EuroVis Workshop on Visual Analytics (2019)
C. Turkay and T. von Landesberger (Editors)

Visual Analytics of Event Data using Multiple Mining Methods

M. Adnan[1], P. H. Nguyen[2], R. A. Ruddle[1] and C. Turkay[2]

[1]University of Leeds, UK
[2]City, University of London, UK

Abstract

Most researchers use a single method of mining to analyze event data. This paper uses case studies from two very different domains (electronic health records and cybersecurity) to investigate how researchers can gain breakthrough insights by combining multiple event mining methods in a visual analytics workflow. The aim of the health case study was to identify patterns of missing values, which was daunting because the 615 million missing values occurred in 43,219 combinations of fields. However, a workflow that involved exclusive set intersections (ESI), frequent itemset mining (FIM) and then two more ESI steps allowed us to identify that 82% of the missing values were from just 244 combinations. The cybersecurity case study's aim was to understand users' behavior from logs that contained 300 types of action, gathered from 15,000 sessions and 1,400 users. Sequential frequent pattern mining (SFPM) and ESI highlighted some patterns in common, and others that were not. For the latter, SFPM stood out for its ability to action sequences that were buried within otherwise different sessions, and ESI detected subtle signals that were missed by SFPM. In summary, this paper demonstrates the importance of using multiple perspectives, complementary set mining methods and a diverse workflow when using visual analytics to analyze complex event data.

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics;***

1. Introduction

Applications from domains as diverse as security, health, retail and education produce large quantities of event data [RV10, MCB*11, SS13, MA13, RWA*13]. Methods such as frequent itemset mining (FIM), exclusive set intersections (ESI) and sequential frequent pattern mining (SFPM) are often used in visual analytics systems to analyze such data, but each method has fundamental weaknesses. FIM and SFPM are only scalable if users make a priori choices of parameters (e.g., minimum support threshold), and do not distinguish between partial vs. full sets/sequences [FVLV*17, FVLK*17]. ESI is computationally more scalable but ignores frequent subsets of events and is susceptible to noise that often requires excessive user interaction to filter [MLL*13].

This paper investigates how multiple event mining methods may be combined within visual analytics. The paper's contributions are: (1) identifying similarities and differences in the visual analytics requirements for diverse analysis tasks (missing data vs. logfile analysis) from two completely different application domains (healthcare vs. cybersecurity), (2) characterizing the insights that each event mining method provides, and (3) evaluating the benefits of combining multiple methods into one visual analytics workflow. Through two contrasting case studies, we demonstrate how visualization enhances the analysis of event sequences and discuss observations and guiding principles that take initial steps towards a comprehensive event analysis framework for visual analytics tools.

2. Related work

Event mining can be broadly classified into two types: (1) event set mining (e.g., FIM and ESI), and (2) event sequence mining (e.g., SFPM). FIM mines event sets (or itemsets) that meet a user specified minimum support threshold [FVLV*17]. It usually produces a large number of itemsets, which can be reduced by computing closed or maximal itemsets.An itemset is closed if no superset has the same support [PBTL99]. An itemset is maximal if it does not have any superset [UKA04]. PowerSetViewer [MKN*05], FIsViz [LIC08] and FpVAT [LC10] are examples of visual analytics systems that use the above variations of FIM.

ESI mines sets of events (or set intersections) and allows users to compute all non-empty intersections in a dataset by ignoring the subsets of events [AMA*16]. By contrast, FIM often requires a high minimum support threshold to produce a reasonable number of itemsets. There are several visual analytics systems [LGS*14, AR17] and workflows [AR18] that use ESI.

Lastly, SFPM mines the sequences of events that co-occur frequently [FVLK*17]. Depending on the support threshold, the number of resulting patterns can be large. Constraints can be added to reduce the number of patterns such as time [PW14] (duration between two consecutive events) and cycle [NTA*18] (patterns that are cyclic of each other). Many visual analytics systems have applied SPFM, with extensions to make the patterns more manageable and meaningful [WL14, LDDH16, LKD*17, LWD*17].

