Social Sequence Analysis

Methods and Applications

BENJAMIN CORNWELL
Cornell University

CAMBRIDGE
UNIVERSITY PRESS

Contents

List of Figures page xii
List of Tables xv
Preface xvii
Acknowledgment xix

PART I. INTRODUCTION

1 Sequence Analysis in the Social Sciences 3
1.1 Timing and Context of the Book 4
1.2 Contributions 7
1.3 Audience and Scope 9
1.4 Related References 14
1.5 Plan of the Book 16

PART II. THEORETICAL BACKGROUND

2 Theoretical Foundations of Social Sequence Analysis 21
2.1 What Are Social Sequences? 21
2.2 The Structural Origins of Social Sequences 22
2.3 Why Social Sequences Matter 25
2.4 Origins of Social Sequence Analysis Techniques 32
2.4.1 Narrative Positivism 33
2.4.2 The Relational Nature of Social Phenomena 36
2.4.3 The Development of Whole-Sequence Analysis Methods 38
2.4.4 Pivotal Criticisms 41
2.5 The Second Wave of Social Sequence Analysis 45
2.5.1 Sequences as Networks 46
2.5.2 Microsequence Analysis 52
2.6 Looking Ahead 55

PART III. SOCIAL SEQUENCE ANALYSIS CONCEPTS AND TECHNIQUES

3 Sequence Analysis Concepts and Data 59

3.1 Sequence Properties 59

3.1.1 Positions 60

3.1.2 Elements 60

3.1.3 Substructures 61

3.1.4 Subjects 62

3.2 Key Issues in Sequence Data 62

3.2.1 Recurrence 63

3.2.2 Ties and Multidimensionality 65

3.2.3 Gaps 66

3.3 The Sequence Universe 67

3.3.1 Size 67

3.3.2 Boundary Specification 69

3.4 Sequence Data 70

3.4.1 Data Collection 70

3.4.2 Element-Position Sampling 73

3.4.3 Units of Measure 74

3.5 Data Used in This Book 75

3.5.1 The Survey of Health, Ageing, and Retirement in Europe (SHARE) 75

3.5.2 The Correlates of War Global Trade Network 76

3.5.3 The Davis, Gardner, and Gardner Deep South *Study* 78

3.5.4 The Multinational Time Use Study (MTUS) 79

3.5.5 The American Time Use Survey (ATUS) 80

4 Detecting Sequence Structure 83

4.1 Descriptive Analysis 83

4.1.1 Sequence Length 84

4.1.2 Element Frequencies 84

4.1.3 n-*grams* 85

4.1.4 Position Reports 85

4.2 Describing Stochastic Patterns 86

4.2.1 Transition Matrices 86

4.2.2 Markov Chains 91

4.3 Sequential Connection 91

4.3.1 Empirical Illustration: Gender Differences in the Parenthood-Stress Link 93

4.4 Stationarity 94

4.4.1 Empirical Illustration: Stationarity in the World System 96

4.5 Spells 97

4.6 Homogeneity 98
4.7 On Using Summary Statistics and Tests 99
4.8 Visual Aids 100
4.8.1 Transition Bubble Graphs 100
4.8.2 State Transition Diagrams 101
4.8.3 Sequence Index Plots 102
4.8.4 State Distribution Graphs 104
4.8.5 Tempograms 104
4.8.6 Sequence-Network Diagrams 105
4.8.7 On the Use of Color and Grayscale 105
4.9 Looking Ahead 108

5 Whole-Sequence Comparison Methods 109
5.1 Sequence Alignment 110
5.1.1 Sequence Alignment Operations 110
5.1.2 Operation Costs and Distance 111
5.2 Classical Optimal Matching (OM) 111
5.2.1 Finding the Optimal Solution 113
5.2.2 The Needleman–Wunsch Algorithm 114
5.3 Basic Operation Cost Regimes 115
5.3.1 Levenshtein Distance 115
5.3.2 Levenshtein II Distance 115
5.3.3 Hamming Distance 116
5.4 Considerations in Setting Operation Costs 116
5.4.1 Order versus Timing 116
5.4.2 Time Warping 118
5.5 Criticisms of Classical OM 119
5.5.1 Theory/Method Fit 120
5.5.2 Arbitrary Operation Costs 120
5.6 Improvements on Classical OM 122
5.6.1 Variable Substitution Costs 122
5.6.2 Distance Normalization 124
5.6.3 Reference Sequence Comparison 125
5.6.4 Spell-Adjusted Distances 126
5.6.5 The Dynamic Hamming Method 127
5.7 Detecting Whole Sequence Patterns 129
5.7.1 Hierarchical Cluster Analysis 130
5.7.2 Identifying Sequence Classes 133
5.7.3 Assessing Validity and Reliability 136
5.7.4 Describing Sequence Classes 139
5.8 Recent Advances in Sequence Comparison 143
5.8.1 Multidimensional Sequence Analysis 144
5.8.2 Two-Stage Optimal Matching (2SOM) Analysis 146
5.8.3 Transition Sequence Analysis 147

