

Sequence Analysis & Extensions to the Dyadic Context

Allison Dunatchik[†]

December 7, 2022

QM Methodology Working Group, Fall 2022

[†]Materials partly based SA lecture by Xi Song



ELSEVIER

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch



Sequence analysis: Its past, present, and future

Tim F. Liao^{a,*}, Danilo Bolano^b, Christian Brzinsky-Fay^c, Benjamin Cornwell^d,
Anette Eva Fasang^e, Satu Helske^f, Raffaella Piccarreta^b, Marcel Raab^g,
Gilbert Ritschard^h, Emanuela Struffolinoⁱ, Matthias Studer^h



Using Sequence Analysis to Quantify How Strongly Life Courses Are Linked

Tim F. Liao

University of Illinois

Agenda

- ▶ What is sequence analysis?
 - ▶ Motivation
 - ▶ History
 - ▶ Typical approaches
 - ▶ Recent developments

- ▶ Extending sequence analysis to the polyadic context
 - ▶ Method proposed in Liao 2021
 - ▶ Fasang and Raab 2014
 - ▶ R codes using TraMineR and seqpolyads

What Is Sequence Analysis?

Sequence analysis is the analysis of **categorical** sequences of events to model entire life paths.

In contrast to typical linear regression approaches, which focus on particular event states in comparison to others, SA focuses on **entire life processes**.

- ▶ The **timing** of events;
- ▶ The **duration** of time spent in various states;
- ▶ The **order** in which events occur.

SA and Life Course Research

In analysing trajectories of categorical states, SA is closely tied to the core theoretical ambitions of **life course research**.

- ▶ Trajectories of categorical states, not just metric outcomes, are of central importance in studying the life course.
- ▶ Life courses exemplify *process outcomes*, defined by Andrew Abbot as "... long-run stabilities established by myriads of individual events... it is the whole walk that is the outcome."
(Abbot 2001, p176)

SA and Life Course Research

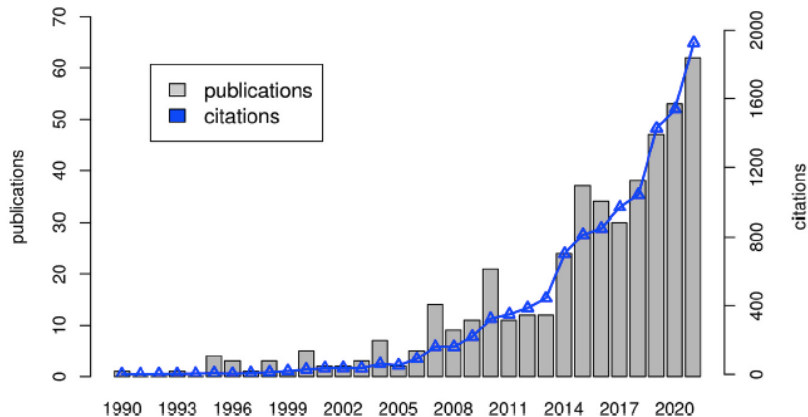


Figure 1: Journal publication and citation trends of SA applications in Social Science, 1990–2021, web of science (Liao et al 2022)

SA and Life Course Research

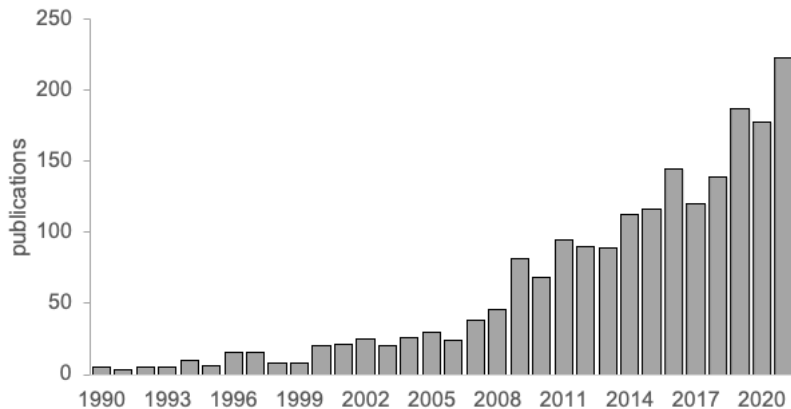


Figure 2: Journal publication trends for "Life Course Perspective" in Social Science, 1990–2021, web of science (own analysis)

