the set of matching-graphs M_{ω_l}, $|M_{\omega_l}^{in}|$ represents the number of matching-graphs in the set that lie inside the circle, and $|M_{\omega_l}|$ is the total number of matching-graphs in the set.

Another quantitative measure, *Out-Diff*, is calculated to assess the extent to which matching-graphs exceed the circle boundary. It is defined as the average relative outbound difference for the matching-graphs that lie outside the circle, as given by the equation:

$$\text{Out-Diff} = \frac{1}{|M_{\omega_l}^{out}|} \sum_{m \in M_{\omega_l}^{out}} \frac{d_m - r_m}{d_m} \times 100 \tag{3.2}$$

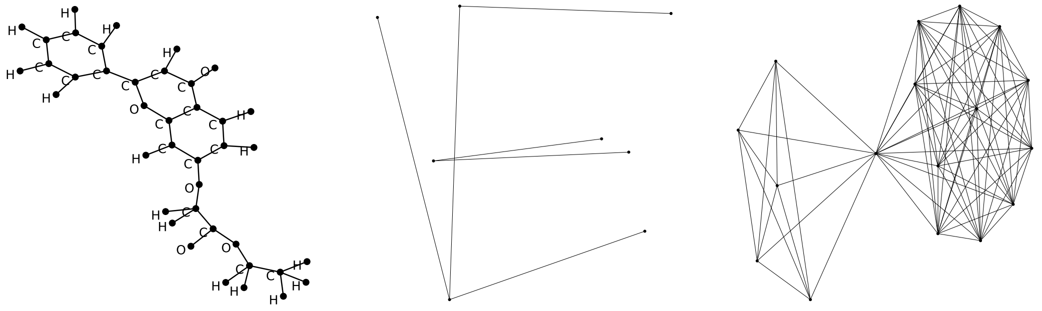
The terms d_m and r_m correspond to the total distance from the midpoint to the matching-graph and the radius of the circle for each matching-graph m, respectively.

Measures

We define several measures to assess the properties of the graphs within a class, both for the original graphs and for the matching-graphs. These measures include the average node size (∅ Nodes) and edge size (∅ Edges), as well as the average node degree (∅ deg). Applying the algorithm to the dataset may change these quantities, and we quantify this change by calculating the relative percentage change ($\Delta(\%)$). For clarity, the original graphs will be denoted as og and the matching-graphs as mg.

3.2 Datasets

We use five datasets representing chemical compounds in different contexts. The AIDS dataset, derived from the IAM Graph Repository [26] and originally collected by the National Cancer Institute (NCI), includes compounds that effectively protect human cells against HIV (Confirmed Active) and those that do not (Confirmed Inactive). The Mutagenicity (MUTA) dataset [26] classifies compounds into mutagenic and non-mutagenic categories. Other datasets in this category, including NCI1, PTC(MR) and COX-2, are taken from Morris et al. [17]. The NCI1 dataset consists of compounds from anticancer screens, divided into those that inhibit the growth of lung cancer (active) and those that do not (inactive). The PTC(MR) dataset contains potentially carcinogenic compounds, while the COX-2 dataset contains COX-2 inhibitors, both active and inactive. The nodes in these datasets represent atoms, labelled by their chemical symbols.



(a) COX-2 (b) LETTER (c) IMDB

Figure 3.2: Example graph representations: (a) molecule from COX-2, (b) letter ”E” from LETTER and (c) actor/actress - movie connections from IMDB.

The IMDB dataset [17] is a film collaboration network, where nodes symbolise actors/actresses and edges indicate their involvement in film. Each graph represents a film. The nodes are unlabelled.

The LETTER dataset [26] contains graphs representing artificially distorted line drawings of 15 straight-line letters (A, E, F, H, I, K, L, M, N, T, V, W, X, Y, Z). Nodes represent line endpoints and edges illustrate the line drawings, with nodes labelled with their corresponding x and y coordinates.

Figure 3.2 illustrates three example graphs, (a) is a molecule from the dataset COX-2, (b) is the representation of the letter "E" from the dataset LETTER and (c) is a graph from the dataset IMDB.

Table 3.1 presents an overview of the datasets. Each row represents a separate dataset, detailing the name, total size, the number of distinct classes, type of graph representation, node labels, as well as the average number of nodes and edges per graph.

Dataset	Size	# Classes	Type	Node labels	∅ Nodes	∅ Edges
AIDS	2000	2	Molecule	Atom	15.7	16.2
MUTA	4337	2	Molecule	Atom	30.3	30.8
NCI1	4110	2	Molecule	Atom	29.9	32.3
IMDB-B	1000	2	Actor/Actress - Movie	-	19.8	96.5
COX-2	466	2	Molecule	Atom	41.2	43.4
PTC-MR	344	2	Molecule	Atom	14.6	14.7
LETTER-H	2250	15	Letter drawings	(x,y) coordinate	4.7	4.5

Table 3.1: Summary of dataset characteristics.

3.3 Results

In this chapter, we present the results of our research. To streamline the presentation and interpretation of the results, we grouped the datasets based on results for the Subway method into (NCI1 and MUTA) and (PTC(MR) and AIDS). For this method, we will present detailed results for COX-2, PTC(MR), NCI1, IMDB and LETTER. For Pepway, we grouped (MUTA, PTC(MR), NCI1 and COX-2) together and will present detailed results for COX-2. For additional datasets, we have provided tables and plots in the appendix A.

3.3.1 Subway

In the following section, we will analyse the matching-graphs generated through the Subway method, which was described in Section 2.3.1. This method generates subgraphs that aim to either represent the core parts of the original graphs (center

Class	**-1**				**1**			
Selection	**center**		**span**		**center**		**span**	
Pruning	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**
∅ Nodes								
og	40.7	40.1	42.8	45.0	39.2	37.1	41.4	44.6
mg	15.1	38.3	36.8	42.3	15.1	34.1	33.3	41.5
Δ (%)	-63.0	-4.6	-14.0	-5.9	-61.4	-8.1	-19.6	-6.9
∅ Edges								
og	43.1	42.2	45.2	47.5	41.2	39.1	43.4	46.8
mg	8.6	40.3	37.9	44.5	9.0	36.0	33.8	43.4
Δ (%)	-79.9	-4.6	-16.0	-6.3	-78.1	-7.8	-22.2	-7.3
∅ deg								
og	2.1	2.1	2.1	2.1	2.1	2.1	2.1	2.1
mg	1.2	2.1	2.1	2.1	1.2	2.1	2.0	2.1
Δ (%)	-45.8	0.0	-2.4	-0.5	-43.3	0.5	-3.3	-0.5
In-Circle (%)	40.0	58.8	22.5	28.8	6.2	61.2	7.5	38.8
Out-Diff (%)	48.2	4.0	74.9	31.4	38.6	15.8	75.9	28.2

Table 3.2: Comparison of original and matching graph characteristics for different selection and pruning approaches for dataset COX-2.

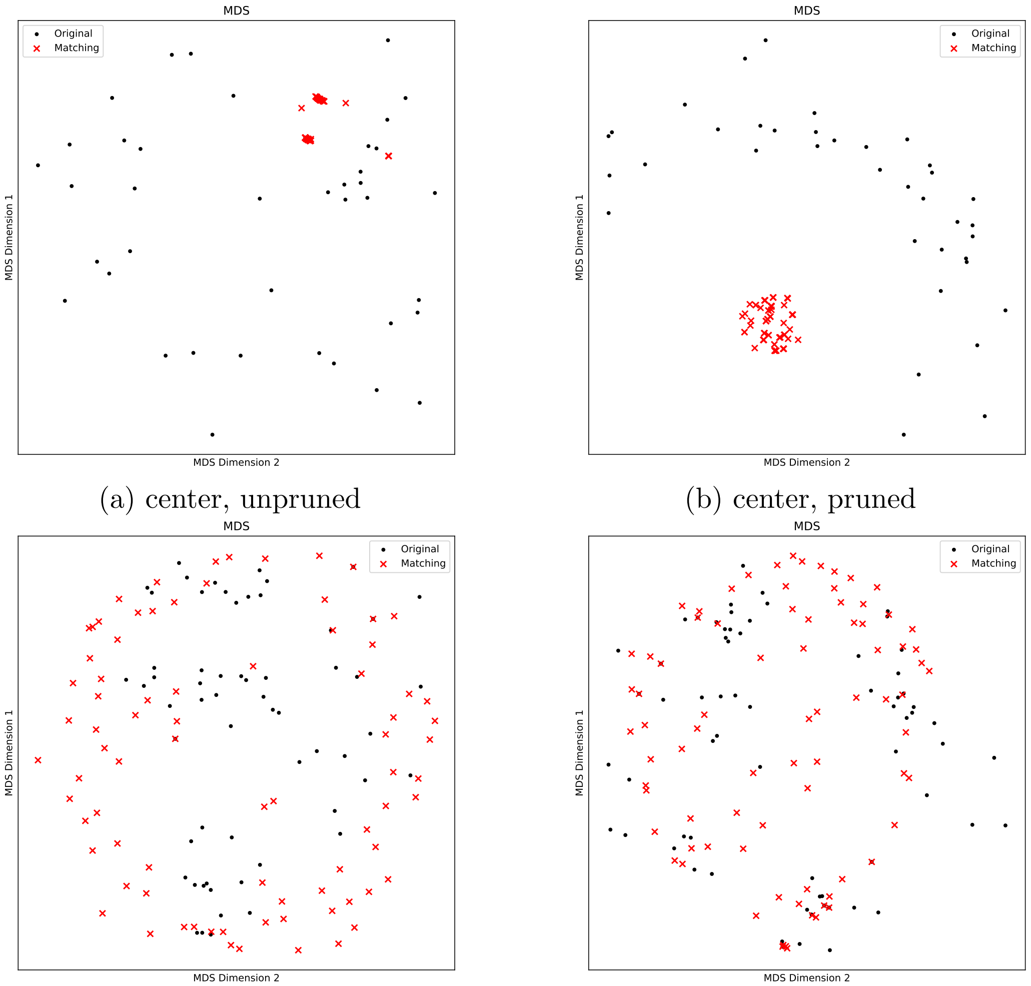
selection) or the diverse parts of the original graphs (spanning selection). It also uses a pruning mechanism to enhance the accuracy of the representation. For this, we will use the matching-graphs created from following dataset: COX-2, PTC(MR), NCI1, IMDB and LETTER.

COX-2

In Table 3.2, we can see the characteristics of the original and matching-graphs for the COX-2 dataset using either center or spanning (*span*) selected graphs that are pruned (*pr*) or unpruned (*unpr*).

Average Measures: We can see that the matching-graphs have fewer average nodes, edges, and degrees than the original graphs, indicating a significant simplification of the graph structure during the matching process. The largest percentage reduction is observed in the average number of nodes (-63.0%) and edges (-79.9%) under the pruning condition. In the case of the average degree, the most drastic reduction reductions (-45.8%) are not as drastic as nodes and edges. This might be due to the fact that the degree of a node could be more resilient to the effects of pruning, especially when nodes and edges are equally removed in a balanced ratio.

Pruning: For both classes and each selection method, the In-Circle percentages are consistently lower for pruned matching graphs compared to their unpruned counterparts. For class 1, prune center selected (40.0%) is lower than unpruned center selected (58.8%) and pruned spanning selected (22.5%) is also lower than unpruned spanning selected (28.8%). For class 1, it is more obvious with pruned center selected (6.2%) being lower than unpruned center selected (61.2%) and pruned



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

(d) spanning, pruned

Figure 3.3: MDS plots for class -1 of the dataset COX-2. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

spanning selected (7.5%) being lower than unpruned spanning selected (38.8%). This observation suggests that the unpruned matching graphs tend to be more centrally located than those that have been pruned. Conversely, we observe that the Out-Diff percentages are consistently higher for pruned matching graphs than for the unpruned ones, that is for class -1 center pruned (48.2%) greater than spanning pruned (22.5%) and center unpruned (58.8%) being greater than spanning unpruned (28.8%). This same behaviour is present in class 1. This demonstrates that when matching graphs do fall outside the inbound circle, pruned matching graphs have a larger portion lying outside this circle than unpruned ones.

Selection: When investigating the relative differences between center and spanning selection in the context of pruned and unpruned conditions, several trends emerge. Both in the pruned and unpruned state, center consistently demonstrates a higher In-Circle percentage than spanning across both classes, that is for class -1:

40.0% vs 22.5% and 55.8% vs 28.8% and for class 1: 61.2% vs 38.8%. However, an exception occurs within class 1, where the In-Circle percentage of pruned graphs under center (6.2%) is slightly lower than that under spanning (7.2%). The reason for this outlier in the otherwise consistent pattern is not immediately apparent from the current data. Furthermore, when comparing the Out-Diff percentage between center and spanning, we consistently observe that spanning produces a higher Out-Diff percentage for both pruned and unpruned graphs, irrespective of the class, that is for class -1: 48.2% vs 74.9% and 4.0% vs 31.4% and for class 1: 38.6% vs 75.9% and 15.8% vs 28.2%. This pattern validates our assumptions, that graphs chosen via the spanning method to represent diverse features, tend to disperse more broadly and often exceed the bounds of the inbound circle more than their center-selected counterparts.

