The Skew Spectrum of Graphs

Risi Kondor RISI@GATSBY.UCL.AC.UK
Gatsby Computational Neuroscience Unit, UCL, 17 Queen Square, London, WC1N 3AR, U.K.

Karsten M. Borgwardt KMB51@CAM.AC.UK
Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, U.K.

Abstract

The central issue in representing graph-structured data instances in learning algorithms is designing features which are invariant to permuting the numbering of the vertices. We present a new system of invariant graph features which we call the skew spectrum of graphs. The skew spectrum is based on mapping the adjacency matrix of any (weigted, directed, unlabeled) graph to a function on the symmetric group and computing bispectral invariants. The reduced form of the skew spectrum is computable in $O(n^3)$ time, and experiments show that on several benchmark datasets it can outperform state of the art graph kernels.

1. Introduction

After real valued vectors and strings, the third most fundamental type of data instance in machine learning are graphs. In addition to application domains such as bioinformatics (Sharan & Ideker, 2006), chemoinformatics (Bonchev & Rouvray, 1991), social networks (Kumar et al., 2006), etc., where information is presented as a graph from the start, graphs are also used to capture the relationships between the different parts of segmented images in computer vision (Harchaoui & Bach, 2007), and to capture grammatical structure in language (Collins & Duffy, 2002). Graphs may be directed or undirected, weighted or unweighted, and their vertices may be labeled, partially labeled or unlabeled. In each of these cases, the challenge is to represent graphs in a way that preserves their structure, but is insensitive to spurious transformations, such as changing the (arbitrary) numbering of their vertices.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

Given a graph $\mathcal{G}$, the two main lines of research that have emerged to address the above problem focus respectively on (a) designing an explicit feature mapping $\mathcal{G} \mapsto (q_1, q_2, \ldots, q_k)$; and (b) designing a kernel $k(\mathcal{G}_1, \mathcal{G}_2)$. Proponents of the first approach exploit global invariant properties of $\mathcal{G}$, such as the eigenvalues of its graph Laplacian, or local invariant properties, such as the number of occurrences in $\mathcal{G}$ of a library of small subgraphs. In contrast, proponents of the kernel approach use various intuitions about simultaneous random walks and diffusion on product graphs (Gärtner, 2003).

The new method that we present in this paper belongs in the first of the above two categories, but is distinguished from prior work (with the exception of (Shawe-Taylor, 1993)) by its algebraic character. In this regard, it is related to the recent line of papers (Kondor et al., 2007; Huang et al., 2008; Kondor, 2007a) introducing concepts from non-commutative harmonic analysis to machine learning. The mathematical foundations of our work are Kakarala's seminal results on the bispectra of functions on compact groups (Kakarala, 1993; Kakarala, 1992), and the recent discovery of a unitarily equivalent, but computationally more attractive set of invariants called the skew spectrum (Kondor, 2007b). We show how these general theories can be harnessed to construct graph invariants, and examine in detail their computational properties.

Experiments on standard datasets of chemical compounds show that the skew spectrum of graphs is competitive with the state of the art in graph features, and in some cases outperforms all other methods. A major advantage of the skew spectrum is that since it is an explicit feature mapping, it can be applied as a preprocessing step, and hence scales linearly with the number of examples. The computational complexity of computing the (reduced) skew spectrum of a single graph of n nodes scales with n^3. Uniquely amongst the graph invariants used in machine learning, the skew spectrum has a fixed number of scalar components (85

for the complete skew spectrum and 49 for its reduced version), resulting in a very compact representation. This does not stop the skew spectrum form remaining competitive both in speed and representational accuracy up to about $n = 300$.

For those technical details of the skew spectrum which could not be squeezed into this conference paper we refer the reader to the accompanying report (Kondor, 2008).

2. Graph Invariants

In this paper $\mathcal{G}$ will be a directed weighted graph of n vertices. We represent $\mathcal{G}$ by its adjacency matrix $A \in \mathbb{R}^{n\times n}$, where $[A]_{i,j} \in \mathbb{R}$ is the weight of the edge from vertex i to vertex j. Unweighted graphs are special cases satisfying $[A]_{i,j} \in \{0, 1\}$, while undirected graphs are special cases satisfying $A^\top = A$. We assume that A has no self-loops, i.e., $[A]_{i,i} = 0$ for $i = 1, 2, \ldots, n$.

Recall that a **permutation** of n objects is a bijective map $\pi\colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$. Permuting the labels on the vertices of $\mathcal{G}$ by π results in a new adjacency matrix A^π with entries

$$[A^\pi]_{\pi(i),\pi(j)} = [A]_{i,j}, \tag{1}$$

but A and A^π both represent the same graph $\mathcal{G}$. A function $q(A)$ is called a **graph invariant** if it is invariant to relabelings of this kind, i.e., if $q(A) = q(A^\pi)$ for any permutation π. Our objective is to construct a system $(q_1, q_2, \ldots, q_k)$ of graph invariants which capture as much information about $\mathcal{G}$ as possible, yet can be computed economically.

2.1. Reduction to Left-translation Invariance

Our approach is based on the fact that permutations form a **group**. This means that if for a pair of permutations σ_1 and σ_2, we define their product $\sigma_3 = \sigma_2\sigma_1$ by composition of maps, i.e., $\sigma_3(i) = \sigma_2(\sigma_1(i))$, then the following axioms are satisfied:

1. for any two permutations σ_1 and σ_2, the product $\sigma_2\sigma_1$ is also a permutation;
2. for any three permutations σ_1, σ_2 and σ_3, $\sigma_1(\sigma_2\sigma_3) = (\sigma_1\sigma_2)\sigma_3$;
3. The identity $e(i) = i$ is a permutation;
4. For any permutation σ, there is an inverse permutation σ^{-1} satisfying $\sigma\sigma^{-1} = \sigma^{-1}\sigma = e$.

