TimeStitch: Interactive Multi-focus Cohort Discovery and Comparison

Peter J. Polack, Jr.* Shang-Tse Chen† Minsuk Kahng‡ Moushumi Sharmin§ Duen Horng Chau¶

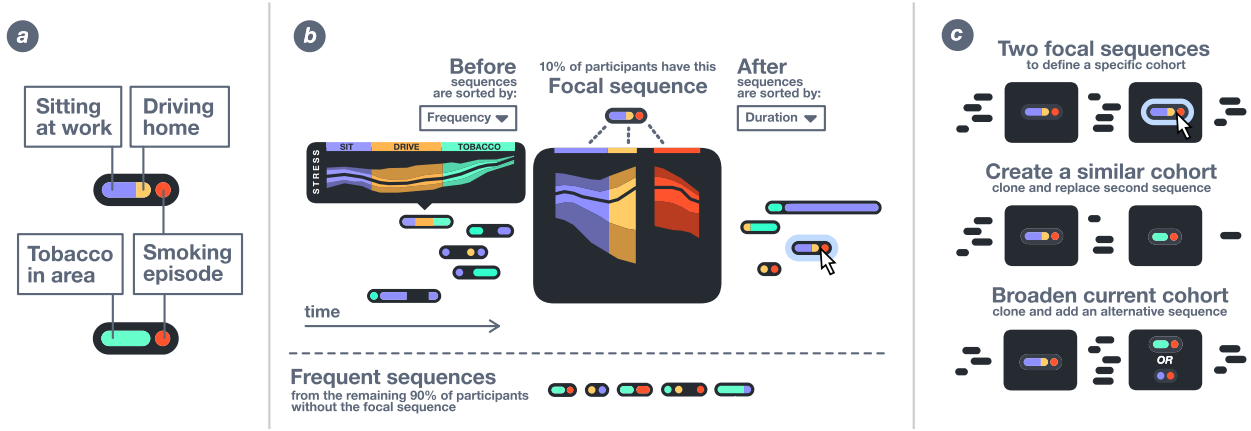

Figure 1: Using TimeStitch to explore events surrounding smoking lapse and to build and compare cohorts of abstinent smokers. (a) Event sequences frequently performed by abstinent smokers, ending with smoking episodes. (b) Selecting the *work→drive →smoke* sequence defines a *cohort* (subset) of abstinent smokers that have this *focal* sequence, and yields a timeline of the common frequent event sequences *before* and *after* the *focal* sequence. (c) Top: cohort defined by two *focal* sequences; Center: top cohort cloned and modified, shown alongside one another vertically; Bottom: center cohort cloned and *generalized* to include more abstinent smokers by allowing two possible second focal sequences through an "OR" operation.

Abstract

Whereas event-based timelines for healthcare enable users to visualize the chronology of events surrounding events of interest, they are often not designed to aid the discovery, construction, or comparison of associated cohorts. We present TimeStitch, a system that helps health researchers discover and understand events that may cause abstinent smokers to lapse. TimeStitch extracts common sequences of events performed by abstinent smokers from large amounts of mobile health sensor data, and offers a suite of interactive and visualization techniques to enable cohort discovery, construction, and comparison, using extracted sequences as interactive elements. We are extending TimeStitch to support more complex health conditions with high mortality risk, such as reducing hospital readmission in congestive heart failure.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interaces (GUI)

1 Introduction

Through the advent of commercial wearable devices, sources of health data are becoming increasingly ubiquitous, fundamentally changing how health researchers, who previously relied solely on sparse self-reported data, study complex health conditions. The Mobile Sensor Data-to-Knowledge (MD2K) research center aims to leverage the widespread use of these sensors to discover the underlying factors associated with health risks and diseases. As a core part of MD2K, this research focuses on developing tools to help health researchers discover and make sense of events that may cause abstinent smokers to lapse, a critical research problem as cigarette smoking is the leading, but preventable cause of death in the United States, responsible for 1 in 5 deaths annually.

