Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

A Property Testing Framework for the Theoretical Expressivity of Graph Kernels

Nils M. Kriege, Christopher Morris, Anja Rey, Christian Sohler
TU Dortmund University, Dortmund, Germany
{nils.kriege,christopher.morris,anja.rey,christian.sohler}@tu-dortmund.de

Abstract

Graph kernels are applied heavily for the classification of structured data. However, their expressivity is assessed almost exclusively from experimental studies and there is no theoretical justification why one kernel is in general preferable over another. We introduce a theoretical framework for investigating the expressive power of graph kernels, which is inspired by concepts from the area of property testing. We introduce the notion of *distinguishability* of a graph property by a graph kernel. For several established graph kernels we show that they cannot distinguish essential graph properties. In order to overcome this, we consider a kernel based on k-disc frequencies. We show that this efficiently computable kernel can distinguish fundamental graph properties. Finally, we obtain learning guarantees for nearest neighbor classifiers in our framework.

1 Introduction

Linked data arises in various domains such as chem- and bioinformatics, social network analysis and pattern recognition. Such data can naturally be represented by graphs. Therefore, machine learning on graphs has become an active research area of increasing importance. The prevalent approach to classify graphs is to design kernels on graphs in order to employ standard kernel methods such as support vector machines. Consequently, in the past two decades a large number of graph kernels have been proposed, see, e.g., [Vishwanathan *et al.*, 2010]. Most graph kernels decompose graphs and add up the pairwise similarities between their substructures following the seminal concept of convolution kernels [Haussler, 1999]. Here, substructures may be walks [Gärtner *et al.*, 2003] or certain subgraphs [Ramon and Gärtner, 2003; Shervashidze *et al.*, 2009]. Considering the large number of available graph kernels and the wealth of available benchmark data sets [Kersting *et al.*, 2016], it becomes increasingly difficult to perform a fair experimental comparison of kernels and to assess their advantages and disadvantages for specific data sets. Indeed, current experimental comparisons cannot give a complete picture and are of limited help to a practitioner who has to choose a kernel for a particular application.

Graph kernels are developed with the (possibly conflicting) goals of being efficiently computable and capturing the topological information of the input graphs adequately. Newly proposed graph kernels are often justified by their ability to take structural graph properties into account that were ignored by previous kernels. Yet, to the best of our knowledge, this argument has not been formalized. Moreover, there is no theoretical justification why certain kernels perform better than others, but merely experimental evaluations. We address this by introducing a theoretical framework for the analysis of the expressivity of graph kernels motivated by concepts from property testing, see, e.g., [Goldreich, 2017]. We consider normalized kernels, which measure similarity in terms of angles in a feature space. We say that a graph kernel *identifies* a property if no two graphs are mapped to the same normalized feature vector unless they both have or both do not have the property. A positive angle between two such feature vectors can be helpful to classify the property. As the graph size increases, on the one hand, this angle can become very small (dependent on the graph size), which is hindering when applying this knowledge to a learning setting. On the other hand, we observe that a constant angle between any two feature vectors of two graphs with complementing properties can only rarely be the case, since only a marginal change in a graph's features can change its property. If a graph can be edited slightly to obtain a property, it can, however, be viewed as close enough to the property to be ignored. Thus, in the sense of property testing, it is desirable to differentiate between the graph set far away from a property and the property itself, which motivates the following concept. We say that a graph kernel *distinguishes* a property if it guarantees a constant angle (independent of the graph size) between the feature vectors of any two graphs, one of which has the property and the other is far away from doing so. We study well-known graph kernels and their ability to identify and distinguish fundamental properties such as connectivity.

The significance of our framework is demonstrated by addressing several current research questions. In the graph kernels literature it has been argued that many kernels take either local or global graph properties into account, but not both [Kondor and Pan, 2016; Morris *et al.*, 2017]. Recent property testing results, however, suggest that under mild assumptions local graph features are sufficient to derive global properties [Newman and Sohler, 2013]. We consider a graph kernel based on local k-discs which can, in contrast to previous kernels, distinguish global properties such as planarity in bounded-degree graphs. For a constant dimensional feature space, we obtain learning guarantees for kernels that distinguish the class label property.

1.1 Related Work

We summarize related work on graph kernels, graph isomorphism, and property testing.

Gärtner *et al.* [2003] and Kashima *et al.* [2003] simultaneously proposed graph kernels based on random walks, which count the number of walks two graphs have in common. Since then, random walk kernels have been studied intensively, see, e.g., [Sugiyama and Borgwardt, 2015; Vishwanathan *et al.*, 2010; Kriege *et al.*, 2014]. Borgwardt and Kriegel [2005] have introduced kernels based on shortest paths; Costa and De Grave [2010] based on neighborhood subgraphs. Recently, graph kernels using matchings [Kriege *et al.*, 2016] and geometric embeddings [Johansson and Dubhashi, 2015] have been proposed. Furthermore, spectral approaches were explored [Kondor and Pan, 2016]. A different line in the development of graph kernels focused on scalable graph kernels, see, e.g., [Shervashidze *et al.*, 2011; Morris *et al.*, 2016; Hido and Kashima, 2009].

There are few works which investigate graph kernels from a theoretical viewpoint. Gärtner *et al.* [2003] introduced the concept of a *complete* graph kernel as a graph kernel with an injective feature map. The concept of completeness is too strict for the comparison of graph kernels and none of the numerous graph kernels proposed for practical applications actually is complete. Two measures of expressivity of kernels from statistical learning theory where proposed and applied to graph kernels [Oneto *et al.*, 2017] . However, these measures are not specific to graph structured data and cannot be interpreted in terms of distinguishable graph properties. The ability of the Weisfeiler–Lehman test to recognize non-isomorphic graphs has been studied extensively and the class of identifiable graphs characterized recently [Kiefer *et al.*, 2015; Arvind *et al.*, 2015].

