Global graph kernels using geometric embeddings

Fredrik D. Johansson FREJOHK@CHALMERS.SE
Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

Vinay Jethava JETHAVA@CHALMERS.SE
Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

Devdatt Dubhashi DUBHASHI@CHALMERS.SE
Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

Chiranjib Bhattacharyya CHIRU@CSA.IISC.ERNET.IN
Indian Institute of Science, Bangalore 560012 Karnataka, India

Abstract

Applications of machine learning methods increasingly deal with graph structured data through kernels. Most existing graph kernels compare graphs in terms of features defined on small subgraphs such as walks, paths or graphlets, adopting an inherently local perspective. However, several interesting properties such as girth or chromatic number are global properties of the graph, and are not captured in local substructures. This paper presents two graph kernels defined on unlabeled graphs which capture global properties of graphs using the celebrated Lovász number and its associated orthonormal representation. We make progress towards theoretical results aiding kernel choice, proving a result about the separation margin of our kernel for classes of graphs. We give empirical results on classification of synthesized graphs with important global properties as well as established benchmark graph datasets, showing that the accuracy of our kernels is better than or competitive to existing graph kernels.

1. Introduction

Graph kernels (Gärtner et al., 2003; Vishwanathan et al., 2010) have been used in diverse fields including Computational biology (Schölkopf et al., 2004), Chemistry (Mahé & Vert, 2009), Information retrieval (Hermansson et al., 2013) etc. Design of graph kernels has primarily been motivated from capturing similar structural properties of graphs (Borgwardt & Kriegel, 2005). Searching for structural similarities in a pair of graphs are often computationally expensive. This has led to an interesting line of research (Feragen et al., 2013; Shervashidze et al., 2009; 2011) which explores graph kernels with lower computational complexity.

From a conceptual standpoint, most existing graph kernels compare features of small subgraphs extracted from the original graph. This leads to an inherently local perspective, which may fail to capture global properties of graphs. Further, as Shervashidze et al. (2009) identified, *"There is no theoretical justification on why certain types of subgraphs are better than others"*.

Moreover, it is known that there are graph properties which cannot be captured by studying only local structures, such as small subgraphs. Perhaps the most celebrated result on this topic is Erdős' seminal proof of existence of graphs with high *girth* (the length of the smallest cycle) and high chromatic number (Alon & Spencer, 1992, p. 41-42) – graphs for which all small-sized subgraphs will be trees. Another example is low density parity check (LDPC) codes (Richardson & Urbanke, 2008) which are constructed from particular low-density bipartite graphs. A key property governing the performance of the code is the girth of the graph (Bayati et al., 2009). Lastly, Devroye et al. (2011) describes a method for detecting dependencies among a set of random vectors by transforming the real valued data into graphs. Posed as a problem of hypothesis testing, the solution is given by the *clique number* of the graph.

This paper attempts to address the gap between existing

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

"local" kernels (i.e. kernels defined in terms of small size subgraphs of the original graph) and global graph properties.

Main contributions We present two novel graph kernels designed to capture global properties of unlabeled graphs. The kernels are based on the Lovász number (Lovász, 1979), famous for its central role in combinatorial optimization and graph theory (Goemans, 1997). The first kernel, which we call the Lovász ϑ kernel (Lo-ϑ), explicitly leverages the orthonormal representation of graphs associated with Lovász number, see Section 3.1. The second kernel, which we call the SVM-ϑ kernel (SVM-ϑ), builds on a recent alternative characterization of Lovász ϑ (Jethava et al., 2014), enabling faster computation with known error bounds for classes of graphs, see Section 4. We derive sampling approximations for efficient computation of both kernels, with results for the sample complexities.

We evaluate our kernels empirically using graph classification, see Section 5. We compare the accuracy of our kernels with that of state-of-the-art graph kernels on both synthesized and benchmark datasets. The synthetic datasets are designed to test the kernels' ability to capture global properties of the graphs. We show that on real-world benchmark datasets, we produce results better or competitive with existing graph kernels.

Lastly, we take initial steps towards theoretical justification of kernel choice proving a result bounding the separation margin for both kernels in the task of classifying random and planted clique graphs.

2. Graph kernels

This section reviews prior work on graph kernels. Most existing graph kernels are *R-convolution* kernels (Shervashidze et al., 2011; Vishwanathan et al., 2010). Let χ and χ' be spaces and $k : \chi' \times \chi' \to \mathbb{R}$ a positive semi-definite kernel. The R-convolution kernel for points $x, y \in \chi$, associated with finite subsets $\chi'_x \subseteq \chi'$ and $\chi'_y \subseteq \chi'$ is defined as (Haussler, 1999)

$$K(x,y) = \sum_{(x',y') \in \chi'_x \times \chi'_y} k(x',y') \ . \tag{1}$$

Graph kernels decompose graphs into particular sets of subgraphs and compare features of the subgraphs (Shervashidze et al., 2011). For example, the shortest-path kernel (abbreviated SP) (Borgwardt & Kriegel, 2005) compares features of the shortest paths between all pairs of nodes in two graphs. The random walk kernel (RW) (Gärtner et al., 2003) counts (weighted) numbers of walks of every length in graphs. The graphlet kernel (GL) (Shervashidze et al., 2009) estimates and compares graphlet spectra, distributions of subgraphs of sizes 3,4,5. Shervashidze et al. (2009) argue that the graphlet spectrum is similar to a sufficient statistic for a graph. Subtree kernels compare subtree patterns, matchings between neighbors of pairs of nodes (Ramon & Gärtner, 2003; Shervashidze & Borgwardt, 2009; Mahé & Vert, 2009). A graph kernel is thus characterised by a selection of subgraphs and a set of features of them. Commonly, graph kernels are partitioned into two groups, one concerned with *labeled* or *weighted* graphs, where nodes and/or edges are equipped with attributes, and the other *unlabeled* graphs. In this paper, we consider only *unweighted*, *unlabeled* graphs.