The group of permutations of n objects is called the **symmetric group** over n letters and is denoted $\mathbb{S}_n$.

To find graph invariants we begin by mapping A to a function $f\colon \mathbb{S}_n \to \mathbb{R}$, defined as

$$f(\sigma) = A_{\sigma(n),\sigma(n-1)}. \tag{2}$$

Note that this is a very special type of function on $\mathbb{S}_n$ in that it is constant on each block of permutations

$$S_{i,j} = \{\, \sigma \in \mathbb{S}_n \mid \sigma(n) = i,\ \sigma(n-1) = j \,\}. \tag{3}$$

For $k < n$, identifying $\mathbb{S}_k$ with the subgroup of permutations permuting $1, 2, \ldots, k$ amongst themselves and leaving $k + 1, \ldots, n$ fixed, the above blocks, of which there are $n(n-1)$ in total, each have the form $\sigma\,\mathbb{S}_{n-2} = \{\, \sigma\tau \mid \tau \in \mathbb{S}_{n-2} \,\}$, and are called **left** $\mathbb{S}_{n-2}$–**cosets**.

Defining f as in (2) ensures that under relabeling it transforms in a transparent fashion. Specifically, if f' is the function corresponding to A^π, then

$$f'(\pi\sigma) = A^\pi_{(\pi\sigma)(n),(\pi\sigma)(n-1)} = A_{\sigma(n),\sigma(n-1)} = f(\sigma). \tag{4}$$

In general, a function $g\colon \mathbb{S}_n \to \mathbb{R}$ related to f by $g(\sigma) = f(\pi^{-1}\sigma)$ is called the **left-translate** of f by π, and is denoted f^π. Equation 4 tells us that $f' = f^\pi$, reducing the problem of constructing graph invariants to finding left-translation invariant features of functions on $\mathbb{S}_n$.

2.2. Invariant Matrices

Now consider the weighted sum of matrices

$$\widehat{f}_\rho = \sum_{\sigma\in\mathbb{S}_n} f(\sigma)\,\rho(\sigma), \tag{5}$$

where $\rho(\sigma)$ is a system of complex valued matrices satisfying

$$\rho(\sigma_2\,\sigma_1) = \rho(\sigma_2)\,\rho(\sigma_1) \qquad \sigma_1, \sigma_2 \in \mathbb{S}_n,$$

as well as the unitarity condition $\rho(\sigma^{-1}) = (\rho(\sigma))^{-1} = \rho(\sigma)^\dagger$. Such systems of matrices are called **unitary matrix representations** of $\mathbb{S}_n$. Changing variables from σ to $\sigma' = \pi^{-1}\sigma$ shows that

$$\begin{aligned}\widehat{f}^\pi_\rho = \sum_{\sigma\in\mathbb{S}_n} f(\pi^{-1}\sigma)\,\rho(\sigma) &= \sum_{\sigma'\in\mathbb{S}_n} f(\sigma')\,\rho(\pi\sigma') \\ &= \sum_{\sigma'\in\mathbb{S}_n} f(\sigma')\,\rho(\pi)\,\rho(\sigma') = \rho(\pi)\,\widehat{f}_\rho,\end{aligned}$$

which suggests that (5) is a good starting point for constructing left-translation invariants of f. For example, the matrix $\widehat{a}_\rho = \widehat{f}^\dagger_\rho \cdot \widehat{f}_\rho$ is invariant because

$$\begin{aligned}\widehat{a}^\pi_\rho = \widehat{f}^{\pi\,\dagger}_\rho \cdot \widehat{f}^\pi_\rho = (\rho(\pi)\widehat{f}_\rho)^\dagger(\rho(\pi)\widehat{f}_\rho) = \\ \widehat{f}^\dagger_\rho\,\rho(\pi)^\dagger\,\widehat{\rho}(\pi)\,\widehat{f}_\rho = \widehat{f}^\dagger_\rho \cdot \widehat{f}_\rho = \widehat{a}_\rho.\end{aligned} \tag{6}$$

The question we face is how to construct such invariants in a systematic way with minimum redundancy, yet maximum representational power.

3. Irreps and the Fourier Transform

It is easy to see that if $\rho_1 : \mathbb{S}_n \to \mathbb{C}^{d\times d}$ is a unitary representation of $\mathbb{S}_n$, and T is any $d \times d$ unitary matrix, then $\rho_2(\sigma) = T\,\rho_1(\sigma)\,T^\dagger$ is also a unitary representation. Such pairs of representations are said to be **equivalent**. Once we have computed (5) with $\rho = \rho_1$, computing it again with $\rho = \rho_2$ will not lead to additional invariants, since $\widehat{f}_{\rho_2} = T\widehat{f}_{\rho_1}T^\dagger$.