Goldreich *et al.* [1998] formally established the study of property testing, where a central aim is to decide with high probability in sublinear time whether a property is satisfied or whether it is far from doing so. Goldreich and Ron [2002] initiated a growing line of research of property testers in the bounded degree graph model. For a recent overview, see, e.g., the textbook by Goldreich [2017].

1.2 Our Contribution

We propose a theoretical framework for comparing the expressiveness of kernels on bounded-degree graphs. Within this framework we obtain the following results:

- The shortest path kernel cannot guarantee a constant angle between connected and disconnected graphs of arbitrary size (see Proposition 3.2), but distinguishes connectivity in the considered framework (see Theorem 4.4).
- The random walk kernel and the Weisfeiler–Lehman subtree kernel both fail to identify connectivity, planarity, bipartiteness and triangle-freeness (see Theorems 4.1, 4.2).
- The graphlet kernel can identify triangle-freeness, but fails to distinguish any graph property (see Theorem 4.5).
- We define the k-disc kernel and show that it is able to distinguish connectivity, planarity, and triangle-freeness (see Theorem 5.4).
- We show that the prediction error of the 1-nearest neighbor classifier based on a kernel that distinguishes the class label property can be bounded (see Section 6).

2 Preliminaries

An *(undirected) graph* G is a pair (V, E) with a finite set of *vertices* V and a set of *edges* $E \subseteq \{\{u, v\} \subseteq V \mid u \neq v\}$. We denote the set of vertices and the set of edges of G by $V(G)$ and $E(G)$, respectively. A *walk* in a graph G is a sequence of vertices such that for each pair of consecutive vertices there exists an edge in $E(G)$. A *path* is a walk that contains each vertex at most once; a *cycle* is a walk that ends in the starting vertex. Moreover, $N(v)$ denotes the *neighborhood* of v in $V(G)$, i.e., $N(v) = \{u \in V(G) \mid \{u, v\} \in E(G)\}$. The *$k$-disc* of a vertex v in $V(G)$ is the subgraph induced by all vertices u such that there exists a path of length at most k between u and v. We say that two graphs G and H are *isomorphic* if there exists an edge preserving bijection $\varphi : V(G) \to V(H)$, i.e., $\{u, v\}$ in $E(G)$ if and only if $(\varphi(u), \varphi(v))$ in $E(H)$. The equivalence classes of the isomorphism relation are called *isomorphism types*. We denote the set of graphs on n vertices by $\mathcal{G}_n$.

Let χ be a non-empty set and let $\kappa\colon \chi \times \chi \to \mathbb{R}$ be a function. Then, κ is a *kernel* on χ if there is a Hilbert space $\mathcal{H}_\kappa$ and a mapping $\phi\colon \chi \to \mathcal{H}_\kappa$ such that $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$ for all x and y in χ, where $\langle \cdot, \cdot \rangle$ denotes the inner product of $\mathcal{H}_\kappa$. We call ϕ a *feature map*, and $\mathcal{H}_\kappa$ a *feature space* of the kernel κ. Let $\hat{\kappa}$ be the cosine normalized version of a kernel κ and denote its normalized feature map by $\hat{\phi}$, i.e.,

$$\begin{aligned} \hat{\kappa}(x, y) &= \left\langle \hat{\phi}(x), \hat{\phi}(y) \right\rangle = \left\langle \frac{\phi(x)}{\|\phi(x)\|_2}, \frac{\phi(y)}{\|\phi(y)\|_2} \right\rangle \\ &= \frac{\kappa(x, y)}{\sqrt{\kappa(x, x) \cdot \kappa(y, y)}} \in [-1, 1]. \end{aligned} \tag{1}$$

The normalized kernel $\hat{\kappa}(x, y)$ is equal to the cosine of the angle between $\phi(x)$ and $\phi(y)$ in the feature space. Let $\mathcal{G}$ be the set of all graphs, then a kernel $\kappa\colon \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ is called *graph kernel*.

2.1 Definitions from Property Testing

In this paper we assume the *bounded-degree graph model*. A graph is of *d-bounded degree* if its maximum degree is at most d. In the following d is always independent of the number of vertices n.

Let G and H be two d-bounded degree graphs in $\mathcal{G}_n$. The *edit distance* $\Delta(G, H)$ between G and H is the minimum number of edge modifications, i.e., adding or deleting edges, that has to be performed on G in order to obtain an isomorphic copy of H. A *graph property* is a set $\mathcal{P}$ of graphs that is closed under isomorphism. We denote the set of graphs in $\mathcal{P}$ on n vertices by $\mathcal{P}_n$. Let $\mathcal{P}_n$ be a non-empty graph property. A d-bounded degree graph G with n vertices is *ε-far* from $\mathcal{P}_n$ in the *bounded degree model* if for all d-bounded degree graphs H in $\mathcal{P}_n$ we have $\Delta(G, H) > \varepsilon dn$ for $\varepsilon > 0$. Otherwise, it is *ϵ-close*.

In this paper we study the following graph properties. A graph $G = (V, E)$ is called *connected* if for every two vertices $u, v \in V(G)$ there exists a path from u to v. A graph G is *planar* if there exists an embedding of G in the plane such that no edges cross, it is *bipartite* if $V(G)$ can be partitioned into two sets V_1 and $V_2 \subset V(G)$ such that for each edge $\{u, v\}$, $u \in V_1$ and $v \in V_2$ or vice versa. A graph is *triangle-free* if it does not contain a cycle with three vertices.