Prior Examples

- ▶ Class career sequences (Halpin and Chan 1998)
- ▶ Transitions from school to work (McVicar and Anyadike-Danes 2002)
- ▶ Transition to adulthood (Billari 2001)
- ▶ Employment mobility and career patterns (Abbott and Hrycak 1990; Blair-Loy 1999)
- ▶ Employment patterns (Killewald and Zhuo 2019)
- ▶ Scheduling within dual-earner couples (Lesnard 2008)
- ▶ Partnership and fertility dynamics (Potarca et al. 2010)
- ▶ Ideal and experienced relationship sequences (Frye and Trinitapoli 2015)
- ▶ Retirement patterns (Fasang 2009)
- ▶ Work-family trajectories (Aasave et al. 2007)
- ▶ Linked life-course trajectories (Liao 2021)

What Kind of Research Questions Can Be Answered by a Sequence Analysis?

According to [Abbott \(1990\)](#)

1. To identify **typical** sequential patterns (and population heterogeneity in sequential patterns)
2. To examine **covariates** that predict sequence patterns
3. To examine the **effect** of a given sequence pattern on other outcomes

Five Components of a Typical Sequence Analysis

1. **Visualizing** key sequences via **sequence plots**
2. **Describing** key sequences via **aggregated measures**
3. **Comparing** sequences via **distance measures**
4. **Grouping** sequences into **clusters**
5. **Associating** patterns with other variables within **regression models**

1. Visualizing Sequences

Discovering and plotting (representative) sequences

- ▶ Index plot
- ▶ Density plot
- ▶ etc.¹

¹See [panelView](#) by Yiqing Xu, [ggseqplot](#) by Marcel Raab

Visualizing Sequences: Index plots

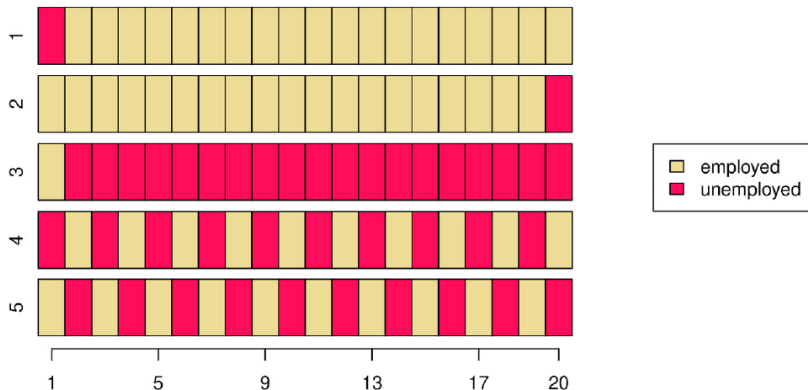


Figure 3: Five Example Sequences (Liao et al 2022)

Visualizing Sequences: Density plots

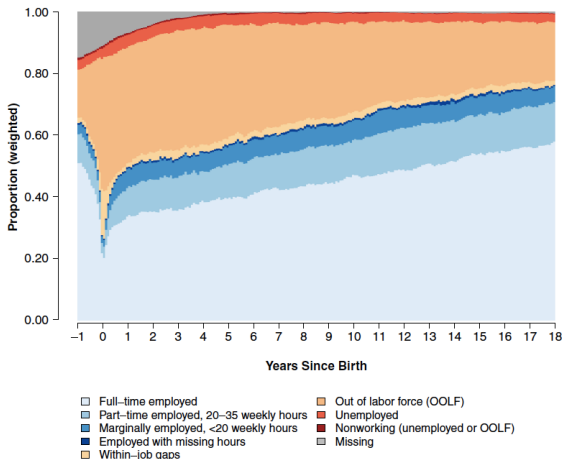


Figure 4: Mothers' Employment Status Distribution after First Birth
(Killewald and Zhuo 2019)