Another potential source of redundancy is reducibility. A representation ρ is said to be **reducible** if for some unitary T it splits in the form

$$\rho(\sigma) = T \left(\begin{array}{c|c} \rho_1(\sigma) & \\ \hline & \rho_2(\sigma) \end{array} \right) T^\dagger \qquad \sigma \in \mathbb{S}_n$$

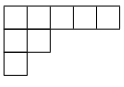
into a direct sum of smaller representations ρ_1 and ρ_2. Once again, $\widehat{f}_\rho$ does not supply any information on top of $\widehat{f}_{\rho_1}$ and $\widehat{f}_{\rho_2}$ because $\widehat{f}_\rho = T(\widehat{f}_{\rho_1} \oplus \widehat{f}_{\rho_2})T^\dagger$.

To avoid these redundancies we will use a *complete set of inequivalent irreducible unitary representations* (**irreps** for short). Such a set we denote by $\mathcal{R}$. The corresponding set of matrices

$$\widehat{f}_\rho = \sum_{\sigma\in\mathbb{S}_n} f(\sigma)\,\rho(\sigma), \qquad \rho \in \mathcal{R}, \tag{7}$$

is called the **Fourier transform** of f, and it provides the basis for generalizing harmonic analysis to non-commutative groups (Diaconis, 1988; Rockmore, 1997). Just as the classical Fourier transforms, $\mathcal{F}\colon f \to (\widehat{f}_\rho)_{\rho\in\mathcal{R}}$ satisfies a generalized form of the translation and convolution theorems. What is most crucial for our present purposes, however, is that (given the appropriate inner products) $\mathcal{F}$ is unitary, and therefore one–to–one: hence, no information is lost in going from f to the set of matrices $(\widehat{f}_\rho)_{\rho\in\mathcal{R}}$.

Several different systems of irreps for $\mathbb{S}_n$ are described in the literature (James & Kerber, 1981). In the interests of saving space, we only describe their general scheme, without going into the details of how to compute the actual representation matrices. In all the major representation schemes the individual irreps $\rho \in \mathcal{R}$ are indexed by **Young diagrams**, which are n boxes arranged in consecutive left-aligned rows satisfying the condition that no row overhangs the row above it. For example,



(8)

is a valid Young diagram for $n = 8$. We will use the letter λ to refer to Young diagrams and write $\lambda \vdash n$ to denote that λ is a Young diagram with n boxes. To simplify notation somewhat we write $\widehat{f}_\lambda$ for $\widehat{f}_{\rho_\lambda}$.

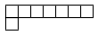
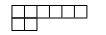
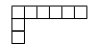
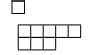
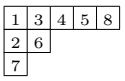
λ		d_λ
(n)		1
$(n-1,1)$		$n-1$
$(n-2,2)$		$\frac{n(n-3)}{2}$
$(n-2,1,1)$		$\frac{(n-1)(n-2)}{2}$
$(n-3,3)$		$\frac{n(n-1)(n-5)}{6}$
$(n-3,2,1)$		$\frac{n(n-2)(n-4)}{3}$

Table 1. The dimensionalities of some representations of $\mathbb{S}_n$. The diagrams are drawn as if $n = 8$, but the formulae hold for general n.

Young diagrams can also be described by listing the number of boxes in each row, for example, the above diagram is $\lambda = (5, 2, 1)$. For concreteness, when we need to draw Young diagrams we will always depict them as if $n = 8$.

Bijectively filling the boxes of a Young diagram with the numbers $1, 2, \ldots, n$ gives a **Young tableau**, and if a tableau satisfies the condition that in each row the numbers increase from left to right and in each column they increase from top to bottom it is called a **standard tableau**. For example,



is a standard tableau of shape $(5, 2, 1)$. The significance of standard tableaux is that they label the individual dimensions of the irrep of the same shape. Hence, we can find the dimensionality of ρ_λ by counting the number of possible standard tableaux of shape λ (Figure 1). An interesting special property of the symmetric group is that all the irreps can be chosen to be real valued. For generality, we retain the complex notation, but note that the actual system of irreps used in our experiments is real, so we could substitute "orthogonal" for "unitary" and $^\top$ for † throughout.

4. The Bispectrum and the Skew Spectrum

Armed with the irreps and non-commutative Fourier transforms, we can now undertake a more systematic study of left-translation invariant features of functions on the symmetric group. For example, (6) leads to the set of invariant matrices

$$\widehat{a}_\lambda = \widehat{f}_\lambda^\dagger \cdot \widehat{f}_\lambda, \qquad \lambda \vdash n,$$

which, by analogy with the analogous quantity in classical signal processing, is called the **power spectrum**

of f. The problem with the power spectrum is that it is very lossy. To see this, one need only consider $\widehat{f}'_\lambda = M_\lambda \widehat{f}_\lambda$ for any sequence of unitary matrices $(M_\lambda)_{\lambda \vdash n}$. The functions f and f' corresponding to these two Fourier transforms may be very different, yet their power spectrum will be the same.

