Algebraic Methods in ML

Symposium and workshop

Jason Morton
Stanford, Math

Guy Lebanon
CSE, Georgia Tech

Risi Kondor
Gatsby Unit, UCL

Themes:

- Algebraic statistics
- Non-commutative harmonic analysis
- Combinatorial problems, e.g., ranking

13.30	**Risi Kondor:** Non-commutative harmonic analysis
14.05	**Guy Lebanon:** Modeling distributions on permutations and partial ranking
14.40	**Jason Morton:** Algebraic models for multilinear dependence
15.15	coffee break
15.25	**Yanxi Liu:** Symmetry Group-based Learning for Regularity Discovery from Real World Patterns
16.00	**Marina Meila:** Estimation and model selection in stagewise ranking: a representation story

Workshop session 1 (Callaghan room @ Westin)

7.30	**Stephen E. Fienberg:** Algebraic statistics for random graph models: Markov bases and their uses
8.05	**Adrian Dobra:** Algebraic statistics and contingency tables
8.40	**Keisuke Yamazaki:** Toric Modification on Mixture Models
9.15	coffee break
9.25	**Vincent Auvray:** Learning Parameters in Discrete Naive Bayes Models by Computing Fibers of the Parametrization ma
10.00	**Paul von Bunau:** Stationary Subspace Analysis

Workshop session 1 (Callaghan room @ Westin)

15.30	**Doru Balcan:** Alternatives to the Discrete Fourier Transform
16.05	**Lek-Heng Lim:** Graph Helmholtzian and rank learning
16.40	**Xiaoye Jiang:** Identity Management On Homogeneous spaces
17.15	coffee break
17.25	**Jonathan Huang:** Exploiting Probabilistic Independence for Permutations
17.55	**Tiberio Caetano:** Consistent structured estimation for weighted bipartite matching

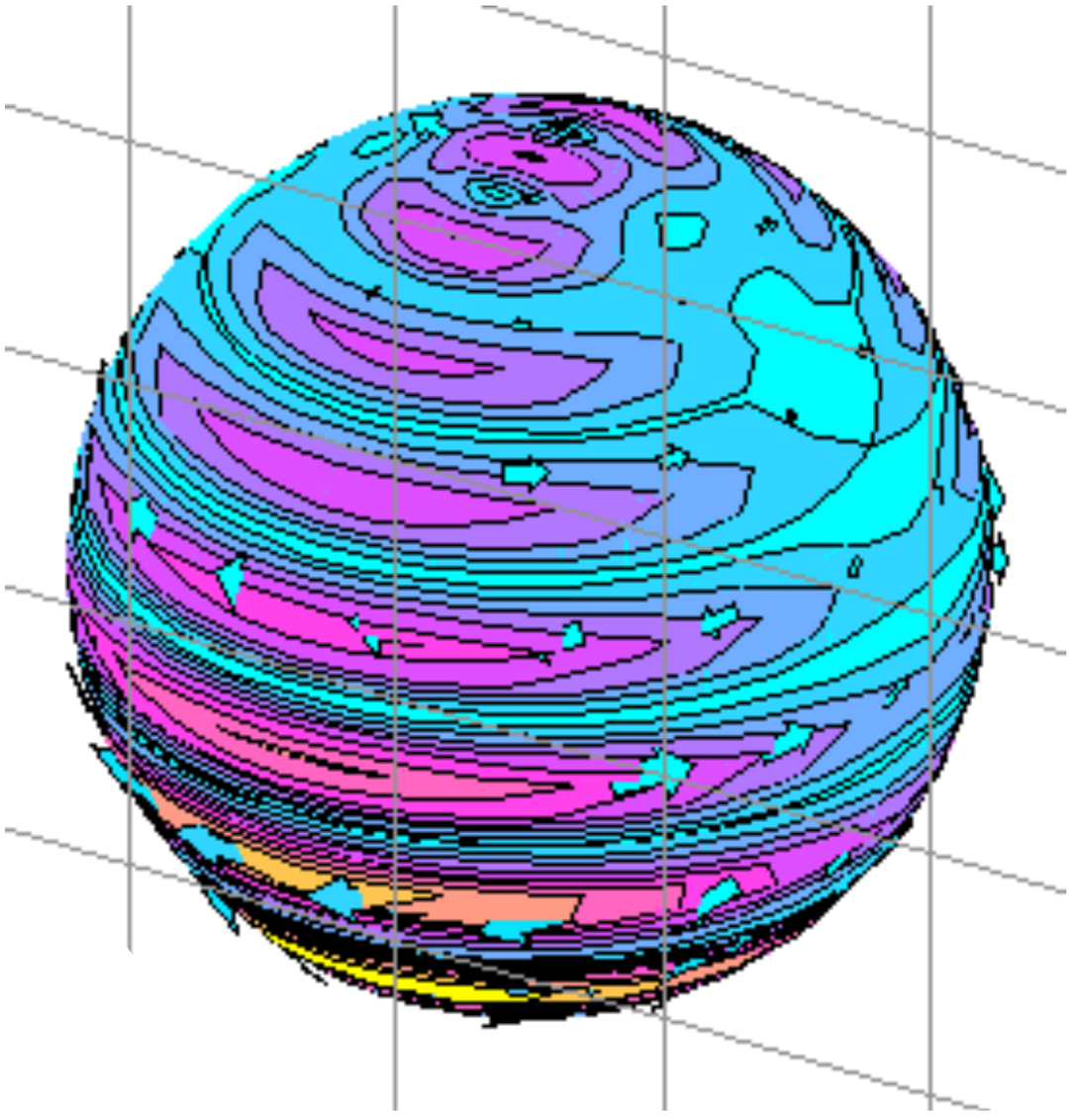
Non-commutative harmonic analysis

Risi Kondor

$$f(x) = \int e^{2\pi i k x} \, \widehat{f}(k) \, dk$$

$$\widehat{f}(k) = \int e^{-2\pi i k x} \, f(x) \, dx$$

$$f : S^2 \to \mathbb{R}$$



$$[1,2,3,4,5] \mapsto c_1$$

$$[1,2,3,5,4] \mapsto c_2$$

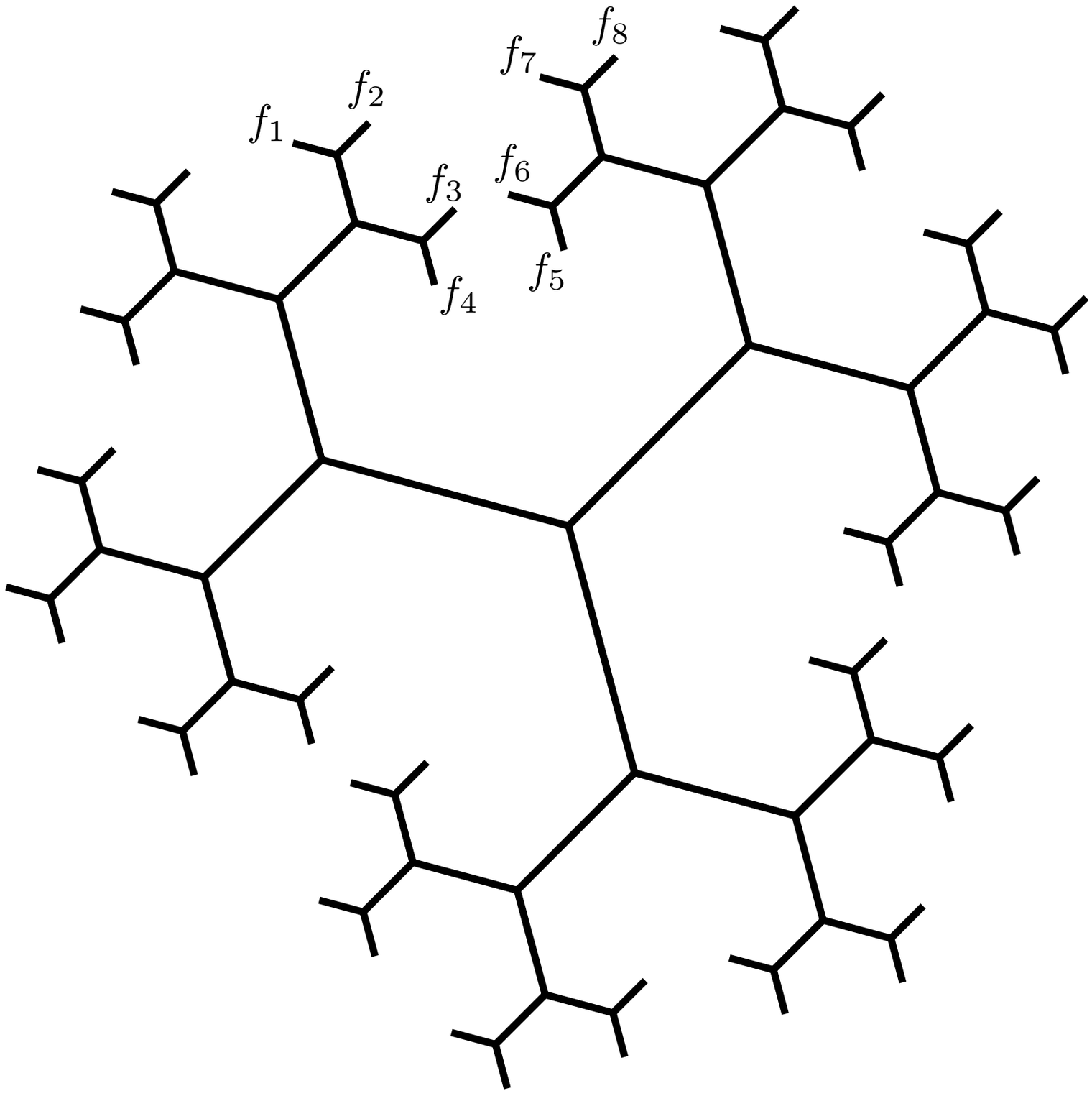
$$[1,2,5,3,4] \mapsto c_3$$

$$[1,5,2,3,4] \mapsto c_4$$

$$[5,1,2,3,4] \mapsto c_5$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$[5,4,3,2,1] \mapsto c_{5!}$$


