L I V E S
W O R K I N G
P A P E R
2 0 1 3 / 29

TITLE

Rendering the order of life events

Research paper

Author

Reto Bürgin and Gilbert Ritschard

http://dx.doi.org/10.12682/lives.2296-1658.2013.29
ISSN 2296-1658

FNSNF
Swiss National Science Foundation

The National Centres of Competence in Research (NCCR) are a research instrument of the Swiss National Science Foundation.

LIVES
Swiss National Centre of Competence in Research

Author

Reto Bürgin and Gilbert Ritschard

Abstract

This article proposes a decorated parallel coordinate plot for longitudinal categorical data, featuring a jitter mechanism revealing the diversity of observed longitudinal patterns and allowing the tracking of each individual pattern, variable point and line widths reflecting weighted pattern frequencies, the rendering of simultaneous events, and different flter options for highlighting typical patterns. The proposed visual display has been developed for describing and exploring the temporal ordering of events, but it can be equally applied to other types of longitudinal categorical data. Alongside the description of the principle of the plot, we demonstrate the scope of the plot with two real applications.

Keywords

Sequence analysis | Event sequences | State sequences | Longitudinal categorical data | Exploratory data analysis | Graphical statistics | Visualization | Multiple time series plot

Author's affiliation

Institute for Demographic and Life Course Studies, University of Geneva

Correspondence to

reto.buergin@unige.ch | gilbert.ritschard@unige.ch

1 Introduction

The study of longitudinal categorical data enjoys great popularity in social sciences, for example to track the behaviour of consumers, to examine the cognitive processing of students or to follow professional careers. Longitudinal categorical data typically track chronologically ordered categorical values of multiple individuals over a time interval. Such data has been classified in a variety of different types. A useful distinction to situate this article is between *state sequences* and *event sequences* (Ritschard, Gabadinho, Studer, & Müller, 2009). A *state*, such as being jobless, defines the situation of an individual which can be measured at any time and which lasts for a certain period. An *event*, such as ending a job or getting married, occurs at a given time point. An event does not persist, but provokes, possibly in conjunction with other events, a state change. Two major aspects differentiate event sequences from state sequences. First, in state sequences the position of the recorded value conveys time information so that we can know duration from the difference between two positions. This is not the case in event sequences. Secondly, while each individual can only be in one single—possibly complex—state at each time point, multiple events can occur simultaneously. State sequences are nicely rendered by stacking colored bars or line segments each of length proportional to the duration in the corresponding state. Since events do not have duration, event sequences are much harder to represent graphically. This article introduces a new plot primarily intended for such event sequences, the aim being to render the ordering of the elements in the sequences including possible simultaneous occurrences (Figure 1, right panel). Although the plot can optionally reflect time, it basically requires only information about the ordering of the observed values and could, therefore, be used for rendering any type of sequences of categorical elements.

The analysis of event sequences can concern three main different aspects: the *timing*, the *sequencing* and the *quantum* of events (Billari, Fürnkranz, & Prskawetz, 2006). As demonstrated by Figure 1, the proposed plot is especially useful for uncovering sequencing patterns. In the right panel plot, the first coordinate gives the first event—or first set of simultaneous events—in the sequence, the second coordinate the next event, and so on. This right panel plot clearly exhibits that, for the 1930-39 birth cohort, the first union most often coincides with the first marriage, and that the first birth most often occurs only after both the leaving home and the marriage. Incorporating the timing information often renders the plot too cluttered to be really useful as shown by the left panel. Since it allows for multiple points—the simultaneous events—on any of the parallel coordinates, the plot is not a true, but a pseudo parallel coordinate representation.

Event data organization An event sequence can be defined as a strictly ordered list of transitions (sometimes known as transactions), where each transition is itself a set of simultaneous unordered events. The necessary information about the order of the events can be stored in vertical tabular form as in Zaki (2001) with one row for each observed event and at least three columns: A first column for the individual's identity label, a second one for the event name (the categorical value) and a third one reporting the rank order of occurrence of the transition the event belongs to. Table 1 shows an example of such a vertical tabular organization where we have also reported the time stamp (Age) from which we derived the rank order of occurrence.

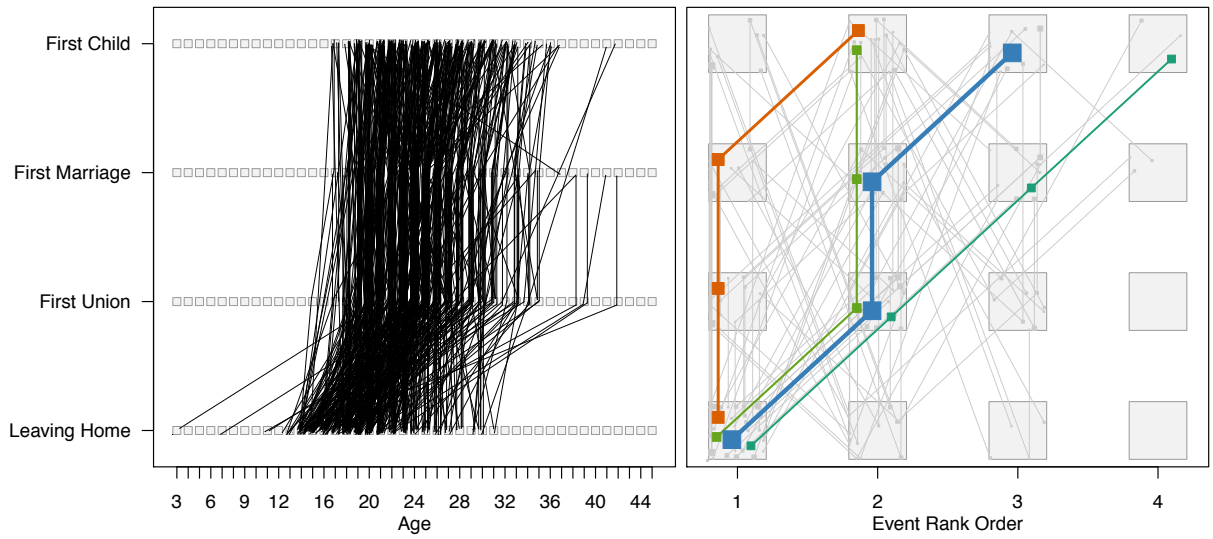

Figure 1: Parallel coordinate plot of Scandinavian family life events of the 1930-39 birth cohort. Left panel, alignment on event time stamps. Right panel, alignment on rank orders of occurrence of the events.