4.1. The Bispectrum

Kakarala realized that the lossiness of the power spectrum can be addressed by forming tensor products of the various Fourier components, and proposed the alternative system of invariant matrices

$$\widehat{b}_{\lambda_1,\lambda_2} = (\widehat{f}_{\lambda_1} \otimes \widehat{f}_{\lambda_2})^\dagger \, C_{\lambda_1,\lambda_2} \Big[\bigoplus_\lambda \widehat{f}_\lambda\Big] \quad \lambda_1, \lambda_2 \vdash n \quad (9)$$

called the **bispectrum** (Kakarala, 1993)[1]. The bispectrum is based on the observation that $\widehat{f}_{\lambda_1} \otimes \widehat{f}_{\lambda_2}$ transforms according to

$$\widehat{f}^\pi_{\lambda_1} \otimes \widehat{f}^\pi_{\lambda_2} = (\rho_{\lambda_1}(\pi) \otimes \rho_{\lambda_2}(\pi)) \cdot (\widehat{f}_{\lambda_1} \otimes \widehat{f}_{\lambda_2}),$$

and that $\rho_{\lambda_1}(\pi) \otimes \rho_{\lambda_2}(\pi)$ is also a representation, although in general not irreducible. The general formula

$$\rho_{\lambda_1}(\sigma) \otimes \rho_{\lambda_2}(\sigma) = C_{\lambda_1,\lambda_2} \Big[\bigoplus_\lambda \rho_\lambda(\sigma)\Big] C^\dagger_{\lambda_1,\lambda_2} \quad (10)$$

telling us how to reduce it into a direct sum of irreps is called the Clebsch-Gordan decomposition, and the C_{λ_1,λ_2} unitary matrices appearing in (10) and (9) are called **Clebsch-Gordan matrices**.

By plugging (10) into (9) it is easy to see that the bispectrum is indeed invariant to left-translation. A much more remarkable fact, proved in (Kakarala, 1992), is that provided the technical condition that each $\widehat{f}_\lambda$ is invertible is satisfied, the bispectrum is also *complete* (or lossless) in the sense that the matrices $(\widehat{b}_{\lambda_1,\lambda_2})_{\lambda_1,\lambda_2 \vdash n}$ uniquely determine f up to translation.

4.2. The Skew Spectrum

Some of the drawbacks of using the bispectrum in practical applications are that (a) computing (9) may involve multiplying together very large matrices; (b) that the Clebsch-Gordan matrices, despite being universal constants, are not generally available in tabuated form; and (c) that for large n they are extremely difficult to compute. To address these concerns, Kondor (2007b) proposed an alternative set of invariants, called the **skew spectrum**, which are unitarily equivalent to the bispectrum, but much more straightforward to compute. The skew spectrum of $f\colon \mathbb{S}_n \to \mathbb{C}$ is defined as the collection of matrices

$$\widehat{q}_{\nu,\lambda} = \widehat{r}^\dagger_{\nu,\lambda} \cdot \widehat{f}_\lambda, \qquad \lambda \vdash n, \quad \nu \in \mathbb{S}_n, \qquad (11)$$

where $(\widehat{r}_{\nu,\lambda})_{\lambda \vdash n}$ is the Fourier transform of the function $r_\nu(\sigma) = f(\sigma\nu)\, f(\sigma)$. In (Kondor, 2007b) it is shown that if for some subgroup H, f is constant on left σH-cosets (as the function defined in (2) is constant on left $\mathbb{S}_{n-2}$-cosets), then it is sufficient to let ν take on just one value from each

$$H\sigma H = \{\, h_1 \sigma h_2 \mid h_1, h_2 \in H \,\}$$

double-coset, since every other component of $\widehat{q}$ will be linearly dependent on these.

5. The Skew Spectrum of Graphs

By the results of Sections 2 and 4, plugging (2) into (11) will give a relabeling invariant representation of any weighted graph $\mathcal{G}$. As it stands, however, this seems of only academic interest, since ν must extend over $n!$ different values for any one of which the combined size of the $(\widehat{q}_{\nu,\lambda})_{\lambda \vdash n}$ matrices is itself $n!$. Moreover, computing each $(\widehat{q}_{\nu,\lambda})_{\lambda \vdash n}$ requires a separate Fourier transform.

The first clue to how these problems may be remedied is provided by the comment at the end of the last section that if we are only interested in linearly independent invariants, then due to the special structure of f, we need only let ν take on one value from each $\mathbb{S}_{n-2}\, \sigma\, \mathbb{S}_{n-2}$ double coset. It is easy to see that there are only 7 such double cosets in $\mathbb{S}_n$, namely

$$\begin{aligned}
S^{n\mapsto n}_{n-1\mapsto n-1} &= \{\, \sigma \in \mathbb{S}_n \mid \sigma(n) = n,\ \ \sigma(n-1) = n-1 \,\} \\
S^{n\mapsto n-1}_{n-1\mapsto n} &= \{\, \sigma \in \mathbb{S}_n \mid \sigma(n) = n-1,\ \ \sigma(n-1) = n \,\} \\
S^{n\mapsto n}_{n-1\mapsto *} &= \{\, \sigma \in \mathbb{S}_n \mid \sigma(n) = n,\ \ \sigma(n-1) \in [n-2] \,\} \\
S^{n\mapsto n-1}_{n-1\mapsto *} &= \{\, \sigma \in \mathbb{S}_n \mid \sigma(n) = n-1,\ \ \sigma(n-1) \in [n-2] \,\} \\
S^{n\mapsto *}_{n-1\mapsto n-1} &= \{\, \sigma \in \mathbb{S}_n \mid \sigma(n) \in [n-2],\ \sigma(n-1) = n-1 \,\} \\
S^{n\mapsto *}_{n-1\mapsto n} &= \{\, \sigma \in \mathbb{S}_n \mid \sigma(n) \in [n-2],\ \ \sigma(n-1) = n \,\} \\
S^{n\mapsto *}_{n-1\mapsto *} &= \{\, \sigma \in \mathbb{S}_n \mid \sigma(n), \sigma(n-1) \in [n-2] \,\}\ ,
\end{aligned} \qquad (12)$$

where $[n-2] = \{1, 2, \ldots, n-2\}$.