Figure 3.3 illustrates the MDS plots for class -1 for different selection and pruning methods. There are two observations that need to be discussed. In center selection, we see that the pruned and unpruned matching-graphs are clustered, showing high similarity among each other. The dissimilarity of the matching-graphs to the original graphs is more evenly distributed in the pruned versions than in the unpruned ones. This might lead to the interpretation that the pruned matching-graphs better represent the core parts of all original graphs than the unpruned matching-graphs. In spanning selection, unpruned matching-graphs appear to be more evenly distribute than pruned ones. This might lead to the interpretation that unpruned matching-graphs can better represent the diversified parts of the original graphs.

The Subway algorithm seems to be suitable for these types of graphs, with pruned center selected matching-graphs well representing the core parts of the original graphs and unpruned spanning selected matching-graphs well representing the diversified parts of the original graphs.

PTC-MR

Table 3.3 presents the characteristics of original and matching graphs for the PTC(MR) dataset.

Average Measures: We can see that the matching-graphs have fewer average nodes, edges, and degrees than the original graphs, indicating a significant simplification of the graph structure during the matching process. The largest percentage reduction is observed in the average number of nodes and edges under the pruning condition (-83.9%, resp. -92.4%) . In the case of the average degree, the reductions are not as drastic as nodes and edges. Center selected matching-graphs tend to have a larger reduction of the node degree (in the range from -28.8% to -83.8%) than spanning selected ones (in the range from -13.8% to -51.2%).

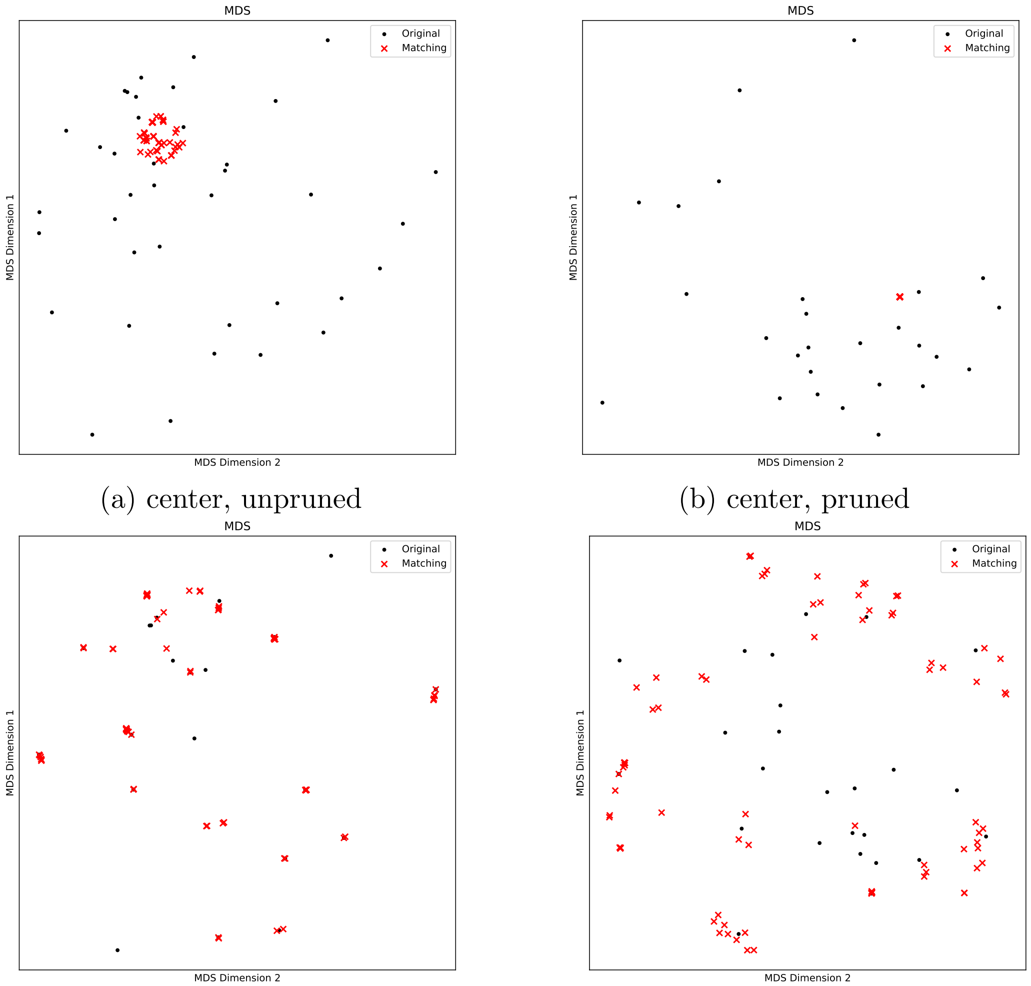
Class	**-1**				**1**			
Selection	**center**		**span**		**center**		**span**	
Pruning	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**
∅ **Nodes**								
og	15.9	14.3	17.0	24.5	12.8	10.8	15.2	24.2
mg	4.0	5.0	15.4	22.4	2.0	4.1	10.2	20.4
Δ (%)	-74.9	-64.9	-9.3	-8.5	-83.9	-61.9	-32.7	-15.4
∅ **Edges**								
og	16.6	14.9	17.7	25.9	13.3	11.1	15.8	26.2
mg	2.0	2.8	15.4	23.5	1.0	2.1	7.8	22.0
Δ (%)	-88.0	-81.2	-12.9	-9.3	-92.4	-81.0	-50.7	-16.3
∅ **Node Deg.**								
og	2.1	2.1	2.1	2.1	2.1	2.0	2.1	2.2
mg	1.0	1.1	2.0	2.1	1.0	1.0	1.5	2.2
Δ (%)	-51.9	-46.4	-4.3	-0.5	-52.2	-50.2	-26.4	-0.9
In-Circle (%)	31.2	48.8	13.8	50.0	28.8	83.8	50.0	51.2
Out-Diff (%)	44.5	20.3	39.6	15.6	30.6	19.7	50.8	16.5

Table 3.3: Comparison of original and matching graph characteristics for different selection and pruning approaches for data set PTC(MR).

Pruning: Pruned matching-graphs consistently have a lower In-Circle percentage than unpruned matching-graphs, that is for class -1: 31.2% vs 48.8% and 13.8% vs 50.0% and class 1: 28.8% vs 83.8% and 50.0% vs 51.2% and a higher Out-Diff percentage, that is for class -1: 44.5% vs 20.3% and 39.6% vs 15.6% and for class 1: 30.6% vs 19.7% and 50.8% vs 16.5%. This may indicate that pruned matching-graphs tend to lie outside the circle more often.

Selection: The In-Circle percentage for pruned and unpruned matching-graphs regarding selection is not consistent. In Class -1, pruned center selected matching-graphs (31.2%) are higher than spanning selected (13.8%). In Class 1, pruned center selected matching-graphs (28.8%) are lower than spanning selected matching-graphs (50.0%). For Class 1, we observe a contrary pattern: the pruned matching-graphs from the center selection have a lower In-Circle percentage (28.8%) than the ones from the spanning selection (50.0%). The Out-Diff is inconsistent as well. In Class -1, Out-Diff is consistently higher for center selected graphs (44.5% vs 39.6% and 20.3% vs 15.6%). In Class 1, this is not the case, prune center (30.6%) is lower than prune spanning (50.8%).

Figure A.7 illustrates the MDS plots for class -1 for different selection and pruning methods. There are two observations that need to be discussed. First, in center selection, unpruned matching-graphs are tightly clustered. This means that they are very similar to each other in reference to structure. The pruned matching-graphs seem to be all the same subgraph, which is shown as one single red x in the plot. This leads to the interpretation that while the unpruned matching-graphs capture variety of core parts of the original graphs, the pruned matching-graphs seem to



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

(d) spanning, pruned

Figure 3.4: MDS plots for class -1 of the dataset PTC(MR). Black dot markers refer to original graphs and red x markers refer to matching-graphs.

capture only one distinct subgraph that is believed to best represent the core parts.

In spanning selection, pruned matching-graphs are more evenly distributed than unpruned ones. This leads to the interpretation that pruned matching-graphs can have a wider range regarding the diversified structure and therefore more subgraph possibilities than unpruned graphs.

NCI1

Table 3.4 presents the characteristics of original and matching graphs for the NCI1 dataset.

Average Measures: For the average number of nodes and edges, it can be observed that in the matching-graphs, they are consistently lower than in the original graphs, indicating a successful reduction of complexity in the resulting matching-graphs. This reduction is particularly noticeable for pruned graphs, which show a

Class	0				1			
Selection	**center**		**span**		**center**		**span**	
Pruning	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**
∅ Nodes								
og	27.1	28.3	24.9	41.0	31.1	35.6	37.2	64.3
mg	9.6	13.5	12.4	36.7	10.7	16.5	17.1	57.9
Δ (%)	-64.8	-52.2	-50.1	-10.5	-65.5	-53.6	-54.0	-10.0
∅ Edges								
og	29.5	30.3	26.8	44.7	33.8	38.3	40.2	69.7
mg	6.3	9.0	7.6	39.4	6.8	11.6	13.6	61.9
Δ (%)	-78.7	-70.4	-71.9	-11.7	-79.8	-69.7	-66.1	-11.2
∅ Node Deg.								
og	2.2	2.1	2.2	2.2	2.2	2.2	2.2	2.2
mg	1.3	1.3	1.2	2.2	1.3	1.4	1.6	2.1
Δ (%)	-39.9	-37.9	-43.5	-1.4	-41.3	-34.9	-25.9	-1.4
In-Circle (%)	12.5	58.8	18.8	46.2	3.8	72.5	6.2	50.0
Out-Diff (%)	60.5	21.2	55.7	8.3	63.4	14.6	67.3	11.4

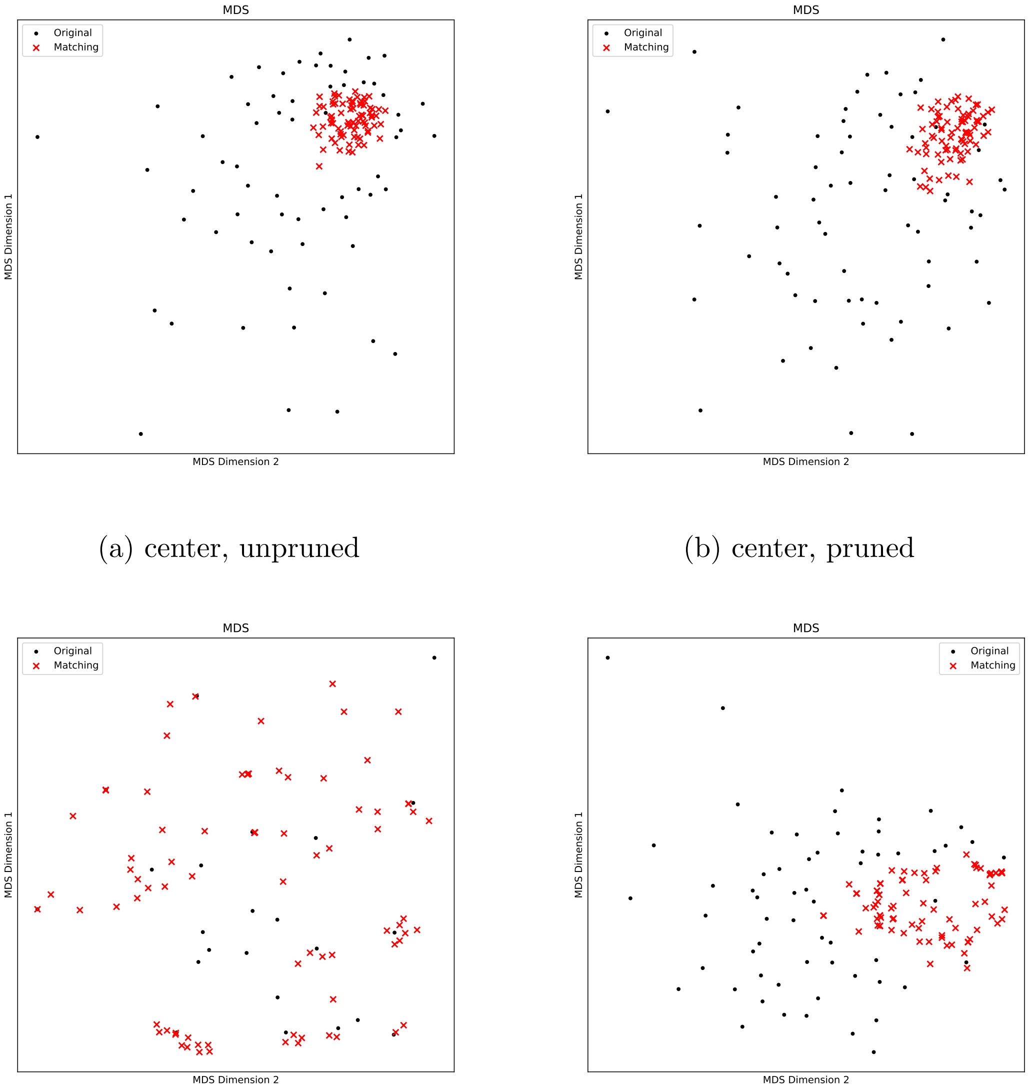
Table 3.4: Comparison of original and matching graph characteristics for different selection and pruning approaches for data set NCI1.

considerable decrease in the average number of nodes (in the range from -50.1% to -64.5%) and edges (in the range from -66.1% to -79.8%) when compared with the original graphs. The node degree is generally lower in the matching-graphs than in the original graphs. Pruning causes the highest reduction in node degree (in the range from -25.9% to -43.5%).