The example data in Table 1 reports the events amongst *graduating university*, *first employment*, *first unemployment*, *first marriage*, and *first child* experienced between age 15 and 45 by individual 1. To describe such a sequence we will also use the notation (graduating university)$^1-$(first employment, first marriage)$^2-$(first child)^3 where the terms inside parentheses are events, each whole parenthesis defines a transition, and the superscripts stand for the rank order of occurrence.

By ranking from the beginning of the sequence we implicitly left align the event sequences. More specifically, we upwardly number time points in which one or more events occurred, starting with giving a 1 to the earliest of those time points. For individual 1 in Table 1 there are three event-occurrence time points: 25, 27 and 30. Correspondingly, 'graduating university' is assigned order position 1, both 'first employment' and 'first marriage' are assigned a 2 and 'first child' is assigned a 3. Different alignments could be considered as well such as right aligning on the last event of each sequence, or aligning on the occurrence of a given event, 'first employment' for instance. For *right-aligned* order positions, we would assign order 1 to the time of occurrence of the last event and number the time points backwards.

Individual	**Event**	**Age**	**Rank**
1	graduating university	25	1
1	first employment	27	2
1	first marriage	27	2
1	first child	30	3
2	...	...	...

Table 1: Extract of a fictitious event sequence data set.

Related methods Methods for analyzing the ordering of events vary among disciplines. A long-standing method in socio-demographic life-course studies is treating a selection of predefined sequence patterns as categorical values of a dependent variable and studying their relationships with explanatory variables by means of log-linear regression models (Hogan, 1978; Marini, 1984) or classification trees (Billari, 2005). Multistate models (Willekens, 2006) and especially flowgraphs (Huzurbazar, 2004) are useful for exhibiting the relational structure between events including possible feedback events. The data-mining community has emphasized the search of frequent sub-sequences and association rules (Agrawal & Srikant, 1995; Mannila, Toivonen, & Inkeri Verkamo, 1997). Moen (2000) and Studer, Müller, Ritschard, and Gabadinho (2010) developed dissimilarity measures for event sequences which account for the ordering of events. Such dissimilarity measures give access to the whole palette of dissimilarity-based methods, such as cluster analysis (Reynolds, Richards, de la Iglesia, & Rayward-Smith, 2006), self organizing maps (Massoni, Olteanu, & Rousset, 2009) or discrepancy analysis (Anderson, 2001; Studer, Ritschard, Gabadinho, & Müller, 2011).

This article is focusing on a graphical method. The intention is to provide a simple, understandable plot for exploring the diversity of distinct sequence patterns of a target population. The plot should facilitate the identification of standard patterns, including simultaneity of events, while reflecting at the same time all the diversity of the observed patterns. The plot should also be helpful for group comparisons.

Since we consider categorical sequences, let us first stress the possibilities of representing sequences with plots for categorical variables. Bar, mosaic or association plots (Friendly, 2000; Hartigan & Kleiner, 1984) are helpful to render distributions of categorical data and highlight the association between pairs of categorical variables. By cross tabulating event occurrences with the order position, such plots can visualize how events are distributed among the successive positions but do not render the individual sequence patterns and their diversity. Alternatively, by considering the event occurring at each successive position as a categorical variable, a set of sequences can be seen as a series of categorical variables and we could resort to some kind of parallel coordinate plots to visualize the successions of events and the relations between the events at the successive positions. Examples of categorical parallel coordinate plots are the plot proposed by (Yang, 2003) for rendering itemsets, the hammock plots (Schonlau, 2003) and parallel sets (Kosara, Bendix, & Hauser, 2006).

Among plots specifically designed for sequence data, there are indeed all nice plots for state sequences (Brzinsky-Fay, Kohler, & Luniak, 2006; Gabadinho, Ritschard, Müller, & Studer, 2011). Those plots require data on the duration of the states and do not apply for sequences made of elements such as events which do not have durations. Three suggestions in the literature can potentially be applied to any kind of sequences including event sequences. The first class of graphics, known as *life lines* or *calendar plots*, arrange color-coded event symbols along horizontal lines (Wang, Plaisant, & Shneiderman, 2010; Wongsuphasawat et al., 2011). The second class of graphics are *directed graphs* (Hébrail & Cadalen, 2000; Huzurbazar, 2004), such as the graphical representation of a flowgraph, which connect event nodes with directed line segments along the event order. The third class of plots make use of the already discussed categorical parallel coordinate principle. Yang (2003), for instance, uses such an approach for sequential patterns. For sequence data, a parallel coordinate plot can be seen as an adaptation of the so called 'spaghetti plot' used for rendering multiple time series. The plot consists in reporting the position in the sequence (or time point) on

the x-axis and assigning a vertical coordinate to each event-category. Each unique sequence pattern is then visualized as a polyline connecting the successive events in the exact order they appear in the sequence. Varying line widths can be used to visualize the support of each event-to-event segment.

None of the existing plots fully fits all of our objectives, i.e., get a plot which can trace the individual sequences and their diversity, can render simultaneous events and allows for easy group comparisons. We found that the principle of categorical parallel coordinates would be, with some crucial extensions, the most convenient way to meet our goals.

The contribution of the article is an extension of the categorical parallel plot (Kosara et al., 2006; Schonlau, 2003; Yang, 2003) to categorical sequence data. The main extending features are the followings: (i) A *translation-arrangement* which allows rendering *simultaneous events* and facilitates the tracing of individual sequences by avoiding overlapping lines; (ii) the merging of *embeddable* sequences; (iii) the accounting of *weights* and *zero-event sequences*; and (iv) filter instruments and criteria to improve the exploratory power of the plot.

