

Soft Patterns: Moving Beyond Explicit Sequential Patterns During Visual Analysis of Longitudinal Event Datasets

David Gotz

Abstract—Reflecting the growing importance of longitudinal and sequential event analysis to many fields, researchers have proposed a variety of visualization methods and interactive visual analysis techniques in recent years. Early work focused on visualizing all permutations, and approach which scales poorly when applied to large-scale datasets with many event types and long sequence lengths. More recently, a variety of proposed techniques use methods such as simplification algorithms, manual user-driven simplification interaction capabilities, pattern mining algorithms, and statistical prioritization to overcome these scalability challenges. However, these methods commonly seek explicit sequential patterns of the form ‘**Event A** before **Event B**’ subject to certain time constraints (e.g., within 60 days) or boolean operators (e.g., ‘**Event A** or **Event B**’). This paper asserts that this approach is not sufficient for certain high-dimensional real-world event analysis applications. Instead, it argues that new research is needed to develop methods and tools that support soft pattern discovery and analysis.

Index Terms—Visual Analytics, Event Analysis, Temporal Event Sequences, Pattern Mining, Interactive Mining, Visualization

INTRODUCTION

Visual analysis methods are often used to examine large sets of event sequences with the aim of discovering patterns of events associated with a given outcome of interest. Because such datasets are often generated from very large numbers of actual observations of complex real-world activities, they have the potential to inform analysts about very complex real-world phenomena. For example, healthcare analysts may be interested in sequences of medications that most often lead to good patient outcomes. In computer security applications, analysts may be interested in finding patterns of behaviour that indicate an attack on a computer network. In finance, analysts may be interested in learning how event-based transaction data be used to detect fraud or guide investments. In marketing, streams of social network data can be used to learn about customer preferences and behavioural responses to campaigns.

Reflecting the emerging importance of this topic, a variety of visualization and visual analytics techniques have been proposed in recent years. Some of these techniques are designed to help users understand common patterns of events. Others are designed to help users associate specific sequences of events to specific outcome measures. However, many of the event analysis techniques being proposed within the visualization literature are often too focused on strict sequence constraints, resulting in missed insights which remain hidden within the complex and nuanced real-world data that analysts often need to understand.

1 FINDING PATTERNS

While a broad range of visualization techniques have been proposed for analysing event sequence data, nearly all research in this area has focused on the development of methods for event pattern discovery. In this context, an *event pattern* is a common (or rare, for outlier detection) motif within the event data being examined by an analyst.

- David Gotz is with the University of North Carolina. E-mail: gotz@unc.edu.

Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis.

Available online at: <http://eventevent.github.io>

1.1 Hard Patterns

The vast majority of event sequence data visualizations have focused on the discovery of explicit sequential patterns of events, taking the form of ‘**Event A** followed by **Event B** followed by **Event C**.’ These event orderings, which are defined using exact sequences of specific event types subject to temporal ordering constraints, are examples of *hard patterns*. Some early visualization methods for event sequence data, as shown in Figure 1, aimed to show all observed hard patterns by visualizing all permutations present in the data [2], [3]. While such approaches scale well to support datasets with large numbers of event sequences, they have difficulty with many real-world datasets where the number of event types are large and the length of event sequences can grow long.

More powerful representations of hard patterns have been developed to overcome some of these challenges. These extensions allow for greater expressivity in how patterns are defined, and support both visual simplification as well as more advanced temporal event query capabilities. For example, this work includes both constraints on the time durations between events (e.g., “within 60 days”), as well as Boolean operators applied to event types (‘**Event A** followed by either **Event B** or **Event C**’) [5]–[7].

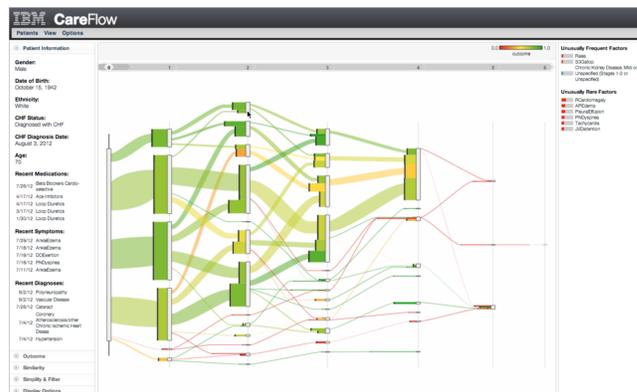


Fig. 1. Visualizations that render all permutations of events, as in this example showing differences in the order of medications given to a cohort of heart failure patients [1], fail to scale given large numbers of events.

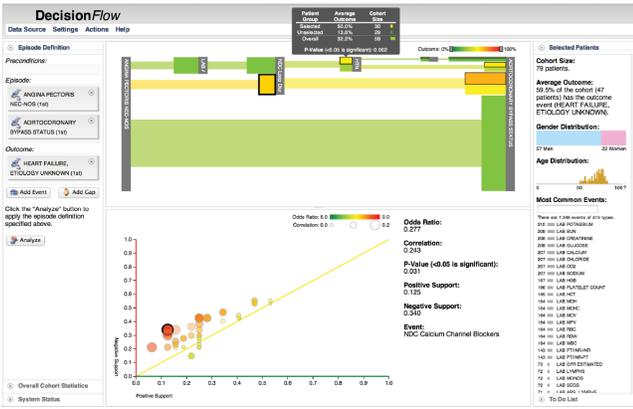


Fig. 2. User interaction and statistical measures can help users manage high-dimensional event sequences, but existing methods (e.g., DecisionFlow [4]) typically place hard constraints on event order within patterns.