Definition 1 *Given a graph $\mathcal{G}$ of n vertices and adjacency matrix A, the **skew spectrum** of $\mathcal{G}$ is defined as the collection of matrices*

$$\widehat{q}_{\nu,\lambda} = \widehat{r}^\dagger_{\nu,\lambda} \cdot \widehat{f}_\lambda, \qquad \lambda \vdash n, \qquad (13)$$

where $r_\nu(\sigma) = f(\sigma\nu)\, f(\sigma)$; f is defined as in (2); and ν takes on one value from each of the double cosets listed in (12).

The second important consequence of the form of (2) is that using the right system of irreps, $\widehat{f}$ becomes very

[1] The exact definition of the bispectrum varies somewhat between authors. However, the various definitions are all unitarily equivalent to each other.

sparse. To be specific, we use **Young's orthonormal representation** (YOR), which has the special property that if σ is restricted to $\mathbb{S}_{n-1}$, then the $\rho_\lambda(\sigma)$ matrices block-diagonalize in the form

$$\rho_\lambda(\sigma) = \bigoplus_{\lambda^-} \rho_{\lambda^-}(\sigma), \qquad \sigma \in \mathbb{S}_{n-1},$$

where λ^- extends over all valid Young diagrams derivable from λ by the removal of a single box. If the pair of standard tableaux t and t' feature n at the same box, then

$$[\rho_\lambda(\sigma)]_{t,t'} = [\rho_{\lambda^-}(\sigma)]_{t\downarrow_{n-1},t'\downarrow_{n-1}}$$

where $t\downarrow_{n-1}$ is the standard tableau that we get from t by removing the box containing n and λ^- is the corresponding Young diagram. If t and t' feature n at different locations, then $[\rho_\lambda(\sigma)]_{t,t'} = 0$. Applying this relation recursively gives that for $\sigma \in \mathbb{S}_k$,

$$[\rho_\lambda(\sigma)]_{t,t'} = \begin{cases} [\rho_{\lambda^-}(\sigma)]_{t\downarrow_k,t'\downarrow_k} & \text{or} \\ 0 \end{cases} \tag{14}$$

depending on whether $k+1,\ldots,n$ are each in the same boxes in t and t' or not.

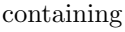
Now letting $\mathbb{S}_n/\mathbb{S}_{n-2}$ be a set of $n(n-1)$ permutations, one from each $\sigma\mathbb{S}_{n-2}$ coset, and defining $h_\sigma\colon \mathbb{S}_{n-2} \to \mathbb{C}$ as $h_\sigma(\tau) = f(\sigma\tau)$, the Fourier transform may be written as

$$\widehat{f}_\lambda = \sum_{\sigma\in\mathbb{S}_n/\mathbb{S}_{n-2}} \sum_{\tau\in\mathbb{S}_{n-2}} f(\sigma\tau)\,\rho_\lambda(\sigma)\,\rho_\lambda(\tau) = \sum_{\sigma\in\mathbb{S}_n/\mathbb{S}_{n-2}} \rho_\lambda(\sigma) \sum_{\tau\in\mathbb{S}_{n-2}} h_\sigma(\tau)\,\rho_\lambda(\tau).$$

Plugging in the appropriate decomposition of ρ_λ into a direct sum of irreps of $\mathbb{S}_{n-2}$ gives

$$\widehat{f}_\lambda = \sum_{\sigma\in\mathbb{S}_n/\mathbb{S}_{n-2}} \rho_\lambda(\sigma) \sum_{\tau\in\mathbb{S}_{n-2}} h_\sigma(\tau) \bigoplus_{\lambda^-} \rho_{\lambda^-}(\tau) = \sum_{\sigma\in\mathbb{S}_n/\mathbb{S}_{n-2}} \rho_\lambda(\sigma) \bigoplus_{\lambda^-} \left[\widehat{h}_\sigma\right]_{\lambda^-}, \tag{15}$$

showing that the Fourier transform over $\mathbb{S}_n$ may be broken down into $n(n-1)$ Fourier transforms over $\mathbb{S}_{n-2}$. This relationship is at the heart of the Clausen-type fast Fourier transforms for $\mathbb{S}_n$ (Clausen, 1989).

For f defined by (2), each h_σ is a constant function, and hence its Fourier transform has a very special form: since in YOR the irrep corresponding to $\lambda = (n)$ is the constant representation $\rho_{(n)}(\sigma) = (1)$, the corresponding $\left[\widehat{h}_\sigma\right]_\lambda$ component will be non-zero, but by unitarity all other components of $\widehat{h}_\sigma$ vanish. Plugging this result into (15) and using (14) shows that only those columns of $\widehat{f}$ may be non-zero which are indexed by standard tableau derivable from  by adding a box containing $n-1$ and another box containing n. Here and in the following, when drawing standard tableau, we only indicate the positions of those numbers in them that are not determined by the "numbers increase from left to right and top to bottom" rule. In addition, we use the symbol ■ to denote n and • to denote $n-1$. We summarize the above in the following theorem.