Organization of the article In the upcoming section, we first explain the basic plot design in detail. Subsequently, we introduce some adjustments to better highlight the interesting order patterns and apply the plot on two real data sets. In the first and main application example, we render family life event data and compare our proposition with the basic parallel coordinate plot. For the second application example, we consider ordinal state sequence data to reveal that the proposed plot can also prove useful for non-event sequences and can render time alignments. Finally, we summarize our findings and discuss the scope and limits of the approach.

2 The decorated categorical parallel coordinate plot

The proposed graphic renders event order patterns as slightly displaced lines in a scatter plot. The line displacement is necessary to unambiguously render each distinct sequence and to avoid that sequences or portions of sequences hide other ones. The rank orders of occurrence are located on the x-axis and the event categories at evenly spaced positions on the y-axis. The coordinate assignment for the event categories is basically arbitrary and could be for instance the alphabetical order. The readability of the solution will, however, most often depend on this coordinate order and could be improved by a suitable choice of this order. A meaningful solution is for example to arrange the event categories in their most frequently observed order of occurrences. This is the solution we adopted in the application in Section 4.1, while in Section 4.2 we simply retain the ordinal order of the states.

The initially empty scatter plot is first covered with small sized light gray rectangular areas at the intersections of whole numbered x and y coordinates. These areas are called *translation zones* and they serve for tracking the *translation arrangement* across grid points. The same translation arrangement is shared by all translation zones. It is built as follows: For each unique event order pattern, a solid square of size proportional to the (possibly weighted) sample frequency of the order pattern is allocated. Next, the whole set of these squares is algorithmically arranged in the zone and optimally resized to fit in the zone.

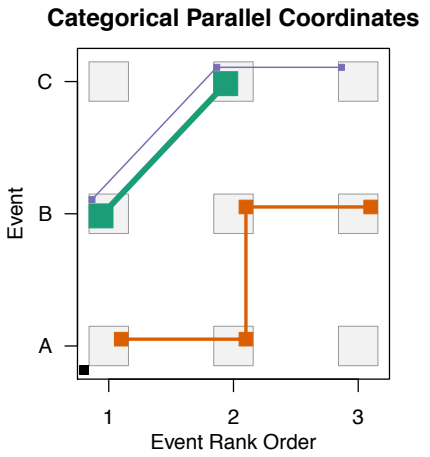

Figure 2: Basic plot of the three ordering patterns among 100 event sequences.

The placing procedure first collects the largest solid square of each sequence pattern and sorts them in decreasing order. Then, starting with the largest of the collected squares, the procedure successively assigns a random location in the remaining free space to each of the squares. In case the remaining space is insufficient, the size of all solid squares are proportionally reduced by the smallest factor allowing to fit all of them in the zone. This gives for each unique event order pattern a unique square area at the same relative position in each concerned translation zone.

The plot is then finalized by successively plotting the unique event order patterns. For each pattern, first its corresponding event-occupied squares in the translation zones are colored and then these squares are connected with line segments along the event order. The widths of the line segments are also adjusted to the sample frequency but they are slightly thinner than the event-squares for readability. Simultaneous events appear as vertical segments. To maintain the line-continuity in these cases, we connect the precedent event with the lowest event and the subsequent event with the highest one of that vertical segment (or optionally conversely). In the exceptional case where a same event would occur several times at a same position the multiple occurrences would be reflected by a 'sunflower' inscribed in the concerned square. Finally, *zero-event sequences*, i.e., empty sequences corresponding to cases which do not experience any event, are reflected by a square outside the bottom-left translation area.

Figure 2 presents the categorical parallel coordinate plot for a fictitious set of 100 event sequences with three unique event order patterns. The lower line represents 21 sequences with a common order pattern which starts with event A at position 1, then has events A and B at position 2 and ends with event B at position 3. The thick line represents 65 identical sequences B-C and the thin line at the top four sequences B-C-C. There are 10 zero-event sequences which are rendered by the black square south-west of the bottom-left translation zone.

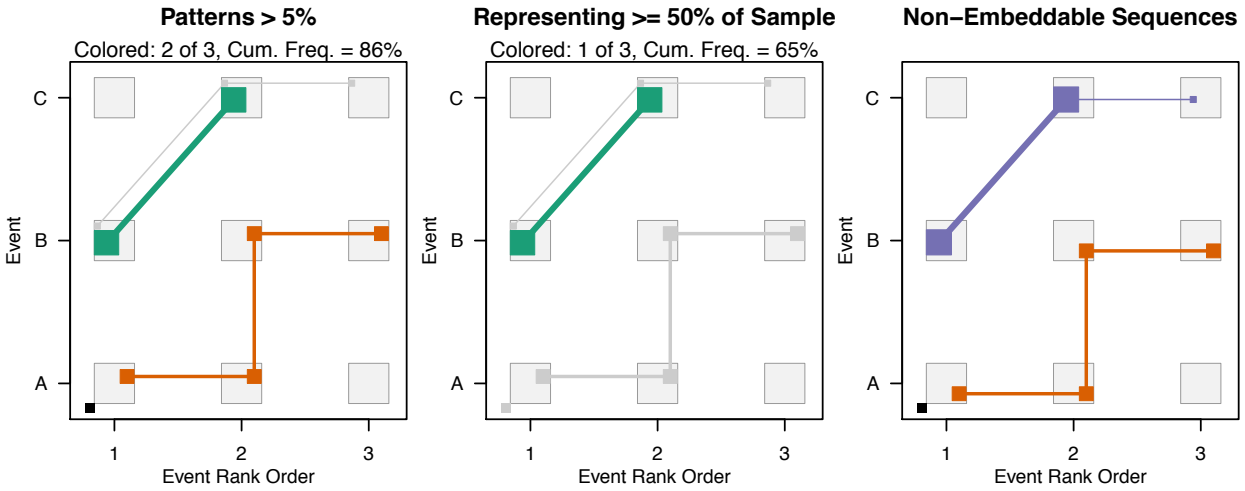

Figure 3: Readability improvements. Highlighting patterns with frequency above a selected threshold (left); Highlighting most frequent patterns until their cumulated frequency exceeds a selected threshold (middle); Embedding shorter sequences into longer ones (right).