Pruning: The In-Circle percentage of pruned matching-graphs compared with unpruned matching-graphs is consistently lower across both classes and for all types of selections, that is for class 0: 12.5% vs 58.8% and 18.8% vs 46.2% and for class 1: 3.8% vs 72.5% and 6.2% vs 50.0%. Conversely, the Out-Diff percentage of pruned matching-graphs is consistently higher than the Out-Diff percentage of unpruned matching-graphs, that is for class 0: 60.5% vs 21.2% and 55.7% vs 8.3% and for class 1: 63.4% vs 14.6% and 67.3% vs 11.4%. This is a strong indication that pruned matching-graphs tend to be located outside the circle.

Selection: There are not large differences in In-Circle percentage between pruned center selected matching-graphs and pruned spanning selected matching-graphs. Unpruned center selected matching-graphs tend to have a higher In-Circle value than unpruned spanning selected matching-graphs, that is for class 0: 58.8% vs 46.2% and for class 1: 72.5% vs 50.0%. For Out-Diff percentages, the only significant difference is in class 0 between unpruned center selection (21.2%) and unpruned spanning selection (11.4%).

Figure 3.5 illustrates the MDS plots for class 0 for different selection and pruning methods. There are two observations that need to be discussed. First, in the center selection, pruned and unpruned matching-graphs build a cluster. The clusters



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

(d) spanning, pruned

Figure 3.5: MDS plots for class 1 of the dataset NCI1. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

are positioned in upper right corner at the edge and not in the middle of the plot. This indicates a high dissimilarity from the matching-graphs to some of the original graphs at the bottom of the plot. This leads to the interpretation that these subgraphs are not able to represent the substructures of all original graphs equally well. The pruning process does not help in this process.

The second observation is in the spanning selection. Here, we see that the pruned spanning selected matching-graphs are evenly distributed over the plot while the pruned counterparts are clustered. This might be an indication for a preferred substructure within the original graphs that represents a diverse part, which appears to recur frequently.

IMDB

Class	0				1			
Selection	**center**		**span**		**center**		**span**	
Pruning	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**
∅ **Nodes**								
og	17.8	18.4	37.8	39.8	18.2	18.9	34.4	33.8
mg	13.7	14.2	29.8	33.4	13.8	13.3	23.6	26.7
Δ (%)	-23.0	-22.8	-21.0	-16.0	-23.9	-29.8	-31.5	-21.2
∅ **Edges**								
og	62.2	84.7	284.2	270.9	50.9	84.8	332.9	279.4
mg	37.3	58.4	138.1	227.0	25.1	40.9	220.3	226.0
Δ (%)	-40.0	-31.0	-51.4	-16.2	-50.7	-51.8	-33.8	-19.1
∅ **deg**								
og	7.0	9.2	15.0	13.6	5.6	9.0	19.3	16.5
mg	5.4	8.2	9.3	13.6	3.6	6.2	18.7	17.0
Δ (%)	-22.0	-10.8	-38.5	-0.2	-35.3	-31.3	-3.3	2.6
In-Circle (%)	72.5	78.8	13.8	52.5	72.5	36.2	22.5	38.8
Out-Diff (%)	28.3	24.6	27.4	20.5	40.5	18.0	12.1	13.5

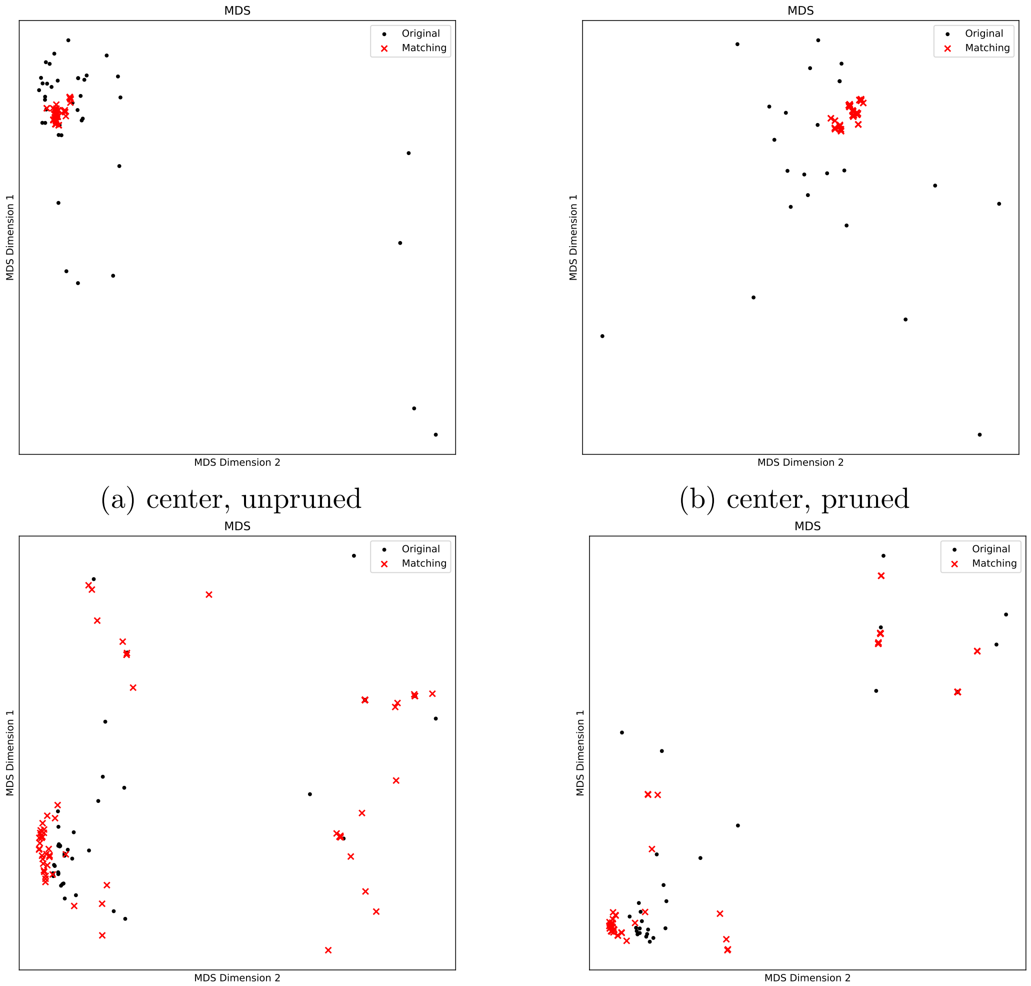
Table 3.5: Comparison of original and matching graph characteristics for different selection and pruning approaches for dataset IMDB.

Table 3.5 presents the characteristics of original and matching graphs for the IMDB dataset.

Average Measures: We notice a consistent reduction in the average number of nodes, edges, and degrees when transitioning from original to matching graphs under all conditions. The largest relative reductions are observed in the average number of edges (in the range from -16.2% to -51.8%). The reduction percentage in edges is always greater than in nodes, indicating that the matching process might be pruning more edges than nodes. The average number of edges and the node degree are high in relation to the other datasets, ranging from 50.9 to 332.9 for edges and 5.6 to 19.3 for node degree.

Pruning: In all conditions, pruned graphs have lower In-Circle percentages than their unpruned counterparts, except for center selection class 1, where the In-Circle percentage for pruned matching-graphs (72.5%) is higher than the percentage for unpruned ones (36.2%). The Out-Diff percentages are generally higher for pruned graphs, except for spanning selection in class 1, where unpruned matching-graphs have a slightly higher Out-Diff percentage (13.5%) than the pruned ones (12.1%).

Selection: Looking at the Out-Diff percentages, center selection produces a larger value than spanning for both pruned and unpruned conditions in both classes with the exception in unpruned matching-graphs in class 1, where center selection (36.2%) is smaller than spanning selection (38.8%). This suggests that pruned



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

(d) spanning, pruned

Figure 3.6: MDS plots for class 1 of the dataset IMDB. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

center selected graphs tend to cluster closer to the center of the circle. The Out-Diff percentage values of center selection are consistently higher than the values for spanning selection for both pruned and unpruned matching-graphs. This suggest that center selected matching-graphs that lie outside of the circle tend to be farther away from the center of the circle than the spanning selected matching-graphs.

Figure 3.6 illustrates the MDS plots for class 1 for different selection and pruning methods. There are two observations that need to be discussed. First, in center selection, the distribution of original and matching-graphs look similar. Pruning does not seem to have an effect on the matching-graphs.

The second observation is in relation to the spanning selection. In this case, the dispersion of matching-graphs is not as uniform as observed in the COX-2 dataset. We see a distinct cluster of original graphs in the left lower corner, indicating a high degree of similarity among them with a few exceptions. Surrounding this cluster of original graphs, a corresponding cluster of matching-graphs has formed. This indicates high similarity between original and matching-graphs.

Both observations could be due to the generally high average node degree of the original graphs in the dataset, ranging from 5.9 to 19.3. This leads to the interpretation that when nodes in a graph are highly interconnected, when pruning or selecting nodes or edges to create matching graphs, the resultant substructures may still retain a high degree of similarity with the original graph due to the sheer number of connections. This could lead to a lack of diversity in the matching-graphs, resulting in more similar graphs and tighter clustering. This pattern could also suggest that graphs with a high degree of nodes are immune to the effects of pruning and the method is not well adjusted to detect the diversified parts of the original graphs.

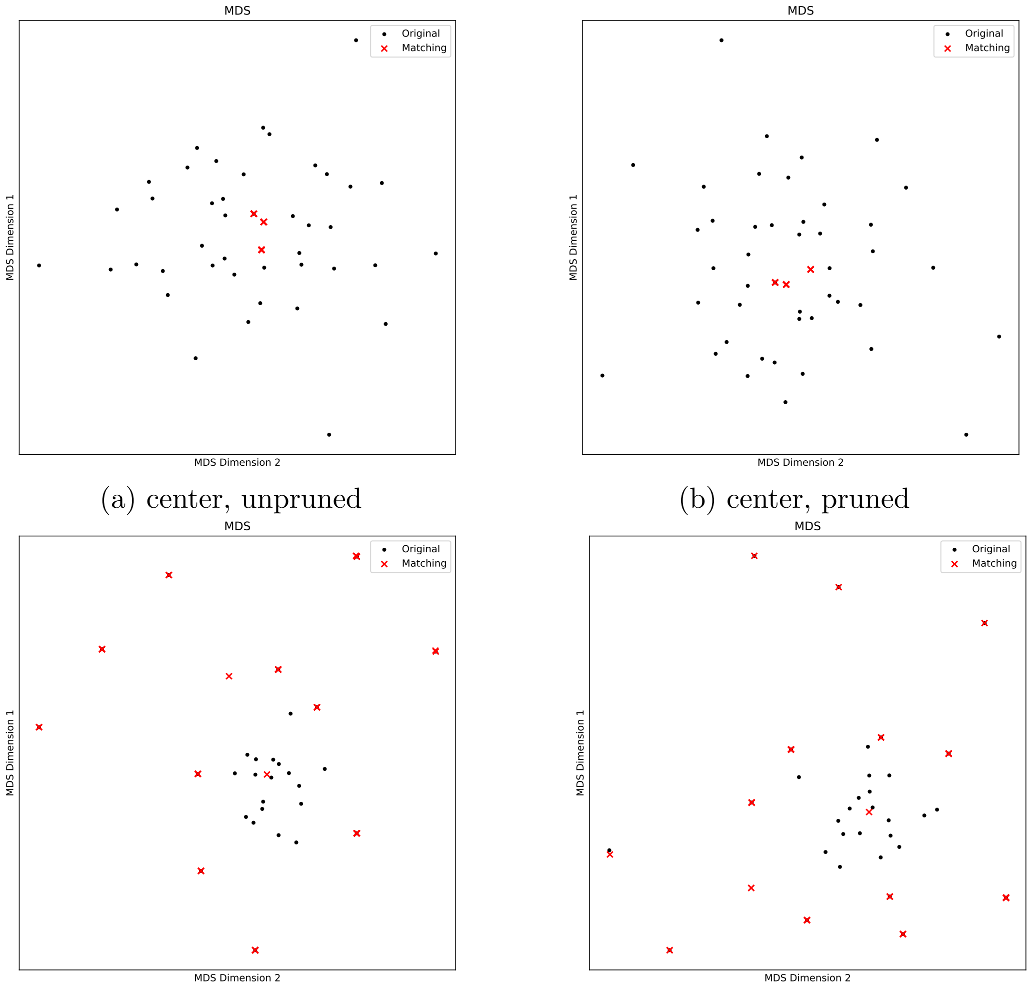
Another reason might be the absence of node labels. This further simplifies the transformation process from one graph to another, as there are fewer distinguishing features to consider. The absence of node labels also reduces the level of diversity across the graph dataset and might limit the variability among matching-graphs as can be seen on the spanning selected plots. Unlabelled graph might therefore also be immune to pruning and selection methods.