f_1
f_2
f_3
f_4
f_5
f_6
f_7
f_8

$\sqrt{\heartsuit} = ?$

$\cos \heartsuit = ?$

$\frac{d}{dx} \heartsuit = ?$

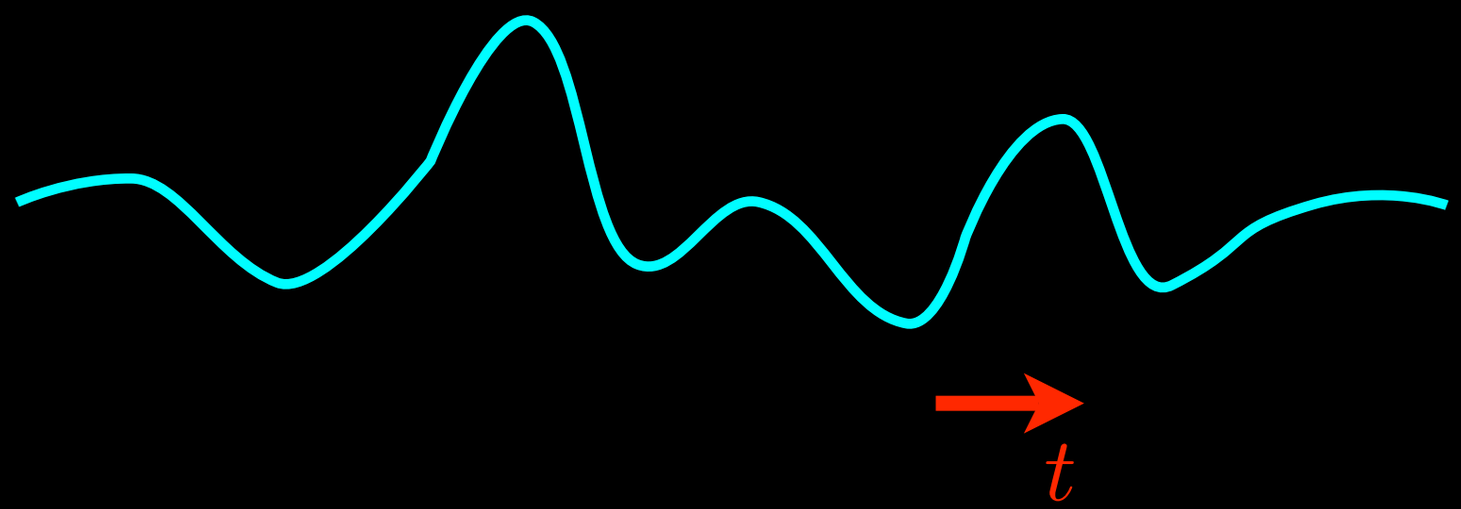
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \heartsuit = ?$

$$F\{\heartsuit\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{it\heartsuit}\, dt = ?$$

My normal approach is useless here.

$$\widehat{f}(k) = \int e^{-i2\pi kx} f(x)\, dx$$

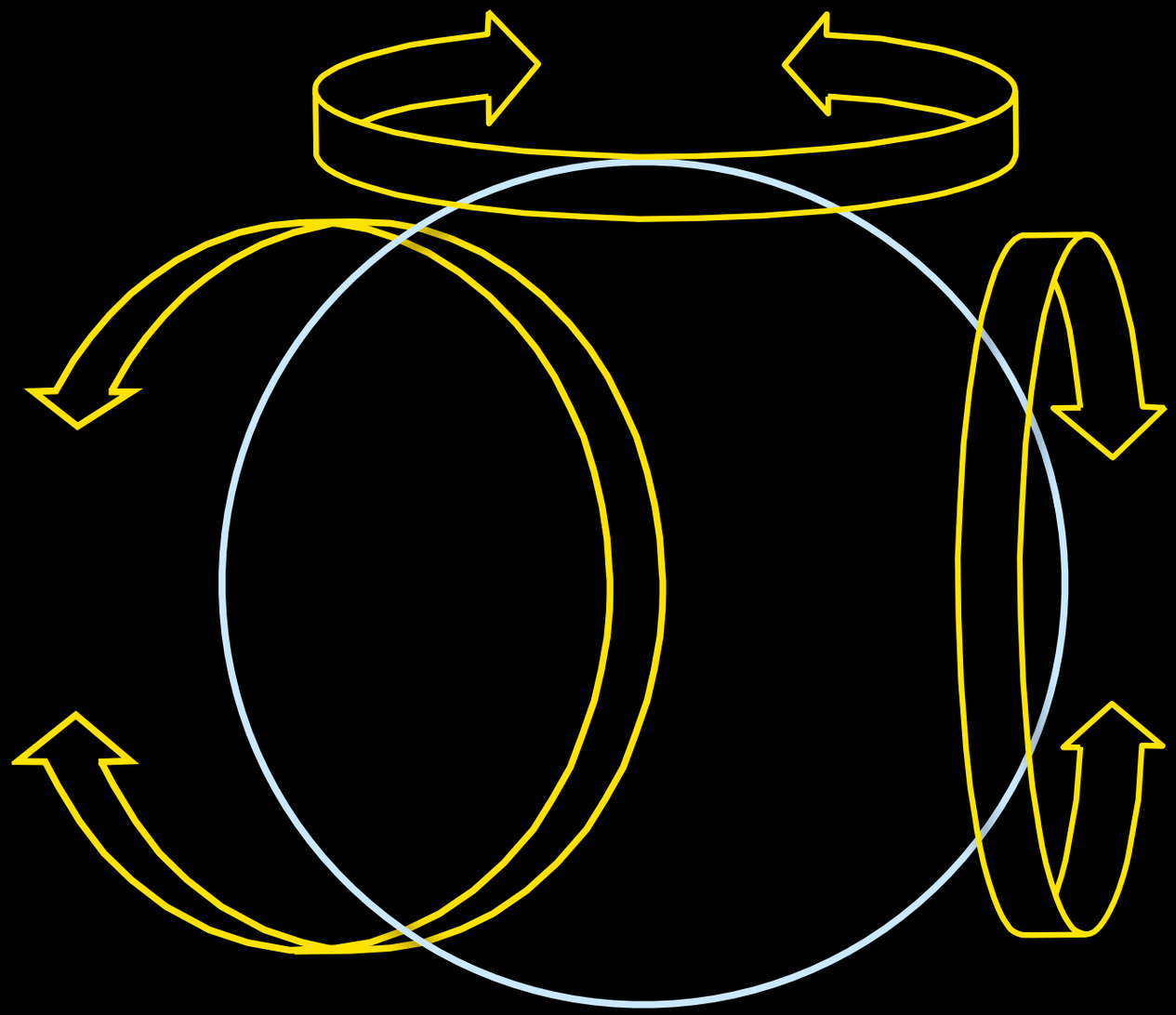
$$f^t(x) = f(x - t)$$



$$\widehat{f^t}(k) = e^{-i2\pi kt}\widehat{f}(k)$$

$$f^R(x) = f(R^{-1}x)$$

$$R \in \mathrm{SO}(3)$$



In general, $f : \mathcal{X} \to \mathbb{R}$ and G is a group acting on $\mathcal{X}$.

$$g_1 g_2 \in G \qquad \forall g_1, g_2 \in G$$

$$g_3(g_2 g_1) = (g_3 g_2) g_1 \qquad \forall g_1, g_2, g_3 \in G$$

$$eg = ge = g \qquad \text{for some } e \in G$$

$$g^{-1} g = g\, g^{-1} = e \qquad \text{for some } g^{-1} \in G$$

$$e^{-i2\pi kx_1}\, e^{-i2\pi kx_2} = e^{-i2\pi k(x_1+x_2)}$$

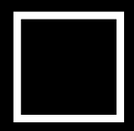
$$\widehat{f}(k) = \int e^{-i2\pi kx} f(x)\, dx$$

$$\rho(x_2)\rho(x_1) = \rho(x_2 x_1)$$

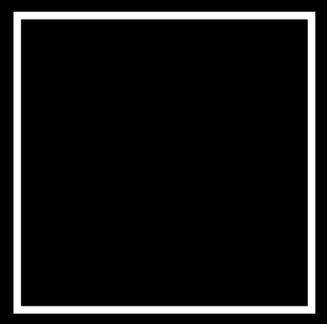
$$\widehat{f}(\rho) = \sum_{x \in G} f(x)\rho(x)$$