3 Improving readability

Generating plots as described so far for full-scaled real data sets we will most often get graphics cluttered with too many lines to be able to distinguish patterns of interest. In this section, we propose two adjustments to improve the plot readability in presence of a high number of distinct patterns and thus facilitate the graphical exploration.

Emphasizing interesting event order patterns In order to emphasize event order patterns of interest, we propose to gray less interesting patterns and lay them in the background. The level of interest will typically be measured by the frequency of the pattern, but could as well be, for example, the inverse frequency if we are interested in atypical patterns, or some measure of the strength of association between the pattern and a target variable such as the sex, birth year or income of the concerned individuals.

Using the frequency, the threshold for coloring event order patterns can be set in two alternative ways: either as a minimum value that should be satisfied to allow for coloring, or as a global cumulated value that should be reached by the whole set of highlighted patterns.

The *first* solution colors event order patterns over a given frequency threshold in intense colors and grays the remaining patterns. This adjustment was applied in the leftmost plot of Figure 3 using a level of 5%. As a result, the previously saturated thin line at the top, which represents 4 out of the 100 sequences, appears in gray. All other patterns keep their original contrast level. In addition, the total number of highlighted patterns and their cumulated frequency are automatically displayed below the plot title to provide some quantitative information.

The *second* solution based on cumulated values consists in coloring patterns in decreasing frequency order until their cumulated frequency reaches the set threshold. As an illustration,

the middle plot in Figure 3 was obtained by setting the threshold for the cumulated value as 50%. In that example, a single pattern is sufficient to represent at least 50% of all sequences. As in the left plot, the total number of highlighted patterns and their cumulated frequency are displayed with the plot title.

The first solution is very general and can be used with any interestingness measure. The second proposition based on cumulated values is less general. Indeed, it requires that the used measure of the interest level can be summed up over patterns, which is not the case, for example, for association measures. We should be careful, therefore, to not use it when cumulated values of the considered interestingness measure do not make sense.

Plotting the non-embeddable event order patterns The second adjustment is a reduction of the number of plotted lines without information loss. This is achieved by drawing only *non-embeddable* event order patterns. An event order pattern S_1 is *embeddable* into a pattern S_2 if S_2 can be transformed into the exact form of S_1 by cutting prefixes or suffixes; i.e., by cutting a starting or ending substring from the sequence S_2. The *non-embeddable* patterns are those unique event order patterns which cannot be embedded into any other one.

The embedding is visualized by adjusting the line widths of shared partial line segments. The rightmost plot in Figure 3 shows that plotting the *non-embeddable* event order patterns reduces the number of lines from three to two for the toy example in Figure 2. The thickest line in Figure 2 is here embedded in the formerly thinnest line. Consequently, the line segment between position 1 and 2, which is shared by $65 + 4 = 69$ sequences, is much thicker than that between positions 2 and 3 which is shared by four out of the 69 sequences only.

The embedding trick raises two difficulties: First, the trick implies a technical ambiguity. Short event order patterns can often be embedded into more than one *non-embeddable* event order candidates. The solution we suggest in that case is to embed the pattern into the most frequent pattern among the available candidates. Doing so, instead of distributing them evenly over all candidates for example, will emphasize the commonness of the shared segments. Second, the interpretation becomes ambiguous when two or more event orders with both different start and end positions are embedded in the same non-embeddable event order pattern. For example, the three sequences A-B-B-*, *-B-B-C and A-B-B-C, where a '*' indicates an empty position, can be merged into the single non-embeddable sequence A-B-B-C with a weight of 2 for the paths A-B and B-C, and a weight of 3 for the path B-B. The same non-embeddable sequence results from the three sequences A-B-B-C, A-B-B-C and *-B-B-* and it is thus not possible to univocally retrieve the original sequences from the non-embedded sequence; hence the ambiguity. Therefore, we recommend to use the embedding adjustment only with either left-aligned or right-aligned sequences. For left-aligned sequences, the embedding should be checked by cutting suffixes only, and for right-aligned sequences by cutting prefixes only.

Combining both adjustments Both adjustments, i.e., plotting only non-embeddable patterns and graying out less interesting patterns can be applied together on a same plot. In that case, when one or more patterns have been embedded in a longer one, the whole non-embeddable event order pattern is highlighted whenever its most frequent segment fulfills

the interestingness constraint. As a consequence, some non-embeddable patterns which do not themselves reach the minimum interest level may be highlighted just because some other patterns were embedded in them.

4 Application

We illustrate the scope of the proposed parallel coordinate plot for group comparison with two applications. The first one demonstrates the interest of the plot for analyzing the changes in the sequencing of family life events over birth cohorts. The second illustration is about marijuana use by U.S. teenagers and shows how the plot can serve for comparing female and male trajectories described by state sequences.

4.1 Family life event histories

As a first application, we consider family-life event sequences of Scandinavians from the 2006 European Social Survey (ESS) Round 3 data and are interested in whether the ordering of family-life events is historically stable or if there are differences between age-groups. The data preparation was the following: First, we extracted the data-subset of the Scandinavian participants (Swedes, Norwegians and Danes) of the two age-groups 1930-1939 and 1950-1959. We then retrieved for each retained person the year of occurrence of the following events: leaving parental home, first union, first marriage and first childbirth, as described in Billari and Liefbroer (2010). Events after age 45 were omitted to allow a consistent comparison between the two age groups. Finally, we left aligned the event sequences to define the order positions. For example, the Danish participant with id 100434 (a female born in 1931) left the parents' house in 1950 and, in 1953, she started to live with a partner, married and gave birth to her first child. Consequently, position 1 was assigned to event *Leaving Home* and position 2 to each of *First Union*, *First Marriage* and *First Child* which occurred the same year. The final used data consists of 1372 individuals and includes a total of 5049 events. To account for the ESS sampling scheme, frequencies and hence line widths will account for the provided *Country* and *Design* weights.