2.2 Graph Kernels

In the following we review four popular graph kernels. First, we describe the *Weisfeiler–Lehman subtree kernel* which is based on the well-known *color refinement algorithm* for isomorphism testing [Babai and Kucera, 1979; Cai *et al.*, 1992], which can be described as follows: Let G and H be graphs, and let l be a label function $V(G) \cup V(H) \to \Sigma$, e.g., $l(v) = |N(v)|$ for v in $V(G) \cup V(H)$. In each iteration $i \geq 0$, the color refinement algorithm computes a new label function $l^i \colon V(G) \cup V(H) \to \Sigma$. In iteration 0 we set $l^0 = l$. Now in iteration $i > 0$, we set $l^i(v) = \mathsf{relabel}((l^{i-1}(v), \mathsf{sort}(\{\!\{l^{i-1}(u) \mid u \in N(v)\}\!\})))$, for v in $V(G) \cup V(H)$, where $\mathsf{sort}(S)$ returns a sorted sequence of the labels in the multiset S and $\mathsf{relabel}$ is a bijection that maps a sequence of labels to a new unique label in Σ, which has not been used in previous iterations. If G and H have an unequal number of vertices labeled σ in Σ, they are not isomorphic. The idea of the Weisfeiler–Lehman subtree kernel [Shervashidze *et al.*, 2011] is to compute the above algorithm for $h \geq 0$ iterations and after each iteration i compute a feature map $\phi^i(G)$ in $\mathbb{R}^{|\Sigma_i|}$ for each graph G, where $\Sigma_i \subseteq \Sigma$ denotes the image of l^i. Each component $\phi^i(G)_{\sigma_j}$ counts the number of occurrences of vertices labeled with σ_j in Σ_i. The overall feature map $\phi(G)$ is defined as the concatenation of the feature maps of all h iterations, i.e.,

$$\left(\phi^0(G)_{\sigma_1^1}, \ldots, \phi^0(G)_{\sigma^1_{|\Sigma_0|}}, \ldots, \phi^h(G)_{\sigma_1^h}, \ldots, \phi^h(G)_{\sigma^h_{|\Sigma_h|}}\right).$$

Then, the Weisfeiler–Lehman subtree kernel for h iterations is $\kappa^h_{\mathrm{WL}}(G, H) = \langle \phi(G), \phi(H) \rangle$.

Secondly, we describe the *shortest path kernel* [Borgwardt and Kriegel, 2005]. Let G be a graph with label function $l \colon V(G) \to \Sigma$ and let $d \colon V(G) \times V(G) \to \mathbb{N}$ denote the shortest path distance function. Then, the feature map ϕ of the shortest path kernel maps a graph to a feature vector, where each component is associated with a triple $(a, b, p) \in \Sigma \times \Sigma \times \mathbb{N}$ and counts the number of shortest paths in G with length p from a vertex with label a to a vertex with label b [Shervashidze *et al.*, 2011]. The shortest path kernel is then defined as $\kappa_{\mathrm{SP}}(G, H) = \langle \phi(G), \phi(H) \rangle$. In our case, ϕ simply maps a graph G to a vector that represents G's frequency of shortest path lengths, since we consider unlabeled, undirected graphs.

The *graphlet kernel* counts the induced subgraphs on k vertices, for $k \in \{3, 4, 5\}$ [Shervashidze *et al.*, 2009]. Note that these subgraphs can be disconnected. Let $\sigma_1, \ldots, \sigma_N$ denote the isomorphism types of graphs on k vertices. For a graph G the kernel computes $\phi(G) = (\phi(G)_{\sigma_1}, \ldots, \phi(G)_{\sigma_N})$, where the component $\phi(G)_{\sigma_i}$ counts the subgraphs of G of type σ_i. The kernel is computed by $\kappa^k_{\mathrm{GR}}(G, H) = \langle \phi(G), \phi(H) \rangle$ for two graphs G and H and graphlet size k.

Finally, the random walk kernel counts the number of common walks of two graphs. The kernel is defined via the *direct product graph* $G \times H$ of two graphs G and H as

$$\kappa^k_{\mathrm{RW}}(G, H) = \sum_{i,j}^{|V_\times|} \left[\sum_{l=0}^{k} \lambda_l A_\times^l\right]_{ij}, \tag{2}$$

with vertex set $V_\times$ and adjacency matrix $A_\times$ of $G \times H$, $k > 0$, $\lambda_0, \ldots, \lambda_k > 0$, and $A^0_\times = \mathbf{I}$. For $k = \infty$ and $\lambda_i = \gamma^i$, $i \in \mathbb{N}$, and γ sufficiently small such that (2) converges, the kernel can be computed by a closed form expression and is referred to as *geometric random walk kernel* [Gärtner *et al.*, 2003].

3 Distinguishable Graph Properties

Let $n \in \mathbb{N}$ be an arbitrary number of vertices. We say that a feature map ϕ can *identify* a graph $G \in \mathcal{G}_n$ (up to isomorphism) if for each other graph $H \in \mathcal{G}_n$ that is not isomorphic to G it holds that $\hat{\phi}(G) \neq \hat{\phi}(H)$.

Definition 3.1. Let $\mathcal{P}$ be a graph property. If a graph kernel $\kappa : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_{\geq 0}$ and for each $n \in \mathbb{N}$, every $G \in \mathcal{P}_n$ and $H \notin \mathcal{P}_n$ satisfy $\hat{\kappa}(G, H) < 1$, we say that $\mathcal{P}$ can be *identified* by κ.

In order for a graph kernel to be able to distinguish a graph property and to use this knowledge in a learning context, a desirable goal is to have a constant angle independent of n. In the strict sense this is, however, rarely the case. In fact, in the following instance a constant difference cannot be achieved.

Proposition 3.2. *For the shortest path kernel, it holds that for each constant c, $0 < c < 1$, there exist some $n \in \mathbb{N}$ and two graphs $G, H \in \mathcal{G}_n$ with G connected, and H not connected such that $\hat{\kappa}_{\mathrm{SP}}(G, H) > 1 - c$.*

