RetGK: Graph Kernels based on Return Probabilities of Random Walks

Zhen Zhang, Mianzhi Wang, Yijian Xiang, Yan Huang, and Arye Nehorai
Department of Electrical and Systems Engineering
Washington University in St. Louis
St. Louis, MO 63130
{zhen.zhang, mianzhi.wang, yijian.xiang, yanhuang640, nehorai}@wustl.edu

Abstract

Graph-structured data arise in wide applications, such as computer vision, bioinformatics, and social networks. Quantifying similarities among graphs is a fundamental problem. In this paper, we develop a framework for computing graph kernels, based on return probabilities of random walks. The advantages of our proposed kernels are that they can effectively exploit various node attributes, while being scalable to large datasets. We conduct extensive graph classification experiments to evaluate our graph kernels. The experimental results show that our graph kernels significantly outperform existing state-of-the-art approaches in both accuracy and computational efficiency.

1 Introduction

Structured data modeled as graphs arise in many application domains, such as computer vision, bioinformatics, and social network mining. One interesting problem for graph-type data is quantifying their similarities based on the connectivity structure and attribute information. Graph kernels, which are positive definite functions on graphs, are powerful similarity measures, in the sense that they make various kernel-based learning algorithms, for example, clustering, classification, and regression, applicable to structured data. For instance, it is possible to classify proteins by predicting whether a given protein is an enzyme or not.

There are several technical challenges in developing effective graph kernels. (i) When designing graph kernels, one might come across the graph isomorphism problem, a well-known NP problem. The kernels should satisfy the isomorphism-invariant property, while being informative on the topological structure difference. (ii) Graphs are usually coupled with multiple types of node attributes, e.g., discrete or continuous attributes. For example, a chemical compound may have both discrete and continuous attributes, which respectively describe the type and position of atoms. A crucial problem is how to integrate the graph structure and node attribute information in graph kernels. (iii) In some applications, e.g., social networks, graphs tend to be very large, with thousands or even millions of nodes, which requires strongly scalable graph kernels.

In this work, we propose novel methods to tackle these challenges. We revisit the concept of random walks, introducing a new node structural role descriptor, the return probability feature (RPF). We rigorously show that the RPF is isomorphism-invariant and encodes very rich connectivity information. Moreover, RPF allows us to consider attributed and nonattributed graphs in a unified framework. With the RPF, we can embed (non-)attributed graphs into a Hilbert space. After that, we naturally obtain our return probability-based graph kernels ("RetGK" for short). Combining with the approximate feature maps technique, we represent each graph with a multi-dimensional tensor and design a family of computationally efficient graphs kernels.

1In the literature, the discrete node attributes are usually called "labels".
Related work. There are various graph kernels, many of which explore the R-convolutional framework \[1\]. The key idea is decomposing a whole graph into small substructures and building graph kernels based on the similarities among these components. Such kernels differ from each other in the way they decompose graphs. For example, graphlet kernels \[24\] are based on small subgraphs up to a fixed size. Weisfeiler-Lehman graph kernels \[23\] are based on subtree patterns. Shortest path kernels \[11\] are derived by comparing the paths between graphs. Still other graph kernels, such as \[28\] and \[9\], are developed by counting the number of common random walks on direct product graphs. Recently, subgraph matching kernels \[17\] and graph invariant kernels \[20\] were proposed for handling continuous attributes. However, all the above R-convolution based graph kernels suffer from a drawback. As pointed out in \[30\], increasing the size of substructures will largely decrease the probability that two graphs contain similar substructures, which usually results in the “diagonal dominance issue” \[13\]. Our return probability based kernels are significantly different from the above ones. We measure the similarity between two graphs by directly comparing their node structural role distributions, avoiding substructures decomposition.

More recently, new methods have been proposed for comparing graphs, which is done by quantifying the dissimilarity between the distributions of pairwise distances between nodes. \[22\] uses the shortest path distance, and \[27\] uses the diffusion distance. However, these methods can be applied only to non-attributed (unlabeled) graphs, which largely limits their applications in the real world.

Organization. In Section 2, we introduce the necessary background, including graph concepts and tensor algebra. In Section 3, we discuss the favorable properties of and computational methods for RPF. In Section 4, we present the Hilbert space embedding of graphs, and develop the corresponding tensor algebra. In Section 3, we discuss the favorable properties of and computational methods for RPF. In Section 5, we present the Hilbert space embedding of graphs, and develop the corresponding tensor algebra. In Section 6, we report the experimental results on 21 benchmark datasets. In the supplementary material, we provide proofs of all mathematical results in the paper.

2 Background

2.1 Graph concepts

An undirected graph \(G\) consists of a set of nodes \(V_G = \{v_1, v_2, ..., v_n\}\) and a set of edges \(E_G \subseteq V_G \times V_G\). Each edge \((v_i, v_j)\) is assigned with a positive value \(w_{ij}\) describing the connection strength between \(v_i\) and \(v_j\). For an unweighted graph, all the edge weights are set to be one, i.e., \(w_{ij} = 1, \forall (v_i, v_j) \in E_G\). Two graphs \(G\) and \(H\) are isomorphic if there exists a permutation map \(\tau : V_G \rightarrow V_H\), such that \(\forall (v_i, v_j) \in E_G\), \((\tau(v_i), \tau(v_j)) \in E_H\), and the corresponding edge weights are preserved.

The adjacent matrix \(A_G\) is an \(n \times n\) symmetric matrix with \(A_G(i, j) = w_{ij}\). The degree matrix \(D_G\) is diagonal matrix whose diagonal terms are \(D_G(i, i) = \sum_{v_j \in V_G} w_{ij}\). The volume of \(G\) is the summation of all node degrees, i.e., \(\text{Vol}_G = \sum_{i=1}^{n} D_G(i, i)\). An \(S\)-step walk starting from node \(v_0\) is a sequence of nodes \(\{v_0, v_1, v_2, ..., v_S\}\), with \((v_s, v_{s+1}) \in E_G\), \(0 \leq s \leq S - 1\). A random walk on \(G\) is a Markov chain \((X_0, X_1, X_2, ..., X_S)\), whose transition probabilities are

\[
\text{Pr}(X_{i+1} = v_{i+1} | X_i = v_i, ..., X_0 = v_0) = \text{Pr}(X_{i+1} = v_{i+1} | X_i = v_i) = \frac{w_{ij}}{D_G(i, i)},
\]

which induces the transition probability matrix \(P_G = D_G^{-1} A_G\). More generally, \(P_G^s\) is the \(s\)-step transition matrix, where \(P_G^s(i, j)\) is the transition probability in \(s\) steps from node \(v_i\) to \(v_j\).

In our paper, we also consider the case that nodes are associated with multiple attributes. Let \(A\) denote a attribute domain. Typically, \(A\) can be a alphabet set or a subset of a Euclidean space, which corresponds to discrete attributes and continuous attributes, respectively.

2.2 Tensor algebra

A tensor \[16\] is a multidimensional array, which has multiple indices\[1\]. We use \(\mathbb{R}^{I_1 \times I_2 \times ... \times I_N}\) to denote the set of tensors of order \(N\) with dimension \((I_1, I_2, ..., I_N)\). If \(U \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_N}\), then \(U_{i_1i_2...i_N} \in \mathbb{R}\), where \(1 \leq i_1 \leq I_1, ..., 1 \leq i_N \leq I_N\).

\[1\] A vector \(\mathbf{u} \in \mathbb{R}^{D_1}\) is a first-order tensor, and a matrix \(A \in \mathbb{R}^{D_1 \times D_2}\) is a second-order tensor.
where

\[
\{v_i\} \in V_G \text{ an } S\text{-dimensional feature called "return probability feature" ("RPF" for short), which describes the "structural role" of } v_i, \text{ i.e.,}
\]

\[
\bar{p}_i = [P^G_1(i, i), P^G_2(i, i), ..., P^G_S(i, i)]^T,
\]

where \(P^G_s(i, i), s = 1, 2, ..., S\), is the return probability of a \(s\)-step random walk starting from \(v_i\). Now each graph is represented by a set of feature vectors in \(\mathbb{R}^S\): \(RPF^S_G = \{\bar{p}_1, \bar{p}_2, ..., \bar{p}_n\}\). The RPF has three nice properties: isomorphism-invariance, multi-resolution, and informativeness.

### 3.1 The properties of RPF

**Isomorphism-invariance.** The isomorphism-invariance property of return probability features is summarized in the following proposition.

**Proposition 1.** Let \(G\) and \(H\) be two isomorphic graphs of \(n\) nodes, and let \(\tau : \{1, 2, ..., n\} \rightarrow \{1, 2, ..., n\}\) be the corresponding isomorphism. Then,

\[
\forall v_i \in V_G, s = 1, 2, ..., \infty, \quad P^G_s(i, i) = P^H_s(\tau(i), \tau(i)).
\]

Clearly, isomorphic graphs have the same set of RPF, i.e., \(RPF^S_G = RPF^S_H, \forall S = 1, 2, ..., \infty\). Such a property can be used to check graph isomorphism, i.e., if \(\exists S, \text{ s.t. } RPF^S_G \neq RPF^S_H\), then \(G\) and \(H\) are not isomorphic. Moreover, Proposition 1 allows us to directly compare the structural role of any two nodes in different graphs, without considering the matching problems.