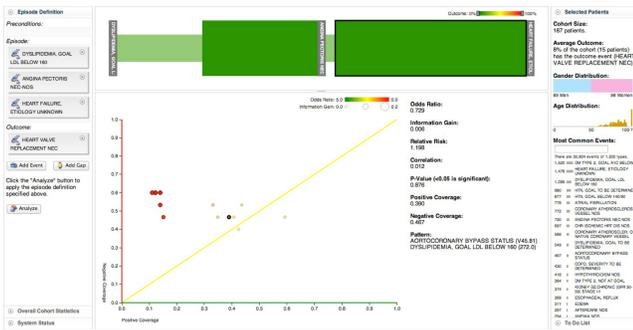


Fig. 3. Mining-based methods [8] also focus on explicit event ordering which is can be too precise for many real-world use cases.

In combination with these representations, other work on pattern discovery within high-dimensional event datasets has explored statistically-driven manual pattern discovery (e.g., [4]) as well as automated pattern mining algorithms (e.g., [8], [9]). Given the combinatorial explosion of possible sequences in such cases, these techniques are useful for prioritizing interesting patterns for user consideration. Examples are shown in Figures 2 and 3. However, these techniques are still based on the concept of hard patterns.

1.2 Soft Patterns

In many real-world applications, however, hard patterns are not sufficient. For example, consider medical analyses based on event data from electronic health records. A patient may need to see two different kinds of specialists for a complex condition. The exact time/order of occurrence between those visits, and the exact time/order of occurrence for lab tests and medications that those specialists order for their patients, may have no important relation to patient outcome. Moreover, they may vary widely between patients due to external factors such as doctors' schedule availability.

Yet time and order cannot simply be ignored. Time and patient progression over time is indeed a critical element of a patient's journey through diagnosis and treatment. Differences in event order or progressions over time spans of hours may not be important, while time spans of months or years may be critical. At the same time, for some events exact order is critical. The difference between death before arrival at the emergency room and after can be enormously important.

It can be argued, therefore, that in many cases, the strict constraints in many existing methods can be too restrictive. This can result, at times, in visual analyses that are too sensitive to small changes in event sequence which have little true domain meaning. Instead, more research is needed to explore visual analyses of *soft patterns*: groups of temporally proximate events which unfold over time, but potentially without the strict ordering and temporal

constraints used in much of the prior visualization work for event sequence analysis.

2 FUTURE RESEARCH DIRECTIONS

Signs of early work toward support for soft patterns can already be seen in the literature. For example, Wongsuphasawat et al. have described an approach to similarity-based search [10], allowing users to retrieve event sequences despite missing, extra, or temporally swapped events. However, this relaxes the hard pattern concept in only small ways. In cases where many events happen at the same time, or as bursts of activity within narrow time periods, counting swapped positions or other similar measures are still too limited.

We argue that a far larger research effort is required to understand how soft patterns should be defined and used within advanced event analysis systems. More specifically, we believe that progress is required in several different aspects of soft pattern analysis including:

- **Data constructs for soft patterns.** New representations to express soft pattern constraints in combination with more traditional explicit constraints on order and timing.
- **Visual representations for soft patterns.** New ways to visually represent event pathways given the presence of soft patterns.
- **Computational and interactive methods for soft pattern construction.** New methods for soft pattern mining, interactive discovery, and exploration.
- **Evaluation.** Case studies and experimental results measuring the utility of soft pattern tools and algorithms, and their applicability to real-world use cases.

A comprehensive research agenda which successfully addresses these core issues would have significant potential to impact a wide range of application domains.

REFERENCES

- [1] A. Perer and D. Gotz, "Visualizations to Support Patient-Clinician Communication of Care Plans," in *ACM CHI Workshop on Patient-Clinician Communication*, Paris, France, 2013.
- [2] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman, "LifeFlow: visualizing an overview of event sequences," in *Proceedings of the 2011 annual conference on Human factors in computing systems*, New York, NY, USA, 2011, pp. 1747–1756.
- [3] K. Wongsuphasawat and D. Gotz, "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2659–2668, 2012.
- [4] D. Gotz and H. Stavropoulos, "DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1783–1792, 2014.
- [5] M. Monroe, R. Lan, J. Morales del Olmo, B. Shneiderman, C. Plaisant, and J. Millstein, "The challenges of specifying intervals and absences in temporal queries: a graphical language approach," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2013, pp. 2349–2358.
- [6] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, "Temporal Event Sequence Simplification," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2227–2236, 2013.
- [7] Christian Partl, Samuel Gratzl, Marc Streit, Anne Mai Wassermann, Hanspeter Pfister, Dieter Schmalstieg, and Alexander Lex, "Pathfinder: Visual Analysis of Paths in Graphs," in *Proc. of EuroVis*, 2016.
- [8] D. Gotz, F. Wang, and A. Perer, "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *J. Biomed. Inform.*, Jan. 2014.
- [9] A. Perer and F. Wang, "Frequency: Interactive Mining and Visualization of Temporal Frequent Event Sequences," in *Proceedings of the 19th International Conference on Intelligent User Interfaces*, New York, NY, USA, 2014, pp. 153–162.
- [10] K. Wongsuphasawat, C. Plaisant, M. Taieb-Maimon, and B. Shneiderman, "Querying event sequences by exact match or similarity search: Design and empirical evaluation," *Interact. Comput.*, vol. 24, no. 2, pp. 55–68, Mar. 2012.