Proof. Let, for each $n \in \mathbb{N}$, G be a path with n vertices, and let H consist of a path with $n - 1$ vertices and one isolated vertex. Note that H is not connected, whereas adding one edge to H is enough to transform it into a connected graph which is isomorphic to G. The feature vectors for G and H counting the number of vertex pairs with distances 1 to $n - 1$ are $\phi = (n - 1, n - 2, \ldots, 1) \in \mathbb{R}^{n-1}$ and $\psi = (n - 2, n - 3, \ldots, 1, 0) \in \mathbb{R}^{n-1}$, respectively. Additionally, in H there are $n - 1$ vertex pairs that are not connected. It can be computed that $\langle \hat{\phi}, \hat{\psi} \rangle^2 = 1 - \frac{3}{4n^2 - 8n + 3}$. Assume there is a constant c, $0 < c < 1$, such that, for each $n \in \mathbb{N}$, it holds that $\langle \hat{\phi}, \hat{\psi} \rangle \leq 1 - c$, then there would be a constant $c' = (1 - c)^2$, $0 < c' < 1$ such that $c' \geq 1 - \frac{3}{4n^2 - 8n + 3}$ which does not hold for a large choice of n. Thus, for each constant c there exists an $n \in \mathbb{N}$ such that, for the graphs G and H as chosen above, $\hat{\kappa}(G, H) > 1 - c$ holds. □

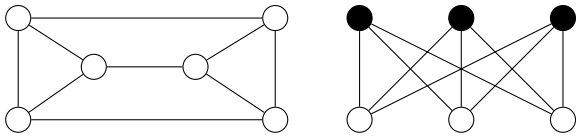
Note that both graphs in the proof of Proposition 3.2 have a maximum degree of 2. Therefore, the statements holds if any degree bound $d \geq 2$ is required.

In order to be able to achieve an angle independent of the graph size, we suggest to employ the notion of a graph being ε-far from a property as used in property testing. We aim to obtain a constant[1] angle between the feature vectors of two graphs whenever one graph has a certain property and the other is ε-far from having that property. In this context we define *distinguishability* of a graph property by a graph kernel as follows. Note that distinguishablity implies identifiability.

Definition 3.3. In the *bounded-degree graph model*, a graph property $\mathcal{P}$ is called *distinguishable* by a graph kernel $\kappa : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_{\geq 0}$, if for every $\varepsilon > 0$, $d \in \mathbb{N}$, there exists some $\delta = \delta(\varepsilon, d) > 0$ such that for every $n \in \mathbb{N}$, every $G \in \mathcal{P}_n$, and every graph H that is ε-far from $\mathcal{P}_n$, we have $\hat{\kappa}(G, H) \leq 1 - \delta$.

Note that this notion does not guarantee an accurate learning algorithm in general. Consider the isomorphism kernel that is 1 for two isomorphic graphs and 0 otherwise. It distinguishes every property, but as a classifier will not generalize to unseen data. Nevertheless, we obtain some learning guarantees, see Section 6.

[1] By constant we refer to a value independent of the input size n, which, however, can depend on ε or d.



(a) Graph G (b) Graph $H = K_{3,3}$

Figure 1: Counterexample for the proof of Theorems 4.1, 4.2, and 4.4

4 Properties Distinguishable by Popular Graph Kernels

In this section we study the identifiability and distinguishability of the random walk, the Weisfeiler–Lehman subtree, the shortest path, and the graphlet kernel. Table 1 sums up these results in comparison to the k-disc kernel studied in Section 5.

Both, the feature maps of a random walk kernel and the Weisfeiler–Lehman subtree kernel cannot identify a regular graph. In fact, each regular graph in $\mathcal{G}_n$ for some $n \in \mathbb{N}$, maps to the same feature vector. In particular, for the random walk kernel, the number of walks of length ℓ starting in a vertex of a regular graph with degree d is d^ℓ. Hence, for two regular graphs with degrees d and d', respectively, it holds that $\kappa^k_{\mathrm{RW}}(G, H) = |V_\times| \sum_{\ell=0}^{k} \lambda_\ell (d \cdot d')^\ell$, independently from the adjacency matrix of the product graph. For the Weisfeiler–Lehman subtree kernel, two regular graphs with the same degree obtain the same feature vector due to [Arvind *et al.*, 2015]. Therefore, as soon as for some graph property $\mathcal{P}$ and $n \in \mathbb{N}$ there exists one regular graph in $\mathcal{P}_n$ and another regular graph in $\mathcal{G}_n \setminus \mathcal{P}_n$, both kernels cannot identify and, thus, not distinguish the graph property.

Theorem 4.1. *The random walk kernel cannot identify connectivity, planarity, bipartiteness, or triangle freeness.*

Proof. A cycle with six vertices and two triangles with three vertices, both regular graphs, are a counterexample to the distinguishability of connectivity. Furthermore, consider the graphs G and H as illustrated in Figure 1. Note that G is planar, but not bipartite, and contains triangles, whereas H is not planar, but bipartite, and triangle-free. □

By the same arguments we obtain the following.

Theorem 4.2. *The Weisfeiler–Lehman subtree kernel cannot identify connectivity, planarity, bipartiteness, or triangle freeness.*

Next, we attend to a positive result regarding connectivity and the shortest path kernel. We will make use of the following technical lemma throughout proofs in this paper.

Lemma 4.3. *Let $n, r \in \mathbb{N}$, $x \in \mathbb{R}^r_{\geq 0}$, and $\varepsilon > 0$. For a non-empty subset of indices $S \subseteq \{1, \ldots, r\}$ such that $\sum_{i \in S} |x_i| = \eta > 0$ the following holds for every $y \in \mathbb{R}^r_{\geq 0}$ with $y_i = 0$ for each $i \in S$: $\langle x, y \rangle / \|x\|_2 \|y\|_2 \leq \sqrt{1 - \eta^2 / |S| \cdot \|x\|_2^2}$.*

Proof. Without loss of generality, let $S = \{1, \ldots, s\}$. For each $y \in \mathbb{R}^r_{\geq 0}$ with $y_i = 0$, $1 \leq i \leq s$, it holds that

$$\frac{\langle x, y \rangle}{\|x\|_2 \, \|y\|_2} = \left\langle \left(\frac{x_{s+1}}{\|x\|_2}, \ldots, \frac{x_r}{\|x\|_2} \right), \left(\frac{y_{s+1}}{\|y\|_2}, \ldots, \frac{y_r}{\|y\|_2} \right) \right\rangle,$$

which, by the Cauchy–Schwarz inequality is at most

$$\frac{\sqrt{\sum_{i=s+1}^{r} x_i^2}}{\|x\|_2} \cdot \frac{\|(y_{s+1}, \ldots, y_r)\|_2}{\|y\|_2} = \sqrt{1 - \frac{\sum_{i=1}^{s} x_i^2}{\|x\|_2^2}}.$$

Moreover, since $\sum_{i=1}^{s} x_i^2 \geq \frac{1}{s} \cdot \left(\sum_{i=1}^{s} x_i\right)^2 = \frac{1}{s}\eta^2$, this is at most $\sqrt{1 - \eta^2 / s\|x\|_2^2}$. □

Theorem 4.4. *The shortest path kernel*

1. *cannot identify planarity, bipartiteness, or triangle freeness;*
2. *can distinguish connectivity.*

Proof. 1. While in general two regular graphs may have different feature vectors, the graphs in Figure 1 also serve as a counterexample here. In both cases the shortest path feature vector are equal, as there are nine shortest paths of length 1 and six of length 2, each.