The graph kernels listed above are inherently local in their perspective, repeatedly comparing subgraphs disjoint from the rest of the graph. A kernel comparing *all* subgraphs would constitute a *complete* kernel k, such that $k(x,\cdot) = k(x',\cdot)$ implies $x = x'$ (Gärtner et al., 2003). While such a kernel is highly expressive, it can be shown that constructing a complete graph kernel is NP-hard (Gärtner et al., 2003). As a result, designing a kernel is a trade-off between expressivity and efficiency (Ramon & Gärtner, 2003).

It is hitherto unknown whether global properties of graphs, such as girth or chromatic number are captured, even approximately, by existing kernels.

3. Orthonormal labellings of graphs

In this section, we define the new Lovász ϑ kernel and review the concepts of orthonormal labellings and Lovász number.

Recall that an orthonormal representation, or labelling, of a graph $G = (V,E)$ consists of a set of unit vectors $U_G := \{\mathbf{u}_i \in \mathbb{R}^p : \|\mathbf{u}_i\| = 1\}_{i \in V}$ where each node $i \in V$ is assigned a unit vector $\mathbf{u}_i$ such that $(i,j) \notin E \implies \mathbf{u}_i^\top \mathbf{u}_j = 0$. We emphasize that the orthonormal representation U_G captures global graph properties since $\mathbf{u}_i$ satisfy a global set of constraints, encoding independences in the entire graph.

It is instructive to consider a subset of vertices $B \subseteq V$ and their corresponding representation $U_{G|B} \subseteq U_G := \{\mathbf{u}_i \in U_G : i \in B\}$. We note that $U_{G|B}$ not only captures the edges encoded in the subgraph of G induced by B, denoted $G[B]$, but is also consistent for the whole graph in that it satisfies orthogonality constraints for *the whole of* G. This would not have been the case had we first isolated the induced subgraph $G[B]$ and then taken its orthonormal representation. In general, $U_{G|B} \neq U_{G[B]}$. Thus, an orthonormal representation U_G and, more importantly, its subset $U_{G|B}$ capture properties of the graph that $G[B]$ does not. In the sequel, we focus on orthonormal representations that capture global properties accurately and concisely, and can be computed efficiently.

An interesting orthonormal representation is associated with the Lovász number (Lovász, 1979). Commonly denoted $\vartheta(G)$, Lovász number has had great impact on combinatorial optimization, graph theory and approximation algorithms, to the extent that Goemans remarked: *It seems all roads lead to ϑ!* (Goemans, 1997). $\vartheta(G)$ has been used to derive fast approximation algorithms for max k-cut (Frieze & Jerrum, 1997), graph coloring (Karger et al., 1998; Dukanovic & Rendl, 2008) and planted clique problems (Feige & Krauthgamer, 2000).

It is well known that $\vartheta(G)$ has strong connections to global properties such as the chromatic number $\chi(G)$ and the clique number $\omega(G)$. See Knuth (1993) for a comprehensive discussion of $\vartheta(G)$ and its characterizations. One definition of Lovász number is given below.

Definition 1. *(Lovász, 1979) For a graph $G = (V, E)$,*

$$\vartheta(G) = \min_{\mathbf{c}, U_G} \max_{i \in V} \frac{1}{(\mathbf{c}^\top \mathbf{u}_i)^2} , \tag{2}$$

where the minimization is taken over all orthonormal representations U_G and all unit vectors $\mathbf{c}$.

Geometrically, $\vartheta(G)$ is thus defined by the smallest cone enclosing a valid orthonormal representation U_G.

It is well-known that $\vartheta(G)$ can be computed to arbitrary precision in polynomial time, by means of solving a semi-definite program (Lovász, 1979), in contrast to $\omega(G)$ and $\chi(G)$, which are both NP-hard to compute.

3.1. The Lovász ϑ graph kernel

We proceed to define the first of our graph kernels, namely the Lovász ϑ kernel (abbreviated henceforth as $\texttt{Lo}-\vartheta$), which compares graphs based on the orthonormal representation U_G associated with $\vartheta(G)$. Henceforth, whenever referring to an orthonormal representation of G, unless otherwise stated, we refer to Lovász's representation defined by the maximizer of (2), and denote it $U_G = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$.

We begin by defining the notion of *Lovász value* of a subset of nodes $B \subseteq V$, which represents the angle of the smallest cone enclosing the set of vectors $U_{G|B}$. Formally,

Definition 2 (Lovász value)**.** *The* Lovász value *of $G[B]$, the subgraph of $G = (V, E)$ induced by $B \subseteq V$, is defined by,*

$$\vartheta_B(G) = \min_{\mathbf{c}} \max_{\mathbf{u}_i \in U_{G|B}} \frac{1}{(\mathbf{c}^\top \mathbf{u}_i)^2} ,$$

where $U_{G|B} := \{\mathbf{u}_i \in U_G \,|\, i \in B\}$ and U_G is the maximizer of (2)*. Note that in general $\vartheta_B(G) \neq \vartheta(G[B])$.*

We state a trivial, but important, result for $\vartheta_B(G)$.