2. Describing Sequences via Aggregated Measures

- ▶ **Individual longitudinal** characteristics of sequences
 - ▶ length
 - ▶ time in each state
 - ▶ turbulence
 - ▶ complexity
 - ▶ etc.
- ▶ **Sequence transversal** characteristics by age point
 - ▶ transversal state distribution
 - ▶ modal state
- ▶ **Other** aggregated characteristics
 - ▶ transition rates
 - ▶ average duration in each state
 - ▶ sequence frequency

3. Comparing Sequences using Distance Measures

- ▶ Distance measures aim to quantify the extent to which two individuals followed dissimilar trajectories.
- ▶ Distances are calculated for all possible pairs of individuals in the data.
- ▶ There are a variety of different distance measures available, each of which carries its own assumptions and may be more or less sensitive to event **timing**, **duration** and **order**.

Optimal Matching (OM)

OM is the most commonly used distance measure.

- ▶ Measures dissimilarity between sequences by computing the minimum 'cost' required to transform one of the sequences into the other (via substitutions, insertions or deletions (indels) of elements of a sequence).

Key criticisms of OM (see e.g., Wu 2000; Levine 2000):

- ▶ Transformation costs impose important assumptions that may be disconnected sociological theory
 - ▶ e.g. what is the cost of transitioning from employment to unemployment? Are the costs of transitioning from employment truly symmetric?
- ▶ Low sensitivity to the ordering of events in a sequence.

Advances in SA: Alternative Distance Measures

Many new distance measures have been proposed to try to overcome the limitations of OM.

- ▶ In an extensive review of distance measures, Studer and Ritschard (2016) provide an assessment of the distance measure most sensitive to each of the three analysis sequence dimensions.
 - ▶ **Timing:** Hamming distance
 - ▶ **Duration:** OM
 - ▶ **Order:** SVRspell, OM of transitions and OM of spells

4. Grouping Sequences in to Clusters

After constructing a matrix of pairwise dissimilarities, researchers often use cluster analysis to identify groups of sequences with similar characteristics.

- ▶ Hierarchical clustering algorithms (e.g. Ward algorithm)
- ▶ Partitioning around medoids algorithm (PAM)

4. Grouping Sequences in to Clusters

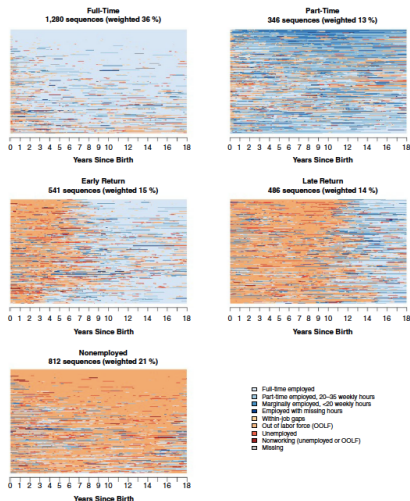


Figure 5: Mothers' employment sequences after first birth by cluster, (Killewald and Zhuo 2019)

5. Associations and Regressions

- ▶ Use clusters as outcomes: Multinomial regression to examine the characteristics that predict belonging to one cluster over another
- ▶ Use clusters as independent variables: Examine how cluster membership predicts some other future outcome.

Advances in SA: Extending SA to the Polyadic Context

- ▶ The life course principle of **linked lives** emphasizes that one individual's life can be and most often is embedded in the lives of family members and close relations.
- ▶ We may be interested in measuring the **intergenerational association** of life course patterns between parents and children or the **intrageneration connections** between spouses or siblings.
- ▶ Recent developments in SA seek to allow researchers to examine the connection between sequences within **dyads** or **polyads**.