5.8.4 Nonalignment Techniques 148
5.8.5 Beyond Cluster Analysis 150

PART IV. NEW DIRECTIONS IN SOCIAL SEQUENCE ANALYSIS

6 Network Methods for Sequence Analysis 155
6.1 Theoretical Rationale 156
6.1.1 Networks of Sequences 156
6.1.2 Sequences as Bases of Affiliations 160
6.2 Network Concepts and Terms 163
6.2.1 Network Matrices and Components 163
6.2.2 Directed and Undirected Networks 164
6.2.3 One- and Two-Mode Networks 165
6.2.4 Subject Comembership and Element Overlap 165
6.3 Sequence-Network Construction 167
6.3.1 Nonrecurrent Sequence Networks 167
6.3.2 Recurrent Sequence Affiliation Networks 170
6.3.3 Directed Sequence Networks 171
6.3.4 Multidimensional Sequence Networks 172
6.4 Visualization 172
6.4.1 Relationships among Subjects 173
6.4.2 Two-Mode Ordered Event Networks 174
6.4.3 Two-Mode Recurrent Sequences 176
6.4.4 Directed Sequence Networks 179
6.5 Structural Measures for Sequence Networks 184
6.6 Identification of Subject/Element Subsets 189
6.6.1 Subsets in Sequence Affiliation Networks 190
6.6.2 Element Subsets in Directed Sequence Networks 193
6.6.3 Sequence Motifs 195
6.7 Event Structure Analysis 196
6.8 Statistical Network Models for Sequential Events 200
6.8.1 Exponential Random Graph (p) Models* 200
6.8.2 ERGMs for Ordered Nonrecurrent Sequence Networks 201
6.9 Dynamic Sequence Networks 202
6.9.1 Subject- and Element-Level Change 203
6.9.2 Sequence Subset Evolution 206
6.9.3 Whole Sequence-Network Evolution 207
6.10 Conclusion 209

Social Microsequence Analysis 210
7.1 An Exemplary Sequence Context 211
7.2 The Elements of Social Microsequences 214
7.2.1 Relationship to Microinteraction Sequences 215

7.2.2 Elements and Positions 215
7.2.3 Data Collection and Availability 216
7.3 Nonnetwork Sequence Approaches 217
7.3.1 OM Sequence Classification 220
7.3.2 Transition and Switching Analysis 222
7.4 Sequence-Network Approaches 226
7.4.1 Synchrony 227
7.4.2 Measurement of Synchrony 230
7.4.3 Identifying Sources of Synchrony 238
7.4.4 Routine 241
7.4.5 Measurement and Visualization of Routine 244
7.5 Next Steps 251

PART V. CONCLUSIONS

8 The Promise of Social Sequence Analysis 255
8.1 Limitations 256
8.2 Future Research 257
8.2.1 Routine and Routinization 257
8.2.2 Sequence Networks, Network Sequences 258
8.2.3 Sequential Statistical Inference 259
8.2.4 Data Collection 261

Appendix A Recent Whole-Sequence Pattern Analyses 263
Appendix B Linkage Criteria for Agglomerative Hierarchical Clustering 270
References 275
Index 303

Figures

1.1 How to read this book *page* 12

4.1 Frequency state transition matrix (N) 87

4.2 Probability state transition matrix (P) 88

4.3 Frequency state transition matrix for SHARELIFE life events data 90

4.4 Probability state transition matrix for SHARELIFE life events data 90

4.5 Transition bubble graph showing probabilities of first-time life-course transitions among SHARELIFE respondents (*N* = 2,191) 101

4.6 State transition diagram graph showing sequential relationships among first-time life-course events among SHARELIFE respondents (*N* = 2,191) 102

4.7 Sequence index plot showing the number of SHARELIFE respondents (*N* = 2,191) who reported experiencing first-time life-course events in different sequence orders 103

4.8 State distribution graph showing the percent of SHARELIFE respondents (*N* = 2,191) reporting a given element at each sequence position 105