Lemma 1. *Let $G = (V, E)$. Then, for any subset $B \subset V$, with $H = G[B]$ the subgraph of G induced by B,*

$$\vartheta(H) \leq \vartheta_B(G) \leq \vartheta_V(G) = \vartheta(G) .$$

The proof is left to the supplementary material (Johansson et al., 2014).

Our goal is to develop a graph kernel capturing global properties of graphs, guided by the intuition that features of subgraphs should be placed in context of the whole graph. To this end, we define a graph kernel on the Lovász value. As opposed to subgraph features used in existing graph kernels, the Lovász value encapsulates information from outside the subgraph by adhering to the global set of orthonormality constraints (specified by the edge set E). We note that using $\vartheta(H)$ as a feature of $H = G[B]$ does not fulfil this property, as $\vartheta(H)$ does not use information from outside H.

We now present the formal definition of Lovász ϑ kernel in terms of the Lovász values of subgraphs.

Definition 3 (Lovász ϑ kernel)**.** *The Lovász ϑ kernel on two graphs, G, G', with a positive semi-definite kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, is defined as*

$$K(G, G') = \sum_{B \subseteq V} \sum_{\substack{C \subseteq V' \\ |C| = |B|}} \frac{1}{Z_{|B|}} \cdot k(\vartheta_B, \vartheta'_C) , \tag{3}$$

with $\vartheta_B = \vartheta_B(G)$, $\vartheta'_C = \vartheta_C(G')$, and $Z_d = \binom{n}{d}\binom{n'}{d}$.

The kernel k is referred to as the *base kernel*. We state the following important property.

Lemma 2. *The Lovász ϑ kernel, as defined in* (3)*, is a positive semi-definite kernel.*

Proof sketch. The kernel in (3) is an R-convolution kernel (Haussler, 1999), see (1). For a complete proof, see the supplementary (Johansson et al., 2014).

As K in (3) is a p.s.d kernel, we can represent it as an inner product $\langle \varphi, \varphi' \rangle$ in a reproducing kernel Hilbert space. Choosing k to be the linear kernel $k(x, y) = xy$, φ can be written explicitely, with its d:th coordinate

$$\varphi(d) = \binom{n}{d}^{-1} \sum_{\substack{B \subseteq V \\ |B| = d}} \vartheta_B(G) . \tag{4}$$

In this case, $\varphi(d)$ represents the average minimum angle of cones enclosing subsets of orthonormal representations of cardinality d. We refer to φ as the feature vector of the kernel.

3.2. Computing the Lovász ϑ kernel

Direct evaluation of the Lovász ϑ kernel as defined in (3) involves two main steps, namely, obtaining the Lovász orthonormal labelling U_G for each graph G, in a set $\mathscr{G}$, by solving the optimization in (2) and subsequently, computing the Lovász value $\vartheta_B(G)$ for *all* subgraphs $B \subseteq G$ of each graph $G \in \mathscr{G}$. Exact computation of the Lovász ϑ kernel is often infeasible, as it involves 2^n computations of minimum enclosing cones using e.g. the algorithm in Welzl (1991). Next, we show that it is sufficient to sample a small number of subsets to get a good approximation of the kernel.

3.2.1. Sample Complexity Bound

We derive an efficient approximation of (3) by evaluating the Lovász value for a smaller number of subgraphs $\mathscr{S} \subset 2^V$ and $\mathscr{S}' \subset 2^{V'}$ respectively for all pairs of graphs G and G'. Let $\mathscr{S}_d$ denote the subset of $\mathscr{S}$ consisting of all sets of cardinality d in $\mathscr{S}$ i.e. $\mathscr{S}_d := \{B \in \mathscr{S} : |B| = d\}$. Then, we define

$$\hat{K}(G,G') = \sum_{B\in\mathscr{S}} \sum_{\substack{C\in\mathscr{S}' \\ |B|=|C|}} \frac{1}{\hat{Z}_{|B|}} \cdot k(\vartheta_B, \vartheta'_C) , \qquad (5)$$

where $\hat{K}$ denotes the approximate value for K in (3) and $\hat{Z}_d = |\mathscr{S}_d||\mathscr{S}'_d|$.

The time complexity of computing $\hat{K}(G,G')$ is, leaving out logarithmic factors and with $s = \max(|\mathscr{S}|, |\mathscr{S}'|)$, $\tilde{O}(n^2|E|\varepsilon^{-1} + s^2 \cdot T(k) + sn)$, where $T(k)$ is the time complexity of computing the base kernel $k(\cdot,\cdot)$ and the first two terms represent the cost of semi-definite program in (2), and, the worst-case complexity of computing the summation in (5) respectively. The last term represents the time complexity of computing the Lovász values. The sampling strategy and choice of base kernel $k(\cdot,\cdot)$ are critical in obtaining a good approximation. We discuss one such scheme below.