**Multi-resolution.** RPF characterizes the "structural role" of nodes with multi-resolutions. Roughly speaking, \(P^G_s(i, i)\) reflects the interaction between node \(v_i\) and the subgraph involving \(v_i\). With an increase in \(s\), the subgraph becomes larger. We use a toy example to illustrate our idea. Fig. 1(a) presents an unweighted graph \(G\), and \(C_1, C_2, \text{ and } C_3\) are three center nodes in \(G\), which play different structural roles. In Fig. 1(b), we plot their \(s\)-step return probabilities, \(s = 1, 2, ..., 200\). \(C_1, C_2, \text{ and } C_3\) have the same degree, as do their neighbors. Thus their first two return probabilities are the same. Since \(C_1\) and \(C_2\) share the similar neighbourhoods at larger scales, their return probability values are close until the eighth step. Because \(C_3\) plays a very different structural role from \(C_1\) and \(C_2\), its return probabilities values deviate from those of \(C_1\) and \(C_2\) in early steps.

In addition, as shown in Fig. 1(b), when the random walk step \(s\) approaches infinity, the return probability \(P^G_s(i, i)\) will not change much and will converge to a certain value, which is known as the stationary probability in Markov chain theory [5]. Therefore, if \(s\) is already sufficiently large, we gain very little new information from the RPF by increasing \(s\).

**Informativeness.** The RPF provides very rich information on the graph structure, in the sense that if two graphs has the same RPF sets, they share very similar spectral properties.

**Theorem 1.** Let \(G\) and \(H\) be two connected graphs of the same size \(n\) and volume \(Vol\), and let \(P^G\) and \(P^H\) be the corresponding transition probability matrices. Let \(\{\lambda_k, \vec{w}_k\}_{k=1}^n\) and \(\{\mu_k, \vec{φ}_k\}_{k=1}^n\) be eigenpairs of \(P^G\) and \(P^H\), respectively. Let \(\tau : \{1, 2, ..., n\} \rightarrow \{1, 2, ..., n\}\) be a permutation map. If \(P^G_s(i, i) = P^H_s(\tau(i), \tau(i))\), \(\forall v_i \in V_G, \forall s = 1, 2, ..., \infty, \text{ i.e., } RPF^G_s = RPF^H_s\), then,
The first conclusion states that the graph structure information contained in both contribute to the scalability of RPF.

Given a graph \( G \), the brute-force computation of \( \text{RPF}_G^S \) requires \((S-1)\) times \( n \times n \) matrix multiplication of \( P_G \). Therefore, the time complexity is \((S-1)n^3\), which is quite high when \( S \) is large.

Since only the diagonal terms of transition matrices are needed, we have efficient techniques. Write

\[
P_G = D_G^{-\frac{1}{2}} A_G D_G^{-\frac{1}{2}} = D_G^{-\frac{1}{2}} (D_G^{-\frac{1}{2}} A_G D_G^{-\frac{1}{2}}) D_G^{-\frac{1}{2}} = D_G^{-\frac{1}{2}} B_G D_G^{-\frac{1}{2}},
\]

(5)

where \( B_G = D_G^{-\frac{1}{2}} A_G D_G^{-\frac{1}{2}} \) is a symmetric matrix. Then \( P_G^s = D_G^{-\frac{1}{2}} B_G^s D_G^{-\frac{1}{2}} \). Let \( \{ (\lambda_k, \bar{u}_k) \}_{k=1}^n \) be the eigenpairs of \( B_G \), i.e., \( B_G = \sum_{k=1}^n \lambda_k \bar{u}_k \bar{u}_k^T \). Then the return probabilities are

\[
P_G^s(i, i) = B_G^s(i, i) = \sum_{k=1}^n \lambda_k^s [\bar{u}_k(i)]^2, \forall v_i \in V_G, \forall s = 1, 2, ..., S.
\]

(6)

Let \( U = [\bar{u}_1, \bar{u}_2, ..., \bar{u}_n] \), let \( V = U \odot U \), where \( \odot \) denotes Hadamard product, and let \( \vec{\Lambda}_s = [\lambda_1^s, \lambda_2^s, ..., \lambda_n^s]^T \). Then we can obtain all nodes' \( s \)-step return probabilities in the vector \( V \vec{\Lambda}_s \). The eigen-decomposition of \( B_G \) requires time \( O(n^3) \). Computing \( V \) or \( V \vec{\Lambda}_s, \forall s = 1, 2, ..., S \), takes time \( O(n^2) \). So the total time complexity of the above computational method is \( O(n^3 + (S + 1)n^2) \).

### 3.2.1 Monte Carlo simulation method

If the graph node number, \( n \), is large, i.e., \( n > 10^5 \), the eigendecomposition of an \( n \times n \) matrix is relatively time-consuming. To make RPF scalable to large graphs, we use the Monte Carlo method to simulate random walks. Given a graph \( G \), for each node \( v_i \in V_G \), we can simulate a random walk of length \( S \) based on the transition probability matrix \( P_G \). We repeat the above procedure \( M \) times, obtaining \( M \) sequences of random walks. For each step \( s = 1, 2, ..., S \), we use the relative frequency of returning to the starting point as the estimation of the corresponding \( s \)-step return probability. The random walk simulation is parallelizable and can be implemented efficiently, characteristics of which both contribute to the scalability of RPF.
4 Hilbert space embeddings of graphs

In this section, we introduce the Hilbert space embeddings of graphs, based on the RPF. With such Hilbert space embeddings, we can naturally obtain the corresponding graph kernels.

As discussed in Section 3, the structural role of each node $v_i$ can be characterized by an $S$-dimensional return probability vector $\bar{p}_i$ (see [3]), and thus a nonattributed graph can be represented by the set $\text{RPF}_G = \{\bar{p}_i\}_{i=1}^n$. Since the isomorphism-invariance property allows direct comparison of nodes' structural roles across different graphs, we can view the RPF as a special type of attribute, namely, "the structural role attribute" (whose domain is denoted as $A_0$), associated with nodes. Clearly, $A_0 = \mathbb{R}^S$.

The nodes of attributed graphs usually have other types of attributes, which are obtained by physical measurements. Let $A_1, A_2, ..., A_L$ be their attribute domains. When combined with RPF, an attributed graph can be represented by the set $\{(\bar{p}_i, a_i^1, a_i^2, ..., a_i^L)\}_{i=1}^n \subseteq A_0 \times A_1 \times ... \times A_L$ (denoted as $\times_{i=0}^L A_i$). Such a representation allows us to consider both attributed and nonattributed graphs in a unified framework, since if $L = 0$, the above set just degenerates to the nonattributed case. The set representation forms an empirical distribution $\mu \in \mathcal{P}(\mathbb{R}^S \times A)$ on $\mathcal{G}$, which can be embedded into a reproducing kernel Hilbert space (RKHS) by kernel mean embedding [10].

Let $k_l$, $l = 0, 1, ..., L$ be a kernel on $A_l$. Let $H_l$ and $\phi_l$ be the corresponding RKHS and implicit feature map, respectively. Then we can define a kernel on $\mathcal{A}$ through the tensor product of kernels [26], i.e., $k = \otimes_{l=0}^L k_l, k(\bar{p}, \bar{q}, a^1, b^1, a^2, b^2, ..., a^L, b^L) = k_0(\bar{p}, \bar{q}) \prod_{l=1}^L k_l(a^l, b^l)$. Its associated RKHS, $\mathcal{H}$, is the tensor product space generated by $\mathcal{H}_l$, i.e., $\mathcal{H} = \otimes_{l=0}^L \mathcal{H}_l$. Let $\phi : \mathcal{A} \rightarrow \mathcal{H}$ be the implicit feature map. Then given a graph $G$, we can embed it into $\mathcal{H}$ in the following procedure,

$$G \rightarrow \mu_G \rightarrow m_G, \text{ and } m_G = \int \phi \text{d} \mu_G = \frac{1}{n} \sum_{i=1}^n \phi((\bar{p}_i, a_i^1, ..., a_i^L)).$$

(7)

4.1 Graph kernels (I)

An important benefit of Hilbert space embedding of graphs is that it is straightforward to generalize the positive definite kernels defined on Euclidean spaces to the set of graphs.

Given two graphs $G$ and $H$, let $\{\Delta_i^G\}_{i=1}^n$ and $\{\Delta_j^H\}_{j=1}^m$ be the respective set representations ($\Delta_i^G = (\bar{p}_i, a_i^1, a_i^2, ..., a_i^L)$ and likewise $\Delta_j^H$). Let $K_{GG}$, $K_{HH}$, and $K_{GH}$ be the kernel matrices, induced by the embedding kernel $k$. That is, they are defined such that $(K_{GG})_{ij} = k(\Delta_i^G, \Delta_j^G)$, $(K_{HH})_{ij} = k(\Delta_i^H, \Delta_j^H)$, and $(K_{GH})_{ij} = k(\Delta_i^G, \Delta_j^H)$.