2. Let $n, d \in \mathbb{N}$, $\varepsilon > 0$, and $H \in \mathcal{G}_n$ be ε-far from being connected. For the shortest path feature vector $\psi = (\psi_1, \ldots, \psi_{n-1})$ and each connected graph $G \in \mathcal{G}_n$ with shortest path feature vector $\phi = (\phi_1, \ldots, \phi_{n-1})$, it holds that $\langle (\psi_1, \ldots, \psi_{n-1}), (\phi_1, \ldots, \phi_{n-1}) \rangle = \langle (\psi_1, \ldots, \psi_{n-1}, \eta), (\phi_1, \ldots, \phi_{n-1}, 0) \rangle$, where η denotes the number of disconnected vertex pairs in H. By Lemma 4.3 it holds that

$$\langle \hat{\phi}, \hat{\psi} \rangle \leq \sqrt{1 - \eta^2 / \|\psi\|_2^2}. \tag{3}$$

Assume that $n > 4/\varepsilon d$. Otherwise with $\eta \geq 1$ and $\|\psi\|_2 \leq n^2 \leq (4/\varepsilon d)^2$, it holds that (3) is at most 1 minus a constant. Now, it is known that there are more than $\varepsilon dn/2$ connected components of which at least $\varepsilon dn/4$ have a size smaller than $4/\varepsilon d$ [Goldreich and Ron, 2002]. At least one vertex in such a small component is disconnected from each vertex outside the component, that is, η is at least $1/2 \cdot \varepsilon dn/4 \cdot (n - 4/\varepsilon d) = \varepsilon n^2 d/8 - n/2$. Moreover, with $\langle \psi, \psi \rangle \leq \left(n(n-1)/2 - \eta\right)^2 + \eta^2$ we obtain that $1 - \eta^2 / \|\psi\|_2^2 \leq 1/2 + \zeta$, for some $\zeta > 0$ independent of n, which implies that (3) is smaller than 1 by a constant strictly between 0 and 1. □

Finally, we consider the graphlet kernel κ^k_{GR}. Although graphlet kernels appear to be rather expressive, from the considered properties they can only identify triangle-freeness. We can find counterexamples for connectivity, bipartiteness and planarity. For distinguishability there is one general obstacle, namely the fact that graphlets do not have to be connected. Each graph with n vertices and bounded degree d has at least $1 - d(k-1)k/2(n-1)$ k-graphlets with k independent vertices. Thus, for each constant $\delta > 0$ and each $G, H \in \mathcal{G}_n$ it holds that $\hat{\kappa}^k_{\mathrm{GR}}(G, H) > 1 - \delta$. Therefore, we obtain the following.

Theorem 4.5. *The graphlet kernel can identify triangle-freeness for $k \geq 3$, but unless the graphlet size k depends on the graph size, it cannot identify connectivity, bipartiteness, or planarity. Moreover, the graphlet kernel cannot distinguish any graph property.*

Property	Graph Kernel				
	WEISFEILER–LEHMAN	RANDOM WALK	SHORTEST PATH	GRAPHLET	k-DISC
Connectivity	✗	✗	✓	✗	✓
Planarity	✗	✗	✗	✗	✓
Bipartitness	✗	✗	✗	✗	
Triangle-freeness	✗	✗	✗	•	✓

Table 1: Distinguishability of graph properties for the random walk, shortest path, graphlet, k-disc and Weisfeiler–Lehman subtree kernel. Key: ✓ distinguishable, • identifiable (but not distinguishable), and ✗ not identifiable.

5 Graph Kernels that Distinguish Graph Properties

While we have observed that established graph kernels often cannot distinguish basic properties, we aim to find a graph kernel that can distinguish fundamental properties and is efficiently computable. Based on ideas by Costa and De Grave [2010] and Newman and Sohler [2013], we define a histogram $\mathrm{hist}_G(k)$ of the numbers of different k-discs, and the frequency vector $\mathrm{freq}_G(k) = \mathrm{hist}_G(k)/n$ of $G \in \mathcal{G}_n$. Consider the following graph kernel.

Definition 5.1. Given two graphs G and H, the *k-disc graph kernel* is defined by $\kappa_{\mathrm{KD}}(G,H) = \langle \mathrm{freq}_G(k), \mathrm{freq}_H(k)\rangle$.

A significant difference between the k-disc kernel and the graphlet kernel is that a k-disc is a connected subgraph, while a graphlet may be disconnected. For d-bounded-degree graphs, the k-disc kernel can be computed in time linear in the graph size. It can even be approximated in constant time, see, e.g., [Newman and Sohler, 2013].

Theorem 5.4 comprises the main results of this section about graph properties distinguishable by the k-disc kernel. From property testing studies, see, e.g., [Newman and Sohler, 2013], we often obtain information about the 1-norm of the distance between the frequency vectors of a graph ε-far from a property and all graphs satisfying the property. In order to translate these facts to a positive angle between the frequency vectors, we need the following two lemmas. Firstly, it can be seen that for two normalized real vectors with at least one index at which the entries differ by at least a constant positive value, their (standard) inner product is strictly less than 1.

