

# Contents

## Preface

## Acknowledgements

## List of abbreviations

## Introduction

### References

## 1 Multivariate analysis techniques for representing graphs and contingency tables

### 1.1 Quadratic placement problems for weighted graphs and hypergraphs

#### 1.1.1 Representation of edge-weighted graphs

#### 1.1.2 Representation of hypergraphs

#### 1.1.3 Examples for spectra and representation of simple graphs

### 1.2 SVD of contingency tables and correspondence matrices

### 1.3 Normalized Laplacian and modularity spectra

### 1.4 Representation of joint distributions

#### 1.4.1 General setup

#### 1.4.2 Integral operators between $L^2$ spaces

#### 1.4.3 When the kernel is the joint distribution itself

#### 1.4.4 Maximal correlation and optimal representations

### 1.5 Treating nonlinearities via reproducing kernel Hilbert spaces

#### 1.5.1 Notion of the reproducing kernel

#### 1.5.2 RKHS corresponding to a kernel

#### 1.5.3 Two examples of an RKHS

#### 1.5.4 Kernel – based on a sample – and the empirical feature map

### References

## 2 Multiway cuts and spectra

### 2.1 Estimating multiway cuts via spectral relaxation

#### 2.1.1 Maximum, minimum, and ratio cuts of edge-weighted graphs

#### 2.1.2 Multiway cuts of hypergraphs

### 2.2 Normalized cuts

### 2.3 The isoperimetric number and sparse cuts

### 2.4 The Newman–Girvan modularity

#### 2.4.1 Maximizing the balanced Newman–Girvan modularity

2.4.2 Maximizing the normalized Newman–Girvan modularity	81
2.4.3 Anti-community structure and some examples	84
2.5 Normalized bicuts of contingency tables	88
References	91
<b>3 Large networks, perturbation of block structures</b>	<b>96</b>
3.1 Symmetric block structures burdened with random noise	96
3.1.1 General blown-up structures	99
3.1.2 Blown-up multipartite structures	109
3.1.3 Weak links between disjoint components	112
3.1.4 Recognizing the structure	114
3.1.5 Random power law graphs and the extended planted partition model	121
3.2 Noisy contingency tables	124
3.2.1 Singular values of a noisy contingency table	127
3.2.2 Clustering the rows and columns via singular vector pairs	129
3.2.3 Perturbation results for correspondence matrices	132
3.2.4 Finding the blown-up skeleton	138
3.3 Regular cluster pairs	142
3.3.1 Normalized modularity and volume regularity of edge-weighted graphs	142
3.3.2 Correspondence matrices and volume regularity of contingency tables	150
3.3.3 Directed graphs	156
References	157
<b>4 Testable graph and contingency table parameters</b>	<b>161</b>
4.1 Convergent graph sequences	161
4.2 Testability of weighted graph parameters	164
4.3 Testability of minimum balanced multiway cuts	166
4.4 Balanced cuts and fuzzy clustering	172
4.5 Noisy graph sequences	175
4.6 Convergence of the spectra and spectral subspaces	177
4.7 Convergence of contingency tables	182
References	187
<b>5 Statistical learning of networks</b>	<b>189</b>
5.1 Parameter estimation in random graph models	189
5.1.1 EM algorithm for estimating the parameters of the block-model	189
5.1.2 Parameter estimation in the $\alpha$ and $\beta$ models	192
5.2 Nonparametric methods for clustering networks	197
5.2.1 Spectral clustering of graphs and biclustering of contingency tables	199
5.2.2 Clustering of hypergraphs	201

5.3 Supervised learning	203
References	205
<b>Appendix A Linear algebra and some functional analysis</b>	<b>207</b>
A.1 Metric, normed vector, and Euclidean spaces	207
A.2 Hilbert spaces	209
A.3 Matrices	217
References	233
<b>Appendix B Random vectors and matrices</b>	<b>235</b>
B.1 Random vectors	235
B.2 Random matrices	239
References	245
<b>Appendix C Multivariate statistical methods</b>	<b>246</b>
C.1 Principal component analysis	246
C.2 Canonical correlation analysis	248
C.3 Correspondence analysis	250
C.4 Multivariate regression and analysis of variance	252
C.5 The $k$ -means clustering	255
C.6 Multidimensional scaling	257
C.7 Discriminant analysis	258
References	261

plan to use existing knowledge on graph spectra and experiences in analyzing social data in the form of hypergraphs, weighted graphs, and contingency tables. At the end of the 1980s, together with my PhD advisor, Gábor Tusnády, we used spectral methods for a binary clustering problem, where the underlying matrix turned out to be the generalization of the graphs' Laplacian to hypergraphs. Then we defined the Laplacian for multigraphs and edge-weighted graphs, and went beyond the expanders by investigating gaps within the spectrum and used eigenvectors corresponding to some structural eigenvalues to find clusters of vertices. We also considered minimum multiway cuts with different normalizations that were later called ratio- and normalized cuts. In the 1990s, spectral clustering became a fashionable area and a lot of papers in this topic appeared, sometimes redefining or modifying the above notions, sometimes having a numerical flavor and suggesting algorithms without rigorous mathematical explanation.

At the turn of the millennium, thanks to the spread of the World Wide Web and the human genome project, there was a rush to investigate evolving graphs and random situations different from the classical Erdős–Rényi one. Graph theorists, for example, László Lovász and co-authors, considered convergence of graph sequences and notable graph parameters, bringing – without their voition – statistical concepts into this discrete area. They also started to consider noisy graph and contingency table sequences, and testability of some balanced versions of the already defined minimum multiway cut densities. Meanwhile, physicists introduced other measures