Two Approaches to Dyadic SA

There are several approaches to dyadic (or polyadic) SA that have been proposed (see Liao 2021 for discussion). Two key approaches include:

- ▶ Multichannel SA
 - ▶ Extends the standard OM SA approach to consider multiple linked sequences.
- ▶ Liao 2021 method
 - ▶ Generates two measures (U and V) that quantify how similar sequences are within a polyad by comparing the distances (using any distance measure) between members of an observed polyad against a set of randomly generated polyads.

Multichannel SA for Analysis of Dyads

Fasang and Raab (2014) apply multichannel SA to examine the association between parent-child family formation sequences using the Longitudinal Survey of Generations.

1. Extend OM to calculate pairwise distances between each dyad in the data.

Extend OM to calculate pairwise distances between each dyad in the data

Age	16	17	18	19	20
Dyad <i>A</i>	[MNC MNC]	[MNC MNC]	[M1C M1C]	[M1C M1C]	[M2C M2C]
Dyad <i>B</i>	[MNC SNC]	[MNC SNC]	[M1C SNC]	[M1C SNC]	[M2C SNC]

Figure 6: Two hypothetical dyadic intergenerational sequences A and B from ages 16-20 (Fasang and Raab 2014)

Multichannel SA for Analysis of Dyads

Fasang and Raab (2014) apply multichannel SA to examine the association between parent-child family formation sequences using the Longitudinal Survey of Generations.

1. Extend OM to calculate pairwise distances between each dyad in the data.
2. Use cluster analysis to group dyads into three pre-hypothesized ideal types of dyads: Strong transmission, Moderated transmission and Contrast.

Use cluster analysis to group dyads into three ideal types

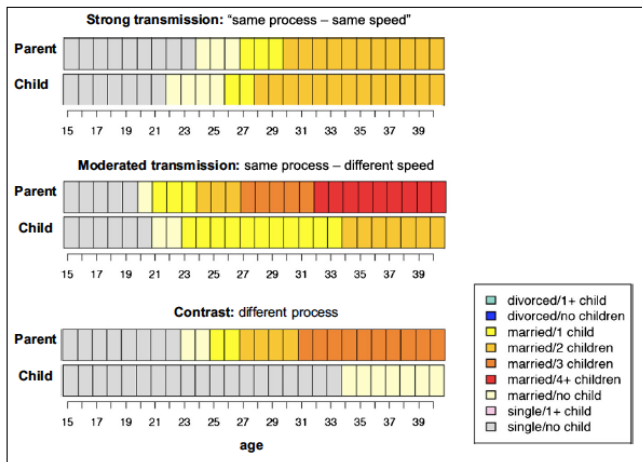


Figure 7: Representative Sequences of Intergenerational Family Formation Clusters (Fasang and Raab 2014)

Multichannel SA for Analysis of Dyads

Fasang and Raab (2014) apply multichannel SA to examine the association between parent-child family formation sequences using the Longitudinal Survey of Generations.

1. Extend OM to calculate pairwise distances between each dyad in the data.
2. Use cluster analysis to group dyads into three pre-hypothesized ideal types of dyads: Strong transmission, Moderated transmission and Contrast.
3. Use regression analysis to examine the factors that predict Strong transmission vs Moderated transmission vs Contrast.

Liao Method for SA of Dyads

Aims to quantify how similar sequences are within a polyad by comparing the distances between members of an observed polyad against a set of randomly generated polyads.

- ▶ U quantifies how much greater (in terms of the distance measure chosen) the members of a polyad resemble one another compared with members of randomly generated polyads.
- ▶ V quantifies the degree to which polyads are linked in terms of how much observed polyads outperform randomized ones.
- ▶ U and V can then be used in regression analysis to examine what factors are associated with greater intergenerational transmission of family formation.