The decorated categorical parallel coordinate plot—without embedding—of the event order patterns is shown in Figure 4. Event orders of cohort 1930-1939 are on the left panel and event orders of cohort 1950-1959 on the right panel. Since line widths and point sizes are adjusted for the within group weighted frequencies, the two cohorts can be compared even though there are much less individuals in cohort 1930-1939 than in cohort 1950-1959. The order of the y-alphabet—the events considered—was set so that most line-patterns monotonically increase along the position axis. To emphasize the most frequent event order patterns, all patterns that do not represent a minimum of 5% of the individual sequences in the group are grayed out. Since all frequencies are weight-adjusted, this threshold applies indeed to weight-adjusted pattern frequencies.

The two plots in Figure 4 widely differ. For the older cohort, there are only four highlighted lines. Three of them run very steeply because they each have two or three simultaneous events. The most frequent pattern is represented by the thickest line and it corresponds to individuals who first left parental home, then later started a union and married in the same

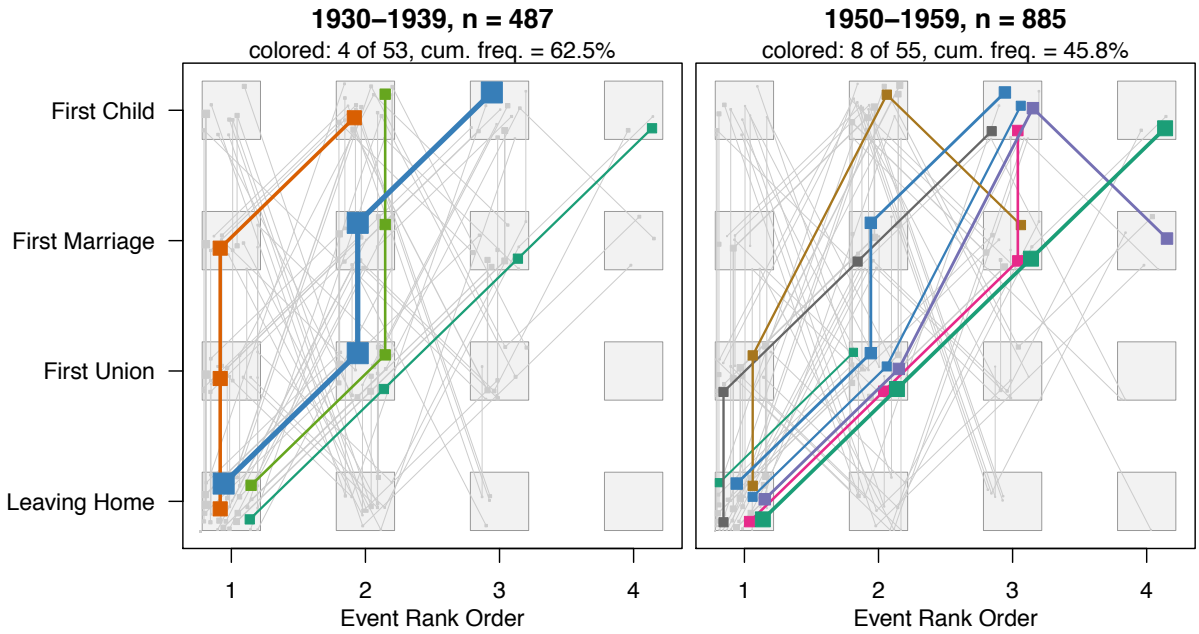

Figure 4: Cohort comparison of Scandinavian family life event orders. Highlighted lines describe order patterns with a weighted frequency above 5%.

year, and later gave birth to a first child. Although we can observe very diverse patterns—the grayed lines—the ordering of life events is clearly dominated by the four colored patterns which represent together 62.5% of all sequences.

The sequences followed by the cohort 1950-1959 are much less standardized. We observe more diversity among the frequent patterns in the right panel. There are eight patterns with a within group frequency above 5% and only two of them were already highlighted standards in the plot for cohort 1930-1939. None of them imposes itself in a proportion equivalent to the most frequent blue pattern of the older cohort. The most frequent pattern for cohort 1950-1959 is the green diagonal which corresponds to first leaving parental home, later starting a union, again later marrying and later again having the first childbirth. The eight highlighted patterns represent together 45.8% of all sequences, that is, much less than the 62.5% represented by the four common patterns of the older cohort. In three of the eight emerging standards—newly highlighted patterns—the first child is not proceeded by the first marriage. The plot thus renders how the norms in the organization of life trajectories changed across cohorts: For the younger generation, getting married is no longer considered as a prerequisite for giving childbirth.

Zero-event sequences play only a marginal role in these data. There are only three cases in cohort 1930-1939 and one case in the 1950-1959 cohort which do not experience any event. Hence, the points representing them on the bottom left become visible only with strong zoom.