We choose the linear kernel $k(x,y) = xy$ as the base kernel in (5) with its explicit feature representation φ given by (4). Furthermore, we choose the sets $\mathscr{S}_d$ by sampling uniformly at random s_d subsets independently for each cardinality d and let $\mathscr{S} = \mathscr{S}_1 \cup \ldots \cup \mathscr{S}_n$ for each graph $G \in \mathscr{G}$.[1]

Let $\hat{\varphi}(d)$ denote the random variable given as

$$\hat{\varphi}(d) = \frac{1}{s_d} \sum_{B\in\mathscr{S}_d} \vartheta_B(G)$$

where $\mathscr{S}_d$ denotes a superset of s_d subsets of vertices $B^{(1)}, \ldots, B^{(s_d)} \subseteq V$ each of size d (i.e. $|B^{(i)}| = d$) chosen uniformly at random. We can then state the following result,

Theorem 1. *For graphs of n nodes, each coordinate $\varphi(d)$ of the feature vector of the linear Lovász ϑ kernel can be estimated by $\hat{\varphi}(d)$ such that*

$$Pr[\hat{\varphi}(d) \geq (1+\varepsilon)\varphi(d)] \leq O(1/n)$$
$$Pr[\hat{\varphi}(d) \leq (1-\varepsilon)\varphi(d)] \leq O(1/n)$$

using $s_d = O(n \log n/\varepsilon^2)$ samples.

Proof sketch. We apply a multiplicative Chernoff bound on ϑ_{V_r} of sampled subsets V_r. For a complete proof, see the supplementary material (Johansson et al., 2014).

This allows us to compute the approximate the linear Lovász ϑ kernel accurately using $\hat{K}(G,G') = \langle\hat{\varphi}, \hat{\varphi}'\rangle$ where $\hat{\varphi}$ is defined analogous to (4).

3.3. Signal subgraphs

Motivated by problems arising from the study of networks in the brain, Vogelstein et al. (2013) introduced a framework for graph classification based on *signal-subgraphs*, subgraphs that have common properties within a class of graphs, but differ between classes. Devroye et al. (2011) had earlier considered a hypothesis testing problem arising in applications such as remote sensing and argued that it could be modelled as a *planted clique* problem in a random geometric graph. This is a classical problem in the theory of random graphs and algorithms (Feige & Krauthgamer, 2000; Alon et al., 1998) with many applications such as cryptography (Juels & Peinado, 2000) and has connections to data mining problems such as epilepsy prediction (Iasemidis et al., 2001). In the classical planted clique problem, a hidden clique of $\Theta(\sqrt{n})$ vertices is planted into a random graph and the goal is for an algorithm to identify it. In a more general version, the planted subgraph could have significantly higher or lower density compared with the underlying random graph. Such planted models are natural special cases of the general framework of Vogelstein et al. (2013) . In their brain networks setting, a denser subgraph could correspond to a subset of neurons that have significantly higher (or lower) connectivity compared to the rest of the network. With this in mind, we consider the question of classifying planted subgraph models with different densities.

Classifying planted clique graphs We let $G(n,p)$ denote the random graph of n nodes, where every edge is present, randomly and independently, with probability p. Further, we let $G(n,p,k)$ denote the graph formed by sampling a random $G(n,p)$ graph and planting a clique of size k within.

We focus now on using the Lovász ϑ kernel for classification of $G(n,p)$ and $G(n,p,k)$ as two different classes. We give a result showing that the two classes of graphs are linearly separable with reasonably large margin in the feature space of the linear Lovász ϑ kernel.

[1] We note that repeatedly sampling pairs of subsets, one for each graph, is not guaranteed to give a positive semi-definite kernel.

Lemma 3. *There exist, with high probability, $Pr \geq 1 - O(1/n)$, a linear separator in linear Lovász ϑ kernel space, separating $G(n,p)$ and $G(n,1-p,k)$ graphs, $k = 2t\sqrt{\frac{n(1-p)}{p}}$, where $p(1-p) = \Omega(n^{-1}\log^4 n)$, with margin*

$$\gamma \geq (t-c)\sqrt{\frac{n(1-p)}{p}} - o(\sqrt{n}) \ ,$$

for some constant c, and large enough $t \geq 1$.

Proof. The proof is left to the supplementary material (Johansson et al., 2014).

These results indicate that the Lovász ϑ kernel is a good candidate kernel for problems related to the detection of large cliques. Similar results can be proved for the more general problem of classifying planted subgraphs with different densities using the results of Jethava et al. (2014, Section 4.3).

4. The SVM-ϑ kernel on graphs

In this section we define the new SVM-ϑ kernel and introduce the concept of SVM-ϑ (Jethava et al., 2014).

State-of-the-art algorithms for computing Lovász number have time complexities $O(n^5 \log n \cdot \varepsilon^{-2})$ (Chan et al., 2009) and $O(n^2 m \log n \cdot \varepsilon^{-1} \log^3(\varepsilon^{-1}))$ (Iyengar et al., 2011), where n and m are the number of nodes and edges respectively and ε the error. These methods are prohibitively slow for most applications.

An alternate characterization of $\vartheta(G)$ was given by (Jethava et al., 2014), who showed that for a graph $G = (V,E)$, such that $|V| = n$,

$$\vartheta(G) = \min_{\kappa \in L} \omega(\kappa)$$

with $\omega(\kappa)$ the kernel one-class SVM,

$$\omega(\kappa) = \max_{\substack{\alpha_i > 0 \\ i=1,\ldots,n}} 2\sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa_{ij} \qquad (8)$$

and S_n^+ the set of $n \times n$ positive semi-definite matrices,

$$L := \{\kappa \in S_n^+ | \kappa_{ii} = 1, \forall i, \kappa_{ij} = 0, (i,j) \notin E\} \ .$$

With slight abuse of notation, from now on, we let α_i denote the *maximizers* of (8). We give a particularly interesting choice of κ below.