LETTER

Letter	**I**				**E**			
Selection	**center**		**span**		**center**		**span**	
Pruning	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**
∅ Nodes								
og	2.4	2.4	2.1	2.2	6.5	6.8	6.7	6.9
mg	2.0	2.0	2.1	2.2	6.0	6.0	5.6	6.6
Δ (%)	-18.0	-16.7	-1.0	-0.5	-8.1	-11.1	-16.5	-4.1
∅ Edges								
og	1.4	1.4	1.1	1.2	6.6	6.4	6.5	6.6
mg	1.0	1.0	1.1	1.2	4.0	4.9	4.4	6.4
Δ (%)	-30.6	-28.6	-2.7	-0.8	-39.8	-23.4	-32.7	-3.0
∅ deg								
og	1.2	1.2	1.0	1.1	2.0	1.9	2.0	1.9
mg	1.0	1.0	1.0	1.1	1.3	1.6	1.6	2.0
Δ (%)	-15.3	-14.5	-1.0	-0.9	-34.5	-14.2	-19.4	1.0
In-Circle (%)	62.5	62.5	67.5	56.2	65.0	62.5	60.0	40.0
Out-Diff (%)	31.3	29.8	0.3	0.0	38.9	41.0	48.7	7.1

Table 3.6: Comparison of original and matching graph characteristics for different selection and pruning approaches for the letters I and E from the dataset LETTER.

Table 3.6 presents the characteristics of original and matching graphs for the LETTER dataset for letter ”I” and ”E”.



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

(d) spanning, pruned

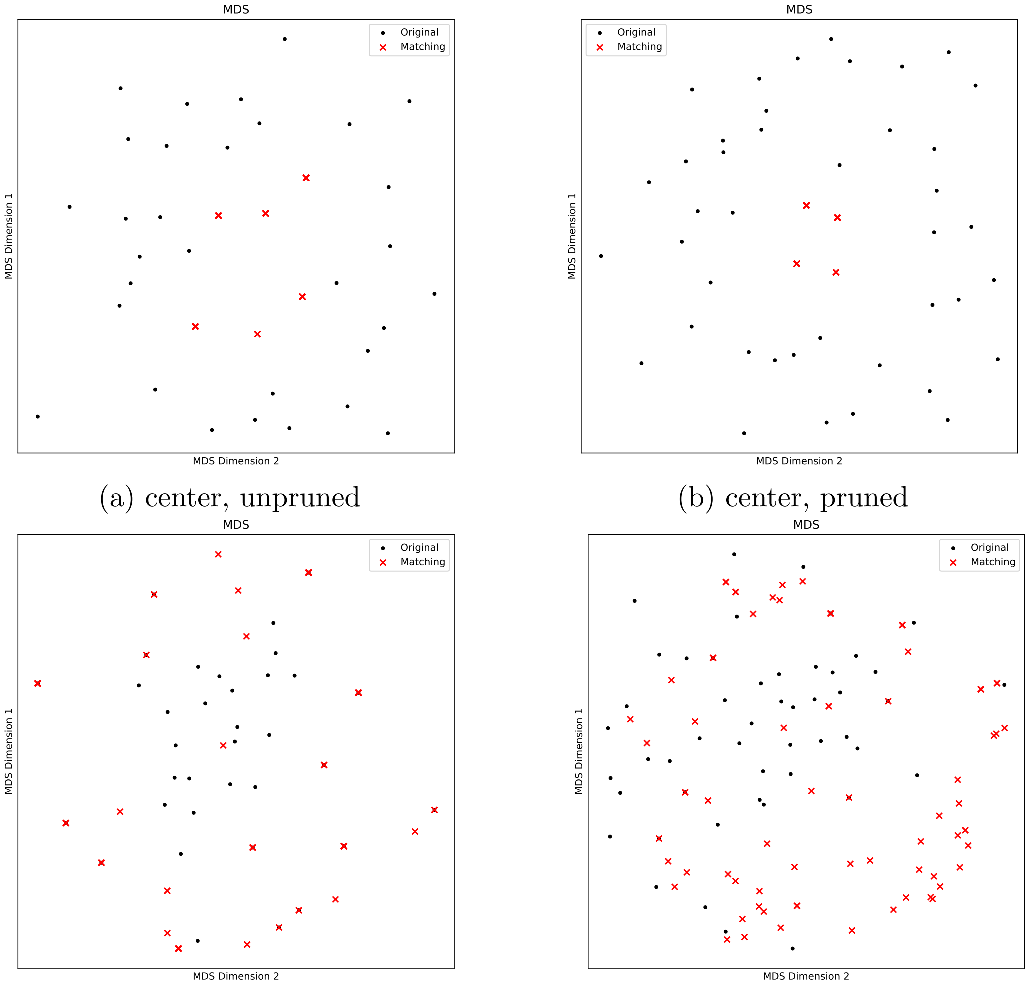
Figure 3.7: MDS plots for letter "I" of the dataset LETTER. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

Average Measures: In the cases of both letters "I" and "E", the average nodes, edges, and degrees of the matching-graphs are consistently lower than the original graphs, indicating that information is lost during the matching process. Pruning tends to intensify this reduction, where node reduction ranges from -1.0% to -18.0% and edge reduction from -2.7 to -39.8. The average number of nodes and edges are low in relation to other datasets, ranging from 2.2 to 6.9 for nodes and 1.1 to 6.6 for edges.

Pruning: For In-Circle, it appears that pruning has little impact in center selection. However, spanning selection tends to lead to a higher In-Circle percentage for pruned matching-graphs. For letter "E", pruned matching-graphs have a significantly higher Out-Diff percentage (48.7%) than the unpruned ones (7.1%).

Selection: The only noticeable pattern seen in the table is that for center selection in letter "I", the matching-graphs have consistently higher Out-Diff percentages than their counterparts in spanning selection, that is 31.3% vs 0.3% and 29.8% vs 0.0%. This is contradictory to our expectation that center selected matching-graphs are less dispersed than spanning selected ones.

Figure 3.7 illustrates the MDS plots for Letter ”I” for different selection and pruning methods. There are three observations that need to be discussed. First, in center selection, the distribution of original and matching-graphs look similar, which indicates that pruning has little effect in creating the core parts of the original graphs. In spanning selection, the distribution of original and matching-graphs looks also similar. Both show original graphs in the middle of the distribution and matching-graphs dispersed. A key observation here is the distinct nature of the distribution on the matching-graphs. This indicates that there are only few possible subgraph configurations that represent the diversified parts of the original graphs.



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

(d) spanning, pruned

Figure 3.8: MDS plots for letter ”E” of the dataset LETTER. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

Figure 3.8 illustrates the MDS plots for Letter "E" for different selection and pruning methods. For center selected matching-graphs, we can see that the pruned ones form a tighter cluster than the unpruned ones. This might indicate that due to the higher average number of nodes and edges in the letter "E" than in "I", the algorithm is able to perform better.

In spanning selection, we see that in the unpruned scenario, the original graphs are more centrally located than in the pruned scenario. This could indicate that the unpruned matching-graphs represent a wider range of diversity than the pruned ones. The pruned matching-graphs seem to be more evenly distributed than the unpruned one. This may indicate that there are more possibilities for subgraphs than in the unpruned scenario due to the nature of the pruning process.

Subway Conclusions

The effectiveness of the Subway method, particularly the pruning and selection techniques, varies significantly depending on the specific dataset. For example, the method excels at simplifying complex molecular graphs such as those observed in the COX-2, PTC(MR) and NCI1 datasets. In contrast, the pruning technique isn't as effective for smaller graphs, such as those in the LETTER dataset, or graphs with a high degree of nodes and absence of node labels, such as those in the IMDB dataset.

The selection methods showed a similar variability in performance. While they worked well on the COX-2, PTC(MR), NCI1 and LETTER datasets, they did not perform as well on the IMDB dataset. The high node-degree and the absence of node labels could be reasons for the bad performance. The combination of both may result in a large set of highly similar matching-graphs, leading to the noticeable clustering observed in the MDS plots.

The absence of node labels also reduces the level of diversity across the graph dataset and might limit the variability among matching-graphs as can be seen on the spanning selected plots. Unlabelled graph might therefore also be immune to pruning and selection methods.

3.3.2 Pepway

In this part of our work, we will examine the matching-graphs produced by the Pepway method, which was described in detail in Section 2.3.2. Unlike Subway, Pepway uses sequences from the edit path to generate unique data structures as matching-graphs, rather than forming subgraphs of the original graphs. The unique advantage of this approach is the generation of novel structures that capture different features of the original graphs. However, due to the similar appearances of the MDS plots and the non-optimization of the Circle-Bound measures for this method, our analysis of Pepway matching-graphs will focus exclusively on the COX-2 dataset. The data in the tables were obtained by considering groups of matching-graphs that were generated using the same probability parameter p, but optimized using different α values. The mean values for various metrics were then calculated from these groups.

COX-2

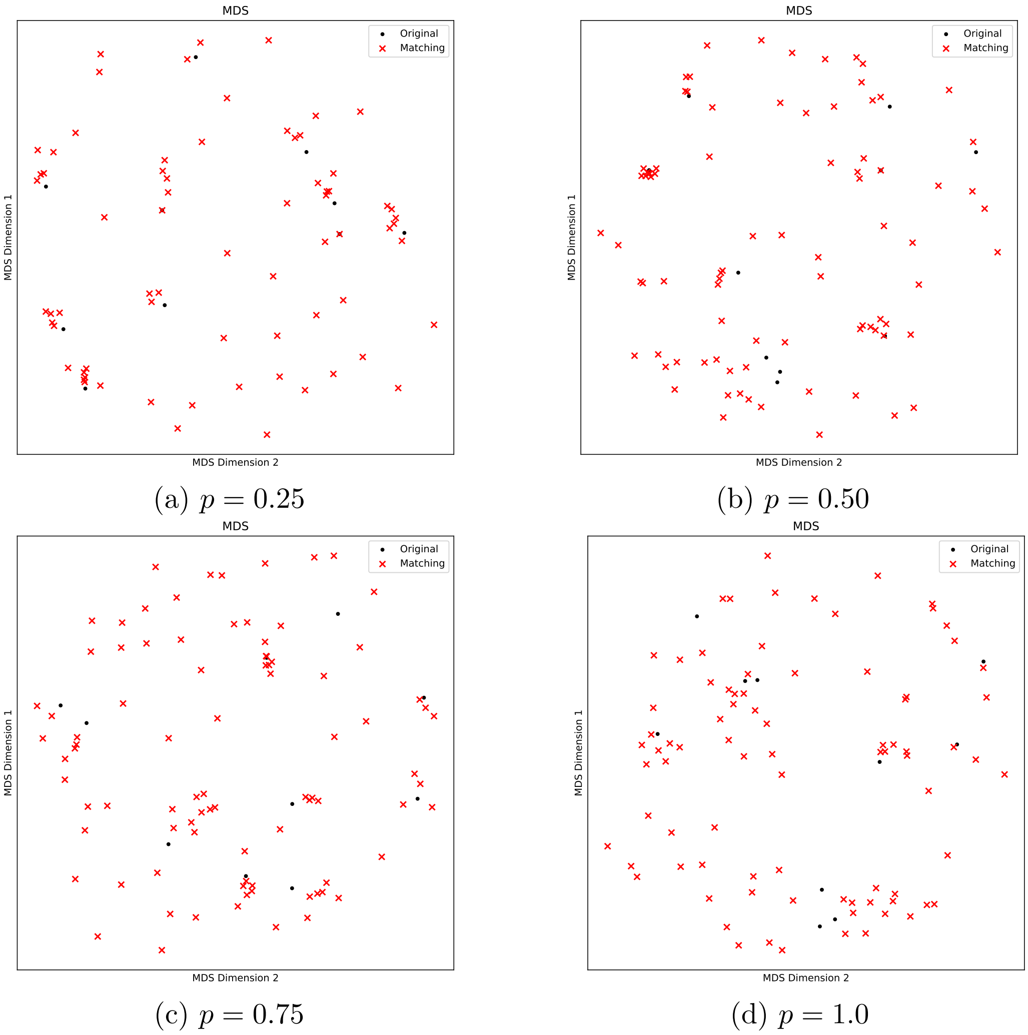
Class	**-1**				**1**			
p	**0.25**	**0.5**	**0.75**	**1.0**	**0.25**	**0.5**	**0.75**	**1.0**
∅ Nodes								
og	41.1±2.5	41.4±1.9	40.8±0.7	40.0±1.2	39.9±1.3	40.4±1.8	40.9±0.8	39.6±1.2
mg	40.5±2.7	40.2±1.7	38.7±0.6	37.5±0.6	39.6±1.2	39.1±1.5	38.5±0.7	37.0±0.8
Δ (%)	-1.4±0.7	-2.9±0.7	-5.2±1.7	-6.1±2.6	-0.7±0.6	-3.0±1.0	-5.9±1.3	-6.7±1.1
∅ Edges								
og	43.4±2.7	43.6±2.0	42.9±0.7	42.1±1.3	41.9±1.3	42.4±1.8	43.0±0.9	41.7±1.2
mg	42.5±2.9	42.1±1.9	40.3±0.6	39.2±0.8	41.4±1.2	41.0±1.6	40.1±0.8	38.5±0.8
Δ (%)	-2.0±0.9	-3.4±0.9	-6.1±1.7	-7.0±2.6	-1.2±0.4	-3.3±1.0	-6.8±1.6	-7.7±1.0
∅ deg								
og	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0
mg	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0	2.1±0.0
Δ (%)	-0.6±0.2	-0.6±0.6	-1.0±0.0	-0.9±0.4	-0.5±0.3	-0.5±0.3	-1.0±0.7	-1.0±0.4
In-Circle (%)	53.8±34.2	64.4±18.3	51.7±12.7	45.8±11.9	60.4±24.3	67.3±25.3	40.5±9.6	45.3±17.9
Out-Diff (%)	27.4±14.0	30.6±7.3	30.2±10.1	34.6±9.7	16.2±8.5	23.4±9.8	34.1±9.1	35.1±1.9

Table 3.7: Comparison of original and matching graph characteristics for the Pepway method for dataset COX-2. Probability is denoted as p.