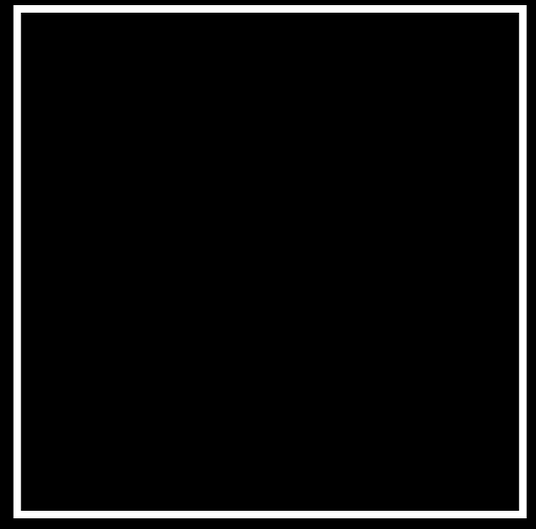
$\rho \colon G \to \mathbb{C}^{d \times d}$ is called a representation

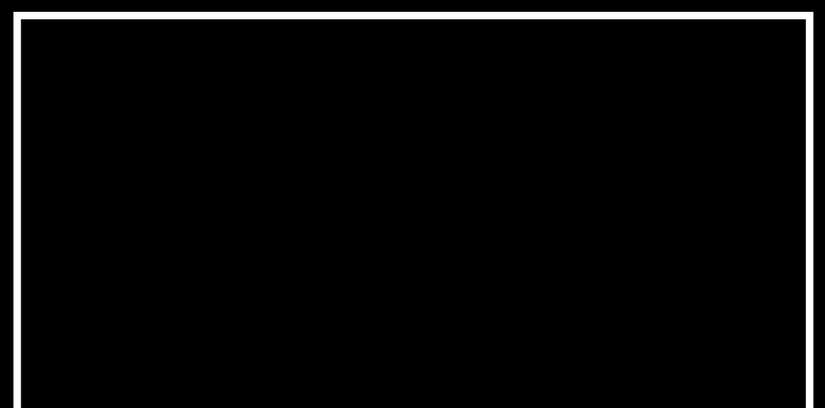
Irreducible Representations of $\mathrm{SO}(3)$

$\rho_0(R)$ 

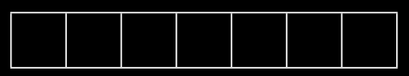
$$[\rho_l(\theta,\phi,\psi)]_{m,m'} = e^{-im'\psi}\, Y_l^m(\theta,\phi)$$

$\rho_1(R)$ 

$\rho_2(R)$ 

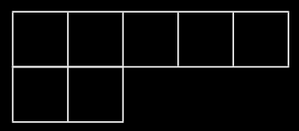
$\rho_3(R)$ 

Irreducible representations of $\mathbb{S}_n$

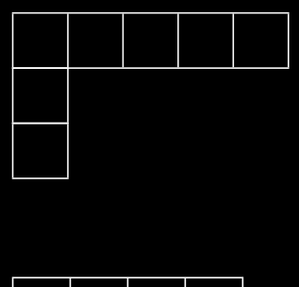


$d = 1$

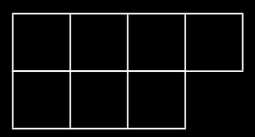
$d = n - 1$



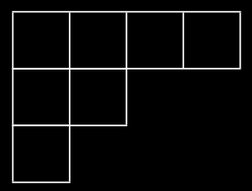
$d = n(n-3)/2$



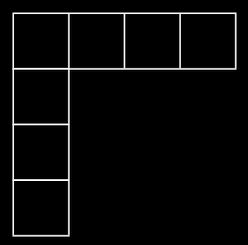
$d = (n-1)(n-2)/2$



$d = n(n-1)(n-5)/6$



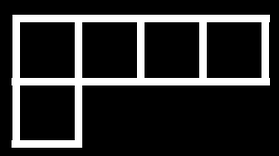
$d = n(n-2)(n-4)/3$



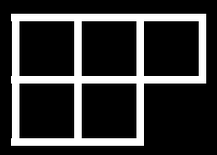
$d = (n-1)(n-2)(n-3)/6$

$$\sigma = (2, 1, 3, 4, 5) \in \mathbb{S}_5$$

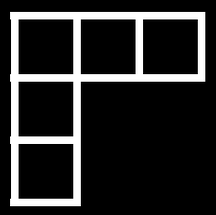
$$\rho_1(\sigma) = \begin{pmatrix} 1 \end{pmatrix}$$



$$\rho_2(\sigma) = \begin{pmatrix} -0.5 & -0.289 & -0.204 & -0.791 \\ 0.866 & -0.167 & -0.118 & -0.456 \\ 0 & 0.943 & -0.0833 & -0.323 \\ 0 & 0 & 0.968 & -0.25 \end{pmatrix}$$



$$\rho_3(\sigma) = \begin{pmatrix} 0.25 & -0.433 & 0.433 & -0.75 & 0 \\ -0.433 & -0.25 & -0.75 & -0.433 & 0 \\ -0.433 & -0.25 & 0.25 & 0.144 & -0.816 \\ 0.75 & -0.144 & -0.433 & 0.0833 & -0.471 \\ 0 & 0.816 & 0 & -0.471 & -0.333 \end{pmatrix}$$



$$\rho_4(\sigma) = \begin{pmatrix} 0.333 & 0.236 & 0 & 0.913 & 0 & 0 \\ -0.471 & 0.0417 & 0.217 & 0.161 & 0.839 & 0 \\ 0.816 & -0.0722 & 0.125 & -0.28 & 0.484 & 0 \\ 0 & -0.484 & -0.28 & 0.125 & 0.0722 & 0.816 \\ 0 & 0.839 & -0.161 & -0.217 & 0.0417 & 0.471 \\ 0 & 0 & 0.913 & 0 & -0.236 & 0.333 \end{pmatrix}$$

$$\widehat{f}(\rho) = \sum_{x \in G} f(x)\, \rho(x) \qquad f(x) = \frac{1}{|\,G\,|} \sum_{\rho \in \mathcal{R}} d_\rho \,\mathrm{tr}\left[\widehat{f}(\rho)\, \rho(x^{-1})\right]$$

1. Linearity: $\widehat{f+g} = \widehat{f} + \widehat{g}$
2. Unitarity: $\langle f, g \rangle = \langle f, \widehat{g} \rangle$
3. Left-translation: $\widehat{f^z(\rho)} = \rho(z)\, \widehat{f}(\rho)$
4. Convolution: $\widehat{f * g}(\rho) = \widehat{f}(\rho)\, \widehat{g}(\rho)$
5. The individual components correspond to different levels of smoothness.

Group algebra: $\{\boldsymbol{e}_x\}_{x\in G}$ $\boldsymbol{e}_x \cdot \boldsymbol{e}_y := \boldsymbol{e}_{xy}$

$$\mathcal{F}\colon\ \mathbb{C}G \overset{isom.}{\longrightarrow} \bigoplus_{\rho\in\mathcal{R}} \mathrm{GL}(d_\rho)$$

"Group Representations in Probability and Statistics"
IMS, 1988

Persi Diaconis

Multi-object tracking with representations of the symmetric group [K., Howard, Jebara, AISTATS 2007]

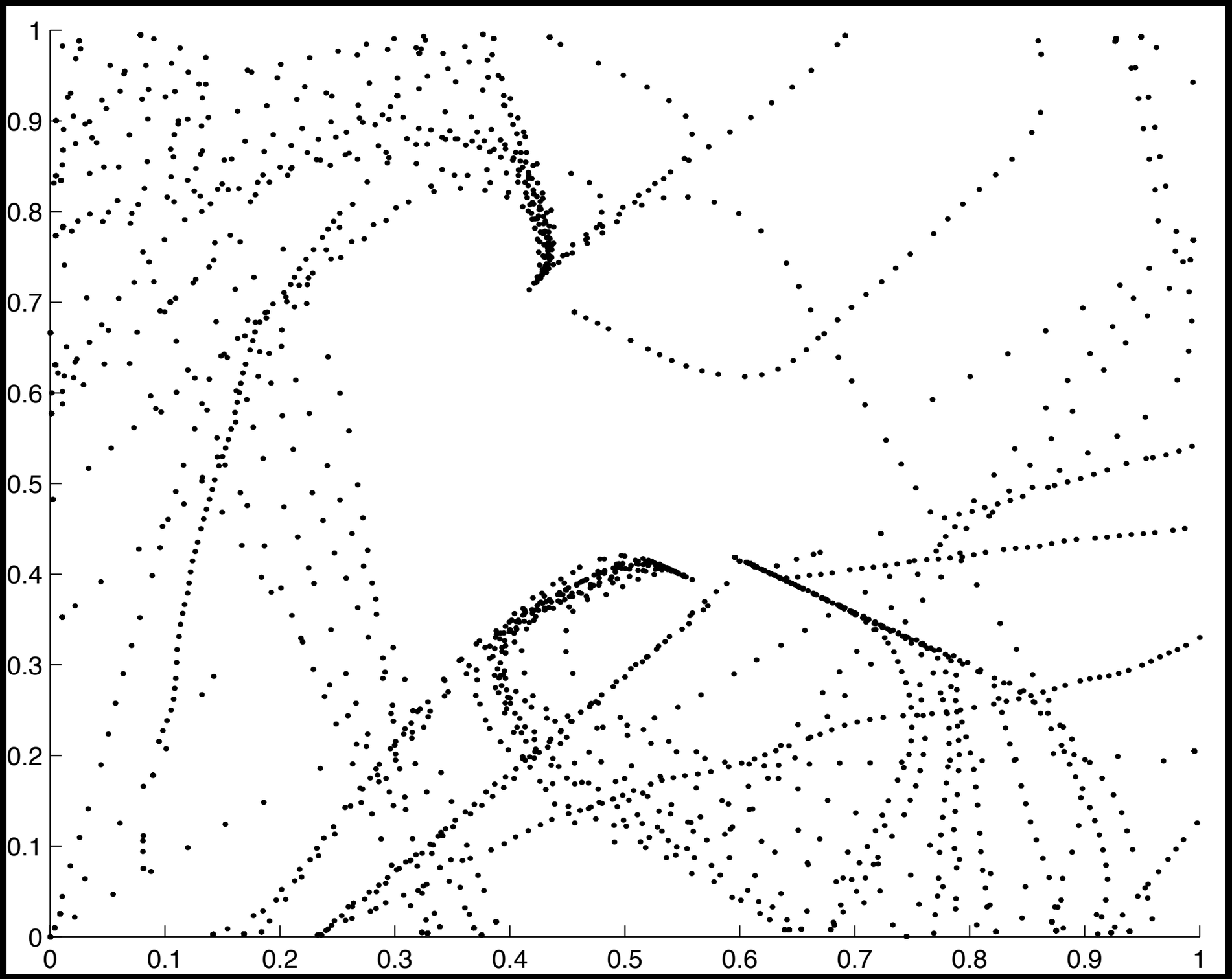
Efficient inference for distributions on permutations [Huang, Guestrin, Guibas, NIPS 2007]