The superiority of our proposition over the basic parallel coordinate plot appears clearly when we compare the basic plot of sequences of the 1950-1959 cohort shown in the left panel of Figure 5 with the plot in the right panel of Figure 4. In the basic plot, the plotted

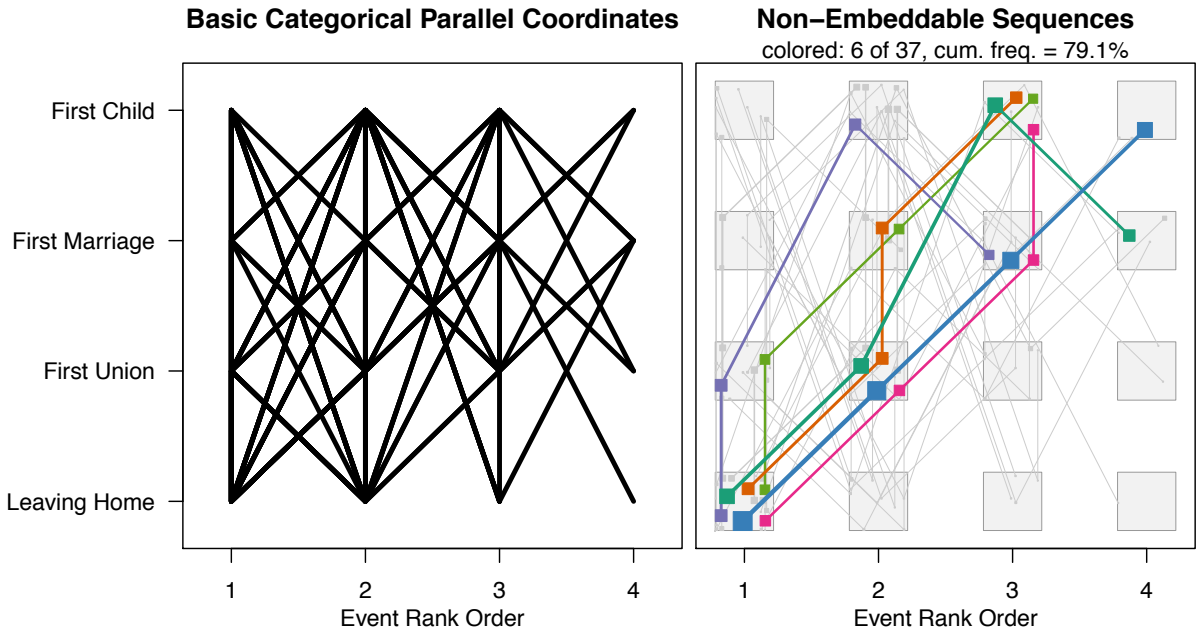

Figure 5: Alternative plots of the 1950-1959 cohort. Left panel: basic parallel coordinate plot. Right panel: non-embeddable event order patterns.

lines overlap, which makes it impossible to track single patterns. Even worse, basic parallel coordinates could be misleading regarding patterns actually not observed. For example, the pattern (First Union, First Child)^1 − (Leaving Home, First Marriage)^2 is not present in the data set while the plotted line segment may suggest it is. This problem does not occur with our proposition because, as can be seen in Figure 4, the distinct sequence patterns are jittered and can be univocally tracked by following up the corresponding small squares at identical position in the translation zones.

In Figure 4, the plot for the Scandinavian 1950-1959 cohort looks somewhat cluttered because of the many patterns satisfying the minimum frequency condition. A way to treat this problem is by plotting the non-embeddable event order patterns only. The resulting plot is shown in the right panel of Figure 5. In that plot, the pattern (Leaving Home)^1−(First Union)^2−(First Child)^3, for example, has been embedded into the pattern (Leaving Home)^1−(First Union)^2−(First Child)^3−(First Marriage)^4, and both patterns are visualized by a same single line. The method reduces the total number of lines from 55 to 37 and the number of highlighted patterns from 8 to 6. Due to these changes, the square points within the translation zones have been arranged differently and the widths of the event-squares and line segments differ from those in Figure 4.

4.2 Marijuana use among U.S. teenagers

The aim of this second illustrative application is to demonstrate the potential of our plot for rendering state sequences. The difference with event sequences is that the position in a state sequence conveys time information and that simultaneous states cannot occur. In this

application the x-axis reports ages.

We consider data about the use of marijuana taken from Lang, McDonald, and Smith (1999) and based on the first five annual waves (1976-1980) of the U.S. National Youth Survey (Elliott, Huizinga, & Menard, 1989). The data concern adolescents aged 13 at the first wave (1976) and report adolescents' marijunana-use state at the successive ages between 13 and 17 years old. The marijuana use is a categorical ordinal variable with three levels ('never', 'no more than once a month', 'more than once a month') obtained by Lang et al. (1999) by collapsing the nine levels of the original marijuana-use scale.

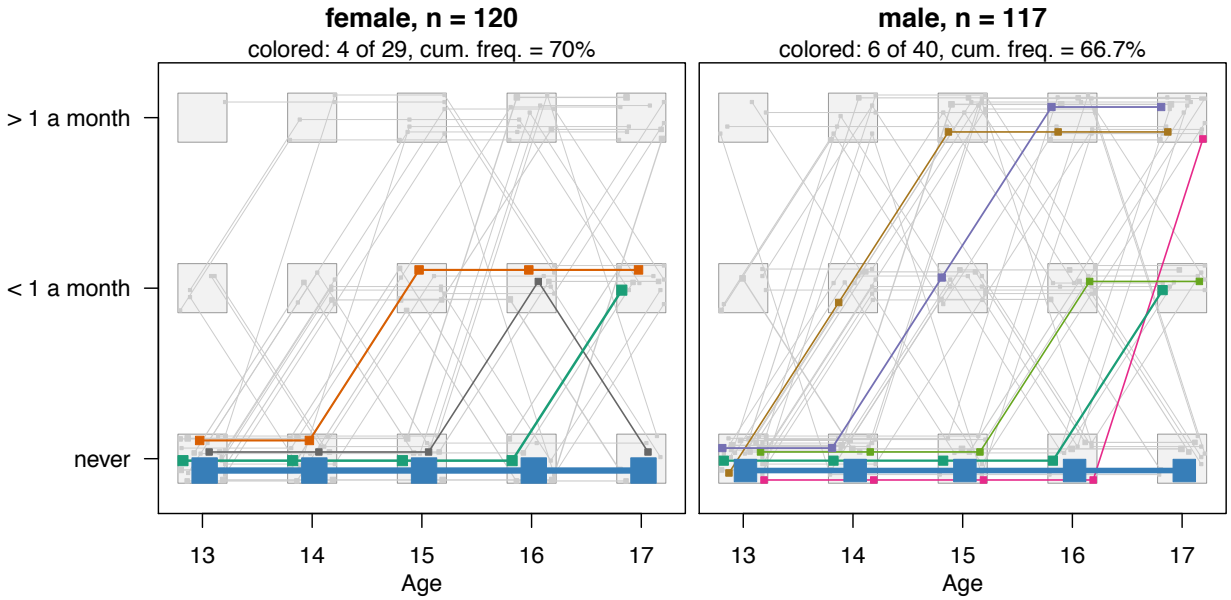

Figure 6: Marijuana use of U.S. teenagers between ages 13 to 17. The trajectories shared by at least three adolescents in the group are highlighted in different colors.