**Proposition 2.** Let $\mathcal{G}$ be the set of graphs with attribute domains $A_1, A_2, ..., A_L$. Let $G$ and $H$ be two graphs in $\mathcal{G}$. Let $m_G$ and $m_H$ be the corresponding graph embeddings. Then the following kernels are positive definite graph kernels defined on $\mathcal{G} \times \mathcal{G}$.

$$K_1(G, H) = (c + (m_G, m_H)_{\mathcal{H}})^d = (c + \frac{1}{n_{GH}^d} \bar{T}_{m_G}^T K_{GH} \bar{T}_{m_H})^d, c \geq 0, d \in \mathbb{N},$$

(8a)

$$K_2(G, H) = \exp(-\gamma \|m_G - m_H\|^p_{\mathcal{H}}) = \exp(-\gamma \text{MMD}^p(\mu_G, \mu_H)), \gamma > 0, 0 < p \leq 2,$$

(8b)

where $\text{MMD}(\mu_G, \mu_H) = (\frac{1}{n_{GH}^T} \bar{T}_{m_G} K_{GH} \bar{T}_{m_H} + \frac{1}{n_{GH}^T} \bar{T}_{m_H} K_{HH} \bar{T}_{m_H} - \frac{2}{n_{GH}^T} \bar{T}_{m_G} K_{GH} \bar{T}_{m_H})^{1/2}$ is the maximum mean discrepancy (MMD) [10].

**Kernel selection.** In real applications, such as bioinformatics, graphs may have discrete labels and (multi-dimensional) real-valued attributes. Hence, three attributes domains are involved in the computation of our graph kernels: the structural role attribute domain $A_0$, the discrete attribute domain $A_l$, and the continuous attribute domain $A_c$. For $A_l$, we can use the Delta kernel $k_l(a, b) = I_{a=b}$. For $A_0$, $A_c$, which are just the Euclidean spaces, we can use the Gaussian RBF kernel, the Laplacian RBF kernel, or the polynomial kernel.

5 Approximated Hilbert space embedding of graphs

Based on the above discussions, we see that obtaining a graph kernel value between each pair of graphs requires calculating the inner product or the $L_2$ distance between two Hilbert embeddings
(see (8a) and (8b)), both of which scale quadratically to the node numbers. Such time complexity precludes application to large graph datasets. To tackle the above issues, we employ the recently approximated explicit feature maps [21].

For a kernel \( k_l \) on the attribute domain \( A_l, l = 0, 1, \ldots, L \), we find an explicit map \( \hat{\phi} : A_l \rightarrow \mathbb{R}^{D_l} \), so that

\[
\forall a, b \in A_l, \langle \hat{\phi}(a), \hat{\phi}(b) \rangle = \hat{k}_l(a, b), \text{ and } \hat{k}_l(a, b) \rightarrow k_l(a, b) \text{ as } D_l \rightarrow \infty. \tag{9}
\]

The explicit feature maps will be directly used to compute the approximate graph embeddings, by virtue of tensor algebra (see Section 2.2). The following theorem says that the approximate explicit graph embeddings can be written as the linear combination of rank-one tensors.

**Theorem 2.** Let \( G \) and \( H \) be any two graphs in \( \mathcal{G} \). Let \( \{(\hat{p}_l, a_1^l, a_2^l, \ldots, a_{L}^l)\}_{l=1}^{n_G} \) and \( \{(\hat{q}_j, b_1^j, b_2^j, \ldots, b_{L}^j)\}_{j=1}^{n_H} \) be the respective set representations of \( G \) and \( H \). Then their approximate explicit graph embeddings, \( \hat{m}_G \) and \( \hat{m}_H \), are tensors in \( \mathbb{R}^{D_0 \times D_1 \times \ldots \times D_L} \), and can be written as

\[
\hat{m}_G = \frac{1}{n_G} \sum_{i=1}^{n_G} \hat{\phi}_0(\hat{p}_i) \circ \hat{\phi}_1(a_1^l) \circ \ldots \circ \hat{\phi}_L(a_L^l), \quad \hat{m}_H = \frac{1}{n_H} \sum_{j=1}^{n_H} \hat{\phi}_0(\hat{q}_j) \circ \hat{\phi}_1(b_1^j) \circ \ldots \circ \hat{\phi}_L(b_L^j). \tag{10}
\]

That is, as \( D_0, D_1, \ldots, D_L \rightarrow \infty \), we have \( \langle \hat{m}_G, \hat{m}_H \rangle_T \rightarrow \langle m_G, m_H \rangle_\mathcal{H} \).

### 5.1 Graph Kernels (II)

With approximate tensor embeddings (10), we obtain new graph kernels.

**Proposition 3.** The following functions are positive definite graph kernels defined on \( \mathcal{G} \times \mathcal{G} \).

\[
\hat{K}_1(G, H) = \langle c + \text{vec}(\hat{m}_G), \text{vec}(\hat{m}_H) \rangle_T, c \geq 0, d \in \mathbb{N}, \tag{11a}
\]

\[
\hat{K}_2(G, H) = \exp(-\gamma \|\hat{m}_G - \hat{m}_H\|_T^p) = \exp(-\gamma \|\text{vec}(\hat{m}_G) - \text{vec}(\hat{m}_H)\|_F^p), \gamma > 0, 0 < p \leq 2.. \tag{11b}
\]

Moreover, as \( D_0, D_1, \ldots, D_L \rightarrow \infty \), we have \( \hat{K}_1(G, H) \rightarrow K_1(G, H) \) and \( \hat{K}_2(G, H) \rightarrow K_2(G, H) \).

The vectorization of \( \hat{m}_G \) (or \( \hat{m}_H \)) can be easily implemented by the Kronecker product, i.e., \( \text{vec}(\hat{m}_G) = \frac{1}{n_G} \sum_{i=1}^{n_G} \hat{\phi}_0(\hat{p}_i) \otimes \hat{\phi}_1(a_1^l) \otimes \ldots \otimes \hat{\phi}_L(a_L^l) \). To obtain above graph kernels, we need only to compute the Euclidean inner product or distance between vectors. More notably, the size of the tensor representation does not depend on node numbers, making it scalable to large graphs.

**Approximate explicit feature map selection.** For the Delta kernel on the discrete attribute domain, we directly use the one-hot vector. For shift-invariant kernels, i.e., \( k(\tilde{x}, \tilde{y}) = k(\tilde{x} - \tilde{y}) \), on Euclidean spaces, e.g., \( A_0 \) and \( A_c \), we make use of random Fourier feature map [21], \( \hat{\phi} : \mathbb{R}^d \rightarrow \mathbb{R}^D \), satisfying \( \langle \hat{\phi}(\tilde{x}), \hat{\phi}(\tilde{y}) \rangle \approx k(\tilde{x}, \tilde{y}) \). To do this, we first draw \( D \) i.i.d. samples \( \omega_1, \omega_2, \ldots, \omega_D \) from a proper distribution \( p(\omega) \). (Note that in this paper, we use \( p(\omega) = \frac{1}{(\sqrt{2\pi})^D} \exp(-\frac{\|\omega\|_2^2}{2}). \)) Next, we draw \( D \) i.i.d. samples \( b_1, b_2, \ldots, b_D \) from the uniform distribution on \( [0, 2\pi] \). Finally, we can calculate \( \hat{\phi}(\tilde{x}) = \sqrt{\frac{2}{D}} \left[ \cos(\omega_1^T \tilde{x} + b_1), \ldots, \cos(\omega_D^T \tilde{x} + b_D) \right]^T \in \mathbb{R}^D \).

### 6 Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of our graph kernels. We run all the experiments on a laptop with an Intel i7-7820HQ, 2.90GHz CPU and 64GB RAM. We implement our algorithms in Matlab, except for the Monte Carlo based computation of RPF (see Section 3.2.1), which is implemented in C++.

#### 6.1 Datasets

We conduct graph classification on four types of benchmark datasets [15]. (i) Non-attributed (unlabeled) graphs datasets: COLLAB, IMDB-BINARY, IMDB-MULTI, REDDIT-BINARY, REDDIT-MULTI(5K), and REDDIT-MULTI(12K) [29] are generated from social networks. (ii) Graphs with
We compare our graph kernels with many state-of-the-art graph classification algorithms: (i) the Weisfeiler-Lehman subtree kernel (WL) [23], (ii) the deep graph convolutional neural network (DGCNN) [31], (iii) the graphlet count kernel (GK) [24], (iv) deep graph kernels (DGK) [29], (v) PATCHY-SAN convolutional neural network (PSCN) [19], (vi) deep graph convolutional neural network (DGCCN) [31], (vii) graph invariant kernels (GIK) [20], and (viii) hashing Weisfeiler-Lehman graph kernels (HGK(WL)) [18].

6.2 Experimental setup

We demonstrate both the graph kernels (I) and (II) introduced in Section 4.1 and Section 5.1, which are invariant kernels (GIK) [20], and (viii) hashing Weisfeiler-Lehman graph kernels (HGK(WL)) [18].

6.2 Experimental setup

We demonstrate both the graph kernels (I) and (II) introduced in Section 4.1 and Section 5.1, which are invariant kernels (GIK) [20], and (viii) hashing Weisfeiler-Lehman graph kernels (HGK(WL)) [18].

6.2 Experimental setup

We demonstrate both the graph kernels (I) and (II) introduced in Section 4.1 and Section 5.1, which are invariant kernels (GIK) [20], and (viii) hashing Weisfeiler-Lehman graph kernels (HGK(WL)) [18].