DOI: 10.2312/eurova.20191126

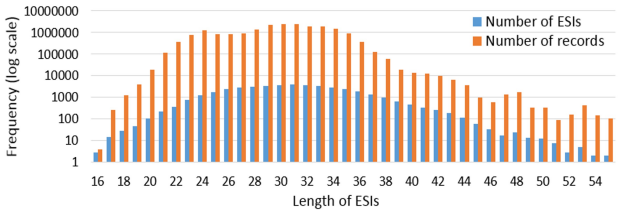

Figure 1: *Distribution of the number of ESIs and the associated records across their length. The unusual wave pattern indicates the presence of local patterns within specific ESI lengths, e.g., 51-55.*

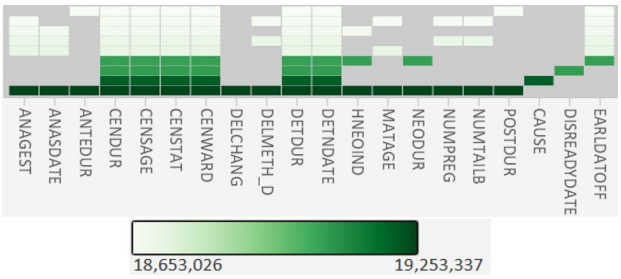

Figure 2: *Maximal itemsets with support $\geq$ 90%. Each row is an itemset. The most infrequent itemset appears in 18.6 million records.*

3. Two contrasting case studies

This paper uses two case studies that aim to: (1) comprehensively identify the patterns of missingness in electronic health records (admitted patient care (APC) data from NHS hospitals), and (2) understand people's behavior when using a software application [NTA*17]. The case studies demonstrate how, by integrating multiple event mining methods with visual analytics, we made breakthroughs in understanding that would not have been possible if we had only used a single method.

A key difference between the two case studies is the order of elements in each data record. That order is irrelevant for the analysis of missing data, but inherently relevant when analysing the actions that people perform when using software. The analysis in both studies can be approached using frequent pattern mining, FIM for Study 1 and SFPM for Study 2. Even though the data in the second study is ordered, we think it could be useful to relax this property and apply an orderless method. Therefore, we use ESI in both studies. In Study 1, values that are missing from a given field form a set, a data record is an element, and a combination of fields that are missing together in one or more records is a set intersection. In Study 2, all occurrences of a given user interface action form a set, each user session is an element, and actions that occur in one or more sessions is a set intersection.

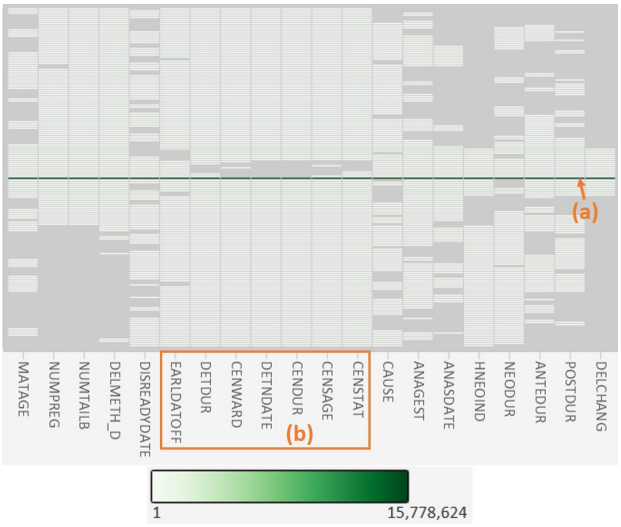

Figure 3: *All 162 ESIs from 20 fields extract. The most frequent ESI occurs in 15.7 million records and involves all fields (a). Seven fields are almost always missing together (b).*

3.1. Missing data in health records

This case study used an extract of APC data that had 20,724,064 records, and 65 fields that were missing 1–20,721,474 values. In total, there were 615,951,572 missing values.

3.1.1. Analysis and discussion

We analyzed the patterns of missingness in four steps. First, we computed the ESIs for the 65 fields, which generated 43,219 intersections of length 16 to 55. High-level statistics showed that the missingness has an unusual wave pattern (see Figure 1), but it is not feasible to visualize the composition of so many intersections.The most frequent intersection has 24 fields and 537,857 records, but accounts for only 12,908,568 (2.1%) of the missing values.