The skew spectrum of graphs [K., Borgwardt, ICML 2008]

Characteristic kernels on groups and semigroups [Fukumizu, Sriperumbudur, Gretton, Scholkopf, NIPS 2008]

Inferring rankings under constrained sensing [Jagabathula, Shah, NIPS 2008]

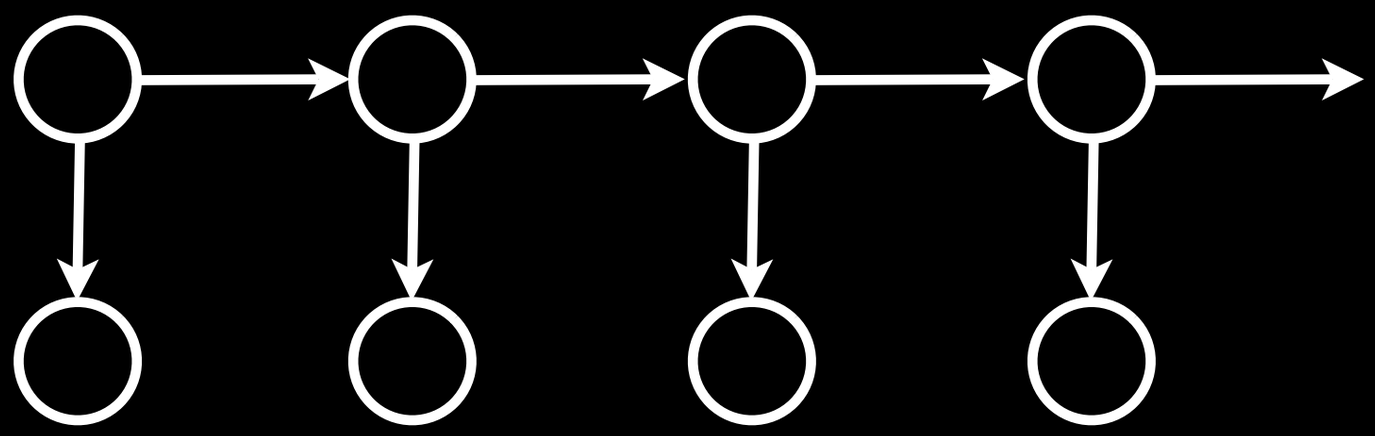
1. Band-limited approximations



http://www4.passur.com/jfk.html

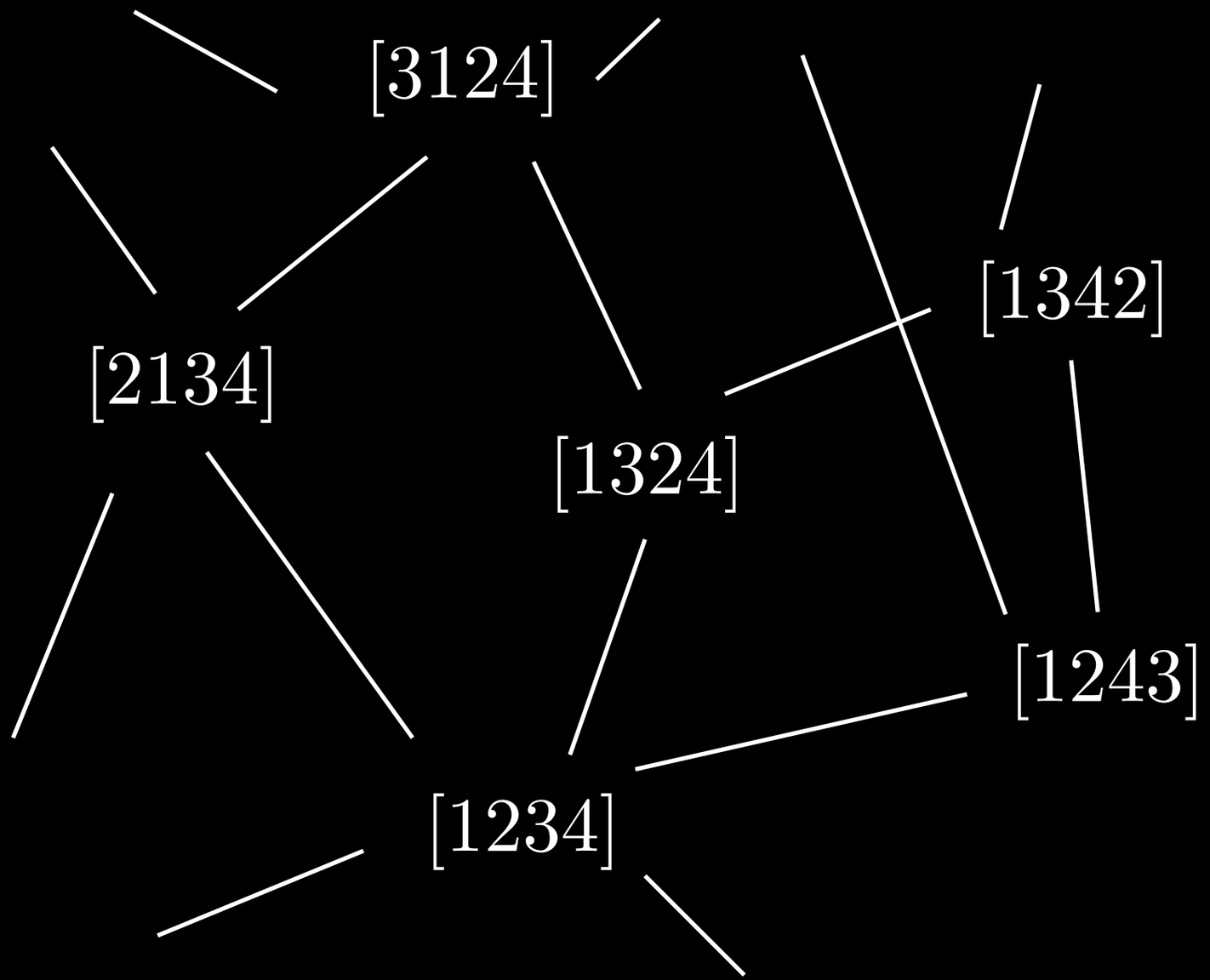
n	$n!$
2	2
3	6
4	24
5	120
6	720
7	5040
8	40320
9	362880
10	3628800
11	39916800
12	$4.8 \cdot 10^8$
$\vdots$	$\vdots$
15	$1.3 \cdot 10^{12}$
$\vdots$	$\vdots$
20	$2.4 \cdot 10^{18}$

$$p_t(x_1, x_2, \ldots, x_n, \sigma)$$



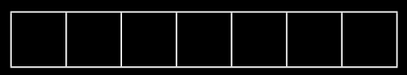
What does it mean for a function over permutations to be smooth?