Table A.7 presents the characteristics of original and matching graphs for the COX-2 dataset created by Pepway.

Average Measures: If p represents the probability of how much of an edit path is applied to the original graph, then as p increases, we should expect to see more modifications in the original graphs. Both the average number of nodes and edges tend to decrease from the original graphs to the matching graphs across both classes. Furthermore, this decrease is accentuated as the probability p increases from 0.25 to 1.0, that is for the change in node size for class -1 read from left to right (from 0.25 to 1.0): -1.4±0.7 vs -2.9±0.7 vs -5.2±1.7 vs -6.1±2.6 (in percentages) and for edges: -2.0±0.9 vs -3.4±0.9 vs -6.1±1.7 vs -7.0±2.6 (in percentages). The same trend is seen in class 1. This pattern indicates that the matching process tends to produce graphs that are generally smaller than the original graphs, and the size reduction becomes more pronounced with increasing p.

Probability: We can observe that the In-Circle percentages do not strictly decrease with an increase in p. For instance, for class -1, In-Circle percentage



(a) $p = 0.25$

(b) $p = 0.50$

(c) $p = 0.75$

(d) $p = 1.0$

Figure 3.9: MDS plots for class -1 of the dataset COX-2 for different probability p. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

initially increases from 53.8% at $p = 0.25$ to 64.4% at $p = 0.5$, then it decreases to 51.7% at $p = 0.75$ and further down to 45.8% at $p = 1.0$. For class 1, the trend is also not linear, with In-Circle percentage initially increasing from 60.4% at $p = 0.25$ to 67.3% at $p = 0.5$ before dropping sharply to 40.5% at $p = 0.75$ and then slightly increasing to 45.3% at $p = 1.0$.

However, the high standard deviations in In-Circle percentage suggest significant variability in these measurements, which makes it more difficult to draw clear conclusions from the data.

It's worth noting that high standard deviations can also make it more difficult to compare means across different p values, as the overlap in the ranges of possible values can be quite large. This should be taken into account when interpreting the results and considering the impact of different p values on the similarity between original and matching graphs.

Regarding the Out-Diff percentages, we can observe a generally increasing trend with increasing p for both classes. This would be expected since applying more of the edit path should result in more differences between the original and matching graphs.

Overall, while the expected trend of increasing dissimilarity with higher p is observed in the Out-Diff percentage, it's not strictly followed in the In-Circle percentage, suggesting that the relationship between p and the similarity between original and matching-graphs could be more complex than initially assumed.

Figure 3.9 illustrates the MDS plots for class -1 of dataset COX-2 for different p values. The plots are arranged according to probability p. There is one consistent pattern across all plots.

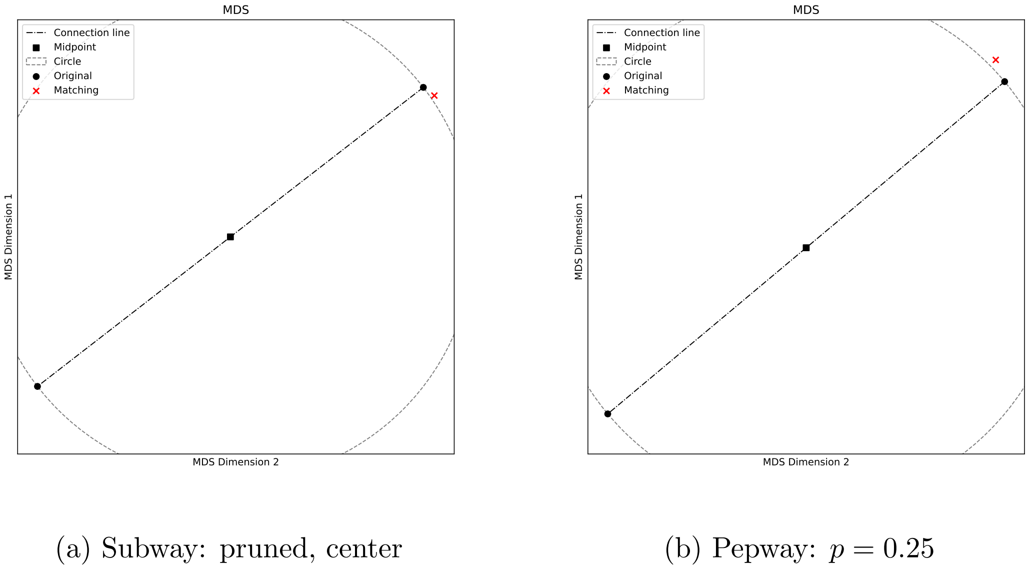
For certain original graphs, a distinct cluster of matching-graphs can be found in their immediate proximity. This clustering might suggest that the matching-graphs have undergone a short partial edit path and therefore are still close to one of the original graphs.

3.3.3 Discussion

The visualizations from the MDS plots seem to highlight the underlying patterns and the structural relationships between the original and the matching-graphs effectively. Therefore, this analysis method proved useful in interpreting the matching-graphs of both Subway and Pepway methods.

The In-Circle measure, as currently defined, may be too restrictive and not fully capture the relationship between original and matching graphs. We observed several instances where a matching-graph is in close proximity to an original graph but falls outside the defined circle, as can be seen in Figure 3.10. These cases are not included in the In-Circle percentage, which can lead to an underestimation of the extent to which matching-graphs resemble their original counterparts.

In the context of the Pepway method, the analytical measures such as In-Circle and Out-Diff appear to be less effective. The matching-graphs created by Pepway are new data structures, not simply subgraphs of the original graphs. This characteristic further complicates the understanding and analysis of these matching-graphs and their relationship to their respective original graphs. Therefore, it may be necessary to develop alternative analytical strategies that are more suitable to fully understand their structure.



(a) Subway: pruned, center

(b) Pepway: $p = 0.25$

Figure 3.10: Two Circle-Bound examples of matching-graphs from class -1 of dataset COX-2. (a) is generated by Subway and is pruned and center selected and (b) is generated by Pepway with $p = 0.25$. The red x marker refers to the matching-graphs, the black dot markers refer to the original graphs.

Chapter 4

Conclusion and Future Work

We used several datasets to evaluate the effectiveness of Subway in graph simplification and one dataset to explore Pepway. We employed MDS to identify patterns of similarity between graphs visually. Our investigation considered two key measures, In-Circle and Out-Diff, to evaluate the efficiency of the methods. In conclusion, our work has shown that the utility of pruning and selection techniques in Subway is context dependent, and is particularly effective in simplifying molecular graphs, as illustrated by the COX-2, PTC(MR) and NCI1 datasets. However, for smaller graphs, such as those in the LETTER dataset, or graphs with a high degree of nodes and absence of node labels, such as those in the IMDB dataset, pruning is less effective.

The selection methods also showed varying efficacy, performing well on COX-2, PTC(MR), NCI1 and LETTER datasets but falling short on the IMDB dataset. We found that the lack of node labels in the IMDB graphs may contribute to the increased similarity between the original and matching graphs. The lack of labels, coupled with the high node degree degree, appears to limit the potential variability introduced by the pruning and selection techniques and results in noticeable clusters.

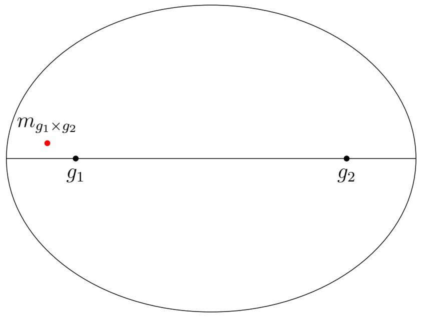


Figure 4.1: An illustration of an ellipse with original graphs g_1 and g_2 as focal points. The matching graph $m_{g_1 \times g_2}$ moves slightly away from g_1, but is still inside the ellipse and considered inside the boundary.

Our research showed that the used analytical measures In-Circle and Out-Diff are less effective for the Pepway method due to the reason that the matching-graphs created by Pepway are new data structure. On the MDS plots however, we saw clusters of matching-graphs near certain original graphs, implying some degree of similarity and pattern preservation.

As a part of future work, we propose exploring other geometric forms, such as ellipses, as potential boundaries to measure structural similarity. In this concept, the original graphs would be positioned at the focal points of the ellipse as shown in Figure 4.1. This modification would allow us to capture the matching-graphs that are close to one original graph but slightly farther from the other.

This approach could provide a more flexible and inclusive boundary, capturing a larger set of matching-graphs that are closely related to their original counterparts but might be missed by the current In-Circle measure.

Appendix A

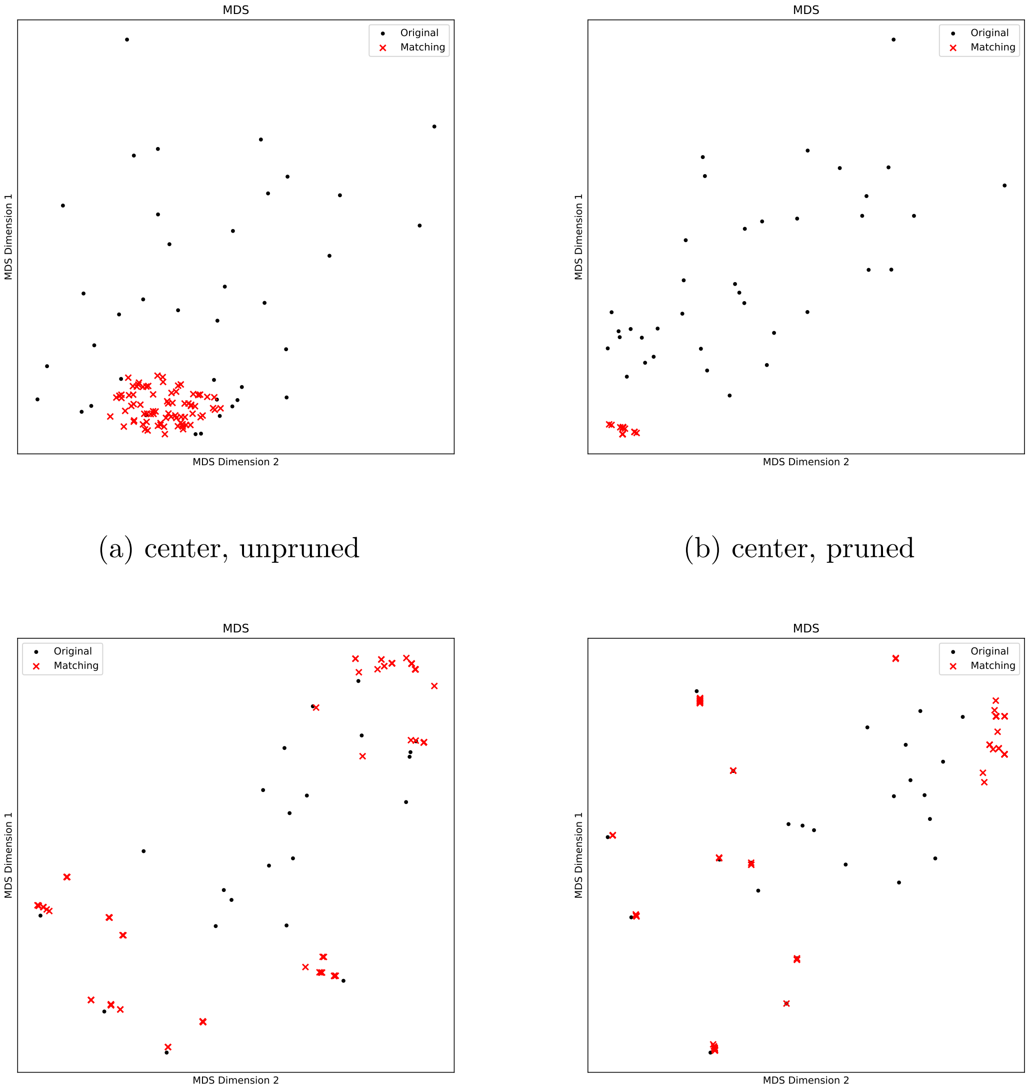
Tables and MDS Plots

A.1 Subway

AIDS

Class	**a**				**i**			
Selection	**center**		**span**		**center**		**span**	
Pruning	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**	**pr**	**unpr**
∅ Nodes								
og	38.5	31.8	54.1	57.6	10.3	9.6	10.4	10.8
mg	5.9	20.2	38.4	43.3	3.0	7.4	8.5	10.2
Δ (%)	-84.6	-36.5	-29.1	-24.8	-70.8	-23.2	-18.5	-6.1
∅ Edges								
og	40.8	33.5	57.2	61.3	10.5	9.3	10.6	11.0
mg	3.0	17.3	39.9	45.6	2.0	6.2	6.2	10.7
Δ (%)	-92.6	-48.4	-30.2	-25.6	-80.9	-33.5	-40.9	-2.6
∅ Node Deg.								
og	2.1	2.1	2.1	2.1	2.0	1.9	2.0	2.0
mg	1.0	1.7	2.1	2.1	1.3	1.7	1.5	2.1
Δ (%)	-52.4	-19.0	-1.9	-1.4	-34.8	-13.5	-27.6	3.4
In-Circle (%)	0.0	43.8	0.0	22.5	8.8	71.2	42.5	48.8
Out-Diff (%)	55.1	22.4	70.2	22.2	58.3	42.2	34.8	10.1

Table A.1: Comparison of original and matching graph characteristics for different selection and pruning approaches for dataset AIDS.