Theorem 1 *If f is defined as in (2), then the only non-zero entries of $\widehat{f}$ in YOR are:*

1. *the single scalar component $\widehat{f}_{(n)}$;*
2. *the*  *column of $\widehat{f}_{(n-1,1)}$;*
3. *the*  *column of $\widehat{f}_{(n-1,1)}$;*
4. *the*  *column of $\widehat{f}_{(n-2,2)}$;*
5. *the*  *column of $\widehat{f}_{(n-2,1,1)}$.*

This remarkable sparsity is the key to computing the skew spectrum of graphs efficiently. At the same time it is rather disappointing, since it manifestly destroys the invertibility of the $\widehat{f}_\lambda$ matrices required for Kakarala's completeness result. The $\widehat{r}_{\nu,\lambda}$ matrices are also column sparse, but their sparsity pattern is somewhat more complicated, so we leave describing it to (Kondor, 2008).

Equation (13) only yields non-zero elements in $\widehat{q}_{\nu,\lambda}$ where a non-zero row of $\widehat{r}^\dagger_{\nu,\lambda}$ meets a non-zero column of $\widehat{f}_\lambda$. By the above, this happens at only a constant number of row/column combinations. The exact result, derived in (Kondor, 2008), is the following.

Theorem 2 *Using YOR and an appropriate choice of $\{\nu\}$ double coset representatives, the skew spectrum of $\mathcal{G}$ has at most 85 non-zero scalar components.*

6. Computational Considerations

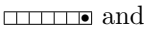
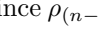
The computational properties of the skew spectrum are closely related to the structural results of the previous section. In particular, it is repeated applications of Clausen decompositions similar to (15) together with the sparsity of YOR that yields an efficient algorithm to compute $\widehat{q}$. In contrast to the previous section, we now employ a two-level factorization $\sigma = \sigma_1\sigma_2\tau$, where $\tau \in \mathbb{S}_{n-2}$, $\sigma_2 \in \mathbb{S}_{n-1}/\mathbb{S}_{n-2}$, and $\sigma_1 \in \mathbb{S}_n/\mathbb{S}_{n-1}$. As before, we have $n(n-1)$ functions $h_{\sigma_1\sigma_2}\colon \mathbb{S}_{n-2} \to \mathbb{C}$ defined $h_{\sigma_1\sigma_2}(\tau) = f(\sigma_1\sigma_2\tau)$, and by (2) each of these

is a constant function equal to $[A]_{\sigma_1\sigma_2(n),\,\sigma_1\sigma_2(n-1)}$. However, now we will also have intermediate functions $g_{\sigma_1}\colon \mathbb{S}_{n-1} \to \mathbb{C}$ defined $g_{\sigma_1}(\tau) = f(\sigma_1\tau)$. We then have the following results.

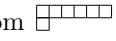
Lemma 1 *Each $\widehat{g}_{\sigma_1}$ can be computed from A in $O(n^2)$ scalar operations.*

Proof. Similarly to (15), we can relate the Fourier transform of g_{σ_1} to the Fourier transforms of $(h_{\sigma_1\sigma_2})_{\sigma_2}$ by

$$[\widehat{g}_{\sigma_1}]_\lambda = \sum_{\sigma_2 \in \mathbb{S}_{n-1}/\mathbb{S}_{n-2}} \rho_\lambda(\sigma_2) \bigoplus_{\lambda^-} \left[\widehat{h}_{\sigma_1\sigma_2}\right]_{\lambda^-}.$$

Since each $\widehat{h}_{\lambda_1\lambda_2}$ is confined to the one dimensional component $[\widehat{h}_{\sigma_1\sigma_2}]_{(n-2)}$, the only non-zero columns of $\widehat{g}_{\sigma_1}$ will be the ones indexed by standard tableaux derivable from  by addition of the single box , namely and  . The first one of these is trivial to compute, since $\rho_{(n-1)}(\sigma_2) \equiv (1)$, collapsing the above sum to

$$[\widehat{g}_{\sigma_1}]_{(n-1)} = \sum_{\sigma_2 \in \mathbb{S}_{n-1}/\mathbb{S}_{n-2}} \left[\widehat{h}_{\sigma_1\sigma_2}\right]_{(n-2)}.$$

This is a sum of $n-1$ scalars, so it can be computed in $O(n)$ time. Computing the second component involves taking the direct sum $M_{\sigma_1\sigma_2} = \bigoplus_{\lambda^-} \left[\widehat{h}_{\sigma_1\sigma_2}\right]_{\lambda^-}$, where λ^- extends over the two diagrams $(n-2)$ and $(n-3,1)$ derivable from  by removing a box. However, $\left[\widehat{h}_{\sigma_1\sigma_2}\right]_{(n-3,1)} = 0$, so $M_{\sigma_1\sigma_2}$ has only one non-zero entry. For given σ_2, multiplying $\rho_{(n-2,1)}(\sigma_2)$ with $M_{\sigma_1\sigma_2}$ thus requires $n-2$ operations. We are summing over $(n-1)$ possible values of σ_2, so the total time complexity is $(n-1)(n-2)$. ■

Lemma 2 *$\widehat{f}$ can be computed from the intermediate transforms $(\widehat{g}_{\sigma_1})_{\sigma_1 \in \mathbb{S}_n/\mathbb{S}_{n-1}}$ in $O(n^3)$ operations.*

The proof of Lemma 2 is similar to that of Lemma 1, but also involves considerations of the sparsity of the YOR matrices. Unfortunately, space limitations prevent us from providing a proof of this result. Putting the two lemmas together gives the following theorem.