Lemma 5.2. *Let x, y be two vectors in $\mathbb{R}^n_{\geq 0}$ for some $n \in \mathbb{N}$ with $\|x\|_2 = \|y\|_2 = 1$. Let $\zeta > 0$ be an arbitrarily small real value. If there exists some i, $1 \leq i \leq n$ such that $|x_i - y_i| \geq \zeta$, then $\langle x, y\rangle \leq 1 - \zeta^2/2$.*

Proof. It holds that $\langle x, y\rangle = 1/2(\|x\|_2^2 + \|y\|_2^2 - \|x-y\|_2)$

$$= 1 - \tfrac{1}{2}(\underbrace{(x_i - y_i)^2}_{\geq \zeta^2} + \textstyle\sum_{j=1, j\neq i}^{n} \underbrace{(x_j - y_j)^2}_{\geq 0}) \leq 1 - \tfrac{\zeta^2}{2}.$$

Thus, the angle between x and y is positive and constant. □

Secondly, we need to take care of the fact that the studied frequency vectors are normalized with respect to their 1-norm. However, since the number of different k-discs is independent of the number of vertices in the bounded-degree model, we can show the following lemma for two frequency vectors with a positive distance with respect to their 1-norm.

Lemma 5.3. *Let ϕ and ψ be two vectors in $\mathbb{R}^n_{\geq 0}$ with $\|\phi\|_1 = \|\psi\|_1 = 1$ and $\|\phi - \psi\|_1 \geq \eta$. Then, there exists an index i, $1 \leq i \leq n$, and a real value $\zeta > 0$ such that $|\phi_i/\|\phi\|_2 - \psi_i/\|\psi\|_2| \geq \zeta$.*

Proof. Let wlog $\|\phi\|_2 \geq \|\psi\|_2$. Moreover, let s denote the number of positive entries in both, ϕ and ψ. $\|\phi - \psi\|_1 \geq \eta$ implies the existence of a j: $\phi_j - \psi_j \geq \frac{\eta}{s}$. Case 1: If $\|\phi\|_2 - \|\psi\|_2 \leq \frac{\eta}{2s^{3/2}}$, then $\frac{\phi_j}{\|\phi\|_2} - \frac{\psi_j}{\|\psi\|_2}$ is at most

$$\geq \underbrace{\|\psi\|_2}_{\geq \frac{1}{\sqrt{s}}} \underbrace{(\phi_j - \psi - j)}_{\geq \frac{\eta}{s}} - \underbrace{\psi_j}_{\leq 1-\eta} \underbrace{(\|\phi\|_2 - \|\psi\|_2)}_{\leq \frac{\eta}{2s^{3/2}}} \geq \frac{\eta}{2s^{3/2}}.$$

Case 2: $\|\phi\|_2 - \|\psi\|_2 > \frac{\eta}{2s^{3/2}}$. Let R denote the subset of indices i such that $\phi_i/\|\phi\|_2 > \psi_i/\|\psi\|_2$. Observe, that $0 \leq |R| < s$. We split the sum $\|\hat{\phi} - \hat{\psi}\|_1 = \sum_{i=1}^n |\phi_i/\|\phi\|_2 - \psi_i/\|\psi\|_2|$ up into indices $i \in R$ and $i \notin R$. Since $\|\psi\|_1 = \|\phi\|_1 = 1$, this is equal to

$$\frac{1}{\|\psi\|_2} - \frac{1}{\|\phi\|_2} + 2\sum_{i\in R} \underbrace{(\phi_i/\|\phi\|_2 - \psi_i/\|\psi\|_2)}_{\geq 0} > \frac{\eta}{2s^{3/2}}.$$

Thus, there exists an index i, $1 \leq i \leq n$, such that $|\phi_i/\|\phi\|_2 - \psi_i/\|\psi\|_2| > \eta/2s^{5/2}$. With $0 < \zeta \leq \eta/2s^{5/2}$ in both cases, this completes the proof. □

Finally, we can proof the main theorem of this section.

Theorem 5.4. *For the k-disc graph kernel, it holds that*

1. *connectivity is distinguishable for $k \geq 4/\varepsilon d$,*
2. *triangle-freeness for $k \geq 1$, and*
3. *for each $\varepsilon > 0$, $d \in \mathbb{N}$ there exists some $k \in \mathbb{N}_+$ such that distinguishability is satisfied for planarity.*

Proof. 1. Let $H \in \mathcal{G}_n$ be a graph with bounded degree d that is ε-far from being connected for some $\varepsilon > 0$. By [Goldreich and Ron, 2002] we know that the number of connected components of a size smaller than $4/\varepsilon d$ is at least $\varepsilon dn/4$. For each vertex in such a small component, the full component is found as a k-disc of size $4/\varepsilon d$ in H. For each connected graph G, the frequency of such a small component is 0, since each k-disc covers at least $4/\varepsilon d$ vertices. Therefore, $\|\mathrm{freq}_G(k) - \mathrm{freq}_H(k)\|_1 \geq \varepsilon d/4$. The conditions of Lemma 4.3 are satisfied for $x = \mathrm{freq}_H(k)$ and $y = \mathrm{freq}_G(k)$ with S indicating the occurrence of small components. Again, observe that $|S|$ is independent of n. By $\|x\|_2^2 \leq 1$ and $\eta \geq \frac{\varepsilon d}{4}$ we obtain $\langle x,y\rangle/\|x\|_2\|y\|_2 \leq \sqrt{1 - \varepsilon d/4|S|}$, which is a positive constant strictly less than 1.