Figure 6 exhibits the evolution of marijuana use by females and males. Colored patterns are the unique patterns shared by at least three adolescents (3%) in each group. The most frequent trajectory is to never use marijuana between ages 13 to 17 for both genders. Looking at the other patterns including the greyed lines we observe a higher diversity among trajectories followed by males. There are 40 unique trajectories for males versus 29 for females. The plots also reveal, for both groups, a higher tendency to increasing marijuana use than decreasing it. Focusing on the colored lines—most frequent patterns, we observe what is the main conclusion found by Lang et al. (1999), i.e., a higher risk for males to use marijuana. More specifically the plots reveal a tendency for males to start with marijuana use earlier than females.

In this example, all sequences are complete and, therefore, right- and left-aligned. When all sequences are complete, no unique sequence can be embedded in another unique sequence. Plotting only non embeddable sequences would thus produce the same plot. Shifting sequences of different length in order to left or right align them we would lose the time alignment. Thus, the embedding trick is useful for time-aligned sequences only when the sequences are of different length and all of them either start or end at the same time.

5 About the plot usage

The plot has been implemented in the TraMineR package (Gabadinho et al., 2011) for the R environment for statistical computing and graphics (R Core Team, 2012) which can be freely installed from the Comprehensive R Archive Network (CRAN, `http://cran.r-project.org/`). The plot is generated with the `seqpcplot` function which offers a series of arguments for controlling, among others, the placement and widths of the points and lines as well as their coloring, the filtering thresholds and position versus time alignment. A simple argument also permits to highlight a predefined selection of sequence or sub-sequence patterns. The complete list of arguments is documented in the online help file of the `seqpcplot` function where the user also finds several examples. Unlike the tools proposed by Yang (2003) or Wang et al. (2010) for example, our implementation is not interactive, the objective being instead to produce high quality graphics ready for inclusion in publications.

The privileged and default representation is obtained by aligning the successive elements—states or transitions—in each sequences on their rank order of occurrence in the sequence. A possible alternative is to align the states/events on their time of occurrence rather than on the rank order of occurrence. By using time alignment we can render transition times. Practically, however, when the number of time positions increases the resulting graphic may become very cluttered because of the variability in the timing of similarly sequenced events. The left panel in Figure 1, for example, gives the time aligned representation of the Scandinavian family life event data used in Section 4.1. The sequences shown are those for cohort 1930-1939. The time-aligned plot exhibits a high diversity—essentially a timing diversity—of the trajectories which contrasts with the relatively low sequencing diversity shown by the right panel. We learn from the time-aligned plot that leaving home starts about at 14 years old, and that the events first union, first marriage and first child occur since age 17 but become much more frequent after 20 years old. Nevertheless, the plot looks cluttered and other plots such as survival curves or life and calendar lines (Wang et al., 2010; Wongsuphasawat et al., 2011) could be more appropriate for rendering the timing. By transforming event sequences into state sequences we could also resort to plots for state sequences such as index plots (Gabadinho et al., 2011) which explicitly render timing and durations.

Although there are no technical limitations to the scalability of the plot, scalability becomes an issue regarding the usefulness of the plot. The limitation is not that of the total number of sequences but that of the number of unique sequences. The number of unique sequences is intimately linked with the sequence length and the size of the alphabet, i.e., the number of distinct events or states. The larger the alphabet, the less chances we have to find out a significant proportion of sequences sharing a common pattern. The same is true for the sequence length: The longer the sequence, the lower the chances of two sequences following the same pattern. The solution to find out regularities in case of a large alphabet would be to merge close elements of the alphabet. In case of long sequences, the solution could be to use a rougher time granularity which would transform the different sequencings of events occurring in a given laps of time into a unique set of simultaneous events. To give an order of magnitude, the alphabet should not exceed about 10. Likewise, the plot may become hard to read when sequences contain more than 10 distinct successive elements. With shorter sequences we could afford a larger alphabet and reciprocally with a small alphabet we could afford longer sequences.

6 Conclusion

The decorated parallel coordinate plot proposed in this article is a powerful tool for exploring how elements are typically ordered in a set of sequences. The filtering mechanisms which dim out less interesting patterns together with the embedding trick, permit to clearly highlight the most frequent patterns while still rendering the whole diversity of the observed patterns. Although the plot is primarily designed for event sequences where only the position in the sequence rather than the exact time matters, the plot can also render time aligned events and be used with other types of categorical longitudinal data such as categorical panel data for example. The proposed plot offers the following salient features: A pattern jitter mechanism combined with a system of translation zones, point and line widths reflecting pattern frequencies, the possibility to render non-embeddable patterns only, a straightforward way of accounting for weights, the rendering of zero-event sequences, and tunable highlighting of interesting patterns. Unlike the basic parallel coordinate plot as used for instance by Yang (2003), our plot permits to track individual patterns, can render patterns with simultaneous events—a common situation in life-course analysis for example—and reveals all the diversity of the rendered patterns.

To get an idea of the interest of the plot for end users, we presented it to social scientists involved in life-course analysis and to experts from the data mining domain. Life course experts found the display intuitive, identified potential application and suggested, for example, the family-life events application presented in Section 4.1. Experts in data mining provided more mitigated feedback. They agreed about the ease of interpretation of the plot for the examples we provided but criticized its lack of scalability. Indeed, data mining is often concerned with huge data sets consisting of hundred thousands sequences to which the plot would be hardly applicable for the reasons addressed in Section 5.