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

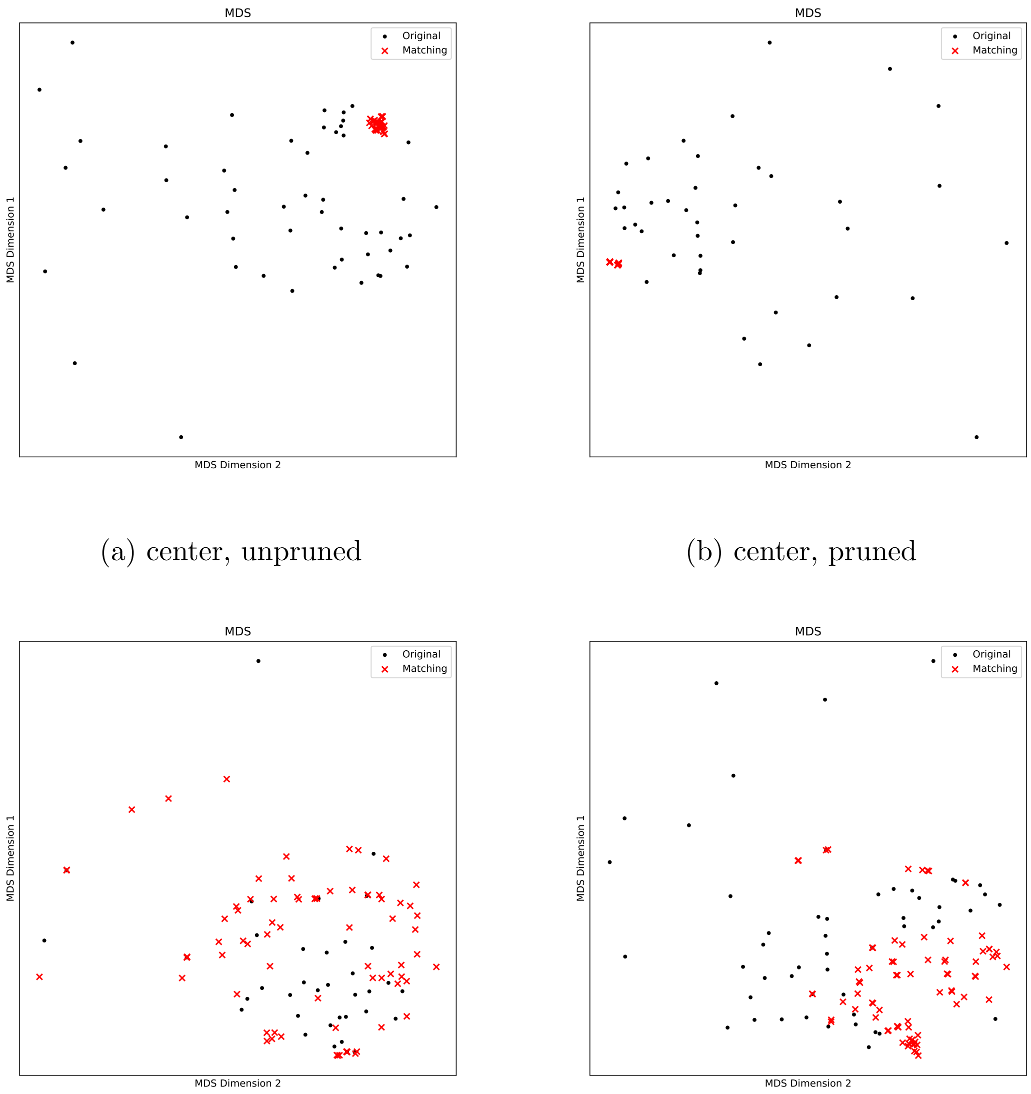
(d) spanning, pruned

Figure A.1: MDS plots for class a of the dataset AIDS. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

MUTA

Class	mutagen				nonmutagen			
Selection	center		span		center		span	
Pruning	pr	unpr	pr	unpr	pr	unpr	pr	unpr
∅ Nodes								
og	31.2	24.8	34.8	49.8	29.2	32.9	35.7	79.9
mg	2.2	3.8	20.1	38.1	2.0	3.9	21.2	64.4
Δ (%)	-92.9	-84.9	-42.4	-23.5	-93.2	-88.2	-40.5	-19.4
∅ Edges								
og	32.3	25.8	37.4	51.3	29.9	33.6	36.5	76.0
mg	1.2	2.4	18.7	39.4	1.0	2.5	18.0	65.7
Δ (%)	-96.3	-90.7	-49.9	-23.3	-96.7	-92.5	-50.8	-13.6
∅ Node Deg.								
og	2.1	2.1	2.1	2.1	2.0	2.0	2.0	1.9
mg	1.1	1.3	1.9	2.1	1.0	1.3	1.7	2.0
Δ (%)	-47.3	-38.3	-12.6	0.5	-51.2	-36.8	-17.2	7.4
In-Circle (%)	2.5	32.5	15.0	37.5	31.2	45.0	15.0	41.2
Out-Diff (%)	40.3	18.2	57.2	26.7	44.7	21.0	57.0	23.7

Table A.2: Comparison of original and matching graph characteristics for different selection and pruning approaches for data set MUTA.



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

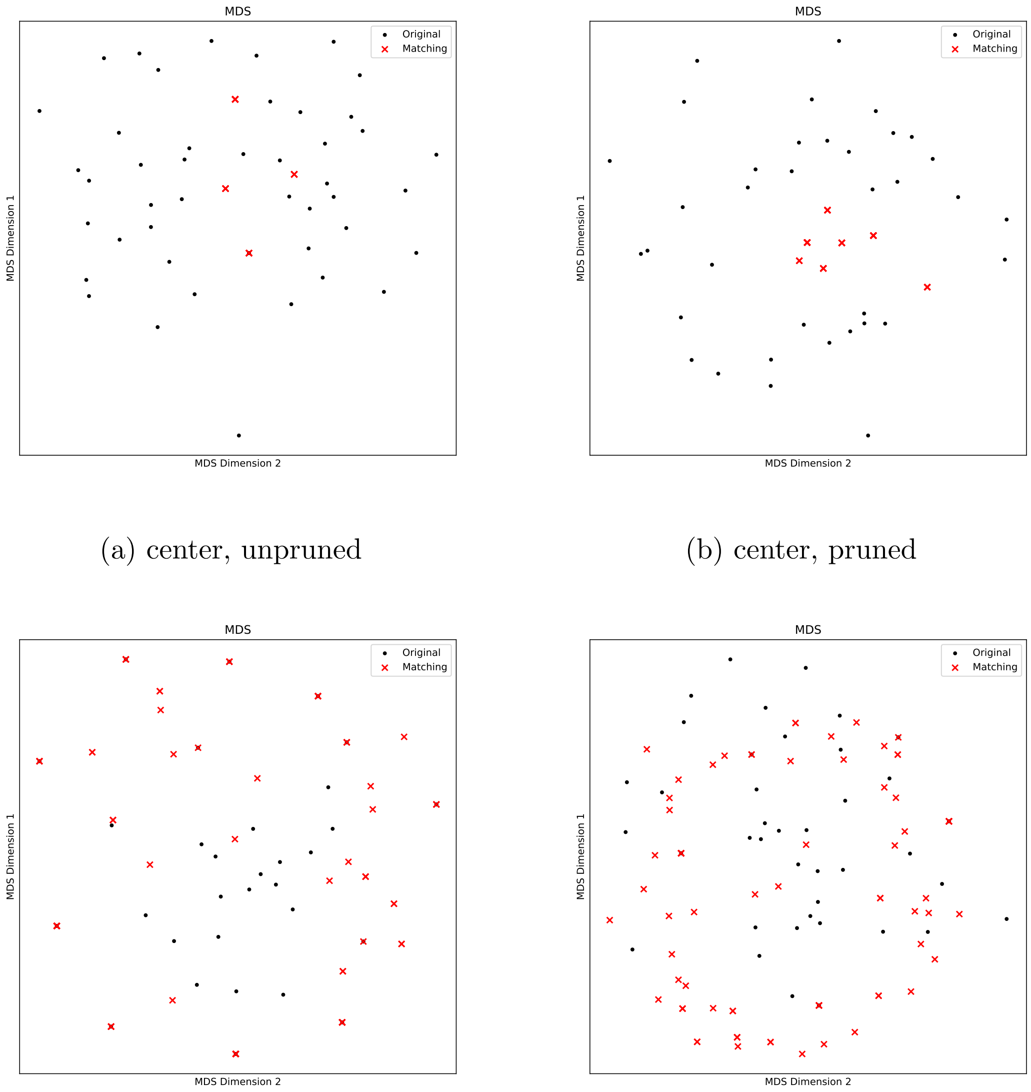
(d) spanning, pruned

Figure A.2: MDS plots for class muta of the dataset MUTA. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

LETTER

Letter	Selection	Pruning	∅ Nodes			∅ Edges			∅ deg			In-Circle (%)	Out-Diff (%)
			og	mg	Δ(%)	og	mg	Δ(%)	og	mg	Δ(%)		
A	center	pr	5.9	4.8	-18	5.4	2.9	-46	1.8	1.2	-35	65	44
		unpr	5.5	5.0	-9	4.8	4.0	-18	1.8	1.6	-10	80	37
	span	pr	5.8	4.9	-16	5.2	3.6	-31	1.8	1.5	-18	46	37
		unpr	6.1	5.8	-6	5.2	5.0	-5	1.7	1.7	0	59	20
E	center	pr	6.5	6.0	-8	6.6	4.0	-40	2.0	1.3	-34	65	39
		unpr	6.8	6.0	-11	6.4	4.9	-23	1.9	1.6	-14	62	41
	span	pr	6.7	5.6	-17	6.5	4.4	-33	2.0	1.6	-19	60	49
		unpr	6.9	6.6	-4	6.6	6.4	-3	1.9	2.0	1	40	7
F	center	pr	5.8	5.0	-13	5.7	3.0	-47	2.0	1.2	-39	69	34
		unpr	5.7	5.0	-12	5.1	3.9	-23	1.8	1.6	-13	72	35
	span	pr	5.8	4.6	-19	5.4	3.5	-35	1.9	1.5	-20	30	42
		unpr	6.3	5.9	-7	5.5	5.0	-9	1.8	1.7	-3	44	19
H	center	pr	6.5	5.9	-9	5.5	3.0	-46	1.7	1.0	-41	59	42
		unpr	6.7	6.0	-10	5.3	3.9	-27	1.6	1.3	-19	75	42
	span	pr	6.7	5.4	-20	5.5	3.5	-37	1.6	1.3	-21	32	43
		unpr	6.9	6.6	-4	5.6	5.4	-4	1.6	1.6	0	61	18
I	center	pr	2.4	2.0	-18	1.4	1.0	-31	1.2	1.0	-15	62	31
		unpr	2.4	2.0	-17	1.4	1.0	-29	1.2	1.0	-15	62	30
	span	pr	2.1	2.1	-1	1.1	1.1	-3	1.0	1.0	-1	68	0
		unpr	2.2	2.2	0	1.2	1.2	-1	1.1	1.1	-1	56	0
K	center	pr	5.9	5.0	-16	5.6	3.0	-46	1.9	1.2	-36	68	42
		unpr	5.9	5.0	-15	5.3	4.1	-22	1.8	1.6	-8	69	36
	span	pr	5.8	4.7	-19	5.1	3.3	-34	1.7	1.4	-18	32	55
		unpr	6.2	5.8	-7	5.3	5.0	-7	1.7	1.7	0	58	8
L	center	pr	3.5	3.0	-15	3.5	2.9	-17	2.0	1.9	-2	54	41
		unpr	3.5	3.0	-14	3.4	3.0	-11	1.9	2.0	4	59	38
	span	pr	3.5	3.2	-11	3.0	2.4	-19	1.7	1.6	-10	42	43
		unpr	3.6	3.6	-2	3.1	3.0	-5	1.7	1.7	-3	44	2
M	center	pr	5.3	5.0	-6	6.4	4.0	-38	2.4	1.6	-34	64	36
		unpr	5.6	5.0	-10	6.3	5.3	-15	2.3	2.1	-6	65	26
	span	pr	5.7	5.0	-11	6.4	4.5	-30	2.3	1.8	-21	49	28
		unpr	5.9	5.6	-5	6.5	6.1	-6	2.2	2.2	-2	52	12