Definition 4 (Luz & Schrijver (2005))**.** *Let A be the adjacency matrix of G, $\rho \geq -\lambda_n(A)$, with $\lambda_n(A)$ the minimum eigenvalue of A, and set*

$$\kappa_{LS}(G) = \frac{A}{\rho} + I \succeq 0$$

Let U be any matrix such that $\kappa_{LS}(G) = U^\top U$.

Jethava et al. (2014) showed that,

$$\omega(\kappa_{LS}(G)) = \sum_{i=1}^{n} \alpha_i$$

where α_i are the maximizers of (8). Further, (Jethava et al., 2014) proved that on families of graphs, referred to by them as SVM-ϑ graphs, $\omega(\kappa_{LS})$ is w.h.p. a constant factor approximation to $\vartheta(G)$,

$$\vartheta(G) \leq \omega(\kappa_{LS}) \leq \gamma\vartheta(G) \ .$$

Important graph families such as Erdős-Rényi random graphs and planted clique graphs have this property.

We proceed to define a new graph kernel called the SVM-ϑ kernel. Inspired by the results of Section 3.3, we seek an SVM-ϑ analogue of the Lovász value, ϑ_B to use as a feature of subgraphs. We note that α_i adheres to the global optimality conditions of (8) defined by the edge set, and thus captures global properties of graphs. Based on this observation, and the connection between $\omega(\kappa)$ and $\vartheta(G)$, we let $\sum_{i \in B} \alpha_i$ serve as an analogue for ϑ_B in (3), when defining our new kernel.

Definition 5. *The SVM-ϑ kernel is defined, on two graphs G, G', with corresponding α, α' maximizers of* (8) *for $\kappa = \kappa_{LS}(G)$, with a positive semi-definite kernel $k : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, as*

$$K(G,G') = \sum_{B \subseteq V} \sum_{\substack{C \subseteq V' \\ |C|=|B|}} \frac{1}{Z_{|B|}} k(\mathbf{1}^\top \alpha_B, \mathbf{1}^\top \alpha'_C) \qquad (9)$$

where $\alpha_B = [\alpha_{B(1)}, \ldots, \alpha_{B(d)}]^\top$ with $d = |B|$, $Z_d = \binom{n}{d}\binom{n'}{d}$ and $\mathbf{1}$ the all one vector of appropriate size.

Lemma 4. *The SVM-ϑ kernel, as defined in* (9)*, is a positive semi-definite kernel.*

Proof sketch. The kernel in (9) is an R-convolution kernel (Haussler, 1999), see (1). For a complete proof, see the supplementary (Johansson et al., 2014).

4.1. Computing the SVM-ϑ kernel

Computation of the SVM-ϑ kernel has the following phases. First, for each graph G, $\kappa_{LS}(G)$ is computed in $O(n^3)$ time (due to the eigenvalue computation). Then, the one-class SVM in (8) is solved, in $O(n^2)$ time (Hush et al., 2006), to obtain $\alpha_i(G)$. This offers a substantial speedup compared to the first step of computing the Lovász ϑ kernel, see Section 3.2. Finally, the kernel is computed as in (9). The dominating factor in the complexity is the summation over all subsets, so in a fashion analogous to Section 3.2, we approximate the SVM-ϑ kernel using sampling.

For the SVM-ϑ kernel with a linear base kernel $k(x,y) = xy$

and its explicit feature representation φ,

$$\varphi(d) = \binom{n}{d}^{-1} \sum_{\substack{B \subseteq V \\ |B|=d}} \sum_{j \in B} \alpha_j(G) ,$$

we can state the following result.

Theorem 2. *For graphs of n nodes, each coordinate $\varphi(d)$ of the feature vector of the linear* SVM-ϑ *kernel can be estimated by $\hat{\varphi}(d)$ such that*

$$Pr[\hat{\varphi}(d) \geq (1+\varepsilon)\varphi(d)] \leq O(1/n)$$
$$Pr[\hat{\varphi}(d) \leq (1-\varepsilon)\varphi(d)] \leq O(1/n)$$

using $s_d = O(n^2 \log n / \varepsilon^2)$ samples.

Proof. We leave the proof to the supplementary material (Johansson et al., 2014).

We observe in practice that a lower number of samples is sufficient for good performance in graph classification. The overall time complexity of the sampled SVM-ϑ kernel is $O(n^3 + s^2 T(k) + sn)$, where s is the number of sampled subgraphs per graph, and $T(k)$ is the time complexity of computing the base kernel k. The first and third term are due to eigenvalue computation and summation of α_i respectively.

4.2. Planted clique graphs

A result about the margin of the SVM-ϑ kernel in classification of planted clique graphs, similar Lemma 3 can be derived. We leave the result and proof to the supplementary material (Johansson et al., 2014).

5. Experiments

We evaluate the Lovász ϑ and SVM-ϑ kernels by performing graph classification of synthetic and benchmark datasets. We report the classification accuracy using 10-fold cross-validation with a C-Support Vector Machine classifier, LIBSVM (Chang & Lin, 2011). All experiments were repeated 10 times and the results averaged, to counter the effects of randomized folds. The SVM parameter C was optimized for each kernel and fold and the best was used for the final accuracy.

5.1. Experimental setup

We compare our kernels to state-of-the-art kernels for unlabeled, unweighted, selected so as to represent three major groups of graph kernels, based on walks, small subgraphs and paths respectively. The chosen walk kernel is the p-random walk kernel, denoted RW, which counts common random walks up to length p (a special case of Gärtner et al. (2003)). p was chosen from $\{1, 10, 100, 1000\}$ and λ with the heuristic of Shervashidze et al. (2011) as the largest power of 10 smaller than the inverse of the squared maximum degree.