Key questions health researchers often want to answer include: Which events or sequences of events commonly precede smoking lapses? What are the lifestyle choices that may trigger such lapses? Are there subsets (cohorts) of abstinent smokers that exhibit similar or different behaviors? How do we discover and define such cohorts? Currently, few tools allow users to fluidly work with large amount of health sensor data to answer these questions in unified ways. In our pilot study, 6 smokers wore the AutoSense Sensor Suite [2] that uses over 40 data streams to record a wide variety of physiological data, e.g., electrocardiogram (ECG), galvanic skin response (GSR), and heart-rate variability (HRV) measurements. After 3 days of data collection, the resulting dataset exceeded 13GB. Most current interactive tools have not be designed to handle datasets of this scale.

We present TimeStitch (Fig. 1), a system designed to address the above challenges by proposing multiple improvements over existing works, which we will detail in Section 2 through a use case scenario. Here, we first highlight our key contributions and discuss them in the context of important works that inspire us.

Focal event Sequences as Interactive Elements for Cohort Discovery. We introduce the idea of using event sequences as units of interactive operation, to address the scalability challenge of working with health sensor data — each person (abstinent smoker) is often associated with tens of sensor streams, from which numer-

*Georgia Tech. e-mail: ppolack@gatech.edu
†Georgia Tech. e-mail: schen351@gatech.edu
‡Georgia Tech. e-mail: kahng@gatech.edu
§University of Memphis. e-mail: msharmin@memphis.edu
¶Georgia Tech. e-mail: polo@gatech.edu

ous kinds of events may be inferred. Existing tools such as EventFlow [4] and LifeFlow [7] were designed to handle relatively few types of events and focused on visual simplification; it is unclear how they may scale to many event types over many sensors. By extracting frequent event sequences from the data and promoting them as interactive elements (Fig. 1a), TimeStitch enables the user to discover temporally correlated events, and use this higher-level knowledge to interactively construct and compare cohorts of interest (Fig. 1b & c). Through TimeStitch, we introduce the concept of *focal* sequences, event sequences of interest (e.g., those involving smoking episodes) that can be selected to show the cohort of people associated with it. TimeStitch displays sequences that are frequently found before and after the *focal* sequence (Fig. 1b), drawing on earlier work such as EventFlow [4], which describes data as events and not event sequences. Additionally unlike EventFlow, TimeStitch supports multiple focal points, which helps the user answer a wider variety of questions, such as "*what are the commonly occuring activities between two smoking lapses?*," as shown in Fig. 1c.

Interaction Techniques and Focal Sequences for Cohort Construction. In our early discussion with doctors and health researchers on the MD2K team, we learned that cohorts are often difficult to define and characterize. Current approaches (e.g., those based on demographics data such as age, gender) may not be sufficient, as they do not take into account activities associated with smoking lapses. TimeStitch aims to fill this gap by offering visual and interactive techniques to allow the user to more easily construct cohorts, test and compare hypotheses, and thereby better understand mHealth data. We have been experimenting with operations that apply to a whole cohort (e.g., cohort cloning), part of a cohort, and specific sequences (e.g., logic operations like *OR* and negation), operations demonstrated in Fig. 1c. While existing works such as CoCo [3] support cohort comparison, they are often limited to two cohorts provided by the user, and there is no support for cohort construction and exploration. Frequence [6] works with multiple cohorts, but does not support cohort construction.

2 TimeStitch Interface Design

Figure 1 presents a scenario of using TimeStitch to discover common event sequences surrounding smoking episodes, and then building these sequences into cohort-representative timelines. Our user Jane is a health researcher studying the aggregated behaviors of 200 abstinent smokers, who wants to discover and understand the common events associated with smoking relapse.