Theorem 3 *The Fourier transform of f as defined in (2) can be computed in $O(n^3)$ operations.*

Proof. Each of the n different $\widehat{g}$ transforms can be computed in $O(n^2)$ operations, followed by the single $O(n^3)$ step of computing $\widehat{f}$ from the $\widehat{g}$'s. ■

Computing $\widehat{r}_\nu$ is unfortunately more costly than computing $\widehat{f}$. An extended version of this paper, which is in preparation, will show that the time complexity of this is $O(n^6)$. While for n less than about 20 this might still be feasible, for the type of experiments on which we wish to validate the skew spectrum it is not a viable option. The following subsection shows that most of the components of $\widehat{q}$ can still be computed in $O(n^3)$ operations.

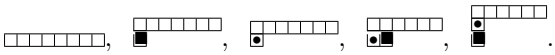
6.1. The Reduced Skew Spectrum

The expensive part of computing $\widehat{r}_\nu$ is computing those columns outside the five listed in Theorem 1. This leads to the idea of simply forcing these columns to be zero.

Definition 2 *Given a graph $\mathcal{G}$ of n vertices and adjacency matrix A, the **reduced skew spectrum** of $\mathcal{G}$ is the collection of matrices*

$$\widehat{q}^*_{\nu,\lambda} = \widehat{r}^*_{\nu,\lambda}{}^\dagger \cdot \widehat{f}_\lambda, \qquad \lambda \vdash n, \tag{16}$$

where f,r, and ν are as in Definition 1, and $\widehat{r}^*_\nu$ denotes the projection of $\widehat{r}_\nu$ to its columns labeled by

 . (17)

Since $\widehat{r}^*_\nu$ is identical to $\widehat{r}_\nu$ except for zeroing out certain columns, $(\widehat{q}^*_\nu)_\nu$ will yield a subset of the 85 scalar invariants in $(\widehat{q}_\nu)_\nu$. For each value of ν, for $\lambda = (n)$ we have one row of $\widehat{r^*}_\nu^\dagger$ meeting one column of $\widehat{f}_\nu$ giving one component; for $\lambda = (n-1,1)$ we have two rows meeting two columns, giving four components, etc. In total the reduced skew spectrum has $7\,(1+4+1+1) = 49$ non-zero scalar components.

The space of functions the Fourier transform of which has the sparsity pattern (16) is exactly the space of functions which are invariant on $\sigma\mathbb{S}_{n-2}$ cosets. This means that for each $\widehat{r}^*_\nu$ there must be a corresponding matrix B_ν related to it the same way that f is related to the adjacency matrix A. These matrices are given by the following theorem, the proof of which we again relegate to a longer publication.

Theorem 4 *For $\widehat{r}^*_\nu$ as defined in Definition 2,*

$$r^*_\nu(\sigma) = [B_\nu]_{\sigma(n),\sigma(n-1)},$$

where the seven possible B_ν matrices corresponding to the seven double cosets listed in (12) are

$$\begin{aligned}
[B_1]_{i,j} &= A_{i,j}\, A_{i,j} \\
[B_2]_{i,j} &= A_{i,j}\, A_{j,i} \\
[B_3]_{i,j} &= \tfrac{1}{n} A_{i,j} \textstyle\sum_{i'=1}^n A_{i',j} \\
[B_4]_{i,j} &= \tfrac{1}{n} A_{i,j} \textstyle\sum_{j'=1}^n A_{i,j'}
\end{aligned}$$

$$[B_5]_{i,j} = \frac{1}{n} A_{j,i} \sum_{i'=1}^{n} A_{i',j}$$
$$[B_6]_{i,j} = \frac{1}{n} A_{j,i} \sum_{j'=1}^{n} A_{i,j'}$$
$$[B_7]_{i,j} = \frac{1}{n(n-1)} A_{i,j} \sum_{i'=1}^{n} \sum_{j'=1}^{n} A_{i',j'}$$

Theorem 4 tells us that the reduced skew spectrum is very simple to compute: simply form the matrices $B_1, \ldots, B_7$, compute the corresponding $\widehat{r}^*_\nu$ the same way as $\widehat{f}$ is computed from A and form the products (16). In total this takes 8 partial Fourier transforms, each of which takes $O(n^3)$ time.

7. Experiments

In our experiments we evaluate the performance of the skew spectrum features on four benchmark datasets of chemical structures of molecules: MUTAG, ENZYMES, NCI1, and NCI109. MUTAG (Debnath et al., 1991) is a dataset of 188 mutagenic aromatic and heteroaromatic nitro compounds. The classification task is to predict for each molecule whether it exerts a mutagenic effect on the Gram-negative bacterium *Salmonella typhimurium*. ENZYMES is a dataset which we obtained from (Borgwardt et al., 2005), and which consists of 600 enzymes from the BRENDA enzyme database (Schomburg et al., 2004). In this case the task is to correctly assign each enzyme to one of the 6 EC top level classes. The average number of nodes of the graphs in this dataset is 32.6 and the average number of edges is 124.3. Finally, we also conducted experiments on two balanced subsets of NCI1 and NCI109, which classify compounds based on whether or not they are active in an anti-cancer screen ((Wale & Karypis, 2006) and `http://pubchem.ncbi.nlm.nih.gov`).

Since in these datasets the number of vertices varies from graph to graph, we set n to be the maximum over the entire dataset and augment each of the smaller graphs with the appropriate number of unconnected "phantom" nodes. The experiments consisted of running SVMs on the above data using the reduced skew spectrum features (linear kernel on these features), the random walk kernel (Gärtner et al., 2003), (with λ set to 10^{-3} on MUTAG/ENZYMES, and 10^{-4} on the NCI datasets for optimal performance), and an equal length shortest-path kernel (Borgwardt & Kriegel, 2005).