Table 1. Average classification accuracy (%). Numbers in bold indicate the best results in each column. The kernels introduced in this paper are Lo−ϑ and SVM−ϑ. [†] Lo−ϑ was run on $M = 100$ graphs. [‡] Lo−ϑ did not finish within 2 days. LDPC$_g$ are synthetic graphs of girth $\geq g$ for $g = 4, 5, 6$ used for low density parity check codes.

KERNELS	ERDOS	LDPC$_4$	LDPC$_5$	LDPC$_6$
SP	61.8	**60.6**	74.2	**96.5**
GL	56.7	50.0	50.0	50.0
RW	58.8	50.0	50.0	54.1
Lo−ϑ	63.2[†]	‡	‡	‡
SVM−ϑ	**66.3**	**60.6**	**75.0**	95.5

For a kernel counting small subgraphs, we use the graphlet kernel, counting all subgraphs of size 4 (Shervashidze et al., 2009), denoted GL. For a kernel on paths, we use the (delta) shortest-path kernel which counts shortest paths of identical length (Borgwardt & Kriegel, 2005), denoted SP. For the SP and GL kernels, we use the publicly available Matlab implementations of Shervashidze & Borgwardt (2012). Note that while some of the kernels used for comparison have variants exploiting labels, these have not been included, as the focus of this paper is unlabeled graphs. We sample the Lovász ϑ kernel using $n \log n$ samples per coordinate, and the SVM-ϑ kernel using $n^2 \log n$ samples. Both kernels were used with either the linear kernel $k(x,y) = xy$ or the radial basis function kernel $k(x,y) = e^{-\|x-y\|_2^2/(2\sigma^2)}$ with σ from the set $[0.01, 10]$.

5.2. Synthesized graphs with global properties

We perform graph classification on a suite of synthesized datasets with known global properties.

Datasets An important family of graphs are graphs of high girth and high chromatic number (Alon & Spencer, 1992). These have the property that all small subgraphs are trees (Erdős, 1959). Such graphs can be constructed by sampling a random Erdős-Rényi $G(n, p)$ graph and repeatedly removing a node from each small cycle until the graph has the desired girth (Alon & Spencer, 1992). The resulting graphs are guaranteed to have at least $n' \geq n/2$ nodes. In this manner, we generate $M = 300$ graphs with $n = 100$ for each $p \in \{0.03, 0.01, \ldots, 0.25\}$, removing all cycles of length ≤ 3, and denote the resulting numbers of nodes and densities, after cycle deletion, by n'_m and p'_m, for $m = 1, \ldots, M$, labeled (+1). Then we generate a set of M random $G(n, p)$ graphs with p and n matching the distribution of p'_m and n'_m, labeled (-1). This dataset is denoted ERDOS. The average accuracy (over varying p) in classifi-

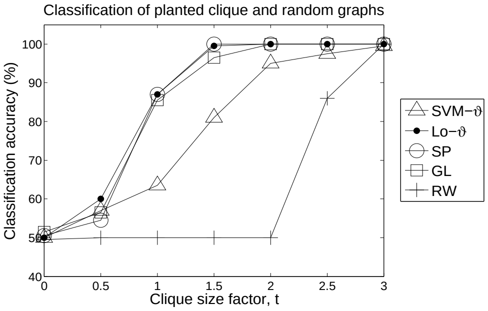

Figure 1. Classification accuracy on the PC dataset of 100 random $G(n,1/2)$ graphs (labeled -1) and $G(n,1/2,k)$ planted clique graphs (+1) with varying clique size. The horizontal axis is the factor t for the clique size $k = 2t\sqrt{n}$.

cation of the two sets of graphs, using a selection of graph kernels, is presented in Table 1.

In information theory, low-density-parity-check (LDPC) codes, are used as error correcting codes, and are constructed using a bipartite graph (Richardson & Urbanke, 2008). It is known that high girth of the graph is a factor contributing to the performance of the code (Bayati et al., 2009). To this end, we generate sets of sparse graphs with girth $g \geq g'$ for $g' \in \{4,5,6\}$. We use the construction of Bayati et al. (2009), which adds one edge at a time, without destroying the girth property. For this experiment, we generate $M = 100$ graphs, with $n = 200$ and 200 edges, labeled (+1), and M random graphs with n nodes and m edges, where a new edge is added with uniform probability until m has been added (-1). These sets are denoted LDPC$_g$, for $g \in \{4,5,6\}$.

We synthesize a third dataset, PC, to evaluate the practical implications of Lemma 3. The dataset consists of $N = 200$ graphs, half of which are random graphs $G(n, \frac{1}{2})$, labeled (-1), and half planted clique graphs $G(n, \frac{1}{2}, k)$ with $k = 2t\sqrt{n}$ and $n = 200$ (+1). Such a set is constructed for each of $t \in \{0, 0.5, \ldots, 3\}$.

Results The results of classifying the ERDOS and LDPC datasets are presented in Table 1. We note that the SVM-ϑ kernel performs well through-out the experiments, as does the SP kernel. These results indicate that SVM-ϑ and SP capture the global property girth, better than GL and RW.

The results of the planted clique experiment are presented in Figure 1. We see that for $t = 3$, all of the kernels give perfect classification. We also see, as expected that the classification rate is 0.5 at $t = 0$, were semi-random and random graphs are equivalent. Lo−ϑ, ksp, and GL all perform well, distinguishing between the two classes of graphs for small t. The results for SVM−ϑ conform with the results of Jethava et al. (2014), who showed empirically that $t = 3$ was the lower limit for perfect distinction. Worst is RW which did not perform well in either experiment.