4.9 Sequence index plot of life event sequences from Figure 4.7 in grayscale 106

4.10 Grayscale sequence index plot of life event sequences from Figure 4.9 stretched vertically 107

5.1 Frequency state transition matrix for world-system positions 123

5.2 Dendrogram for hierarchical clustering of 42 unique world-system-position sequences from 129 countries between 1967 and 2006 133

5.3 Line graph showing the relationship between the number of clusters at given dissimilarity thresholds in the hierarchical clustering of world-system-position sequences 135
5.4 Tempograms showing the proportions of workers who were working at specific time points throughout the day in five weekday clusters, based on analysis of data on 8,997 individuals from the 2008–2010 American Time Use Surveys 141
5.5 Sequence index plots showing sequences of the 129 countries in each of the five world-system-position clusters 143
6.1 Narrative network showing the structure of the identity history of a Nazi 159
6.2 Affiliation matrix (A) showing women's involvement in different events, as reported in the *Deep South* study 168
6.3 Comembership matrix (N) showing overlap between women with respect to involvement in different events in the *Deep South* study 169
6.4 Network showing the strengths of relationships among the women in the Davis, Gardner, and Gardner *Deep South* study 174
6.5 Bipartite network showing which events were attended by which of the women in the *Deep South* study 175
6.6 Hypothetical two-mode sequence network 177
6.7 Two-mode sequence network depicting activity sequences of two parents and their two children between 5 P.M. and midnight on a Wednesday evening in October 2000 178
6.8 Annotated directed activity sequence network for two parents and their two children between 5 P.M. and midnight on a Wednesday evening in October 2000 180
6.9 Node-free, variable-path-width depiction of the activity sequences of a U.K. family 182
6.10 Colorized variable-path-width depiction of the activity sequences of a U.K. family 183
6.11 Directed activity sequence network for two parents and their two children between 5 P.M. and midnight on a Wednesday evening in October 2000, with time-equivalent position-elements tiled along the *x*-axis 184
6.12 Overview of useful structural measures for two-mode sequence networks 186
6.13 Overview of useful structural measures for one-mode sequence networks 187

6.14 Affiliation matrix (A) from the *Deep South* study, rearranged to indicate block membership 192
6.15 Event structure analysis (ESA) diagram showing connections among elements of the labor struggle in the meatpacking industry during the Reagan era 199
6.16 Evolution of a firm's investment network 204
6.17 Three-hour-long snapshots of a U.K. family's two-mode sequence network, 5:00–8:00 P.M. 206
7.1 Annotated microsequence diagram showing how one woman spent the day on a Monday in May 2011 219
7.2 Tempograms showing the percent of respondents in the 2003–2011 ATUS who reported different types of contact at each 5-minute interval between 6 A.M. and midnight, shown separately for each of five clusters ($N = 2{,}853$) 221
7.3 Transition probability matrix showing switches between different types of social contacts in a single workday among working mothers in the 2003–2011 ATUS ($N = 4{,}658$) 226
7.4 Transition probability matrix showing switches between different types of social contacts in a single workday among working fathers in the 2003–2011 ATUS ($N = 6{,}752$) 226
7.5 The position of work activity throughout the day in the Spanish Daily Sequence Network 237
7.6 Annotated sequence motif illustrating points of synchrony in Spain in 2003 (uncommon transitions not shown) 240
7.7 Histogram showing distribution of average degree of routine (in minutes) between pairs of weekdays among respondents in the 2005 Time Use Survey of the Netherlands ($N = 1{,}807$) 250
B.1 The types of relationships between cases that are considered given different linkage criteria for merging separate clusters in hierarchical clustering 271

Tables

4.1 Transition matrices showing frequency of transition from having first child to experiencing first period of major stress among parents in the SHARELIFE data *page* 93

4.2 Transition matrices showing movement within the world system between 1967–1986 and 1987–2006 96

5.1 The relationship between the operation cost regime and the emphasis on features being used as a basis for sequence comparison 118

5.2 Country membership in the five world-system clusters 144

7.1 Extent of interpersonal and generalized synchrony in individuals' activity sequences as reported in the 2002–2003 Spanish Time Use Survey (in minutes and percent), by socio demographic characteristics (*N* = 20,136) 234

7.2 Extent of routine in individuals' activity sequences as reported in the 2005 Time Use Survey of the Netherlands (in minutes and percent), by socio demographic characteristics (*N* = 1,807) 249

A.1 Studies that have analyzed whole-sequence patterns using OM and other methods since 2000 264