Our experimental procedure was as follows. We split each dataset into 10 folds of identical sizes. We then split 9 of these folds again into 10 parts, trained a C-SVM (implemented by LIBSVM (Chang & Lin, 2001)) on 9 parts, and predicted on the 10th part. We repeated this training and prediction procedure for $C \in \{10^{-7}, 10^{-6}, \ldots, 10^{7}\}$, and determined the C reaching maximum prediction accuracy on the 10th part. We then trained an SVM with this best C on all 9 folds (= 10 parts), and predicted on the 10th fold, which acts as an independent evaluation set. We repeated the whole procedure 10 times so that each fold acts as independent evaluation set exactly once. For each dataset and each method, we repeat the whole experiment 10 times and report mean accuracy levels and standard errors in Table 2. In three out of four experiments the skew spectrum beats the other methods, including the shortest-path kernel, which is considered state of the art for graphs of this type. Using a Gaussian RBF kernel instead of the linear kernel yields very similar results.

8. Conclusions

We have presented a new system of graph invariants, called the skew spectrum of graphs, based on a purely algebraic technique. From a mathematical point of view the skew spectrum is interesting because it brings a fundamentally new technique to constructing graph invariants. From a practical machine learning point of view the skew spectrum is interesting because it provides a powerful, yet efficiently computable representation for graph structured data instances.

Acknowledgments

We would like to thank Ramakrishna Kakarala for providing us with a hard copy of his thesis. We would also like to thank Dan Rockmore, Tony Jebara, Rocco Servedio, Maria Chudnovsky and Balázs Szendrői for discussions and the anonymous reviewers for helpful comments.

References

Bonchev, D., & Rouvray, D. H. (Eds.). (1991). *Chemical graph theory: Introduction and fundamentals*, vol. 1. London, UK: Gordon and Breach Science Publishers.

Borgwardt, K. M., & Kriegel, H.-P. (2005). Shortest-path kernels on graphs. *Proc. Intl. Conf. Data Mining* (pp. 74–81).

Borgwardt, K. M., Ong, C. S., Schonauer, S., Vishwanathan, S. V. N., Smola, A. J., & Kriegel, H. P. (2005). Protein function prediction via graph kernels. *Bioinformatics*, *21*, i47–i56.

Chang, C.-C., & Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

	MUTAG	ENZYME	NCI1	NCI109
Number of instances/classes	188/2	600/6	4110/2	4127/2
Max. number of nodes	28	126	111	111
Reduced skew spectrum	**88.61** (0.21)	25.83 (0.34)	**62.72** (0.05)	**62.62** (0.03)
Random walk kernel	71.89 (0.66)	14.97 (0.28)	51.30 (0.23)	53.11 (0.11)
Shortest-path kernel	81.28 (0.45)	**27.53** (0.29)	61.66 (0.10)	62.35 (0.13)

Table 2. Prediction accuracy in percent of the (reduced) skew spectrum features and state of the art graph kernels on four classification benchmarks in 10 repetitions of 10-fold cross-validation. Standard errors are indicated in parentheses. Best results for each datasets are in bold.

Clausen, M. (1989). Fast generalized Fourier transforms. *Theor. Comput. Sci.*, 55–63.

Collins, M., & Duffy, N. (2002). Convolution kernels for natural language. *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J., & Hansch, C. (1991). Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *J Med Chem*, *34*, 786–797.

Diaconis, P. (1988). *Group representation in probability and statistics*, vol. 11 of *IMS Lecture Series*. Institute of Mathematical Statistics.

Gärtner, T. (2003). A survey of kernels for structured data. *SIGKDD Explorations*, *5*, 49–58.

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. *COLT03* (pp. 129–143). Springer.

Harchaoui, Z., & Bach, F. (2007). Image classification with segmentation graph kernels. *Proceedings of CVPR07*.

Huang, J., Guestrin, C., & Guibas, L. (2008). Efficient inference for distributions on permutations. *Proceedings of NIPS07*.

James, G., & Kerber, A. (1981). *The representation theory of the symmetric group*. Addison-Wesley.

Kakarala, R. (1992). *Triple corelation on groups*. Doctoral dissertation, Department of Mathematics, UC Irvine.

Kakarala, R. (1993). A group theoretic approach to the triple correlation. *IEEE Workshop on higher order statistics* (pp. 28–32).

Kondor, R. (2007a). A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum. `http://arxiv.org/abs/cs.CV/0701127`.

Kondor, R. (2007b). The skew spectrum of functions on finite groups and their homogeneous spaces. `http://arxiv.org/abs/0712.4259`.

Kondor, R. (2008). The skew spectrum of graphs. To appear at `http://arxiv.org/`.

Kondor, R., Howard, A., & Jebara, T. (2007). Multi-object tracking with representations of the symmetric group. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*.

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. *KDD* (pp. 611–617).

Rockmore, D. N. (1997). Some applications of generalized FFTs. *Proceedings of the DIMACS workshop on groups and computation*.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., & Schomburg, D. (2004). Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Research*, *32D*, 431–433.

Sharan, R., & Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, *24*, 427–433.

Shawe-Taylor, J. (1993). Symmetries and discriminability in feedforward network architectures. *IEEE Transactions on Neural Networks*, *4*, 816–826.

Wale, N., & Karypis, G. (2006). Comparison of descriptor spaces for chemical compound retrieval and classification. *Proc. of ICDM* (pp. 678–689). Hong Kong.