Table A.3: LETTER A-M



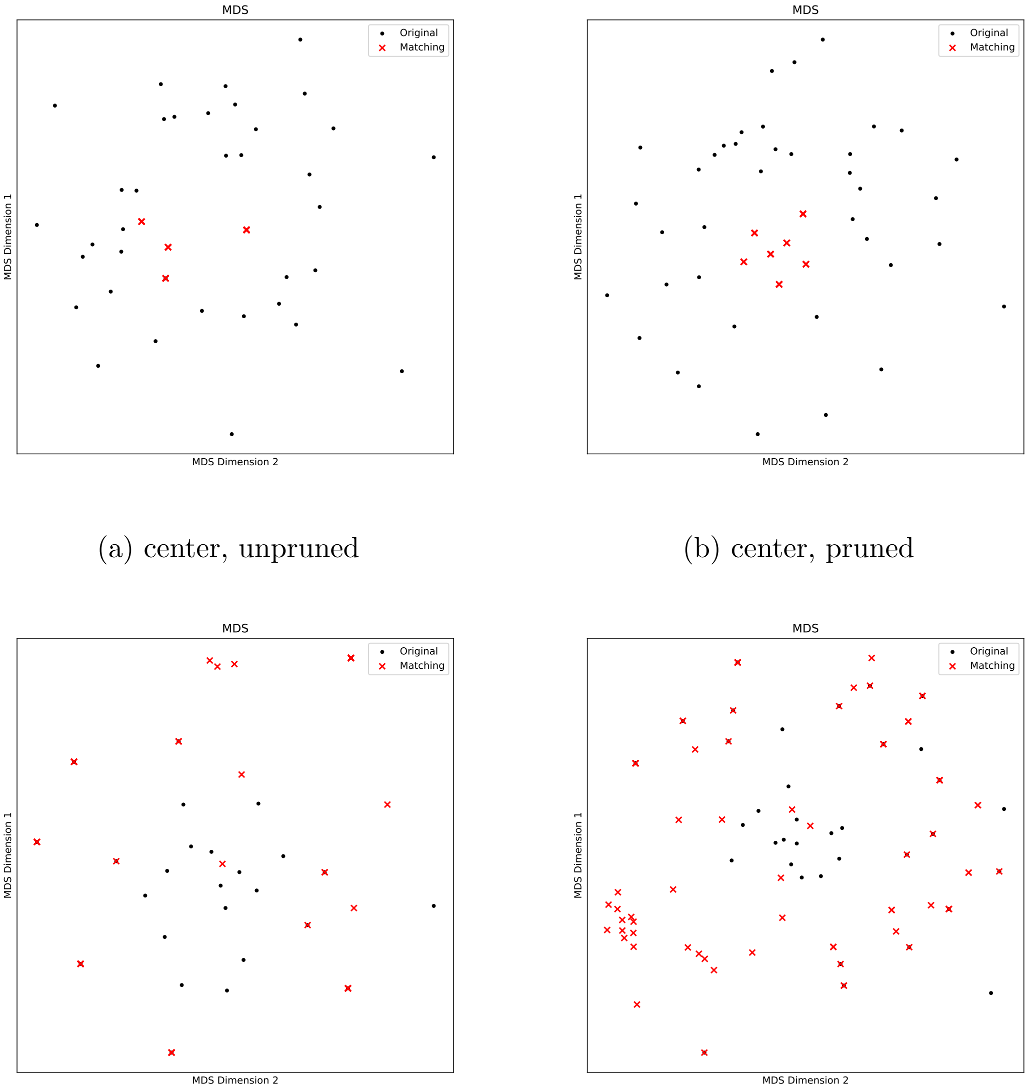
(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

(d) spanning, pruned

Figure A.3: MDS plots for letter "A" of the dataset LETTER. Black dot markers refer to original graphs and red x markers refer to matching-graphs.



(a) center, unpruned

(b) center, pruned

(c) spanning, unpruned

(d) spanning, pruned

Figure A.4: MDS plots for letter "Z" of the dataset LETTER. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

Letter	Selection	Pruning	∅ Nodes			∅ Edges			∅ deg			In-Circle (%)	Out-Diff (%)
			og	mg	Δ (%)	og	mg	Δ (%)	og	mg	Δ (%)		
N	center	pr	4.5	4.0	-11	5.3	4.1	-23	2.4	2.0	-14	69	26
		unpr	4.3	4.0	-7	5.2	5.0	-5	2.4	2.5	3	88	25
	span	pr	4.9	4.5	-8	5.4	4.6	-15	2.2	2.1	-8	68	27
		unpr	5.1	4.9	-3	5.6	5.2	-6	2.2	2.1	-3	55	3
T	center	pr	4.6	4.0	-12	4.0	2.0	-50	1.7	1.0	-43	78	36
		unpr	4.4	4.0	-9	3.9	3.3	-16	1.8	1.6	-7	72	11
	span	pr	4.5	3.9	-13	3.8	2.6	-31	1.7	1.3	-20	35	32
		unpr	4.6	4.3	-5	3.9	3.5	-11	1.7	1.6	-6	44	11
V	center	pr	3.6	3.0	-18	3.3	2.4	-27	1.8	1.6	-12	81	38
		unpr	3.3	3.0	-10	3.2	3.0	-7	1.9	2.0	4	75	24
	span	pr	3.7	3.3	-9	3.2	2.5	-21	1.7	1.5	-12	38	33
		unpr	3.7	3.5	-6	3.1	2.8	-10	1.7	1.6	-5	34	24
W	center	pr	5.8	5.0	-14	6.1	4.0	-35	2.1	1.6	-24	64	47
		unpr	5.7	5.0	-12	6.1	5.0	-18	2.1	2.0	-7	64	41
	span	pr	5.7	4.7	-18	5.7	4.1	-28	2.0	1.8	-12	38	39
		unpr	5.8	5.2	-9	5.8	5.3	-7	2.0	2.0	1	32	11
X	center	pr	5.2	4.0	-23	4.4	2.0	-54	1.7	1.0	-40	61	44
		unpr	5.0	4.0	-20	4.1	2.6	-36	1.7	1.3	-20	59	32
	span	pr	5.1	4.3	-15	4.3	3.0	-30	1.7	1.4	-17	48	35
		unpr	5.4	4.9	-9	4.4	4.0	-10	1.6	1.6	-1	52	16
Y	center	pr	4.7	4.0	-15	4.7	3.0	-36	2.0	1.5	-24	66	44
		unpr	4.5	4.0	-11	4.1	3.3	-19	1.8	1.7	-9	80	31
	span	pr	4.7	4.1	-12	4.4	3.3	-24	1.9	1.6	-14	39	23
		unpr	4.8	4.6	-4	4.5	4.1	-8	1.9	1.8	-4	60	6
Z	center	pr	4.6	4.0	-14	5.0	3.2	-37	2.2	1.6	-26	81	49
		unpr	4.6	4.0	-14	5.2	4.7	-11	2.2	2.3	4	68	36
	span	pr	4.8	4.3	-10	4.8	3.8	-20	2.0	1.8	-12	69	16
		unpr	4.8	4.6	-4	4.6	4.3	-7	1.9	1.8	-4	48	1

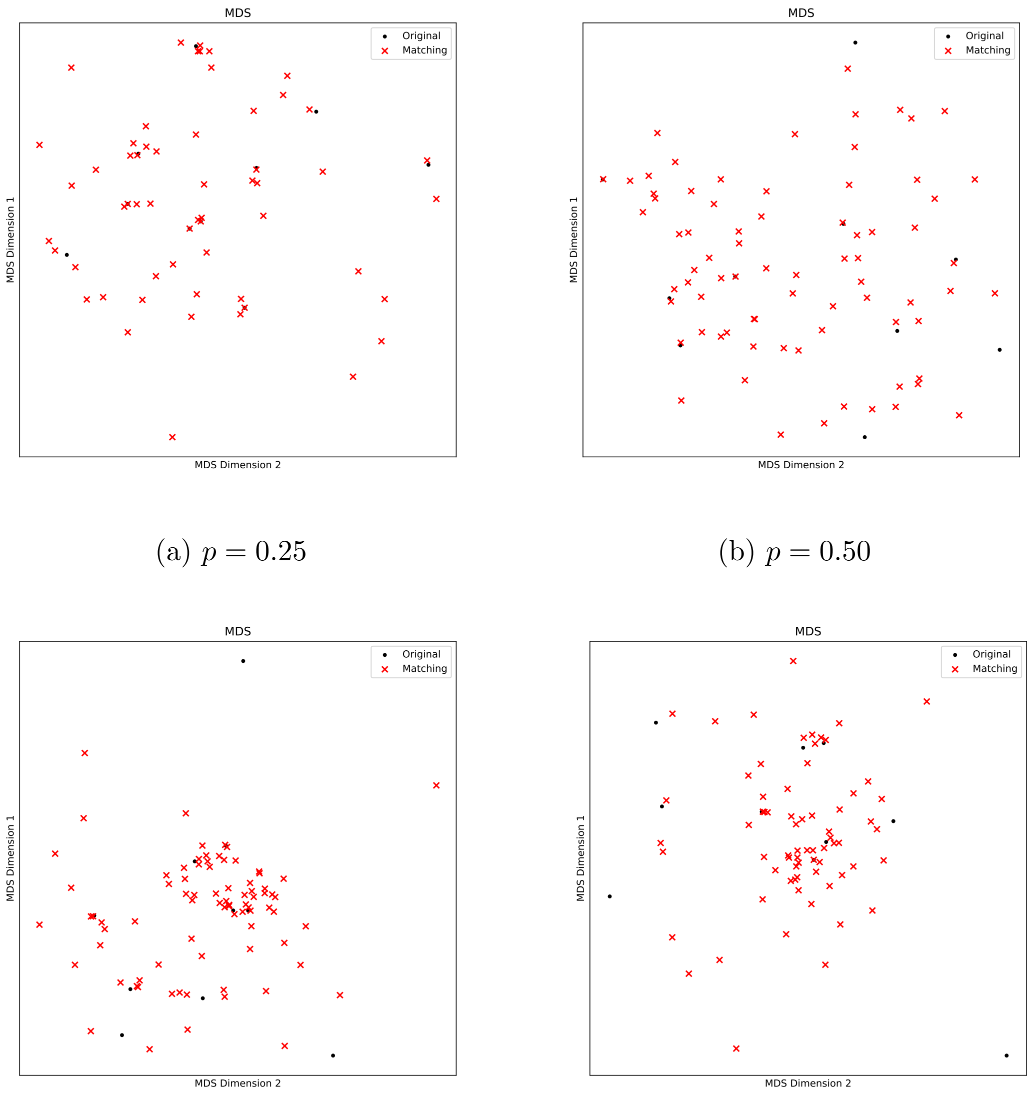
Table A.4: LETTER N-Z

A.2 Pepway

PTC(MR)

Class	**-1**				**1**			
p	**0.25**	**0.5**	**0.75**	**1.0**	**0.25**	**0.5**	**0.75**	**1.0**
∅ Nodes								
og	16.3±2.2	14.4±4.3	15.4±1.8	16.3±2.1	12.3±2.1	13.1±1.7	12.2±3.2	12.1±3.6
mg	16.4±2.4	12.5±3.8	12.0±1.4	11.6±1.1	12.5±2.1	11.4±1.5	9.0±2.3	8.4±2.5
Δ (%)	0±4	-13±1	-22±2	-29±3	1±3	-13±3	-26±4	-31±0
∅ Edges								
og	16.9±2.3	14.9±5.0	15.9±1.9	16.6±2.5	12.7±2.4	13.7±2.0	12.6±3.8	12.6±4.4
mg	16.2±2.4	12.1±4.1	11.3±1.4	11.0±1.1	12.3±2.2	10.9±1.6	8.3±2.4	7.9±2.8
Δ (%)	-4±4	-19±1	-28±2	-34±3	-3±2	-20±5	-34±4	-37±0
∅ deg								
og	2.1	2.1±0.1	2.1±0.1	2.0	2.1±0.1	2.1	2.0±0.1	2.1±0.1
mg	2.0±0.1	1.9±0.1	1.9	1.9	2.0±0.1	1.9	1.8±0.1	1.9±0.1
Δ (%)	-4±1	-7±1	-8±1	-6±1	-5±1	-8±3	-11±1	-9±0
In-Circle (%)	71±18	68±16	66±16	56±4	71±20	69±21	68±12	60±11
Out-Diff (%)	17±7	25±6	24±9	21±5	19±12	29±4	19±7	25±10

Table A.5: Comparison of original and matching graph characteristics for the Pepway method for dataset PTC(MR). Probability is denoted as p.