Calculating U and V in 4 Steps

Consider the example from Fasang and Raab (2014) examining the intergenerational transmission of family formation sequences.

Let S_{ip} and S_{ic} indicate the family formation sequences for the parent (p) and child (c) in the i th of N dyads.

Step 1: Compute a distance vector D_i of the i th polyad, measuring the distance between each member pair within the polyad using a researcher-chosen dissimilarity measure.

For parent-child dyads, this is denoted by:

$$D_i = d(S_{ip}, S_{ic}) \text{ for } i=1 \text{ to } N$$

Calculating U and V in 4 Steps

Step 2: Compute distances between reassigned polyadic sequences randomly drawn from observed polyadic member sets.

For parent-child dyads, this is denoted by:

$$R_t = d(S_{ap}, S_{bc})$$

where a and b are instances of i and each t for $t= 1$ to T represents a cross-polyadic matching of randomly drawn members from each generation.

Calculating U and V in 4 Steps

Step 3: Repeat step 2 T number of times, with T being a large number, preferably $\geq 1,000$.

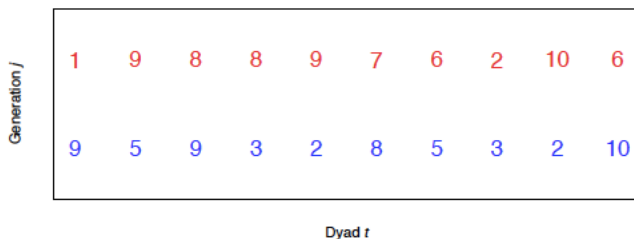


Figure 8: A diagram of 10 randomized dyads of two generations from computing R_t ($T=10$) (Liao 2021)

Calculating U and V in 4 Steps

Step 4: Calculate U_i and V_i .

To calculate U_i , subtract the observed D_i from the mean of R_t .

$$U_i = \frac{\sum_t R_t}{T} - D_i$$

This measure quantifies the extent to which family formation sequences among observed members of parent-child dyads are more similar than those among randomly generated parent-child dyads.

The more positive U_i is, the more similar the observed dyads are compared to randomly generated ones.

Calculating U and V in 4 Steps

Step 4: Calculate U_i and V_i .

Calculate V_i as the proportion out of T times that $D_i < R_t$.

This measure (ranging from 0-1), captures the degree of linkage between parent and child family formation sequences and can be thought of as a test statistic.

A V_i greater than 0.5 suggests that, more often than not, an observed dyad is more similar than randomly generated ones. A V_i of 0.95+ would suggest strong evidence of within-dyad similarity.

Dyadic SA: Application using LSOG Data

Liao (2021) uses LSOG data to re-examine the association between parent-child family formation sequences.

Four distance measures:

1. Hamming distance (timing)
2. CHI2 distance (duration)
3. SVRspell distance (order)
4. OMspell distance (neutral)

Dyadic SA: Descriptive Results

	Mean	SD
Timing U	0.238	4.434
Duration U	0.055	1.179
Order U	0.012	0.729
Timing V	0.479	0.298
Duration V	0.488	0.289
Order V	0.436	0.265

Figure 9: Descriptive statistics for LSOG measures of dyadic distance U and linkage V , using 1,000 simulated dyads (Liao 2021)

Dyadic SA: Descriptive Results

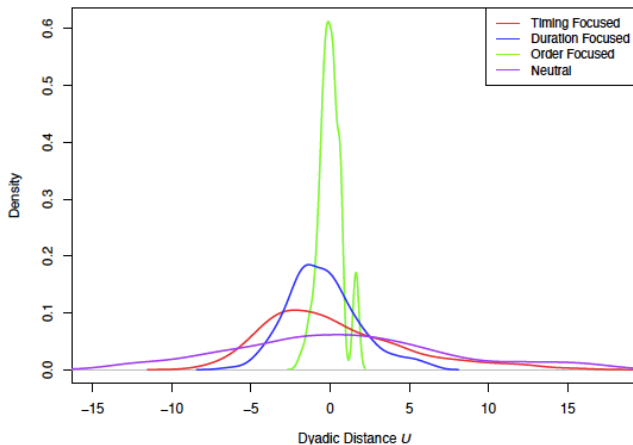


Figure 10: Density plots of LSOG dyadic distances U , using 1,000 simulated dyads (Liao 2021)