Cayley graph:

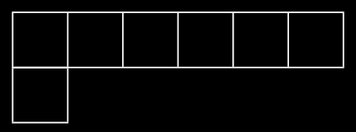


Theorem
The eigenvectors of the Laplacian of the Cayley graph are the vectors $v_{\lambda,i,j}(\sigma) = [\rho_\lambda(\sigma)]_{i,j}$ and the corresponding eigenvalues are

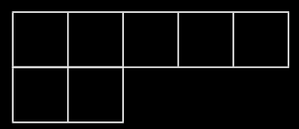
$$\alpha_\lambda = \binom{n}{2}\left(1 - \frac{\operatorname{tr}\left[\rho_\lambda((1,2))\right]}{d_\lambda}\right)$$



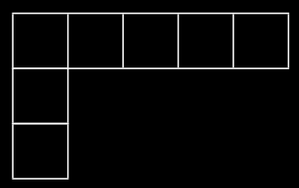
$d = 1$



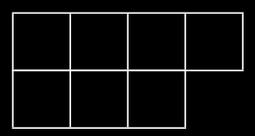
$d = n - 1$



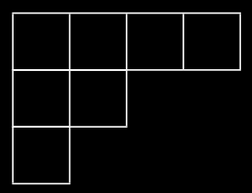
$d = n(n-3)/2$



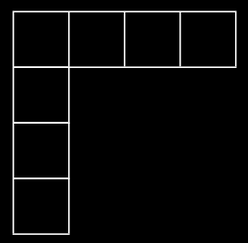
$d = (n-1)(n-2)/2$



$d = n(n-1)(n-5)/6$

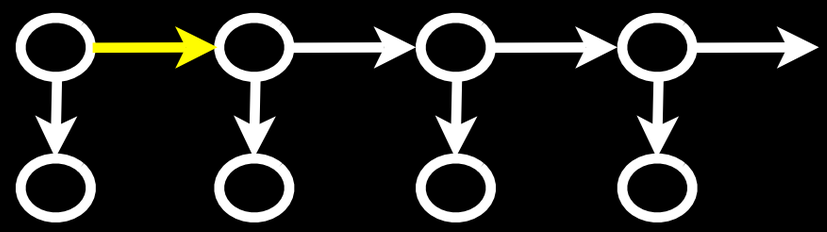


$d = n(n-2)(n-4)/3$



$d = (n-1)(n-2)(n-3)/6$

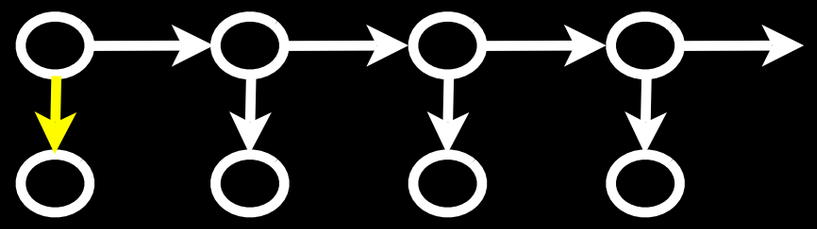
1. Noise/filtering/prediction



$$\widehat{f}_{t+1}(\rho_\lambda) = \widehat{g}(\rho_\lambda) \cdot \widehat{f}_t(\rho_\lambda).$$

$O(D^2)$

2. Observations/conditioning



$$p\,(\,O^\pi_{a\to i}\,|\,\sigma\,) = \begin{cases} \pi & \text{if } \sigma(a) = i, \\ (1-\pi)/(n-1) & \text{if } \sigma(a) \neq i. \end{cases}$$

$$f_{t+1}(\sigma) = \frac{p\,(\,O_{a\to i}\,|\,\sigma\,)\;\; f_t(\sigma)}{\sum_{\sigma'\in} p\,(\,O_{a\to i}\,|\,\sigma'\,)\;\; f(\sigma')}\,.$$

$O(D^2 n)$

3. Inference

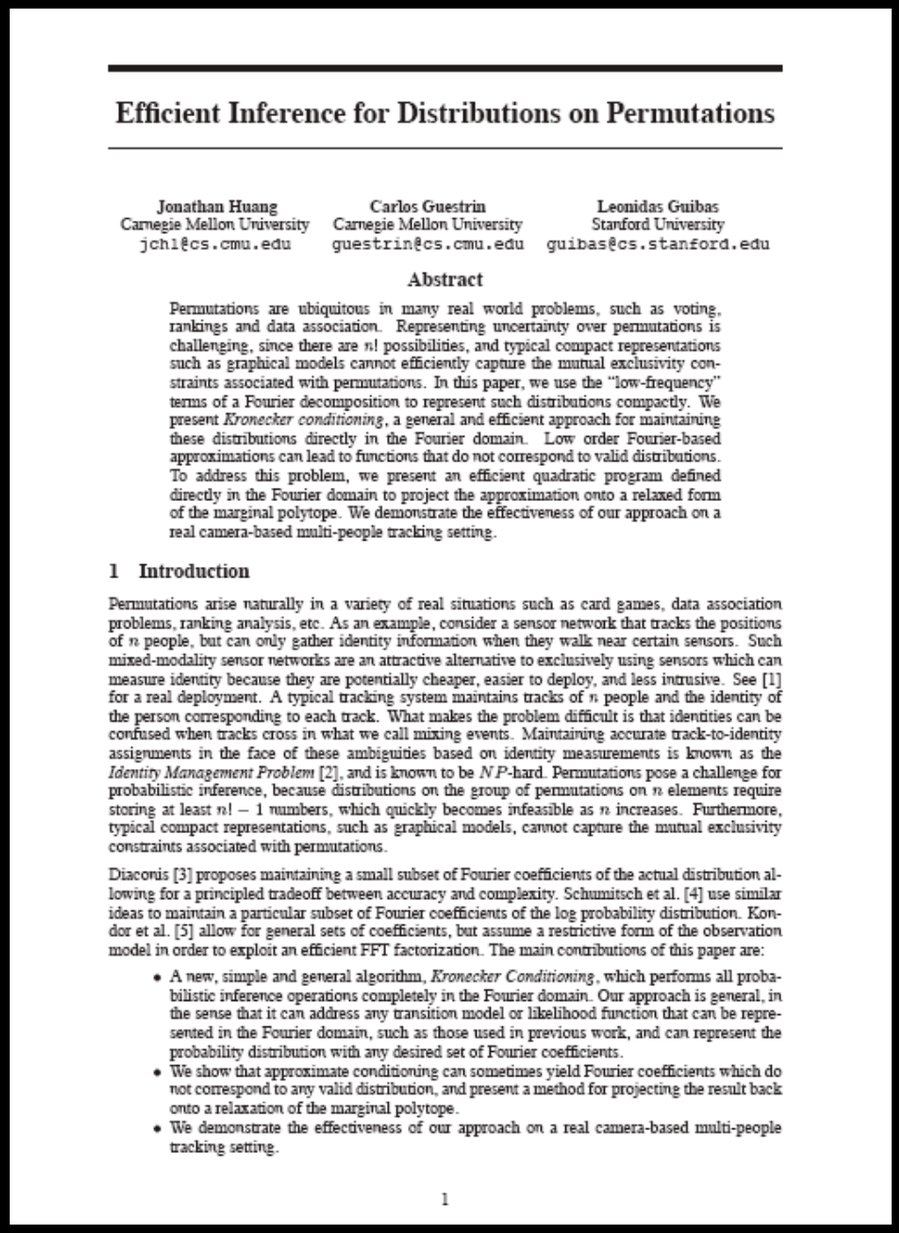
$$\mathbb{P}\{\,\sigma(n) = i\,\} = \sum_{\sigma(a)=i} f(\sigma) = \widehat{f}_{i,n}(\rho_{(n-1)})$$

$O(n^3)$

[K., Howard, Jebara, 2007]

- Enforcing positivity constraints by projection.

- More general, but somewhat more expensive updates requiring Clebsch-Gordan decomposition.



Efficient Inference for Distributions on Permutations

Jonathan Huang
Carnegie Mellon University
jch1@cs.cmu.edu

Carlos Guestrin
Carnegie Mellon University
guestrin@cs.cmu.edu

Leonidas Guibas
Stanford University
guibas@cs.stanford.edu

Abstract

Permutations are ubiquitous in many real world problems, such as voting, rankings and data association. Representing uncertainty over permutations is challenging, since there are $n!$ possibilities, and typical compact representations such as graphical models cannot efficiently capture the mutual exclusivity constraints associated with permutations. In this paper, we use the "low-frequency" terms of a Fourier decomposition to represent such distributions compactly. We present *Kronecker conditioning*, a general and efficient approach for maintaining these distributions directly in the Fourier domain. Low order Fourier-based approximations can lead to functions that do not correspond to valid distributions. To address this problem, we present an efficient quadratic program defined directly in the Fourier domain to project the approximation onto a relaxed form of the marginal polytope. We demonstrate the effectiveness of our approach on a real camera-based multi-people tracking setting.

1 Introduction

Permutations arise naturally in a variety of real situations such as card games, data association problems, ranking analysis, etc. As an example, consider a sensor network that tracks the positions of n people, but can only gather identity information when they walk near certain sensors. Such mixed-modality sensor networks are an attractive alternative to exclusively using sensors which can measure identity because they are potentially cheaper, easier to deploy, and less intrusive. See [1] for a real deployment. A typical tracking system maintains tracks of n people and the identity of the person corresponding to each track. What makes the problem difficult is that identities can be confused when tracks cross in what we call mixing events. Maintaining accurate track-to-identity assignments in the face of these ambiguities based on identity measurements is known as the *Identity Management Problem* [2], and is known to be *NP*-hard. Permutations pose a challenge for probabilistic inference, because distributions on the group of permutations on n elements require storing at least $n! - 1$ numbers, which quickly becomes infeasible as n increases. Furthermore, typical compact representations, such as graphical models, cannot capture the mutual exclusivity constraints associated with permutations.