Step 2 used FIM to try to reveal the patterns. First, we set the minimum support threshold to 50%, but this generated a staggering 75,965,885 frequent itemsets. To reduce this number to a manageable size and find more meaningful, non-redundant itemsets, we increased the minimum support threshold to 90% and computed maximal frequent itemsets. This generated only nine maximal itemsets.

A heatmap illustrates the composition of these maximal itemsets and that they only involve 20 fields (see Figure 2). This important insight enabled us to split the dataset into two extracts. The first includes the 20 fields that are present in the nine maximal itemsets. The second extract includes the remaining 45 fields.

In Step 3, we computed ESIs in the 20 fields extract, which produced only 162 intersections. By visualizing these, we gain two main insights. First, the most frequent intersection (see Figure 3a) occurs in 15,778,624 records and involves all 20 fields (51.2% of the missing values in the whole dataset). Second, a subset of seven fields are missing together in 143 out of the remaining 161 ESIs (see Figure 3b). These 143 ESIs account for another 11.2% of the missingness in the whole dataset.

In Step 4, we computed the ESIs in the second extract of 45 fields. This generated 21,457 intersections, which is half the number that occur when all 65 fields are analyzed together and shows

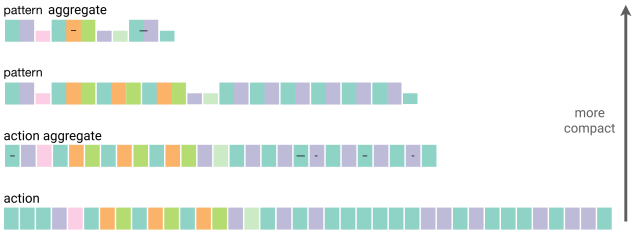

Figure 4: *The process of summarizing a session in a compact way with different event types having distinct colours. From bottom to top: consecutively identical actions are merged, sub-sequences of actions are replaced with mined patterns, and consecutively identical patterns are merged.*

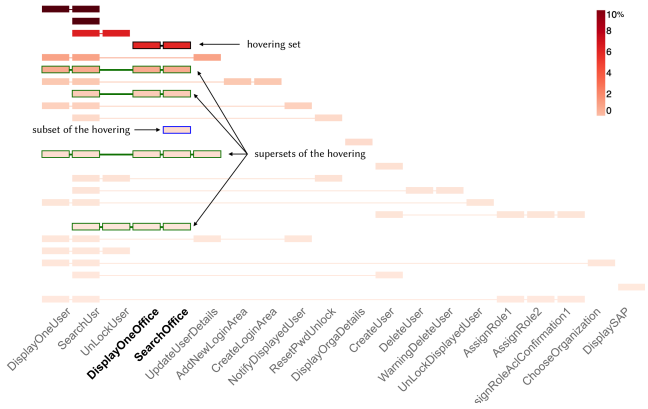

Figure 5: *The 25 most frequent session sets. Each row represents a session set and is colour-coded by its frequency.*

the advantage of dividing the data into primary and secondary groups (the 20 and 45 fields, respectively). The most frequent intersection involves 4 fields (2,572,920 missing values). The 100 most frequent intersections have 121,548,393 missing values (19.7% of the missingness in the whole dataset). Those 100 ESIs involve 2–17 fields, but a more detailed analysis is outside the scope of this paper.

To conclude, our approach of combining the FIM and ESI methods in a single visual analytics workflow allowed us to identify 244 ESIs (144 from Step 3 and 100 from Step 4) that account for a very large proportion of the missing values (82.1%) in our dataset. In further analysis iterations, one could repeat the four steps to explore the remaining 17.9% of the missingness, and keep iterating until it is all accounted for.

3.2. Cybersecurity logfile analysis

This case study discusses the analysis of event sequence data from application logs in a cybersecurity context. The primary goal of the analysis is to gain understanding of user behaviors through the actions they perform. A secondary goal is to explore potentially unusual behaviors. We analyze a dataset spanning 31 days on approximately 15,000 sessions performed by 1,400 users with 300 different action types (such as *SearchUser* and *DisplayOneUser*). Each session comprises a sequence of timestamped actions, with the longest one containing 893 actions and each session containing 15 actions on average.