Dyadic SA: Descriptive Results

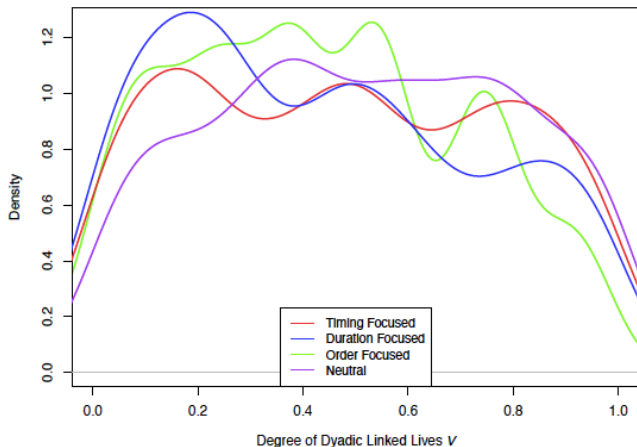


Figure 11: Density plots of LSOG degrees of dyadic linked lives V , using 1,000 simulated dyads (Liao 2021)

Dyadic SA: Regression Results

	Timing		Duration		Order	
	<i>U</i>	<i>V</i>	<i>U</i>	<i>V</i>	<i>U</i>	<i>V</i>
Gender constellation						
Mother-daughter (ref.)						
Father-son	-0.190 (-0.357)	0.002 (0.045)	-0.052 (-0.354)	0.002 (0.062)	0.053 (0.507)	0.016 (0.370)
Mother-son	-0.184 (-0.324)	-0.016 (-0.439)	-0.042 (-0.279)	-0.019 (-0.511)	0.086 (0.857)	0.034 (0.956)
Father-daughter	-0.063 (-0.131)	0.007 (0.216)	-0.026 (-0.197)	0.005 (0.176)	-0.028 (-0.367)	-0.014 (-0.544)
Dyad's age difference	0.525 [†] (7.960)	0.039 [†] (9.563)	0.129 [†] (7.244)	0.038 [†] (9.533)	0.046 [†] (4.078)	0.017 [†] (3.900)
Dyad's average years of education	-0.198 (-1.860)	-0.013 (-1.949)	-0.051 (-1.737)	-0.012 (-1.840)	0.003 (0.132)	0.004 (0.469)
Dyad's difference in years of education	-0.029 (-0.429)	-0.002 (-0.409)	-0.007 (-0.406)	-0.002 (-0.371)	0.033* (2.445)	0.011* (2.226)
Sibling position	-1.913 [†] (-6.640)	-0.143 [†] (-7.401)	-0.482 [†] (-6.275)	-0.138 [†] (-7.440)	-0.114 (-1.953)	-0.042 (-1.851)
Affectual solidarity scale (child-parent)	0.448 (1.931)	0.029* (2.028)	0.118 (1.863)	0.029* (2.056)	0.102* (2.554)	0.036* (2.469)
Constant	-8.664 [†] (-3.757)	-0.197 (-1.381)	-2.104** (-3.247)	-0.181 (-1.309)	-1.482 [†] (-3.687)	-0.146 (-1.016)
<i>R</i> ²	0.201	0.259	0.170	0.262	0.094	0.094

Note: *t* statistics are in parentheses. * $p < 0.05$, † $p < 0.01$.