6.2 Experimental setup

We demonstrate both the graph kernels (I) and (II) introduced in Section 4.1 and Section 5.1, which are invariant kernels (GIK) [20], and (viii) hashing Weisfeiler-Lehman graph kernels (HGK(WL)) [18].
Figure 2: Parameter sensitivity study for RetGK on six benchmark datasets

accuracy of ten repetitions of 10-fold cross-validations with respect to $S$, $D_0$, and $D_c$ in Fig. 2. It can be concluded that RetGKII performs consistently across a wide range of parameter values.

Table 1: Classification results (in %) for non-attributed (unlabeled) graph datasets

| Datasets | WL | GK | DGK | PSCN | RetGK I | RetGK II | RetGKII (MC) |
|----------|----|----|-----|------|---------|----------|-------------|
| COLLAB   | 74.8(0.3) | 72.0(0.3) | 73.1(0.3) | 72.6(2.2) | 81.0(0.3) | 80.6(0.3) | 73.0(3.3) |
| IMDB-BINARY | 70.8(0.5) | 65.9(1.0) | 67.0(0.6) | 71.0(2.3) | 71.8(1.0) | 72.3(0.6) | 71.0(0.6) |
| IMDB-MULTI | 49.3(0.5) | 45.9(0.4) | 44.6(0.5) | 45.2(3.8) | 47.7(0.3) | 46.7(0.6) | 46.7(0.6) |
| REDDIT-BINARY | 68.2(0.2) | 77.3(0.2) | 78.0(0.4) | 86.3(1.6) | 92.6(0.3) | 91.6(0.3) | 90.8(0.2) |
| REDDIT-MULTI(5K) | 71.0(0.5) | 41.0(0.2) | 41.3(0.2) | 49.1(0.9) | 56.1(0.5) | 55.3(0.3) | 54.2(0.5) |
| REDDIT-MULTI(12K) | 32.6(0.3) | 31.8(0.1) | 32.2(0.1) | 41.3(0.4) | 48.7(0.2) | 47.1(0.3) | 45.9(0.2) |
| Total time | 2h3m | – | – | – | 48h14m | 17m14s | 6m9s |

Table 2: Classification results (in %) for graph datasets with discrete attributes

| Datasets | SP | WL | GK | CSM | DGK | PSCN | RetGK I | RetGK II | RetGKII (MC) |
|----------|----|----|----|-----|-----|------|---------|----------|-------------|
| ENZYMES  | 38.6(1.5) | 53.4(0.9) | – | 60.4(0.8) | 53.4(0.9) | 60.4(0.8) | 59.1(1.1) |
| PROTEINS | 73.3(0.9) | 71.0(0.8) | 71.7(0.6) | 75.0(0.9) | 75.0(0.3) | 75.0(0.6) | 75.0(3.3) |
| MUTAG    | 85.2(0.5) | 81.6(1.5) | 81.6(1.5) | 85.8(1.2) | 87.4(2.7) | 89.0(4.4) | 89.0(4.4) |
| DD       | 74.0(0.4) | 85.4(0.3) | 62.3(0.3) | 79.4(0.9) | 76.2(2.6) | 81.6(0.3) | 81.6(0.5) |
| NCI1     | 74.0(0.4) | 85.4(0.3) | 62.3(0.3) | 79.4(0.9) | 76.2(2.6) | 81.6(0.3) | 81.6(0.5) |
| PTC-FM   | 65.3(1.1) | 63.8(1.0) | – | 63.8(1.0) | 63.8(1.0) | 63.8(1.0) | 63.8(1.0) |
| PTC-FR   | 61.6(1.0) | 63.9(1.4) | – | 63.9(1.4) | 63.9(1.4) | 63.9(1.4) | 63.9(1.4) |
| PTC-MM   | 62.9(1.4) | 60.6(1.1) | – | 60.6(1.1) | 60.6(1.1) | 60.6(1.1) | 60.6(1.1) |
| PTC-MR   | 57.8(2.3) | 55.4(2.3) | – | 55.4(2.3) | 55.4(2.3) | 55.4(2.3) | 55.4(2.3) |
| Total time | >24h | 2m27s | – | – | – | 38m4s | 49.9s |

Table 3: Classification results (in %) for graph datasets with continuous attributes

| Datasets | HGK(WL) | RetGK I | RetGKII (MC) |
|----------|---------|---------|-------------|
| ENZYMES  | 63.9(1.1) | 70.0(0.9) | 70.7(0.9) |
| PROTEINS | 74.9(0.6) | 76.0(0.5) | 79.0(0.4) |
| SYNTHEIC | 97.6(0.4) | 97.9(0.3) | 98.0(0.4) |
| Synthie  | 80.3(1.4) | 97.1(0.3) | 96.2(0.3) |
| Total time | – | 45m30s | 40.8s |

Table 4: Classification results (in %) for graph datasets with both discrete and continuous attributes

| Datasets | HGK(WL) | RetGK I | RetGKII (MC) |
|----------|---------|---------|-------------|
| ENZYMES  | 71.7(0.8) | 69.0(0.7) | 72.2(0.8) |
| PROTEINS | 76.0(0.3) | 75.9(0.4) | 78.0(0.3) |
| SYNTHEIC | 79.4(1.2) | 80.1(1.2) | 81.4(0.8) |
| Synthie  | 79.9(1.1) | 80.3(1.1) | 82.0(1.3) |
| Total time | – | 4m17s | 2m51s |

7 Conclusion

In this paper, we introduced the return probability feature for characterizing and comparing the structural role of nodes across graphs. Based on the RPF, we embedded graphs in an RKHS and derived the corresponding graph kernels RetGK I. Then, making use of approximate explicit feature maps, we represented each graph with a multi-dimensional tensor, and then obtained the computationally efficient graph kernels RetGKII. We applied RetGK I and RetGKII to classify graphs, and achieved promising results on many benchmark datasets. Given the prevalence of structured data, we believe that our work can be potentially useful in many applications.
References

[1] Karsten M Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[2] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1):i47–i56, 2005.

[3] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[4] Fan RK Chung. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

[5] Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

[6] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

[7] Paul D Dobson and Andrew J Doig. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology*, 330(4):771–783, 2003.

[8] Aasa Feragen, Niklas Kasenburg, Jens Petersen, Marleen de Bruijne, and Karsten Borgwardt. Scalable kernels for graphs with continuous attributes. In *Advances in Neural Information Processing Systems*, pages 216–224, 2013.

[9] Thomas Gärtner, Peter Flach, and Stefan Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*, pages 129–143. Springer, 2003.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[11] David Haussler. Convolution kernels on discrete structures. Technical report, Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[12] Christoph Helma, Ross D. King, Stefan Kramer, and Ashwin Srinivasan. The predictive toxicology challenge 2000–2001. *Bioinformatics*, 17(1):107–108, 2001.

[13] Jaz Kandola, Thore Graepel, and John Shawe-Taylor. Reducing kernel matrix diagonal dominance using semi-definite programming. In *Learning Theory and Kernel Machines*, pages 288–302. Springer, 2003.

[14] Jeroen Kazius, Ross McGuire, and Roberta Buri. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1):312–320, 2005.

[15] Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel, and Marion Neumann. Benchmark data sets for graph kernels, 2016. [http://graphkernels.cs.tu-dortmund.de](http://graphkernels.cs.tu-dortmund.de)

[16] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[17] Nils Kriege and Petra Mutzel. Subgraph matching kernels for attributed graphs. *arXiv preprint arXiv:1206.6483*, 2012.

[18] Christopher Morris, Nils M Kriege, Kristian Kersting, and Petra Mutzel. Faster kernels for graphs with continuous attributes via hashing. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 1095–1100. IEEE, 2016.

[19] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023, 2016.
[20] Francesco Orsini, Paolo Frasconi, and Luc De Raedt. Graph invariant kernels. In Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence, pages 3756–3762, 2015.

[21] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In Advances in neural information processing systems, pages 1177–1184, 2008.

[22] Tiago A Schieber, Laura Carpi, Albert Díaz-Guilera, Panos M Pardalos, Cristina Masoller, and Martín G Ravetti. Quantification of network structural dissimilarities. Nature communications, 8:13928, 2017.

[23] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. Journal of Machine Learning Research, 12(Sep):2539–2561, 2011.

[24] Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In Artificial Intelligence and Statistics, pages 488–495, 2009.

[25] Jeffrey J Sutherland, Lee A O’Brien, and Donald F Weaver. Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships. Journal of chemical information and computer sciences, 43(6):1906–1915, 2003.

[26] Zoltán Szabó and Bharath K Sriperumbudur. Characteristic and universal tensor product kernels. arXiv preprint arXiv:1708.08157, 2017.

[27] Saurabh Verma and Zhi-Li Zhang. Hunt for the unique, stable, sparse and fast feature learning on graphs. In Advances in Neural Information Processing Systems, pages 87–97, 2017.

[28] S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. Journal of Machine Learning Research, 11(Apr):1201–1242, 2010.

[29] Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1365–1374. ACM, 2015.

[30] Pinar Yanardag and SVN Vishwanathan. A structural smoothing framework for robust graph comparison. In Advances in Neural Information Processing Systems, pages 2134–2142, 2015.

[31] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. 2018.
RetGK: Graph Kernels based on Return Probabilities of Random Walks: Supplementary Material

Zhen Zhang, Mianzhi Wang, Yijian Xiang, Yan Huang, and Arye Nehorai
Department of Electrical and Systems Engineering
Washington University in St. Louis
St. Louis, MO 63130
{zhen.zhang, mianzhi.wang, yijian.xiang, yanhuang640, nehorai}@wustl.edu

Abstract

The supplementary material consists of two parts. In the first part, we prove all the mathematical results presented in the paper. In the second part, we provide the detailed descriptions of the benchmark graph datasets used in the paper.