3.2.1. Sequential frequent pattern mining

We consider a session as a sequence of actions and apply a constraint-based sequential pattern mining method [NTA*18] to extract frequent sub-sequences of actions that occur in the dataset. For our dataset, setting the support as 3% (the minimum portion of sessions containing a pattern) and the time gap as 60 seconds (the maximum duration between two consecutive actions) yields a meaningful and manageable set of frequent patterns.

Semantic patterns: Sub-sequences of actions that frequently occur can be considered as units with higher semantics than individual ones. For example, the actions *SearchUser* → *DisplayOneUser* → *UpdateUserDetails* could indicate some 'user update' activity. Therefore, the set of mined patterns describes, at a high level, what is going on in the dataset.

Session summary: We apply a series of operations to summarize a session in terms of raw actions and mined patterns. Briefly, we replace sub-sequences of actions with mined patterns (if matching) and merge consecutively repeated actions and patterns. Figure 4 shows this summarization and details can be found in the paper by Nguyen et al. [NTA*18]. This compact summary provides a way that could help gain understanding of sessions more quickly. Moreover, by showing the *usual* patterns within a session, we can spot the *unusual* actions that are left alone. This provides clues for further investigation. They may be mistakes, uncommon ways to do tasks, exploratory interaction, or more suspicious activities.

3.2.2. Exclusive set intersections

In this section, we analyze the data using the ESI method and discuss the insights that we gain. First, we consider a session as a set of actions (*session set*) instead of an ordered sequence, sacrificing the temporal information. For example, the session, $a \rightarrow b \rightarrow c \rightarrow a \rightarrow b$, would be converted into the session set $\{a,b,c\}$. Then, we apply ESI to compute all of the distinct session sets.

Visual design of session sets: Figure 5 shows the 25 most frequent session sets, ordered by frequency from top to bottom. Each row represents a session set and is colour-coded by its frequency. Mouse hover over a session set highlights its supersets in green and its subsets in blue. This helps users to explore relationships between session sets. For example, we learn that {*SearchUser*, *UnLockUser*} (5.49%) appears twice as often as {*SearchUser*, *DisplayOneUser*, *UpdateUserDetails*} (2.23%), but it co-occurs less with {*SearchOffice*, *DisplayOffice*}.

Comparison with SFPM: We retrieve the 25 most frequent patterns using SFPM (2.5% support threshold) and compare them with the aforementioned 25 common session sets. We observe that these two sets of patterns do not only share many patterns in common but also complement each other. On the one hand, several actions (in bold) appear in common session sets but are absent in the frequent patterns such as {***DisplaySAP***}, {*SearchUsr*, *Display-*

OneUser, ***ChooseOrganization***}, {*SearchUsr*, *DisplayOneUser*, ***UnLockDisplayedUser***}. This implies that ESI helps to identify subtle signals that are not detected with SFPM. On the other hand, some patterns that occur frequently (such as *MoveOgu* → *AdminOguStep1*) but are absent in the common session sets because they often occur together with other actions as well, thus cannot be detected with ESI as whole sessions.

Distribution of sessions: Because intersections are computed exclusively, session sets are considered as a whole rather than ordered combinations of actions as frequent patterns. This helps reveal the distribution of session sets in the dataset. It is interesting to see that the top 25 session sets account for quite a large portion of the whole dataset (47.47%). Taking the most common session set {*SearchUser*, *DisplayUser*} as an example, there are 121 distinct ordered sequences that are converted into this set. In summary, FIM is good for summarizing the sessions as a whole; whereas, SFPM is more suitable for summarizing frequent patterns occurred within the sessions.

Orderless methods for ordered data: The last observation is an abstraction of the use of ESI for event sequence data. The actions in this case study are timestamped, making it natural to choose an order-aware mining method for revealing additional temporal information. This can help users to understand the order of how an activity is commonly performed. However, there could be multiple ways of performing an activity using the same set of actions (but in different orders) and actions might be repeated unnecessarily due to mistake. Ignoring order loses the information; however, could help reduce noise, thus increasing the reliability of the found patterns. A combination of the two techniques could complement each other.