Diaconis [3] proposes maintaining a small subset of Fourier coefficients of the actual distribution allowing for a principled tradeoff between accuracy and complexity. Schumitsch et al. [4] use similar ideas to maintain a particular subset of Fourier coefficients of the log probability distribution. Kondor et al. [5] allow for general sets of coefficients, but assume a restrictive form of the observation model in order to exploit an efficient FFT factorization. The main contributions of this paper are:

- A new, simple and general algorithm, *Kronecker Conditioning*, which performs all probabilistic inference operations completely in the Fourier domain. Our approach is general, in the sense that it can address any transition model or likelihood function that can be represented in the Fourier domain, such as those used in previous work, and can represent the probability distribution with any desired set of Fourier coefficients.
- We show that approximate conditioning can sometimes yield Fourier coefficients which do not correspond to any valid distribution, and present a method for projecting the result back onto a relaxation of the marginal polytope.
- We demonstrate the effectiveness of our approach on a real camera-based multi-people tracking setting.

1

[Huang, Guestrin, Guibas, 2007]

2. Invariances

Recall: $\widehat{f^z}(\rho) = \rho(z)\widehat{f}(\rho) \qquad f^z(x) = f(z^{-1}x)$

Power spectrum: $\widehat{a}(\rho) = \widehat{f}(\rho)^\dagger \, \widehat{f}(\rho)$

Bispectrum:

$$b(\rho_1, \rho_2) = C^\dagger \left(\widehat{f}(\rho_1) \otimes \widehat{f}(\rho_2)\right)^\dagger C \bigoplus_\rho \widehat{f}(\rho)$$

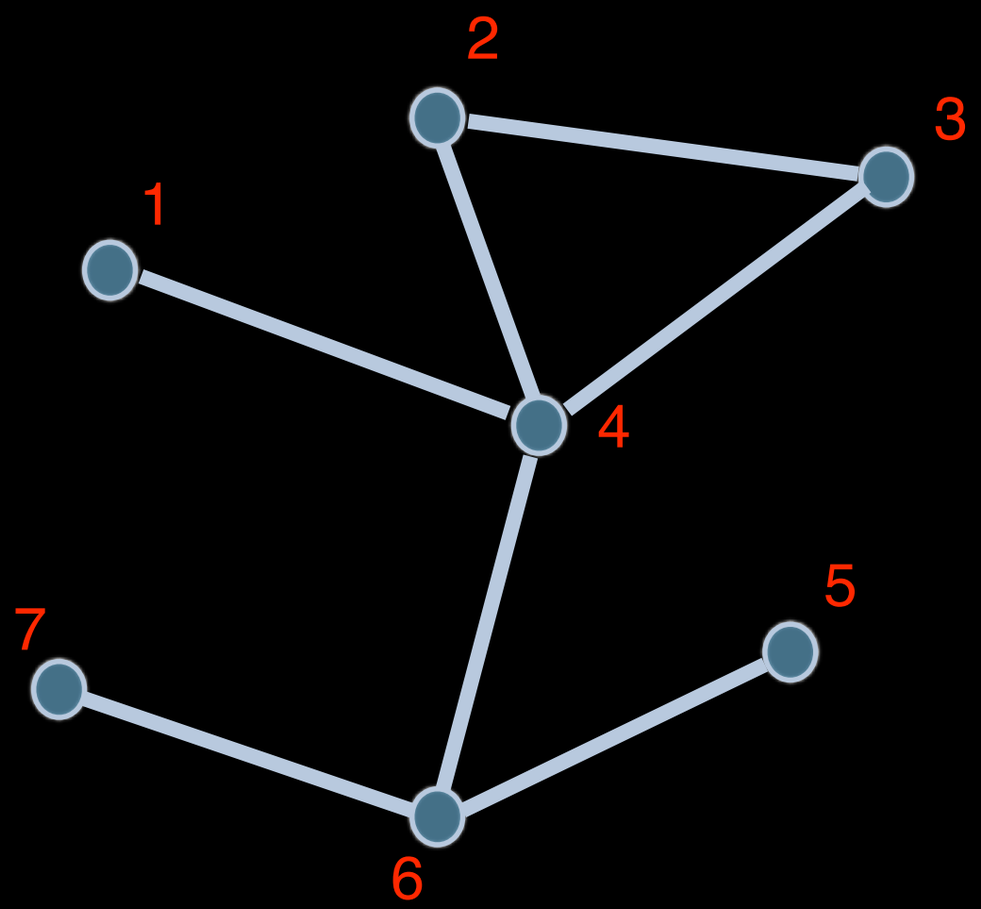
$$\rho_1(z) \otimes \rho_2(z) = C \left[\bigoplus_\rho \rho(z)\right] C^\dagger$$

Skew spectrum:

$$\widehat{q}_z(\rho) = \widehat{r}_z(\rho)^\dagger \cdot \widehat{f}(\rho) \qquad r_z(x) = f(xz)f(x) \qquad z \in G$$

(unitarily equivalent, but easier to compute)

[K. 2008]



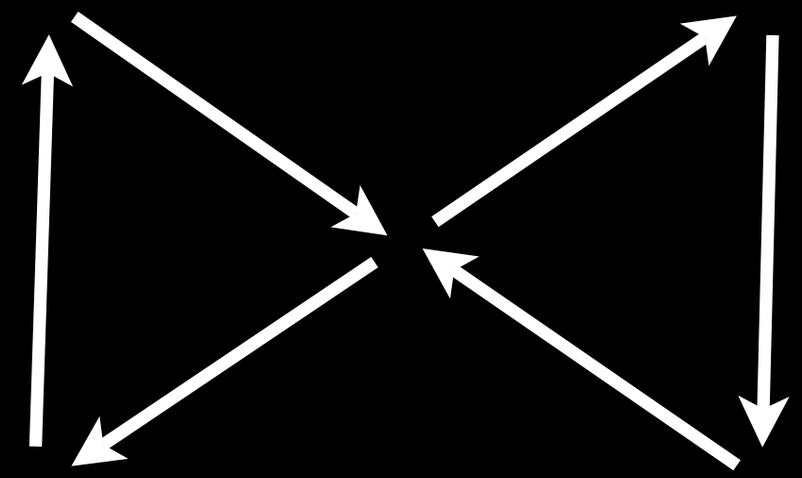
$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

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

[K., Borgwardt, 2008]

$$\widehat{f}(\rho_1) = \boxed{(6)}$$

$$\widehat{f}(\rho_2) = \begin{pmatrix} \boxed{\begin{matrix}\text{-}0.25 \\ \text{-}0.323\end{matrix}} & \boxed{\begin{matrix}\text{-}0.323 \\ \text{-}0.417\end{matrix}} & \begin{matrix}0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0\end{matrix} \\ \begin{matrix}0.913 \\ 0\end{matrix} & \begin{matrix}1.18 \\ 0\end{matrix} & \begin{matrix}0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0\end{matrix} \end{pmatrix}$$



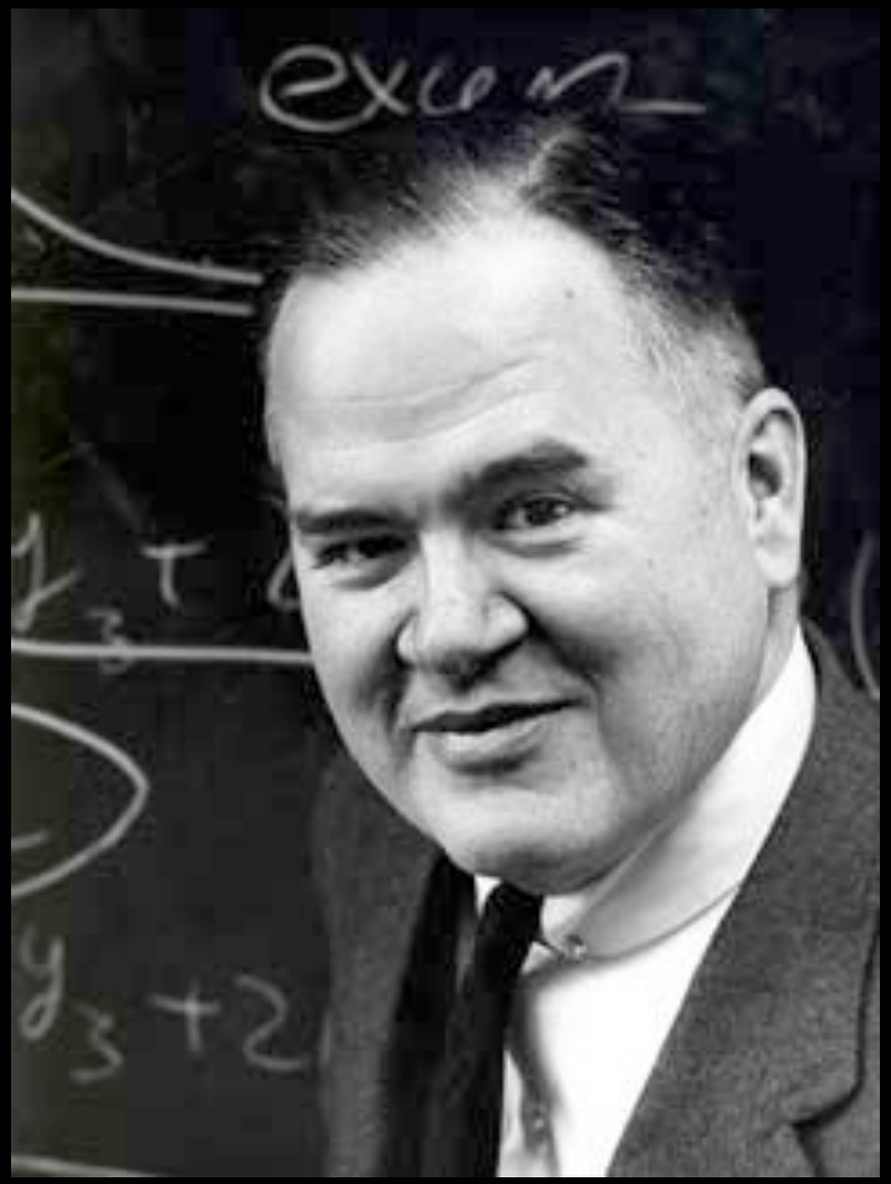
$$\widehat{f}(\rho_3) = \begin{pmatrix} \boxed{\begin{matrix}1.33 \\ 0.471 \\ 0 \\ 0.816 \\ 0\end{matrix}} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} \end{pmatrix}$$