1 Proofs

1.1 Proving Proposition 1

**Proposition 1.** Let $G$ and $H$ be two isomorphic graphs of $n$ nodes, and let $\tau : \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$ be the corresponding isomorphism. Then,

$$\forall v_i \in V_G, s = 1,2,\ldots,\infty, P^s_G(i,i) = P^s_H(\tau(i),\tau(i)). \quad (1)$$

**Proof.** Let $\Pi$ be the permutation matrix induced by $\tau$, i.e., $\Pi(i,j) = \delta_{j = \tau(i)}$, then $\Pi^T \Pi = \Pi \Pi^T = I_n$. Since $G$ and $H$ are isomorphic, we have $A_H(\tau(i),\tau(j)) = A_G(i,j)$, and $D_H(\tau(i),\tau(i)) = D_G(i,i)$, which is equivalent with $A_H = \Pi^T A_G \Pi$ and $D_H = \Pi^T D_G \Pi$.

Then $P_H = D_H^{-1} A_H = (\Pi^T D_G \Pi)^{-1} (\Pi^T A_G \Pi) = \Pi^T P_G \Pi$. So $P_H^s = (\Pi^T P_G \Pi)^s = \Pi^T P_G^s \Pi$, which implies $P_G^s(i,i) = P_H^s(\tau(i),\tau(i))$. \qed

1.2 Proving Theorem 1

**Theorem 1.** Let $G$ and $H$ be two connected graphs of the same size $n$ and volume $\text{Vol}$, and let $P_G$ and $P_H$ be the corresponding transition probability matrices. Let $\{(\lambda_k, \psi_k)\}_{k=1}^n$ and $\{(\mu_k, \phi_k)\}_{k=1}^n$ be eigenpairs of $P_G$ and $P_H$, respectively. Let $\tau : \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$ be a permutation map. If $P_G^s(i,i) = P_H^s(\tau(i),\tau(i))$, $\forall v_i \in V_G, \forall s = 1,2,\ldots,n$, i.e., $\text{RPF}^S_G = \text{RPF}^S_H$, then,

1. $\text{RPF}^S_G = \text{RPF}^S_H, \forall S = n + 1, n + 2, \ldots, \infty$;
2. $\{\lambda_1, \lambda_2, \ldots, \lambda_n\} = \{\mu_1, \mu_2, \ldots, \mu_n\}$;
3. If the eigenvalues sorted by their magnitudes satisfy: $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_m| > 0$, $|\lambda_{m+1}| = \ldots = |\lambda_n| = 0$, then we have that $|\psi_k(i)| = |\phi_k(\tau(i))|, \forall v_i \in V_G, \forall k = 1,2,\ldots,m$.

We first present some useful lemmas.
Lemma 1. (Cayley-Hamilton Theorem, see[11, 16]) Let \( A \) be an \( n \times n \) matrix, and let \( P(\lambda) = \det(\lambda I_n - A) \) be the corresponding characteristic polynomial of \( A \), then \( P(A) = 0 \), i.e.,

\[
A^n + c_{n-1}A^{n-1} + \cdots + c_1A + (-1)^n\det(A)I_n = 0, \tag{2}
\]

where \( c_{n-k} = \frac{(-1)^k}{k!}B_k(s_1, (-1)s_2, 2!s_3, \cdots, (-1)^{k-1}(k-1)!s_k) \), and \( B_k \) is the Bell polynomial and \( s_i = \text{trace}(A^i) \). In particular, \( \det(A) = \frac{1}{n!}B_n(s_1, (-1)s_2, 2!s_3, \cdots, (-1)^{n-1}(n-1)!s_n) \).

Remark 1. We observe that all the coefficients in (2) are determined by \( \text{trace}(A) \), \( \text{trace}(A^2) \), \ldots, \( \text{trace}(A^n) \).

Corollary 1. Let \( A \) and \( B \) be two \( n \times n \) matrices. If \( \text{trace}(A^k) = \text{trace}(B^k), k = 1, 2, \cdots, n \), then \( A \) and \( B \) have the same eigenvalues set.

Proof. Let \( P_A(\lambda) \) and \( P_B(\lambda) \) be the characteristic polynomials of \( A \) and \( B \) respectively, and let \( c_{n-k}^A \) and \( c_{n-k}^B \), \( k = 1, 2, \cdots, n \), be the corresponding coefficients. Since trace(\( A^k \)) = trace(\( B^k \)), \( k = 1, 2, \cdots, n \), we have that \( c_{n-k}^A = c_{n-k}^B \), \( k = 1, 2, \cdots, n-1 \), and \( \det(A) = \det(B) \). Therefore, the roots of \( P_A(\lambda) \) and \( P_B(\lambda) \) are the same, which is equivalent to \( A \) and \( B \) having the same eigenvalues set.

Corollary 2. Let \( A \) and \( B \) be two \( n \times n \) matrices. If \( A^*(i, i) = B^*(i, i), s = 1, 2, \cdots, n, i = 1, 2, \cdots, n, \) then \( A^*(i, i) = B^*(i, i), s = n + 1, n + 2, \cdots, i = 1, 2, \cdots, n \).

Proof. It is easy to obtain \( \text{trace}(A^s) = \text{trace}(B^s), s = 1, 2, \cdots, n \). Then based on the lemma 1, the characteristic polynomials of \( A \) and \( B \) are same. Moreover,

\[
A^n = -c_{n-1}A^{n-1} - c_{n-2}A^{n-2} - \cdots - c_1A + (-1)^n\det(A)I_n. \tag{3}
\]

Multiply \( A^{s-n} \), \( s \geq n + 1 \) on both sides, and we have

\[
A^s = -c_{n-1}A^{s-1} - c_{n-2}A^{s-2} - \cdots - c_1A^{s-(n-1)} + (-1)^n\det(A)A^{s-n}. \tag{4}
\]

Immediately, for any \( i = 1, 2, \cdots, n \),

\[
A^s(i, i) = -c_{n-1}A^{s-1}(i, i) - c_{n-2}A^{s-2}(i, i) - \cdots - c_1A^{s-(n-1)}(i, i) + (-1)^n\det(A)A^{s-n}(i, i). \tag{5}
\]

From the iterative formula (5), we can see that \( A^s(i, i), s = n + 1, n + 2, \cdots \) are uniquely determined by \( \{A(i, i), A^2(i, i), \cdots, A^n(i, i)\} \). Similarly, \( B^s(i, i), s = n + 1, n + 2, \cdots \) are uniquely determined by \( \{B(i, i), B^2(i, i), \cdots, B^n(i, i)\} \). Combining with the fact \( A^s(i, i) = B^s(i, i), s = 1, 2, \cdots, n \), we obtain the desired result.

Lemma 2. (Time-reversible Markov chains, see [2]) If for an irreducible Markov chain with transition matrix \( P \), there exists a probability solution \( \vec{\pi} \) to the “Time-reversibility” set of equations,

\[
\vec{\pi}^T P(i, j) = \vec{\pi}^T P(j, i), \tag{6}
\]

for all pairs of states \( i, j \), then the solution \( \vec{\pi} \) is the unique stationary distribution, i.e.,

\[
\lim_{s \to +\infty} P^s(i, j) = \vec{\pi}_j.
\]

Remark 2. For a connected graph \( G \), the random walk defined on it can be considered as an irreducible Markov chain. We define a probability vector \( \vec{\pi} \) as \( \vec{\pi}_i = \frac{D_G(i, i)}{\text{Vol}_G}, \) where \( \text{Vol}_G \) is the volume of the graph, i.e., \( \text{Vol}_G = \sum_{i=1}^{n} D_G(i, i) \). Then we have,

\[
\vec{\pi}_i P_G(i, j) = \frac{D_G(i, i)}{\text{Vol}_G} \times \frac{w_{ij}}{D_G(i, i)} = \frac{D_G(j, j)}{\text{Vol}_G} \times \frac{w_{ji}}{D_G(j, j)} = \vec{\pi}_j P_G(j, i). \tag{7}
\]

Therefore, \( \vec{\pi} \) defined above is the stationary distribution of the random walk.

Now we prove theorem 1.

Proof. (1) Let \( \Pi \) be the permutation matrix induced by \( \tau \), i.e., \( \Pi(i, j) = \delta_{j=\tau(i)} \). Then we have \( \forall i, j = 1, 2, \cdots, n, P_{\Pi}(\tau(i), \tau(j)) = (\Pi P_{\Pi}(\Pi^T))(i, j) \). Since \( P_{\Pi}^s(i, i) = P_{\Pi}^s(\tau(i), \tau(i)) = (\Pi P_{\Pi}(\Pi^T))^{s}(i, i), \forall i \in V_G, \) and \( \forall s = 1, 2, \cdots, n, \) by Corollary 2, we have
\[ P^*_G(i, i) = (\Pi_P \Pi_T)^*(i, i) = P^*_H(\tau(i), \tau(i)), \forall s = n+1, n+2 \cdots + \infty. \] Now, the first conclusion has been proved.

(2). The second one can be directly concluded from corollary 1.