4. Towards a multi-method event analysis framework

Through the use of a suite of event mining methods, we demonstrate above how the overlapping and contrasting characteristics of the two case studies could be better tackled through visual analytics methods. Here, we first list a number of *observations* emerging from the two case studies:

- By setting a suitable high support threshold, FIM approaches offer significant reductions in data volume while preserving key subsets.
- ESI provides sets of events that co-occur "exclusively", thus highlighting important associations.
- Complementary results are provided when sequences are considered both as ordered and unordered sets.
- Visualization allows quick discovery of patterns that would be non-trivial to compute, e.g., the missingness patterns of seven fields (see Figure 3b).
- Where visual analysis enriched by sequential patterns provides a high-level understanding of the usual and unusual activities (Figure 4), the interactive analysis of exclusive set intersections reveals novel links within common subsets and supersets concurrently (Figure 5).
- Both case studies involved different analysis strategies in how the mining methods are used. While Case Study 1 follows a *pipeline approach*; i.e., output of FIM is the input for ESI, Case Study 2 involved a *parallel approach*; i.e., different techniques are applied in parallel and results compared and contrasted.

Underpinned by the observations above, we discuss a number of guiding principles towards a comprehensive event analysis framework in visual analytics:

Use multiple perspectives: To be effective, visual analytics systems should consider combining methods that handle events in conceptually diverse ways. For instance, ESI and FIM take an orderless approach, whereas SFPM preserves the order of events.

Interchange subset and fullset mining: One fundamental challenge in event analysis is that few sets (or sequences) occur often and many are infrequent (the "long tail"). Separating "core subsets" from the long tail leads to more effective observations.

Adopt diverse workflows: When multiple computational methods are combined, an inevitable diversity emerges in the sequence they are employed during the analysis. To be effective, workflows should adopt strategies in-line with each analytical task, such as the pipeline vs. parallel analysis strategies as discussed above.

5. Conclusion

This paper investigates how the use of multiple event mining methods within visual analytics enhances the reliability and scalability of the analytical process as well as opening up possibilities for novel insights. The two contrasting case studies, that require different ways of thinking due to their inherent differences, enable us to demonstrate how a multi-method approach provides novel findings and perspectives. In the first, we observe how streamlining two methods lead to the identification of a subset of ESIs that are significantly representative of the whole data. In the second, we demonstrate how ESI and SFPM provide complementary sets of patterns with different levels of representations in the data, leading to a more comprehensive and reliable coverage of patterns. We reflect on the observations made and discuss them within the context of a conceptual multi-method analysis framework and take the first steps towards a comprehensive visual analytics approach to analyzing event data that highlights three guiding principles for future visual analytics systems.

Even though the observations in this paper stem from two case studies, they already demonstrate the potential in taking a multi-method visual analytics approach for the analysis of large collections of event data. The results call for further work on extending the scope of our analysis framework through a more systematic review of existing methods and other application domains. More importantly, we highlight the significance and the potential space for novel visual analytics techniques that inherently support such holistic approaches within the analysis of events and their sequences, and call for further research in this area.

6. Acknowledgements

This research was supported by the EPSRC (EP/N013980/1), MRC (ES/L011891/1), ESRC (ES/L011891/1), and the European Commission through the H2020 programme under grant agreement 700692 (DiSIEM). Case Study 1 used a pseudonymized dataset from NHS Digital (NIC-49164-R3G5K). Due to data governance restrictions, the datasets cannot be made openly available.

References

[AMA*16] ALSALLAKH B., MICALLEF L., AIGNER W., HAUSER H., MIKSCH S., RODGERS P.: The state-of-the-art of set visualization. In *Comput. Graph. Forum* (2016), vol. 35, Wiley Online Library, pp. 234–260. 1

[AR17] ALSALLAKH B., REN L.: PowerSet: A comprehensive visualization of set intersections. *IEEE Trans. Vis. Comput. Graph. 23*, 1 (2017), 361–370. 1

[AR18] ADNAN M., RUDDLE R. A.: A set-based visual analytics approach to analyze retail data. In *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA18)* (2018), The Eurographics Association, pp. 37–41. 1

[FVLK*17] FOURNIER-VIGER P., LIN J. C.-W., KIRAN R. U., KOH Y. S., THOMAS R.: A survey of sequential pattern mining. *Data Science and Pattern Recognition 1*, 1 (2017), 54–77. 1