$$\widehat{f}(\rho_4) = \begin{pmatrix} \boxed{\begin{matrix}\text{-}1.67 \\ 1.18 \\ 0 \\ \text{-}0.913 \\ 0 \\ \text{-}2.24\end{matrix}} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} & \begin{matrix}0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0\end{matrix} \end{pmatrix}$$

Proposition [K. & Borgwardt 2008]
The reduced skew spectrum of a graph contains 49 scalar components and may be computed in $O(n^3)$ operations.

	MUTAG	ENZYME	NCI1	NCI109
Number of instances/classes	600/6	188/2	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)

3. Algorithms



John Wilder Tukey
1915-2000

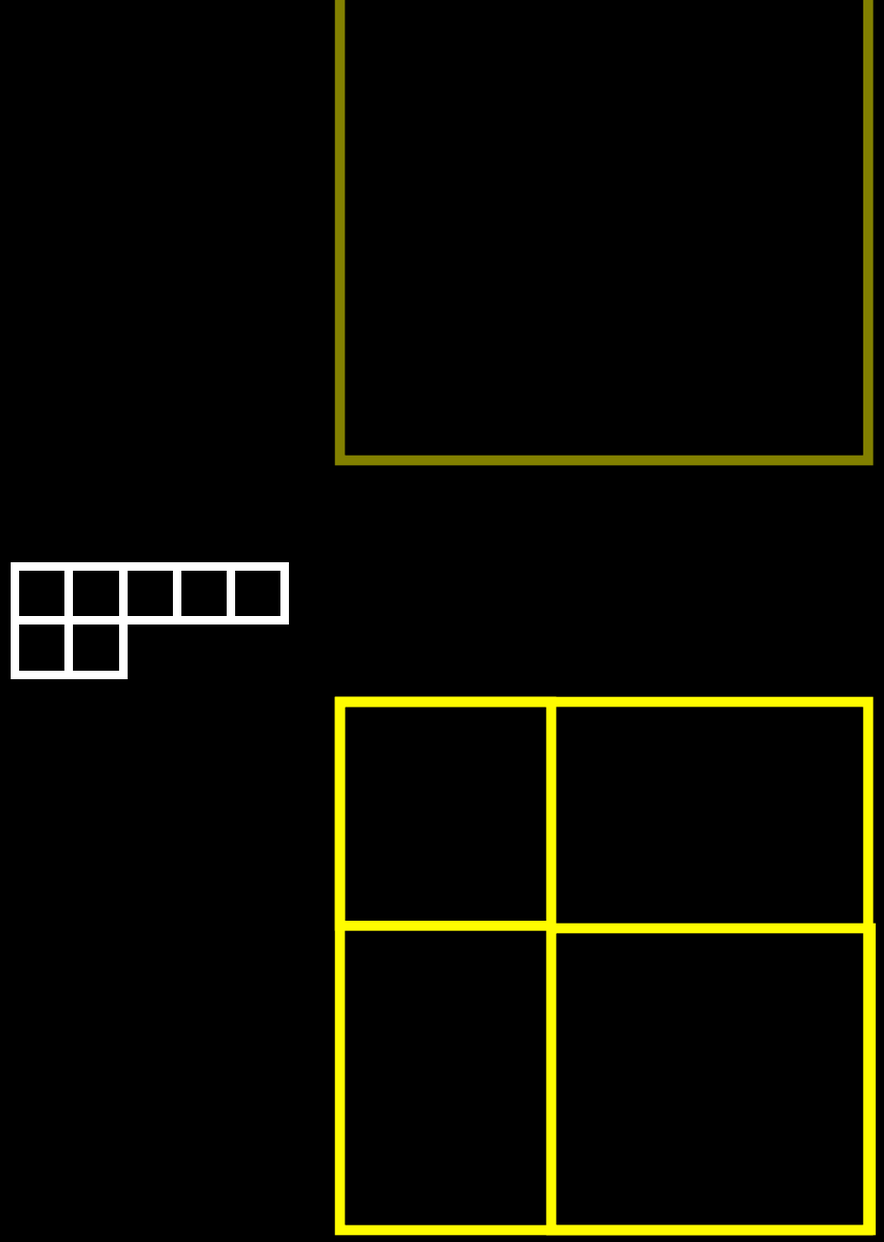
"An algorithm for the machine calculation of complex Fourier series,", James W. Cooley and John W. Tukey, *Math. Comput.* **19**, 297–301 (1965)

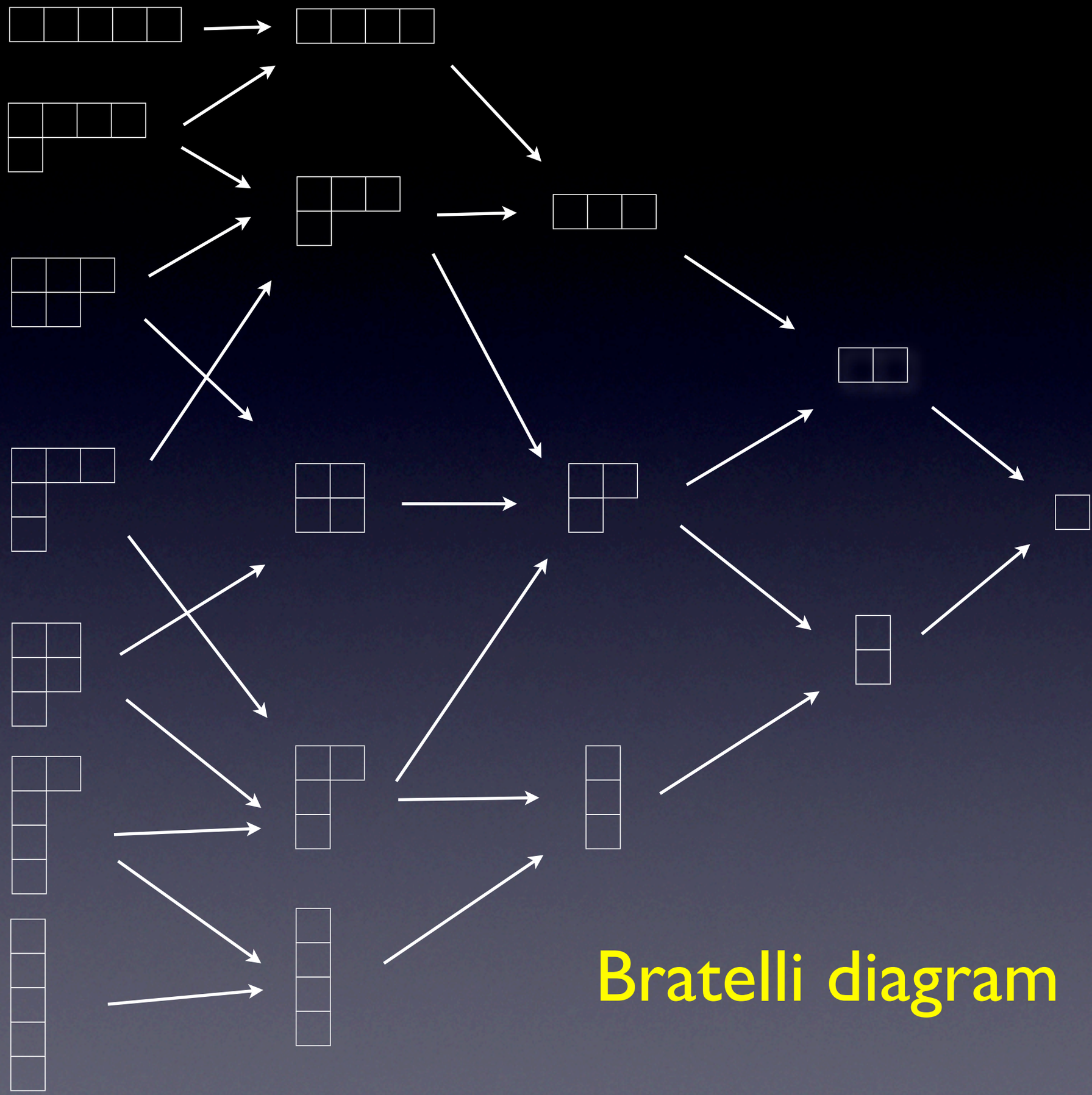
"Fast generalized Fourier transforms" [Clausen, 1989]

FFTs for wreath product groups and many others by Maslen, Rockmore, Healy,...