2. For triangle-freeness, again, we can use similar arguments to [Goldreich and Ron, 2002]. If a graph is ε-far from being triangle-free, there are εdn superfluous edges in H in contrast to any triangle-free graph G. Note that only edges that are part of a triangle are to be removed. Each such edge is shared by two vertices, and there can be at most d edges involved per vertex. That means, that at least $2\varepsilon n$ vertices are incident to a superfluous edge. These vertices hence have k-discs, for each $k \geq 1$, that contain triangles, whereas in G there are no such k-discs. Therefore, $\|\mathrm{freq}_G(k) - \mathrm{freq}_H(k)\|_1 \geq 2\varepsilon$. Since the frequency vectors are always normalised with respect to their 1-norm, the conditions of Lemma 5.3 hold. Thus, there exists an index i, $1 \leq i \leq n$, and a constant $\zeta > 0$ such that $\left| \frac{\mathrm{freq}_G(k)_i}{\|\mathrm{freq}_G(k)\|_2} - \frac{\mathrm{freq}_G(k)_i}{\|\mathrm{freq}_H(k)\|_2} \right| \geq \zeta$. Then, by Lemma 5.2, $\left\langle \frac{\mathrm{freq}_G(k)_i}{\|\mathrm{freq}_G(k)\|_2}, \frac{\mathrm{freq}_G(k)_i}{\|\mathrm{freq}_H(k)\|_2} \right\rangle$ is smaller than 1 by a constant.
3. Benjamini *et al.* [2010] show that for each $\varepsilon > 0$ and degree bound d, there exists a positive integer k independent of n such that for any two graphs $G, H \in \mathcal{G}_n$ with bounded degree d, G planar, H ε-far from being planar, it holds that $\|\mathrm{freq}_G(k) - \mathrm{freq}_H(k)\| \geq 1/k$. Therefore, via Lemmas 5.3 and 5.2, we obtain the claimed result. □

6 A Learning Algorithm

In this section we study a kernel nearest neighbor classifier for graphs and show that its prediction error can be bounded under the assumption that the employed kernel can distinguish the class label property and that all considered graphs either satisfy the property or are ϵ-far from it. We assume the following supervised binary classification problem: Let $\mathcal{Y} = \{0, 1\}$ be the set of possible *class labels*, which represent if a graph has a property or not. We aim to learn a *concept* $c\colon \mathcal{G}_n \to \mathcal{Y}$ such that the *0-1 loss* is minimized. Thereto we receive a *training set* $\{g_1, \ldots, g_m\} \subset \mathcal{G}_n$ and a *test graph* from $\mathcal{G}_n$ sampled i.i.d. according to some unknown distribution, as well as the set of class labels $\{c(g_1), \ldots, c(g_n)\}$ for the training set based on the concept to be learned. We assume in the following that for the considered graph property, G_n only contains graphs that either have the property or are ϵ-far from it.

6.1 Kernel Nearest Neighbor Classification

Based on a training set T of data points in $\mathbb{R}^D$ with known class labels, the k-nearest neighbor classifier (k-NN) assigns a test data point to the class most common among its k nearest neighbors in T. Here, the nearest neighbors are commonly determined based on the Euclidean distance between data points. Kernel nearest neighbor classifiers have been realized by substituting this distance by a kernel metric in Hilbert space, see, e.g., [Yu *et al.*, 2002]. For a kernel κ with feature map ϕ, we consider the 1-NN algorithm using the kernel metric $d_\kappa(x, y) = \|\phi(x) - \phi(y)\|_2 = \sqrt{\kappa(x, x) + \kappa(y, y) - 2\kappa(x, y)}$.

6.2 Learning With Distinguishing Kernels

We again consider the cosine normalized version $\hat{\kappa}$ of a kernel κ and its normalized feature map $\hat{\phi}$. For dimension D of the feature space, the normalized feature map $\hat{\phi}$ assigns graphs to points on the unit sphere S^{D-1}. Let us assume that $\hat{\kappa}$ distinguishes the class label property. Then, there is a δ, such that for all graphs G that have the property and H that are ε-far from it, we have $\hat{\kappa}(G, H) \leq 1 - \delta$ and, consequently, $d_{\hat{\kappa}}(G, H) \geq \sqrt{1 + 1 - 2(1 - \delta)} = \sqrt{2\delta}$. We denote this guaranteed minimum distance by $\Delta = \sqrt{2\delta}$. Consider the spherical cap C within the open ball centered at $\hat{\phi}(G)$ with radius Δ. According to the assumption, every graph H with $\hat{\phi}(H)$ lying on C, must have the same class label.

Proposition 6.1. *Let G be a graph of the training set. Then every graph H with $d_{\hat{\kappa}}(G, H) < \Delta$ is correctly classified by 1-NN where the base set contains all graphs that either have the property or are ε-far from it.*

Proof. Assume H is not correctly classified and $d_{\hat{\kappa}}(G, H) < \Delta$. Due to distinguishability, H must belong to the same class as G. Since H is not correctly classified by 1-NN, there must be a nearest neighbor $N \neq G$ of H with a different class label. Since N is a nearest neighbor of G, we have $d_{\hat{\kappa}}(H, N) \leq d_{\hat{\kappa}}(G, H) < \Delta$, contradicting distinguishability. □

We say that an algorithm (ε, λ)-*learns* a property if a test graph G drawn from the underlying distribution is correctly classified with probability $(1 - \lambda)$ and the base set consists of all graphs that either have the property or are ϵ-far from it.

Theorem 6.2. *Let κ be a kernel that distinguishes the class label property according to Definition 3.3 with some fixed δ and a feature space of dimension D. Let $\Delta = \sqrt{2\delta}$. Let all graphs either satisfy the property or be ϵ-far from it. Assume that the* training set *has cardinality $m \geq {}^{(1+6/\Delta)^D}/_{\lambda} \ln({}^{(1+6/\Delta)^D}/_{\lambda})$ then the 1-NN algorithm (ε, λ)-learns the property.*

Proof (Sketch). We cover the unit sphere with balls of radius $\Delta/3$. It is well-known that such a cover of size $B = (1 + {}^6/_{\Delta})^D$ exists. We observe that if a training example falls into a ball of the cover then by Proposition 6.1 any other example inside this ball is correctly classified. We observe that with probability $1 - \lambda$ every ball with probability mass at least ${}^{\lambda}/_{B}$ contains a training example. The overall probability of the remaining balls is at most λ. Therefore, with probability $1 - \lambda$ the algorithm (ϵ, λ)-learns the property. □

7 Conclusion

The introduced framework provides a starting point, e.g., for investigating other properties and evaluating other graph kernels. We assume promising kernels to be based, e.g., on other property testers, on spectral information, or on the 3-dimensional Weisfeiler–Lehman test. So far we have considered unlabeled, undirected graphs. Since most popular graph kernels are designed for labeled graphs, e.g., nodes are annotated with chemical symbols, future work might consider these as well.