[FVLV*17] FOURNIER-VIGER P., LIN J. C.-W., VO B., CHI T. T., ZHANG J., LE H. B.: A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 7*, 4 (2017), e1207. 1

[LC10] LEUNG C. K.-S., CARMICHAEL C. L.: Fpvat: a visual analytic tool for supporting frequent pattern mining. *ACM SIGKDD Explorations Newsletter 11*, 2 (2010), 39–48. 1

[LDDH16] LIU Z., DEV H., DONTCHEVA M., HOFFMAN M.: Mining, pruning and visualizing frequent patterns for temporal event sequence analysis. In *IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis* (2016). 1

[LGS*14] LEX A., GEHLENBORG N., STROBELT H., VUILLEMOT R., PFISTER H.: UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 1983–1992. 1

[LIC08] LEUNG C. K.-S., IRANI P. P., CARMICHAEL C. L.: Fisviz: a frequent itemset visualizer. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2008), Springer, pp. 644–652. 1

[LKD*17] LIU Z., KERR B., DONTCHEVA M., GROVER J., HOFFMAN M., WILSON A.: Coreflow: Extracting and visualizing branching patterns from event sequences. *Computer Graphics Forum 36*, 3 (2017), 527–538. doi:10.1111/cgf.13208. 1

[LWD*17] LIU Z., WANG Y., DONTCHEVA M., HOFFMAN M., WALKER S., WILSON A.: Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 321–330. 1

[MA13] MAHMOOD T., AFZAL U.: Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In *2013 2nd national conference on Information assurance (ncia)* (2013), IEEE, pp. 129–134. 1

[MCB*11] MANYIKA J., CHUI M., BROWN B., BUGHIN J., DOBBS R., ROXBURGH C., BYERS A. H.: Big data: The next frontier for innovation, competition, and productivity, May 2011. URL: http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation. 1

[MKN*05] MUNZNER T., KONG Q., NG R. T., LEE J., KLAWE J., RADULOVIC D., LEUNG C. K.: Visual mining of power sets with large alphabets. *Department of Computer Science, The University of British Columbia* (2005). 1

[MLL*13] MONROE M., LAN R., LEE H., PLAISANT C., SHNEIDERMAN B.: Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics 19*, 12 (2013), 2227–2236. 1

[NTA*17] NGUYEN P. H., TURKAY C., ANDRIENKO G., ANDRIENKO N., THONNARD O., ZOUAOUI J.: A Visual Analytics Approach for User Behaviour Understanding through Action Sequence Analysis. In *EuroVis Workshop on Visual Analytics* (2017), The Eurographics Association. doi:10.2312/eurova.20171122. 2

[NTA*18] NGUYEN P. H., TURKAY C., ANDRIENKO G., ANDRIENKO N., THONNARD O., ZOUAOUI J.: Understanding user behaviour through action sequences: from the usual to the unusual. *IEEE transactions on visualization and computer graphics* (2018). 1, 3

[PBTL99] PASQUIER N., BASTIDE Y., TAOUIL R., LAKHAL L.: Discovering frequent closed itemsets for association rules. In *International Conference on Database Theory* (1999), Springer, pp. 398–416. 1

[PW14] PERER A., WANG F.: Frequence: Interactive mining and visualization of temporal frequent event sequences. In *International conference on Intelligent User Interfaces* (2014), ACM, pp. 153–162. 1

[RV10] ROMERO C., VENTURA S.: Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40*, 6 (2010), 601–618. 1

[RWA*13] RIND A., WANG T. D., AIGNER W., MIKSCH S., WONGSUPHASAWAT K., PLAISANT C., SHNEIDERMAN B., ET AL.: Interactive information visualization to explore and query electronic health records. *Foundations and Trends® in Human–Computer Interaction 5*, 3 (2013), 207–298. 1

[SS13] SAGIROGLU S., SINANC D.: Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (2013), IEEE, pp. 42–47. 1

[UKA04] UNO T., KIYOMI M., ARIMURA H.: Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *Fimi* (2004), vol. 126. 1

[WL14] WONGSUPHASAWAT K., LIN J.: Using visualizations to monitor changes and harvest insights from a global-scale logging infrastructure at twitter. In *Visual Analytics Science and Technology, 2014 IEEE Conference on* (2014), IEEE, pp. 113–122. 1