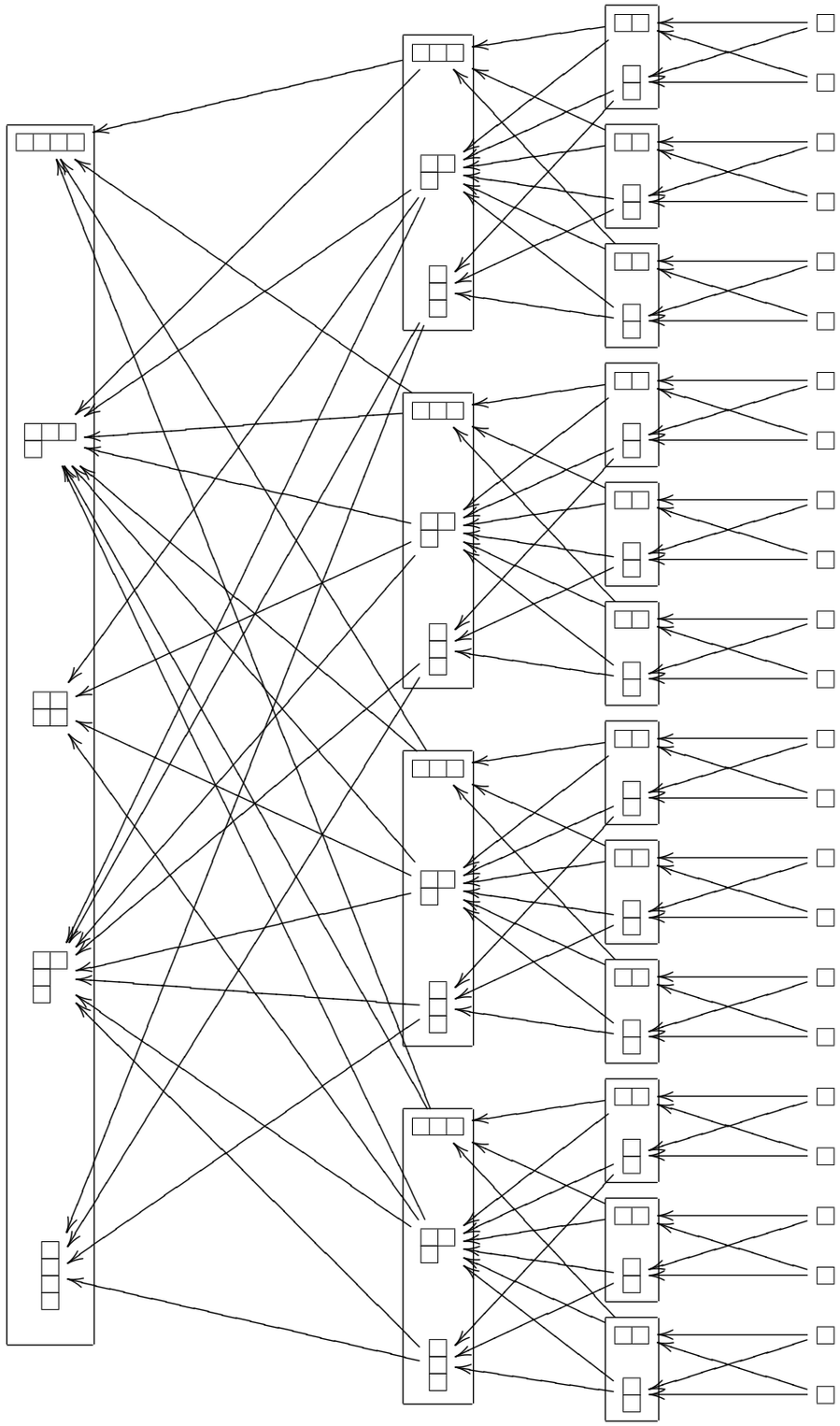
Michael Clausen

$\mathbb{S}_n$



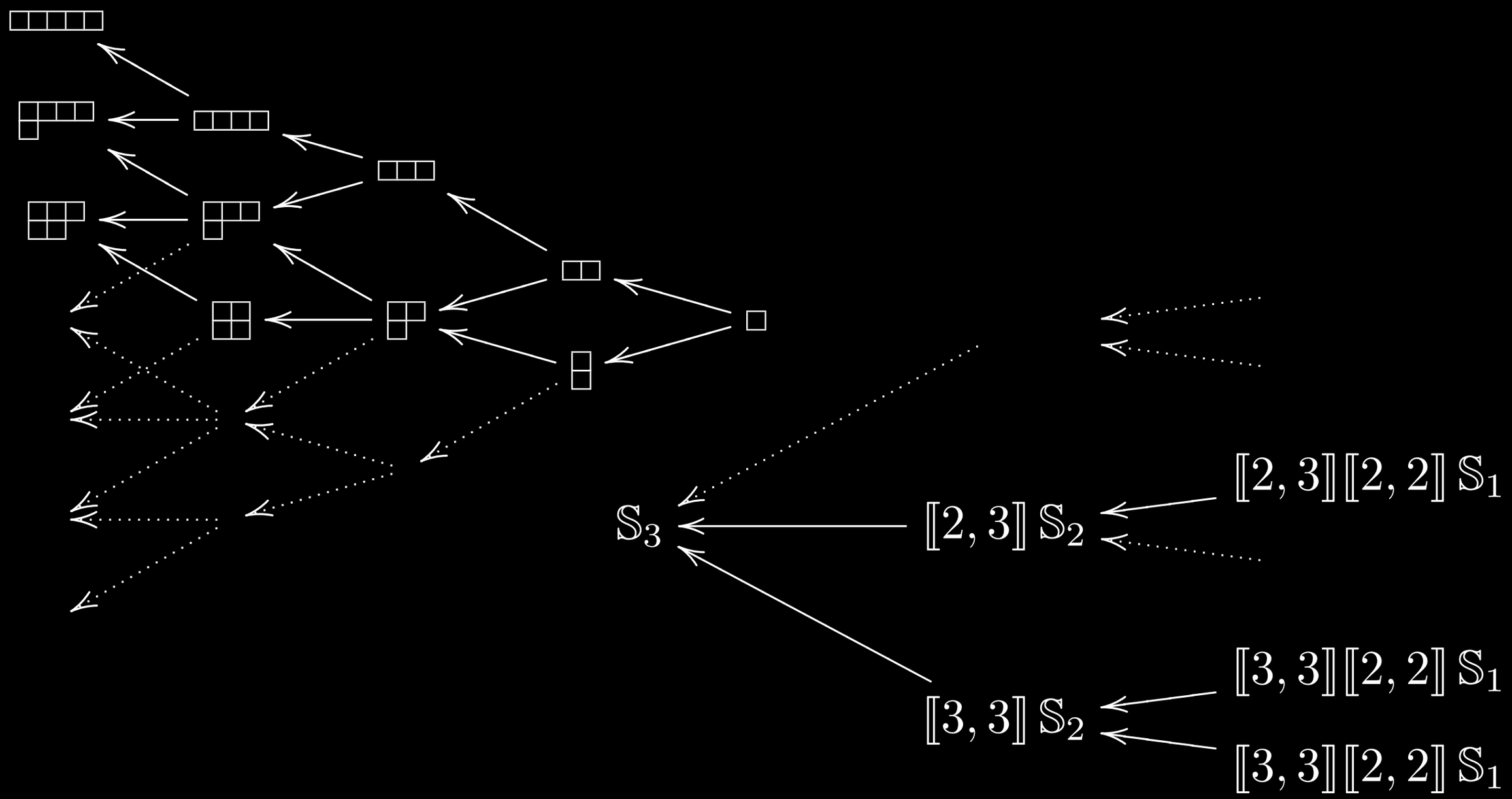


Bratelli diagram



$$n!^2 \to \frac{(n+1)\,n\,(n-1)}{3}\,n!$$

Sparse transforms



SnOB, FFT for the symmetric group

http://www1.cs.columbia.edu/~risi/SnOB/

Google

$S_n$ob

A C++ library for fast Fourier transforms on the symmetric group.

author: Risi Kondor, Columbia University (risi@cs.columbia.edu)

Development version as of August 23, 2006 (unstable!):

Documentation: [ps][pdf]
C++ source code: [directory]
BiBTeX entry: [bib]
Entire package: [tar.gz]

References:

1. Michael Clausen: **Fast generalized Fourier transforms.** Theoretical Computer Science **67(1)**: 55-63, 1989.
2. David K. Maslen and Daniel N. Rockmore: **Generalized FFTs --- a survey of some recent results.** Proceedings of the DIMACS Workshop on Groups and Computation, 1997. [ps]
3. K.-J. Kueh, T. Olson, D. Rockmore and K.-S. Tan: **Nonlinear approximation theory on finite**

http://www.gatsby.ucl.ac.uk/~risi/SnOB

Conclusions

Harmonic analysis generalizes naturally to non-commutative groups of transformations.

Strong connections to well developed branches of mathematics.

Rich algorithmic side: FFTs, etc.