Acknowledgements

This research was supported by the German Science Foundation (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Data Analysis", project A6 "Resource-efficient Graph Mining".

References

[Arvind *et al.*, 2015] V. Arvind, Johannes Köbler, Gaurav Rattan, and Oleg Verbitsky. On the power of color refinement. In *Proceedings of the 20th International Symposium on Fundamentals of Computation Theory*, pages 339–350, 2015.

[Babai and Kucera, 1979] Laszlo Babai and Ludik Kucera. Canonical labelling of graphs in linear average time. In *Proceedings of the 20th Annual Symposium on Foundations of Computer Science*, pages 39–46, 1979.

[Benjamini *et al.*, 2010] Itai Benjamini, Oded Schramm, and Asaf Shapira. Every minor-closed property of sparse graphs is testable. *Advances in Mathematics*, 223(6):2200–2218, 2010.

[Borgwardt and Kriegel, 2005] Karsten M. Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Proceedings of the 5th IEEE International Conference on Data Mining*, pages 74–81, 2005.

[Cai *et al.*, 1992] Jin-yi Cai, Martin Fürer, and Neil Immerman. An optimal lower bound on the number of variables for graph identification. *Combinatorica*, 12(4):389–410, 1992.

[Costa and De Grave, 2010] Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 26th International Conference on Machine Learning*, pages 255–262, 2010.

[Gärtner *et al.*, 2003] Thomas Gärtner, Peter Flach, and Stefan Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Proceedings of 16th Annual Conference on Learning Theory and Kernel Machines*, pages 129–143, 2003.

[Goldreich and Ron, 2002] Oded Goldreich and Dana Ron. Property testing in bounded degree graphs. *Algorithmica*, 32:302–343, 2002.

[Goldreich *et al.*, 1998] Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

[Goldreich, 2017] Oded Goldreich. *Introduction to Property Testing.* Cambridge University Press, 2017.

[Haussler, 1999] David Haussler. Convolution kernels on discrete structures. Technical Report UCS-CRL-99-10, University of California at Santa Cruz, 1999.

[Hido and Kashima, 2009] Shohei Hido and Hisashi Kashima. A linear-time graph kernel. In *Proceedings of the 9th IEEE International Conference on Data Mining*, pages 179–188, 2009.

[Johansson and Dubhashi, 2015] Fredrik Johansson and Dubhashi Dubhashi. Learning with similarity functions on graphs using matchings of geometric embeddings. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 467–476, 2015.

[Kashima *et al.*, 2003] Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the 20th International Conference on Machine Learning*, pages 321–328, 2003.

[Kersting *et al.*, 2016] 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`.

[Kiefer *et al.*, 2015] Sandra Kiefer, Pascal Schweitzer, and Erkal Selman. Graphs identified by logics with counting. In *Proceedings of the 40th International Symposium on Mathematical Foundations of Computer Science*, pages 319–330, 2015.

[Kondor and Pan, 2016] Risi Kondor and Horace Pan. The multiscale laplacian graph kernel. In *Advances in Neural Information Processing Systems*, pages 2982–2990, 2016.

[Kriege *et al.*, 2014] Nils Kriege, Marion Neumann, Kristian Kersting, and Petra Mutzel. Explicit versus implicit graph feature maps: A computational phase transition for walk kernels. In *Proceedings of the 14th IEEE International Conference on Data Mining*, pages 881–886, 2014.

[Kriege *et al.*, 2016] Nils M. Kriege, Giscard. Pierre-Louis, and Richard C. Wilson. On valid optimal assignment kernels and applications to graph classification. In *Advances in Neural Information Processing Systems*, pages 1615–1623, 2016.

[Morris *et al.*, 2016] Christopher Morris, Nils M. Kriege, Kristian Kersting, and Petra Mutzel. Faster kernel for graphs with continuous attributes via hashing. In *Proceedings of the 16th IEEE Int Conference on Data Mining*, pages 1095–1100, 2016.

[Morris *et al.*, 2017] Christopher Morris, Kristian Kersting, and Petra Mutzel. Glocalized Weisfeiler-Lehman graph kernels: Global-local feature maps of graphs. In *Proceedings of 17th IEEE Int Conference on Data Mining*, pages 327–336, 2017.

[Newman and Sohler, 2013] Ilan Newman and Christian Sohler. Every property of hyperfinite graphs is testable. *SIAM Journal on Computing*, 42(3):1095–1112, 2013.

[Oneto *et al.*, 2017] Luca Oneto, Nicol Navarin, Michele Donini, Alessandro Sperduti, Fabio Aiolli, and Davide Anguita. Measuring the expressivity of graph kernels through statistical learning theory. *Neurocomputing*, 2017.

[Ramon and Gärtner, 2003] Jan Ramon and Thomas Gärtner. Expressivity versus efficiency of graph kernels. In *Proceedings of the 1st International Workshop on Mining Graphs, Trees and Sequences*, 2003.

[Shervashidze *et al.*, 2009] Nino Shervashidze, S. V. N. Vishwanathan, Tobias H. Petri, Kurt Mehlhorn, and Karsten M. Borgwardt. Efficient graphlet kernels for large graph comparison. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 488–495, 2009.

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

[Sugiyama and Borgwardt, 2015] Mahito Sugiyama and Karsten M. Borgwardt. Halting in random walk kernels. In *Advances in Neural Information Processing Systems*, pages 1639–1647, 2015.

[Vishwanathan *et al.*, 2010] S. V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

[Yu *et al.*, 2002] Kai Yu, Liang Ji, and Xuegong Zhang. Kernel nearest-neighbor algorithm. *Neural Processing Letters*, 15(2):147–156, 2002.