5.3. Classification of benchmark graphs

We evaluate perform graph classification on a collection of established datasets of varying origin, commonly used for evaluation of graph kernels.

PTC (Predictive Toxicology Challenge) is a set of 417 chemical compound graphs labeled according to their carcinogenic effects on rats and mice (Helma et al., 2001). Those saidd to have *clear* or *some* evidence are labeled (+1) and thos said to have *no* evidence (-1). The dataset is split into groups by male or female rats or mice. A separate classifier was trained for each group and the average accuracy of all four is reported.

MUTAG (Debnath et al., 1991) is a dataset of 188 graphs representing mutagenetic compounds, labeled according to their mutagenic effects. ENZYME is a collection of 600 graphs representing tertiary protein structures collected by (Borgwardt et al., 2005), each labeled with one of the 6 EC top-level classes. NCI$_1$ is a set of 4110 graphs representing a subset of chemical compounds screened for activity against non-small cell lung cancer cell lines (Wale et al., 2008).

Results We report the CPU runtimes for computing each kernel on the benchmark experiments in Table 2, as measured in Matlab R2012a on a 3.4GHz Intel Core i7 with 4 cores and 32GB RAM.

The classfication accuracies of all kernels on the benchmark datasets are presented in Table 2. On PTC, MUTAG, and ENZYME one or both of the kernels presented in this paper perform better than state-of-the-art in terms of accuracy. On NCI$_1$, the Lovász ϑ kernel achieved the second highest accuracy. For all sets except NCI$_1$, the SVM-ϑ kernel and Lovász ϑ kernel performed the best using a RBF base kernel with $\sigma \in [0.1, 1]$. On NCI$_1$, a linear kernel performed better.

The SVM-ϑ kernel showed accuracies better than or competitive to state-of-the-art kernels on all datasets, while also being competitive in terms of runtime.

6. Conclusion

We have defined two graph kernels for unlabeled graphs, the Lovász ϑ and SVM-ϑ kernels, based on the Lovász number and its associated orthonormal representation. The kernels are designed to capture global properties of graphs such as the girth or the clique number. We derive sampling approximations of both kernels with known sample complexity. The kernels are competitive with state-of-the-art

Table 2. Average classification accuracy (%) using 10-fold cross-validation on benchmark datasets. The columns labeled $T_{(\cdot)}$, contain the CPU time used to compute the kernels for each dataset. Numbers in bold indicate the best results in each column. The kernels introduced in this paper are Lo-ϑ and SVM-ϑ.

KERNELS	PTC	MUTAG	ENZYME	NCI$_1$	T_{ptc}	T_{mutag}	T_{enzyme}	T_{nci1}
SP	63.0	87.2	30.5	**67.3**	0.39”	0.2”	1.32”	6.46”
GL	63.1	83.5	26.7	62.9	6.30”	4.5”	42.6”	1’32”
RW	60.6	85.6	21.2	63.1	14.4”	0.4”	24.1”	3’30”
Lo-ϑ	**64.3**	86.2	26.5	65.2	24’31”	6’39”	41’40”	2h 42’
SVM-ϑ	63.8	**87.8**	**33.5**	62.7	1’6”	17.8”	5’7”	3’19”

graph kernels for unlabeled graphs, in terms of accuracy, in several classification tasks, even reaching the highest accuracy on the majority of datasets. The datasets comprise synthesized graphs with important global properties, as well as benchmark graph datasets. We provide a result bounding the separation margin between two classes of graphs in Lovász ϑ kernel space. Future work include designing global kernels which leverage attributes on nodes and edges, and theoretical results about generalization error on classes of graphs.

Acknowledgments

The authors would like to thank Jean-Philippe Vert for fruitful discussions and the ICML reviewers for useful comments. This work is supported in part by the Swedish Foundation for Strategic Research.

References

Alon, N. and Spencer, J.H. *The Probabilistic Method.* Wiley, Chichester, 1992.

Alon, Noga, Krivelevich, Michael, and Sudakov, Benny. Finding a large hidden clique in a random graph. *Random Struct. Algorithms*, 13(3-4):457–466, 1998.

Bayati, Mohsen, Montanari, Andrea, and Saberi, Amin. Generating random graphs with large girth. In Mathieu, Claire (ed.), *SODA*, pp. 566–575. SIAM, 2009.

Borgwardt, Karsten M and Kriegel, Hans-Peter. Shortest-path kernels on graphs. In *Proceedings of ICDM*, pp. 74–81, 2005.

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

Chan, T.-H. Hubert, Chang, Kevin L., and Raman, Rajiv. An sdp primal-dual algorithm for approximating the lovász-theta function. In *Proceedings of ISIT*, pp. 2808–2812, Piscataway, NJ, USA, 2009. IEEE Press.

Chang, Chih-Chung and Lin, Chih-Jen. LIBSVM: A library for support vector machines. 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Debnath, Asim Kumar, de Compadre, Rosa L. Lopez, Debnath, Gargi, Shusterman, Alan J., and Hansch, Corwin. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34:786–797, 1991.

Devroye, Luc, György, András, Lugosi, Gábor, and Udina, Frederic. High-dimensional random geometric graphs and their clique number. *Electron. J. Probab.*, 16:no. 90, 2481–2508, 2011.

Dukanovic, Igor and Rendl, Franz. A semidefinite programming-based heuristic for graph coloring. *Discrete Applied Mathematics*, 156(2):180–189, 2008.