(3). Let \( D_G \) and \( D_H \) be the degree matrices of graph \( G \) and \( H \), respectively. Then by Remark 2,

\[
\frac{D_G(i, i)}{\text{Vol}} = \lim_{s \to +\infty} P_G^*(i, i) = \lim_{s \to +\infty} P_H^*(\tau(i), \tau(i)) = \frac{D_H(\tau(i), \tau(i))}{\text{Vol}}.
\]

So

\[ D_G(i, i) = D_H(\tau(i), \tau(i)), \forall v_i \in V_G. \]

Let \( A_G \) and \( A_H \) be the adjacent matrices of \( G \) and \( H \) respectively, and write \( P_G = D_G^{-1} A_G = D_G^{-1} \frac{1}{2} (A_G + A_G^T) D_G^{-\frac{1}{2}}. \) Let \( B_G = D_G^{-\frac{1}{2}} A_G D_G^{-\frac{1}{2}} \implies P_G = D_G^{-\frac{1}{2}} B_G D_G^{-\frac{1}{2}} = P_G^* \). \( B_G \) is a symmetric matrix, and has the same eigenvalues as \( P_G \). Write the orthonormal eigen-decomposition of \( B_G \) as \( B_G = \sum_{k=1}^{n} \lambda_k u_k u_k^T \), then

\[ P_G^*(i, i) = B_G^*(i, i) = \sum_{k=1}^{n} \lambda_k^* u_k(i)^2, \]

where \( u_k(i) \) denotes the \( i \)th component of the eigenvector \( u_k \). Similarly, we have

\[ P_H^*(\tau(i), \tau(i)) = B_H^*(\tau(i), \tau(i)) = \sum_{k=1}^{n} \mu_k^* \tilde{v}_k(\tau(i))^2 = \sum_{k=1}^{n} \lambda_k^* \tilde{v}_k(\tau(i))^2, \]

where \( \tilde{v}_k(\tau(i)) \) denotes the \( \tau \)th component of \( \tilde{v}_k \), and \( \tilde{v}_k \) is the \( k \)th eigenvector of \( B_H = D_H^{-\frac{1}{2}} A_H D_H^{-\frac{1}{2}}. \) The last equality of (11) holds because \( \{\lambda_1, \lambda_2, ..., \lambda_n\} = \{\mu_1, \mu_2, ..., \mu_n\}. \)

Next, we use mathematical induction to show that \( |u_k(i)| = |\tilde{v}_k(\tau(i))|, \forall v_i \in V_G, \forall k = 1, 2, \cdots, m. \)

**Step1:** For \( k = 1, P_G^* \frac{1}{2} \frac{1}{2} = B_G^* \frac{1}{2} \frac{1}{2} = D_G^* \frac{1}{2} \frac{1}{2} \Leftrightarrow \tilde{u}_1 = \pm D_G^* \frac{1}{2} \frac{1}{2}. \) Similarly, we have \( \tilde{v}_1 = \pm D_H^* \frac{1}{2} \frac{1}{2}. \)

**Step2:** We show that if the first \( k \) eigenvectors satisfy, \( |u_1(i)| = |\tilde{v}_1(\tau(i))|, |u_2(i)| = |\tilde{v}_2(\tau(i))|, \cdots, |u_k(i)| = |\tilde{v}_k(\tau(i))|, \) then \( |u_{k+1}(i)| = |\tilde{v}_{k+1}(\tau(i))|, \forall v_i \in V_G, \forall k = 1, 2, \cdots, m - 2. \)

We suppose that \( \exists v_i^* \in V_G, \mbox{ such that } |u_{k+1}(i^*)| \neq |\tilde{v}_{k+1}(\tau(i^*))|. \) Without loss of generality, we assume that \( u_{k+1}(i^*) - \tilde{v}_{k+1}(\tau(i^*))^2 = \epsilon > 0. \) Write

\[
P_G^*(i^*, i^*) - P_H^*(\tau(i^*), \tau(i^*)) = \sum_{l=1}^{n} \lambda_l^* u_l(i^*)^2 - \sum_{l=1}^{n} \lambda_l^* \tilde{v}_l(\tau(i^*))^2 = \sum_{l=1}^{m} \lambda_l^* u_l(i^*)^2 - \sum_{l=1}^{m} \lambda_l^* \tilde{v}_l(\tau(i^*))^2 = \sum_{l=k+1}^{m} \lambda_l^* u_l(i^*)^2 - \sum_{l=k+1}^{m} \lambda_l^* \tilde{v}_l(\tau(i^*))^2
\]

\[
= \lambda_{k+1}^* \left[ u_{k+1}(i^*)^2 - \tilde{v}_{k+1}(\tau(i^*))^2 \right] - \sum_{l=k+2}^{m} \lambda_l^* \left[ u_l(i^*)^2 - \tilde{v}_l(\tau(i^*))^2 \right]
\]

where the second equality holds because \( |\lambda_1| > |\lambda_2| > \cdots > |\lambda_m| > 0, |\lambda_{m+1}| = \cdots = |\lambda_n| = 0. \)

With the fact that \( |u_{k+1}(i^*)^2 - \tilde{v}_{k+1}(\tau(i^*))^2| \leq 1 \) (since \( 0 \leq u_{k+1}(i^*)^2, \tilde{v}_{k+1}(\tau(i^*))^2 \leq 1 \)) and \( |\frac{\lambda_l}{\lambda_{k+1}}| < 1, \)
we have that there is a positive integer, $M$, such that,

$$
\epsilon - \sum_{l=k+2}^{m} \left( \frac{\lambda_l}{\lambda_{k+1}} \right)^{2M} \left[ \vec{u}_l(i^\ast)^2 - \vec{v}_l(\tau(i^\ast))^2 \right] > 0.
$$

(13)

Therefore, $P_{G}^{2M}(i^\ast, i^\ast) - P_{H}^{2M}(\tau(i^\ast), \tau(i^\ast)) > 0$, contradicting the fact that $P_{G}^0(i, i) = P_{H}^0(\tau(i), \tau(i))$, $\forall v_i \in V_G, \forall s = 1, 2, \ldots, \infty$. So $|\vec{u}_{k+1}(i)| = |\vec{v}_{k+1}(\tau(i))|$, $\forall v_i \in V_G$.

**Step 3:** We show that if $|\vec{u}_1(i)| = |\vec{v}_1(\tau(i))|$, $|\vec{u}_2(i)| = |\vec{v}_2(\tau(i))|$, $\ldots$, $|\vec{u}_{m-1}(i)| = |\vec{v}_{m-1}(\tau(i))|$, then $|\vec{u}_m(i)| = |\vec{v}_m(\tau(i))|$, $\forall v_i \in V_G$. Since

$$
0 = P_{G}^0(i^\ast, i^\ast) - P_{H}^0(\tau(i^\ast), \tau(i^\ast)) = \lambda_m^* \left[ \vec{u}_m(i^\ast)^2 - \vec{v}_m(\tau(i^\ast))^2 \right],
$$

and $\lambda_m^* \neq 0$, we immediately have that $|\vec{u}_m(i^\ast)| = |\vec{v}_m(\tau(i^\ast))|$. Combining all these three steps, we obtain the desired result $|\vec{u}_k(i)| = |\vec{v}_k(\tau(i))|$, $\forall v_i \in V_G$, $\forall k = 1, 2, \ldots, m$.

Since $P_{G} = D_{G}^{-\frac{1}{2}} B_{G} D_{G}^{\frac{1}{2}}$, we have the fact that $(\lambda_k, \vec{u}_k)$ is an eigenpair of $B_{G}$ if and only if $(\lambda_k, D_{G}^{-\frac{1}{2}} \vec{u}_k)$ is an eigenpair of $P_{G}$. The above implies that $\vec{\psi}_k = D_{G}^{\frac{1}{2}} \vec{v}_k$, and similarly $\vec{\varphi}_k = D_{H}^{\frac{1}{2}} \vec{v}_k$. Now, $\forall v_i \in V_G$, $\forall k = 1, 2, \ldots, m$, we have

$$
|\vec{\varphi}_k(\tau(i))| = D_{H}^{\frac{1}{2}}(\tau(i))|\vec{v}_k(\tau(i))| = D_{G}^{\frac{1}{2}}(\tau(i))|\vec{\psi}_k(i)| = |\vec{\psi}_k(i)|.
$$

(15)

\[\square\]

### 1.3 Proving proposition 2

Given two graphs $G$ and $H$, let $\{\Delta_i^G\}_{i=1}^{n_G}$ and $\{\Delta_i^H\}_{i=1}^{n_H}$ be the respective set representations $(\Delta_i^G = (\vec{p}_i, a_1^G, a_2^G, \ldots, a_{p_i}^G)$ and likewise $\Delta_i^H)$. Let $K_{GG}$, $K_{HH}$, and $K_{GH}$ be the kernel matrices, induced by the embedding kernel $k = \otimes_{i=1}^{d} k_i$. That is, they are defined such that $K_{GG}(i,j) = k(\Delta_i^G, \Delta_j^G)$, $K_{HH}(i,j) = k(\Delta_i^H, \Delta_j^H)$, and $K_{GH}(i,j) = k(\Delta_i^G, \Delta_j^H)$.