Figure 12: Regression estimates with robust standard errors correcting for dyadic family clustering (N=391) (Liao 2021)

Assessing Liao's method for polyadic SA

Advantages

- ▶ Flexible
 - ▶ Suitable for all distance measures → able to capture multiple dimensions of similarities (timing, order, duration).
 - ▶ Can be applied to dyads, tetrads, pentads, etc.
- ▶ Provides a method of quantifying the degree similarities between dyads.

Disadvantages

- ▶ However, its difficult to discern the qualitative nature of differences within dyads.
- ▶ The interpretation of U is not intuitive.

Methods of Polyadic SA Continue to Develop...



Research in Social Stratification
and Mobility


Volume 82, December 2022, 100734



Linked labor force trajectories: Empirical evidence from dual- parent families in the United States and Australia

Irma Mooi-Reci ^a  , Tim F. Liao ^b, Matthew Curry ^c

[Show more](#) 

 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.rssm.2022.100734>

[Get rights and content](#)



Estimation Software

For general SA

- ▶ R ([TraMineR](#)) and STATA ([SQ](#)) plugins
- ▶ Handles missing data
- ▶ Handles sample weights
- ▶ Does not require fixed time spacing of measurements

For Liao's polyadic extension

- ▶ R ([seqpolyads](#) function in the [TraMineRextras](#) package)

Thank you!
adunat@sas.upenn.edu

Appendix

OM Transition Costs

	Full-Time Employed	Part-Time Employed, 20-35 Weekly Hours	Marginally Employed, <20 Weekly Hours	Employed With Missing Hours	Within-Job Gaps	Out of Labor Force (OOLF)	Unemployed	Nonworking (unemployed or OOLF)	Missing
Full-Time Employed	0	2	4	2	7	9	8	8.5	5
Part-Time Employed, 20-35 Weekly Hours	2	0	2	2	5	7	6	6.5	5
Marginally Employed, <20 Weekly Hours	4	2	0	2	3	5	4	4.5	5
Employed With Missing Hours	2	2	2	0	5	7	6	6.5	5
Within-Job Gaps	7	5	3	5	0	2	2	2	5
Out of Labor Force (OOLF)	9	7	5	7	2	0	1	0.5	5
Unemployed	8	6	4	6	2	1	0	0.5	5
Nonworking (unemployed or OOLF)	8.5	6.5	4.5	6.5	2	0.5	0.5	0	5
Missing	5	5	5	5	5	5	5	5	0

Figure 13: Substitution Cost Matrix (Killewald and Zhuo 2019)

Multichannel OM Transition Costs

	SNC	MNC	DNC	SC	DC	M1C	M2C	M3C	M4C
Parent Generation									
SNC	0								
MNC	1.87	0							
DNC	2	1.49	0						
SC	2	2	2	0					
DC	2	2	2	1.98	0				
M1C	1.99	1.57	1.94	1.98	1.93	0			
M2C	2	2	2	1.99	1.97	1.66	0		
M3C	2	2	2	2	1.91	2	1.85	0	
M4C	2	2	2	2	1.96	2	2	1.90	0
Child Generation									
SNC	0								
MNC	1.91	0							
DNC	2	1.88	0						
SC	2	2	2	0					
DC	2	2	1.97	2	0				
M1C	2	1.81	1.99	1.92	1.92	0			
M2C	2	2	2	1.98	1.94	1.82	0		
M3C	2	2	2	2	1.96	2	1.95	0	
M4C	2	2	2	2	1.97	2	2	1.94	0

Figure 14: Data-driven Substitution Cost Matrix Based on Transition Rates (Fasang and Raab 2014)

Randomization for Polyadic SA

1. Sequence-conditional random sequence generation: sequences of length l are randomly drawn from the observed set of polyadic members. Using this mechanism preserves the meaningful order of states and is useful when certain states cannot precede certain other states.
2. Sequence-conditional random state generation: sequences of length l are randomly drawn from the whole set of states from the observed sequences under consideration with state replacement within a selected sequence. Each sequence is randomly selected first before a random reshuffle of the states within the selected sequence. This mechanism can be useful for sequences with no logical orders for the list of states.