(a) $p = 0.25$

(b) $p = 0.50$

(c) $p = 0.75$

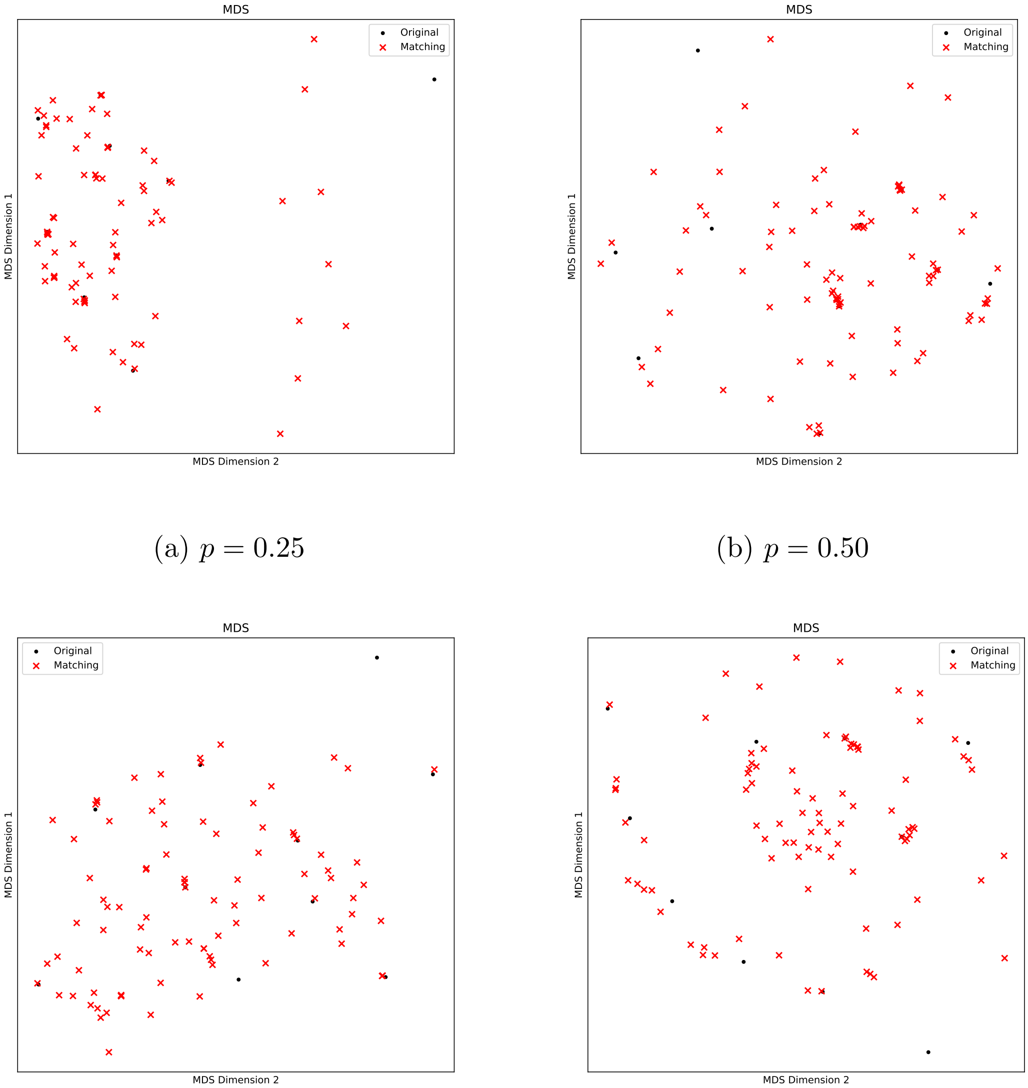
(d) $p = 1.0$

Figure A.5: MDS plots for class 1 of the dataset PTC(MR) for different probability p. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

MUTA

Class	**mutagen**				**nonmutagen**			
p	**0.25**	**0.5**	**0.75**	**1.0**	**0.25**	**0.5**	**0.75**	**1.0**
∅ Nodes								
og	29.5±2.0	32.5±4.7	31.2±8.8	28.2±1.4	32.9±6.0	34.0±4.2	32.3±9.9	29.0±5.6
mg	27.2±2.0	25.2±3.2	23.6±3.4	20.0±2.5	28.9±5.2	27.2±4.7	21.3±4.3	17.4±3.0
Δ (%)	-8±3	-22±5	-22±9	-29±9	-12±3	-20±7	-32±8	-40±1
∅ Edges								
og	30.8±2.5	33.7±4.6	31.4±5.3	30.0±1.5	33.5±6.0	34.8±4.3	32.1±8.3	29.7±5.3
mg	27.8±2.5	24.6±3.1	23.6±3.4	20.1±2.4	28.6±5.1	26.8±5.3	20.4±4.1	16.5±2.8
Δ (%)	-10±3	-27±6	-24±5	-33±9	-14±4	-24±8	-36±6	-44±1
∅ deg								
og	2.1±0.1	2.1	2.1±0.2	2.1	2.0	2.0	2.0±0.1	2.1
mg	2.0±0.1	2.0	2.0	2.0	2.0	2.0±0.1	1.9	1.9
Δ (%)	-2±1	-6±2	-2±8	-6±0	-3±2	-4±2	-5±4	-8±1
In-Circle (%)	64±15	66±21	65±14	55±6	62±16	61±14	56±15	50±8
Out-Diff. (%)	19±14	22±6	26±6	27±4	18±10	22±9	20±6	20±1

Table A.6: Comparison of original and matching graph characteristics for the Pepway method for dataset MUTA. Probability is denoted as p.



(a) $p = 0.25$

(b) $p = 0.50$

(c) $p = 0.75$

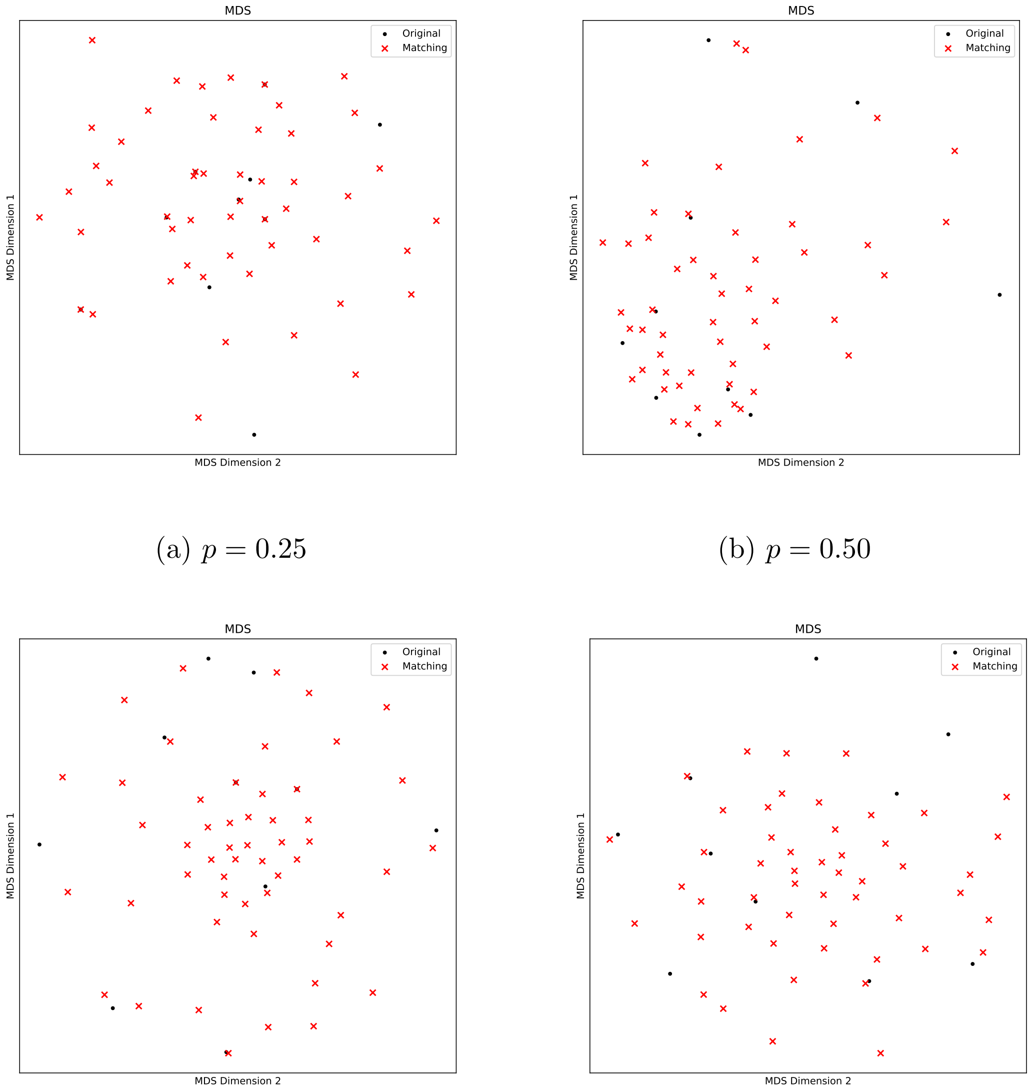
(d) $p = 1.0$

Figure A.6: MDS plots for class mutagen of the dataset MUTA for different probability p. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

NCI1

Class	0				1			
p	**0.25**	**0.5**	**0.75**	**1.0**	**0.25**	**0.5**	**0.75**	**1.0**
∅ Nodes								
og	25.4±5.3	24.9±3.3	26.9±3.7	28.9±2.3	37.8±5.5	36.5±3.8	29.8±3.4	34.6±4.8
mg	26.8±6.2	23.6±2.9	23.0±3.3	22.5±3.4	40.0±5.6	33.5±3.1	24.8±3.2	27.3±3.5
Δ (%)	6±3	-5±2	-15±4	-22±10	6±2	-8±3	-17±4	-21±7
∅ Edges								
og	27.6±5.6	26.8±3.7	29.1±4.1	31.3±3.0	41.2±6.0	39.8±4.8	32.5±3.6	37.3±5.4
mg	27.6±5.5	23.9±2.9	22.9±3.6	21.9±4.5	41.5±5.7	33.7±3.1	24.4±3.6	26.9±4.6
Δ (%)	0±2	-11±3	-21±7	-30±13	1±2	-15±4	-25±4	-28±9
∅ deg								
og	2.2	2.2	2.2	2.2	2.2	2.2±0.1	2.2	2.2
mg	2.1±0.1	2.0	2.0	1.9±0.1	2.1	2.0	2.0	2.0±0.1
Δ (%)	-5±2	-6±2	-8±3	-11±5	-5±1	-7±2	-10±2	-9±4
In-Circle (%)	72±16	79±15	70±27	39±16	60±15	81±20	70±19	50±5
Out-Diff. (%)	27±11	22±14	32±19	41±15	30±16	28±14	26±18	33±14

Table A.7: Comparison of original and matching graph characteristics for the Pepway method for dataset NCI1. Probability is denoted as p.



(a) $p = 0.25$

(b) $p = 0.50$

(c) $p = 0.75$

(d) $p = 1.0$

Figure A.7: MDS plots for class 1 of the dataset NCI1 for different probability p. Black dot markers refer to original graphs and red x markers refer to matching-graphs.

Bibliography

[1] Ertel Wolfgang. *Introduction to artificial intelligence.* Springer, 2017.

[2] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition).* Pearson, 2020.

[3] Elaine Rich and Kevin Knight. *Artificial intelligence (2nd Edition).* McGraw-Hill, 1991.

[4] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[5] Kidiyo Kpalma and Joseph Ronsin. An overview of advances of pattern recognition systems in computer vision. *Vision Systems*, page 26, 2007.

[6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification (2nd Edition).* Wiley, 2001.

[7] Kaspar Riesen. *Structural Pattern Recognition with Graph Edit Distance - Approximation Algorithms and Applications.* Advances in Computer Vision and Pattern Recognition. Springer, 2015.

[8] Jie Liu, Jigui Sun, and Shengsheng Wang. Pattern recognition: An overview. *IJCSNS - International Journal of Computer Science and Network Security*, 6(6):57–61, 2006.

[9] Linda G. Shapiro and Robert M. Haralick. Structural descriptions and inexact matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 3(5):504–519, 1981.

[10] Pierre Mahé, Nobuhisa Ueda, Tatsuya Akutsu, Jean-Luc Perret, and Jean-Philippe Vert. Graph kernels for molecular structure-activity relationship analysis with support vector machines. *J. Chem. Inf. Model.*, 45(4):939–951, 2005.

[11] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *Int. J. Pattern Recognit. Artif.*, 18(3):265–298, 2004.

[12] Wen-Hsiang Tsai and King-Sun Fu. Error-correcting isomorphisms of attributed relational graphs for pattern analysis. *IEEE Trans. Syst. Man Cybern.*, 9(12):757–768, 1979.

[13] Haneen Arafat Abu Alfeilat, Ahmad B. A. Hassanat, Omar Lasassmeh, Ahmad S. Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh Eyal Salman, and V. B. Surya Prasath. Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*, 7(4):221–248, 2019.

[14] Kaspar Riesen and Horst Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image Vis. Comput.*, 27(7):950–959, 2009.

[15] Michael Stauffer, Thomas Tschachtli, Andreas Fischer, and Kaspar Riesen. A survey on applications of bipartite graph edit distance. In *GbRPR*, volume 10310 of *Lecture Notes in Computer Science*, pages 242–252, 2017.

[16] Mathias Fuchs and Kaspar Riesen. A novel way to formalize stable graph cores by using matching-graphs. *Pattern Recognit.*, 131:108846, 2022.

[17] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *CoRR*, abs/2007.08663, 2020.

[18] Tjalling C Koopmans and Martin Beckmann. Assignment problems and the location of economic activities. *Econometrica: journal of the Econometric Society*, pages 53–76, 1957.

[19] Mathias Fuchs and Kaspar Riesen. Matching of matching-graphs - A novel approach for graph classification. In *ICPR*, pages 6570–6576. IEEE, 2020.

[20] Mathias Fuchs and Kaspar Riesen. Augment small training sets using matching-graphs. In *ICPRAI (2)*, volume 13364 of *Lecture Notes in Computer Science*, pages 343–354. Springer, 2022.

[21] Xiaoyi Jiang, Andreas Münger, and Horst Bunke. On median graphs: Properties, algorithms, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(10):1144–1151, 2001.

[22] Ian T. Jolliffe. Principal component analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer, 2011.

[23] Alan J Izenman. *Modern multivariate statistical techniques*, volume 1. Springer, 2008.

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[25] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[26] Kaspar Riesen and Horst Bunke. IAM graph database repository for graph based pattern recognition and machine learning. In *SSPR/SPR*, volume 5342 of *Lecture Notes in Computer Science*, pages 287–297. Springer, 2008.

Erklärung

gemäss Art. 30 RSL Phil.-nat.18

Name/Vorname:	Bardheci Kristjan
Matrikelnummer:	12-722-237
Studiengang:	Informatik

Bachelor ☑ Master □ Dissertation □

Titel der Arbeit:	Visualization of Matching-Graphs An Approach with Multidimensional Scaling
LeiterIn der Arbeit:	PD Dr. Kaspar Riesen

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.
Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.

Stettlen, 2. August 2023

Ort/Datum

Unterschrift