**Proposition 2.** Let $G$ be the set of graphs with attribute domains $A_1, A_2, \ldots, A_L$. Let $G$ and $H$ be two graphs in $G$. Let $m_G$ and $m_H$ be the corresponding mean embeddings. Then the following functions are positive definite graph kernels defined on $G \times G$.

\[
K_1(G, H) = (c + \langle m_G, m_H \rangle) d = (c + \frac{1}{n_{G,H}} \vec{1}_{n_G}^T K_{GH} \vec{1}_{n_H})^d, c > 0, d \in \mathbb{N},
\]

\[
K_2(G, H) = \exp(-\gamma \|m_G - m_H\|_H^p) = \exp\left[-\gamma \text{MMD}^p(\mu_G, \mu_H)\right], \gamma > 0, 0 < p \leq 2,
\]

where $\text{MMD}(\mu_G, \mu_H) = (\frac{1}{n_{G,H}} \vec{1}_{n_G}^T K_{GG} \vec{1}_{n_H} + \frac{1}{n_{H}} \vec{1}_{n_H}^T K_{HH} \vec{1}_{n_H} - \frac{2}{n_{G,H}} \vec{1}_{n_G}^T K_{GH} \vec{1}_{n_H})^2$ is the maximum mean discrepancy (MMD) [6].

**Proof.** (a). We first consider two kernels $K_\alpha(G, H) = \langle m_G, m_H \rangle$ and $K_\beta(G, K) = c$. It can be easily observed that $K_\alpha$ and $K_\beta$ are positive definite graph kernels. Since the sum and multiplication of positive definite kernels are still positive definite, we conclude that (16a) are positive definite.

(b). The positive definiteness of (16b) is obtained from Corollary 3 in [12]. \[\square\]
1.4 Proving theorem 2

**Theorem 2.** Let $G$ and $H$ be two graphs with attribute domains $A_1, ..., A_L$. Let $\hat{\phi}_l : A_l \rightarrow \mathbb{R}^{D_l}$, $l = 0, 1, ..., L$ be the approximate explicit feature maps. Let $\{(\tilde{p}_1, a_1^1, ..., a_1^L)\}_{i=1}^{n_G}$ and $\{(\tilde{q}_j, b_j^1, ..., b_j^L)\}_{j=1}^{n_H}$ be the respective set representations of $G$ and $H$. Then their approximate explicit graph embeddings, $\hat{m}_G$ and $\hat{m}_H$, are tensors in $\mathbb{R}^{D_0 \times D_1 \times ... \times D_L}$, and can be written as

$$
\hat{m}_G = \frac{1}{n_G} \sum_{i=1}^{n_G} \hat{\phi}_0(\tilde{p}_i) \circ \hat{\phi}_1(a_1^i) \circ \cdots \circ \hat{\phi}_L(a_L^i), \quad \hat{m}_H = \frac{1}{n_H} \sum_{j=1}^{n_H} \hat{\phi}_0(\tilde{q}_j) \circ \hat{\phi}_1(b_1^j) \circ \cdots \circ \hat{\phi}_L(b_L^j). \quad (17)
$$

That is, as $D_0, D_1, ..., D_L \rightarrow \infty$, we have $\langle \hat{m}_G, \hat{m}_H \rangle_T \rightarrow \langle m_G, m_H \rangle_T$.

Before we prove theorem 2, we first introduce a lemma about the inner product of multi-dimensional tensors.

**Lemma 3.** Let $U = \tilde{u}^{(0)}(0) \circ \tilde{u}^{(1)}(1) \circ \cdots \circ \tilde{u}^{(L)}(L)$ and $V = \tilde{v}^{(0)}(0) \circ \tilde{v}^{(1)}(1) \circ \cdots \circ \tilde{v}^{(L)}(L)$ be two rank-one tensors in $\mathbb{R}^{D_0 \times D_1 \times ... \times D_L}$. Then we have $\langle U, V \rangle_T = \langle \tilde{u}^{(0)}(0), \tilde{v}^{(0)}(0) \rangle \langle \tilde{u}^{(1)}(1), \tilde{v}^{(1)}(1) \rangle \cdots \langle \tilde{u}^{(L)}(L), \tilde{v}^{(L)}(L) \rangle$.

**Proof.**

$$
\langle U, V \rangle_T = \sum_{i_0=1}^{D_0} \sum_{i_1=1}^{D_1} \cdots \sum_{i_L=1}^{D_L} U_{i_1 i_2 \ldots i_L} V_{i_1 i_2 \ldots i_L} = \sum_{i_0=1}^{D_0} \sum_{i_1=1}^{D_1} \cdots \sum_{i_L=1}^{D_L} \tilde{u}^{(0)}(i_0) \tilde{u}^{(1)}(i_1) \cdots \tilde{u}^{(L)}(i_L) \tilde{v}^{(0)}(i_0) \tilde{v}^{(1)}(i_1) \cdots \tilde{v}^{(L)}(i_L) = \langle \tilde{u}^{(0)}, \tilde{v}^{(0)} \rangle \langle \tilde{u}^{(1)}, \tilde{v}^{(1)} \rangle \cdots \langle \tilde{u}^{(L)}, \tilde{v}^{(L)} \rangle.
$$

Now we prove theorem 2.

**Proof.** First we calculate $\langle \hat{m}_G, \hat{m}_H \rangle_T$.

$$
\langle \hat{m}_G, \hat{m}_H \rangle_T = \langle \frac{1}{n_G} \sum_{i=1}^{n_G} \hat{\phi}_0(\tilde{p}_i) \circ \hat{\phi}_1(a_1^i) \circ \cdots \circ \hat{\phi}_L(a_L^i), \frac{1}{n_H} \sum_{j=1}^{n_H} \hat{\phi}_0(\tilde{q}_j) \circ \hat{\phi}_1(b_1^j) \circ \cdots \circ \hat{\phi}_L(b_L^j) \rangle_T
$$

$$
= \frac{1}{n_G n_H} \sum_{i=1}^{n_G} \sum_{j=1}^{n_H} \langle \hat{\phi}_0(\tilde{p}_i) \circ \hat{\phi}_1(a_1^i) \circ \cdots \circ \hat{\phi}_L(a_L^i), \hat{\phi}_0(\tilde{q}_j) \circ \hat{\phi}_1(b_1^j) \circ \cdots \circ \hat{\phi}_L(b_L^j) \rangle_T
$$

$$
= \frac{1}{n_G n_H} \sum_{i=1}^{n_G} \sum_{j=1}^{n_H} \langle \hat{\phi}_0(\tilde{p}_i), \hat{\phi}_0(\tilde{q}_j) \rangle \langle \hat{\phi}_1(a_1^i), \hat{\phi}_1(b_1^j) \rangle \cdots \langle \hat{\phi}_L(a_L^i), \hat{\phi}_L(b_L^j) \rangle
$$

$$
= \frac{1}{n_G n_H} \sum_{i=1}^{n_G} \sum_{j=1}^{n_H} \hat{k}_0(\tilde{p}_i, \tilde{q}_j) \hat{k}_1(a_1^i, b_1^j) \cdots \hat{k}_L(a_L^i, b_L^j),
$$

where the 3rd equality holds because of lemma 3.
Next we calculate $⟨m_Г,m_Н⟩_Н$.

$$\langle m_G, m_H \rangle_H = \langle \frac{1}{n_G} \sum_{i=1}^{n_G} \phi(p_i, a^1_i, \ldots, a^L_i), \frac{1}{n_H} \sum_{j=1}^{n_H} \phi(q_j, b^1_j, \ldots, b^L_j) \rangle_H$$

$$= \frac{1}{n_G n_H} \sum_{i=1}^{n_G} \sum_{j=1}^{n_H} \langle \phi(p_i, a^1_i, \ldots, a^L_i), \phi(q_j, b^1_j, \ldots, b^L_j) \rangle_H$$

$$= \frac{1}{n_G n_H} \sum_{i=1}^{n_G} \sum_{j=1}^{n_H} k[(p_i, a^1_i, \ldots, a^L_i), (q_j, b^1_j, \ldots, b^L_j)]$$

$$= \frac{1}{n_G n_H} \sum_{i=1}^{n_G} \sum_{j=1}^{n_H} k_0(p_i, q_j) k_1(a^1_i, b^1_j) \cdots k_L(a^L_i, b^L_j),$$

where the last equality holds because of the definition of the embedding kernel $k = \otimes_{i=0}^{L} k_i$.

Since $k_0(p_i, q_j) \rightarrow k_0(p_i, q_j)$, $k_1(a^1_i, b^1_j) \rightarrow k_1(a^1_i, b^1_j)$, $\cdots$, $k_L(a^L_i, b^L_j) \rightarrow k_L(a^L_i, b^L_j)$, as $D_0, D_1, \ldots, D_L \rightarrow \infty$, we conclude that $\langle \hat{m}_G, \hat{m}_H \rangle_T \rightarrow \langle m_G, m_H \rangle_H$. \hfill \qed

### 1.5 Proving proposition 3

**Proposition 3.** Let $\mathcal{G}$ be the set of graphs with attribute domains $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_L$. The following functions are positive definite graph kernels defined on $\mathcal{G} \times \mathcal{G}$.

$$\hat{K}_1(G, H) = (c + \langle \hat{m}_G, \hat{m}_H \rangle_T)^d \in [c + \text{vec}(\hat{m}_G)^T \text{vec}(\hat{m}_H)]^d, c > 0, d \in \mathbb{N}$$

$$\hat{K}_2(G, H) = \exp(-\gamma \|\hat{m}_G - \hat{m}_H\|_F^p) = \exp(-\gamma \|\text{vec}(\hat{m}_G) - \text{vec}(\hat{m}_H)\|_2^p), \gamma > 0, 0 < p \leq 2.$$  

As $D_0, D_1, \ldots, D_L \rightarrow \infty$, we have $\hat{K}_1(G, H) \rightarrow K_1(G, H)$ and $\hat{K}_2(G, H) \rightarrow K_2(G, H)$.