Erdős, P. Graph theory and probability. *Canadian Journal of Mathematics*, 11(1):34, 1959.

Feige, Uriel and Krauthgamer, Robert. Finding and certifying a large hidden clique in a semirandom graph. *Random Structures & Algorithms*, 16(2):195–208, 2000.

Feragen, Aasa, Kasenburg, Niklas, Petersen, Jens, de Bruijne, Marleen, and Borgwardt, Karsten M. Scalable kernels for graphs with continuous attributes. In *NIPS*, pp. 216–224, 2013.

Frieze, Alan M. and Jerrum, Mark. Improved approximation algorithms for max k-cut and max bisection. *Algorithmica*, 18 (1):67–81, 1997.

Gärtner, Thomas, Flach, Peter, and Wrobel, Stefan. On graph kernels: Hardness results and efficient alternatives. *Learning Theory and Kernel Machines*, pp. 129–143, 2003.

Goemans, Michel X. Semidefinite programming in combinatorial optimization. *Math. Program.*, 79:143–161, 1997.

Haussler, David. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz, 1999.

Helma, C., King, R. D., Kramer, S., and Srinivasan, A. The predictive toxicology challenge 2000–2001. *Bioinformatics*, 17 (1):107–108, 2001.

Hermansson, Linus, Kerola, Tommi, Johansson, Fredrik, Jethava, Vinay, and Dubhashi, Devdatt. Entity disambiguation in anonymized graphs using graph kernels. In *Proceedings of CIKM*, pp. 1037–1046. ACM, 2013.

Hush, Don R., Kelly, Patrick, Scovel, Clint, and Steinwart, Ingo. Qp algorithms with guaranteed accuracy and run time for support vector machines. *Journal of Machine Learning Research*, 7:733–769, 2006.

Iasemidis, Leonidas D., Pardalos, Panos M., Sackellares, J. Chris, and Shiau, Deng-Shan. Quadratic binary programming and dynamical system approach to determine the predictability of epileptic seizures. *J. Comb. Optim.*, 5(1):9–26, 2001.

Iyengar, Garud, Phillips, David J., and Stein, Clifford. Approximating semidefinite packing programs. *SIAM Journal on Optimization*, 21(1):231–268, 2011.

Jethava, Vinay, Martinsson, Anders, Bhattacharyya, Chiranjib, and Dubhashi, Devdatt. Lovasz theta function, svms and finding dense subgraphs. *Journal of Machine Learning Research*, 14:3495–3536, 2014.

Johansson, Fredrik D., Jethava, Vinay, Dubhashi, Devdatt, and Bhattacharyya, Chiranjib. Supplementary material. 2014.

Juels, Ari and Peinado, Marcus. Hiding cliques for cryptographic security. *Des. Codes Cryptography*, 20(3):269–280, 2000.

Karger, David R., Motwani, Rajeev, and Sudan, Madhu. Approximate graph coloring by semidefinite programming. *J. ACM*, 45(2):246–265, 1998. Earlier version in FOCS'94.

Knuth, Donald E. *The sandwich theorem*. Stanford University, Department of Computer Science, 1993.

Lovász, László. On the shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25(1):1–7, 1979.

Luz, Carlos J. and Schrijver, Alexander. A convex quadratic characterization of the lovász theta number. *SIAM J. Discrete Math.*, 19(2):382–387, 2005.

Mahé, Pierre and Vert, Jean-Philippe. Graph kernels based on tree patterns for molecules. *Machine Learning*, 75(1):3–35, 2009.

Ramon, Jan and Gärtner, Thomas. Expressivity versus efficiency of graph kernels. In Raedt, Luc De and Washio, Takashi (eds.), *Proceedings of the First International Workshop on Mining Graphs, Trees and Sequences at ECML/PKDD*, pp. 65–74, 2003.

Richardson, Thomas J. and Urbanke, Rüdiger L. *Modern Coding Theory*. Cambridge University Press, 2008. ISBN 978-0-521-85229-6.

Schölkopf, Bernhard, Tsuda, Koji, and Vert, Jean-Philippe. *Kernel methods in computational biology*. The MIT press, 2004.

Shervashidze, Nino and Borgwardt, Karsten. Fast subtree kernels on graphs. In *Proceedings of NIPS*, pp. 1660–1668. 2009.

Shervashidze, Nino and Borgwardt, Karsten. Graph kernels: Code and data. 2012. Software available at `http://webdav.tuebingen.mpg.de/u/karsten/Forschung/research.html?page=research&topic=JMLR10_graphkernels&html=JMLR10`.

Shervashidze, Nino, Vishwanathan, SVN, Petri, Tobias, Mehlhorn, Kurt, and Borgwardt, Karsten M. Efficient graphlet kernels for large graph comparison. In *Proceedings of AISTATS*, 2009.

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

Vishwanathan, SVN, Schraudolph, Nicol N, Kondor, Risi, and Borgwardt, Karsten M. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

Vogelstein, Joshua T., Roncal, William Gray, Vogelstein, R. Jacob, and Priebe, Carey E. Graph classification using signal-subgraphs: Applications in statistical connectomics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1539–1551, 2013.

Wale, Nikil, Watson, IanA., and Karypis, George. Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems*, 14(3):347–375, 2008.

Welzl, Emo. Smallest enclosing disks (balls and ellipsoids). In *Results and New Trends in Computer Science*, pp. 359–370. Springer-Verlag, 1991.