When Jane opns TimeStitch, it runs the PrefixSpan [5] sequence mining algorithm, which works by first retrieving a list of frequent events, then partitioning the data around these 'prefixes,' and appending sequential elements to them recursively. This particular algorithmic strategy of *pattern growing* is well suited for large data sets and is easily extensible to time-series data, both fundamental aspects of our project. Once collected, the mined sequences are represented graphically as TimeStitch's units of interactive operation (Fig. 1a). In this way, Jane can select or drag any event sequence mined from the aggregated participant data.

Jane selects the *work→drive→smoke* event sequence (in Fig. 1a) as the *focal* sequence. In doing so, the *focal* sequence expands to reveal its associated quantitative metrics (like stress) in a colored "stream" [1] of adjoined box-plots. TimeStitch then retrieves the subset of abstinent smokers that contains this *focal* sequence (Fig. 1b). Utilizing the already ordered set of event sequences mined with PrefixSpan, TimeStitch displays the sequences that are found before and after the *focal* sequence (Fig. 1b).

The user can choose to sort these event sequences vertically by any quantitative measure (e.g. duration, sequence length), but by default sequences are sorted by frequency. One of these future behaviors catches Jane's eye (she highlights it with the cursor in Fig. 1b) — it happens to be identical to the *focal* sequence. Evidently, some people within this *focal* cohort repeat the behavior of smoking after driving home from work. Jane decides to investigate more closely.

She drags this found event sequence to the right of the timeline, which creates a second *focal* sequence (Fig. 1c). In doing so, TimeStitch introduces the capacity to reveal sequences of events *between* multiple user-defined sequences. Instead of limiting our analyst to working with only this cohort, TimeStitch additionally displays event sequences that are *not found* in the current cohort (bottom of Fig. 1b). In this way, Jane can explore patterns outside of the current cohort in addition to defining it more specifically.

Noticing another sequence with a smoking episode that follows an event for *tobacco in area*, Jane decides to define a new cohort with this new event sequence specified (in Fig. 1). She could change the current timeline, but for the sake of comparing between cohorts, she *clones* the current timeline to duplicate it into another one just below (Fig. 1c center). By editing this second timeline with the new sequence, comparative differences between the similar cohorts emerge. As a final step, Jane defines a third, broader cohort by compiling multiple sequences into a single *focal* point, demonstrating how TimeStitch can be used to compile multiple event sequences into boolean operations within *focal* points.

3 Conclusions and Ongoing Work

Through the usecase, we demonstrate TimeStitch's key contributions in helping health researchers make sense of events associated with smoking lapse. Specifically, by using a sequence mining algorithm, TimeStitch extracts and visualizes common event sequences, which become interactive elements that the user can directly manipulate to construct and explore cohorts (e.g., through dragging, cloning). By allowing the user to specify multiple focal sequences, cohorts can be incrementally constructed and explored. Currently we are enhancing TimeStitch into a mature prototype for long-term evaluation with health researchers, and We also plan to generalize TimeStitch to support the analysis of other complex health conditions, such as reducing hospital readmission in congestive heart failure patients.

Acknowledgements

Research supported by grant U54EB020404, awarded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) through funds from the trans-NIH Big Data to Knowledge (BD2K) initiative, and by the NSF GRFP under Grant No. DGE-1148903.

References

[1] R. Bade, S. Schlechtweg, and S. Miksch. Connecting time-oriented data and information to a coherent interactive visualization. In *Proc. CHI*. ACM, 2004.

[2] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al'Absi, and S. Shah. Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proc. ACM Sensys*. ACM, 2011.

[3] S. Malik, F. Du, M. Monroe, E. Onukwugha, C. Plaisant, and B. Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proc. IUI*. ACM, 2015.

[4] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12), 2013.

[5] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *ICCCN*. IEEE, 2001.

[6] A. Perer and F. Wang. Frequence: Interactive mining and visualization of temporal frequent event sequences. In *Proc. IUI*. ACM, 2014.

[7] 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 *Proc. CHI*. ACM, 2011.