**Proof.** The positive definiteness of $\hat{K}_1$ and $\hat{K}_2$ can be proved in the same way with Theorem 2. The convergence property can be obtained by Theorem 2. \hfill \qed

### 2 Datasets description

The statistics of the benchmark graph datasets used in the paper are reported in Table 1. Next, we describe in these datasets in detail.

#### 2.1 Non-attributed (unlabeled) graph datasets

**COLLAB** [15] is a scientific collaboration dataset that consists of the ego-networks of 5,000 researchers from three scientific fields: *High Energy Physics*, *Condensed Matter Physics*, and *Astro Physics*. The task is to determine the field of each researcher based on their ego-networks.

**IMDB-BINARY** [15] is a movie collaboration dataset that consists of the ego-networks of 1,000 actors/actresses who played roles in movies in IMDB. In each graph, nodes represent actors/actress, and there is an edge between them if they appear in the same movie. These graphs are derived from the *Action* and *Romance* genres.

**IMDB-MULTI** [15] is generated in a similar way to IMDB-BINARY. The difference is that it is derived from three genres: *Comedy*, *Romance*, and *Sci-Fi*.

**REDDIT-BINARY** [15] consists of graphs corresponding to online discussions on Reddit. In each graph, nodes represent users, and there is an edge between them if at least one of them respond to the other’s comment. There are four popular subreddits, namely, *IAmA*, *AskReddit*, *TrollXChromosomes*, and *atheism*. *IAmA* and *AskReddit* are two question/answer-based subreddits, and *TrollXChromosomes* and *atheism* are two discussion-based subreddits. A graph is labeled according to whether it belongs to a question/answer-based community or a discussion-based community.
REDDIT-MULTI(5K) [15] is generated in a similar way to REDDIT-BINARY. The difference is that there are five subreddits involved, namely, worldnews, videos, AdviceAnimals, aww, and mildlyinteresting. Graphs are labeled with their corresponding subreddits.

REDDIT-MULTI(12K) [15] is generated in a similar way to REDDIT-BINARY and REDDIT-MULTI(5K). The difference is that there are eleven subreddits involved, namely, AskReddit, AdviceAnimals, atheism, aww, IAM, mildlyinteresting, Showerthoughts, videos, todayilearned, worldnews, and TrollXChromosomes. Still, graphs are labeled with their corresponding subreddits.

2.2 Graphs with discrete attributes

MUTAG [3] consists of graph representations of 188 mutagenic aromatic and heteroaromatic nitro chemical compounds. These graphs are labeled according to whether or not they have a mutagenic effect on the Gramnegative bacterium Salmonella typhimurium.

DD [4] consists of graph representations of 1,178 proteins. In each graph, nodes represent amino acids, and there is an edge if they are less than six Angstroms apart. Graphs are labeled according to whether they are enzymes or not.

NCI1 [13] consists of graph representations of 4,110 chemical compounds screened for activity against non-small cell lung cancer and ovarian cancer cell lines, respectively.

PTC [7] consists of graph representations of chemical molecules. In each graph, nodes represent atoms, and edges represent chemical bonds. Graphs are labeled according to carcinogenicity on rodents, divided into male mice (MM), male rats (MR), female mice (FM), and female rats (FR).

2.3 Graphs with continuous attributes

FRANK [8] is a chemical molecule dataset that consists of 2,401 mutagens and 1,936 nonmutagens. Originally, nodes are associated with chemical atom symbols. The most frequent atom symbols are mapped to MNIST digit images. By doing this, the original atom symbols can be recovered through the high dimensional MNIST vectors of pixel intensities, which are treated as the continuous attributes on graphs.

SYNTHETIC [5] consists of 300 random graphs. The continuous node attributes are sampled from the distribution $N(0, 1)$. There are two classes, A and B. Class A has 150 graphs, which are generated by randomly rewiring five edges and permuting ten node attributes. Class B has 150 graphs, which are generated by randomly rewiring ten edges and permuting five node attributes.

Synthie [10] consists of 400 random graphs, all of which are variants of two Erdos-Renyi graphs. The nodes are associated with 15-dimensional continuous attributes. All graphs are divided into four classes. The generation process of these graphs is described in [10].

2.4 Graphs with both discrete and continuous attributes

ENZYMES and PROTEINS [1] consist of graph representations of proteins. Nodes represent secondary structure elements (SSE), and there is an edge if they are neighbors along the amino acid sequence or one of three nearest neighbors in space. The discrete attributes are SSE’s types. The continuous attributes are the 3D length of the SSE. Graphs are labeled according to which EC top-level class they belong to.

BZR, COX2, and DHFR [14], [9] all are chemical compound datasets. Still, in each graph, nodes represent atoms, and edges represent chemical bonds. The discrete attributes correspond to atom types. The continuous attributes are 3D coordinates.

References

[1] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. Bioinformatics, 21(suppl_1):i47–i56, 2005.

[2] Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.
| Datasets          | graph # | class # | average node # | average edge # | discrete attributes | continuous attributes (Dim) |
|------------------|---------|---------|----------------|----------------|---------------------|-----------------------------|
| COLLAB           | 5000    | 3       | 74.49          | 2457.78        | ×                   | ×                           |
| IMDB-BINARY      | 1000    | 2       | 19.77          | 96.53          | ×                   | ×                           |
| IMDB-MULTI       | 1500    | 3       | 13.00          | 65.94          | ×                   | ×                           |
| REDDIT-BINARY    | 2000    | 2       | 429.63         | 497.75         | ×                   | ×                           |
| REDDIT-MULTI(5K) | 4999    | 5       | 508.52         | 594.87         | ×                   | ×                           |
| REDDIT-MULTI(12K)| 11929   | 11      | 391.41         | 456.89         | ×                   | ×                           |
| MUTAG            | 188     | 2       | 17.93          | 19.79          | ✓                   | ×                           |
| DD               | 1178    | 2       | 284.32         | 715.66         | ✓                   | ×                           |
| NCI1             | 4110    | 2       | 29.87          | 32.30          | ✓                   | ×                           |
| PTC-FM           | 349     | 2       | 14.11          | 14.48          | ✓                   | ×                           |
| PTC-FR           | 351     | 2       | 14.56          | 15.00          | ✓                   | ×                           |
| PTC-MM           | 336     | 2       | 13.97          | 14.32          | ✓                   | ×                           |
| PTC-MR           | 344     | 2       | 14.29          | 14.69          | ✓                   | ×                           |
| FRANK            | 4337    | 2       | 16.90          | 17.88          | ✓                   | ✓                           |
| SYNTHEIC         | 300     | 2       | 100            | 196.25         | ✓                   | ✓                           |
| Synthie          | 400     | 4       | 95.00          | 172.93         | ×                   | ✓                           |
| ENZYMES          | 600     | 6       | 32.63          | 64.14          | ✓                   | ✓                           |
| PROTEINS         | 7113    | 2       | 39.06          | 72.82          | ✓                   | ✓                           |
| BZR              | 405     | 2       | 35.75          | 38.36          | ✓                   | ✓                           |
| DHFR             | 467     | 2       | 41.22          | 43.45          | ✓                   | ✓                           |
| COX2             | 467     | 2       | 42.43          | 44.54          | ✓                   | ✓                           |

[3] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

[4] Paul D Dobson and Andrew J Doig. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology*, 330(4):771–783, 2003.

[5] Aasa Feragen, Niklas Kasenburg, Jens Petersen, Marleen de Bruijne, and Karsten Borgwardt. Scalable kernels for graphs with continuous attributes. In *Advances in Neural Information Processing Systems*, pages 216–224, 2013.

[6] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[7] Christoph Helma, Ross D. King, Stefan Kramer, and Ashwin Srinivasan. The predictive toxicology challenge 2000–2001. *Bioinformatics*, 17(1):107–108, 2001.

[8] Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1):312–320, 2005.

[9] Nils Kriege and Petra Mutzel. Subgraph matching kernels for attributed graphs. *arXiv preprint arXiv:1206.6483*, 2012.

[10] Christopher Morris, Nils M Kriege, Kristian Kersting, and Petra Mutzel. Faster kernels for graphs with continuous attributes via hashing. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 1095–1100. IEEE, 2016.

[11] Louis L Pennisi. Coefficients of the characteristic polynomial. *Mathematics magazine*, 60(1):31–33, 1987.

[12] Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[13] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(Sep):2539–2561, 2011.

[14] Jeffrey J Sutherland, Lee A O’brien, and Donald F Weaver. Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships. *Journal of chemical information and computer sciences*, 43(6):1906–1915, 2003.

[15] Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1365–1374. ACM, 2015.
[16] Lotfi Asker Zadeh and Charles A Desoer. Linear system theory: (The) state space approach. 1963.