As already mentioned, the decorated parallel coordinate plot is implemented as a function of the TraMineR R package (Gabadinho et al., 2011) freely downloadable from the Comprehensive R Archive Network (CRAN, `http://cran.r-project.org/`).

References

Agrawal, R., & Srikant, R. (1995). Mining Sequential Patterns. In P. S. Yu & A. L. P. Chen (Eds.), *Proceedings of the international conference on data engeneering (icde), taipei, taiwan* (p. 487-499). IEEE Computer Society.

Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*, 32-46.

Billari, F. C. (2005). Life Course Analysis: Two (Complementary) Cultures? Some Reflections With Examples From The Analysis Of Transition To Adulthood. In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, & E. Widmer (Eds.), *Towards an interdisciplinary perspective on the life course* (p. 267-288). Amsterdam: Elsevier.

Billari, F. C., Fürnkranz, J., & Prskawetz, A. (2006). Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. *European Journal of Population*, *22*(1), 37-65.

Billari, F. C., & Liefbroer, A. C. (2010). Towards a new pattern of transition to adulthood? *Advances in Life Course Research*, *15*, 59–75. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1040260810000407

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence Analysis with Stata. *The Stata Journal*, *6*(4), 435-460.

Elliott, D. S., Huizinga, D., & Menard, S. (1989). *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems*. New York:: Springer-Verlag.

Friendly, M. (2000). *Visualizing Categorical Data*. Cary, NC: SAS Institute.

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011, 4). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, *40*(4), 1–37. Retrieved from http://www.jstatsoft.org/v40/i04

Hartigan, J., & Kleiner, B. (1984). A mosaic of television ratings. *The American Statistician*, *38*, 32–35.

Hébrail, G., & Cadalen, H. (2000). Visualisation et classification automatique de parcours professionnels. In *Actes des XXXIIe Journées de statistique, Fès, Maroc* (p. 458-462).

Hogan, D. (1978). The Variable Order of Events in the Life Course. *American Sociological Review*, *43*, 573-586.

Huzurbazar, A. V. (2004). *Flowgraph Models for Multistate Time-to-Event Data*. New Jersey: Wiley.

Kosara, R., Bendix, F., & Hauser, H. (2006, july-aug.). Parallel Sets: Interactive exploration and visual analysis of categorical data. *Visualization and Computer Graphics, IEEE Transactions on*, *12*(4), 558 -568. doi: 10.1109/TVCG.2006.76

Lang, J. B., McDonald, J. W., & Smith, P. W. F. (1999). Association-Marginal Modeling of Multivariate Categorical Responses: A Maximum Likelihood Approach. *Journal of the American Statistical Association*, *94*(448), 1161-1171. Retrieved from http://www.jstor.org/stable/2669932

Mannila, H., Toivonen, H., & Inkeri Verkamo, A. (1997). Discovery of Frequent Episodes in Event Sequences. *Data Mining and Knowledge Discovery*, *1*, 259-289. Retrieved from http://dx.doi.org/10.1023/A:1009748302351

Marini, M. M. (1984). The Order of Events in the Transition to Adulthood. *Sociology of Education*, *57*(2), pp. 63-84. Retrieved from http://www.jstor.org/stable/2112630

Massoni, S., Olteanu, M., & Rousset, P. (2009). Career-Path Analysis Using Optimal Matching and Self-Organizing Maps. In J. Prìncipe & R. Miikkulainen (Eds.), *Advances in self-organizing maps* (Vol. 5629, p. 154-162). Springer Berlin / Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-02397-2_18

Moen, P. (2000). *Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining*. PhD thesis, University of Helsinki.

R Core Team. (2012). R: A Language and Environment for Statistical Computing (Vol. 1) (Computer software manual No. 2.11.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Reynolds, A., Richards, G., de la Iglesia, B., & Rayward-Smith, V. (2006). Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*, 475-504. Retrieved from http://dx.doi.org/10.1007/s10852-005-9022-1

Ritschard, G., Gabadinho, A., Studer, M., & Müller, N. S. (2009). Converting between

various sequence representations. In Z. Ras & A. Dardzinska (Eds.), *Advances in data management* (Vol. 223, p. 155-175). Berlin: Springer-Verlag. doi: 10.1007/978-3-642-02190-9_8

Schonlau, M. (2003). Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots. In *In proceedings of the section on statistical graphics, american statistical association; 2003, cd-rom.*

Studer, M., Müller, N. S., Ritschard, G., & Gabadinho, A. (2010). Classer, discriminer et visualiser des séquences d'événements. *Revue des nouvelles technologies de l'information RNTI*, *E-19*, 37-48.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy Analysis of State Sequences. *Sociological Methods and Research*, *40*(3), 471-510. doi: 10.1177/0049124111415372

Wang, T. D., Plaisant, C., & Shneiderman, B. (2010). Temporal Pattern Discovery Using Lifelines2. In *Discovery exhibition (discoveryexhibition.org).* Retrieved from http://discoveryexhibition.org/pmwiki.php/Entries/Wang2010

Willekens, F. (2006). *Multistate model for biographic analysis and projection.* The Hague: NIDI.

Wongsuphasawat, K., Gómez, J. A. G., Plaisant, C., Wang, T. D., Taieb-Maimon, M., & Shneiderman, B. (2011). LifeFlow: Visualizing an Overview of Event Sequences. In *Proceedings of the 2011 annual conference on Human Factors in Computing Systems (CHI), Vancouver, Canada, May 7-12, 2011* (p. 1747-1756). New York: ACM.

Yang, L. (2003). Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates. In *Computational science and its applications, iccsa 2003* (p. 21-30). Springer Berlin / Heidelberg.

Zaki, M. J. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, *42*(1/2), 31–60.