ViSeq: Visual Analytics of Learning Sequence in Massive Open Online Courses

Qing Chen, Xuanwu Yue, Xavier Plantaz, Yuanzhe Chen, Conglei Shi,
Ting-Chuen Pong, *Member, IEEE* and Huamin Qu, *Member, IEEE*

Abstract—The research on massive open online courses (MOOCs) data analytics has mushroomed recently because of the rapid development of MOOCs. The MOOC data not only contains learner profiles and learning outcomes, but also sequential information about when and which type of learning activities each learner performs, such as reviewing a lecture video before undertaking an assignment. Learning sequence analytics could help understand the correlations between learning sequences and performances, which further characterize different learner groups. However, few works have explored the sequence of learning activities, which have mostly been considered aggregated events. A visual analytics system called ViSeq is introduced to resolve the loss of sequential information, to visualize the learning sequence of different learner groups, and to help better understand the reasons behind the learning behaviors. The system facilitates users in exploring learning sequences from multiple levels of granularity. ViSeq incorporates four linked views: the projection view to identify learner groups, the pattern view to exhibit overall sequential patterns within a selected group, the sequence view to illustrate the transitions between consecutive events, and the individual view with an augmented sequence chain to compare selected personal learning sequences. Case studies and expert interviews were conducted to evaluate the system.

Index Terms—MOOC, online education, visual learning analytics, event sequence visualization.

1 Introduction

Attracting unprecedented interest worldwide, Massive Open Online Courses (MOOCs) has developed rapidly, producing from less than 10 to more than 9000 by over 800 universities [1]. This modern education wave has drawn millions of people to learn and communicate at an unprecedented scale. With such high enrollment and student-teacher ratios [2], instructors greatly need to receive feedback from students on their learning styles to help improve teaching activities. Compared with traditional classrooms, MOOC learners come from around the world with various backgrounds and motivations, which increases the difficulty for learning behavior analysis.

Fortunately, many MOOC platforms record the web log data of learners, including learner profiles, video clickstream interactions, and forum activities, which shed light on the research on MOOC data analytics of learning behaviors. The web log data contains temporal information of when each student performs which kind of action, such as, viewing a lecture video, posting a question to the forum, and submitting an answer to a problem. Although most MOOCs are carefully designed with a suggested learning sequence, students are unrestricted in following it. The student can adjust the designed sequence to facilitate personalized active learning, which significantly differs from traditional classroom education. For example, some students may skip certain video content and go directly to the assignments and exams in order to gain a certificate, while others may be active in forums but fail to finish any assessments. Thus, this variety in learning sequences is essential for instructors to analyze learning behavior. Furthermore, the sequential data could help characterize different learner groups according to their learning habits and distinguish the intentions of learners in taking MOOCs to better understand the relationship between learning behaviors and outcomes.

The research on MOOC data analytics has increased in the past five years as the MOOC data became available [3], especially in the data mining and machine learning domains. Researchers have applied various data mining approaches in studying learning behaviors, including the early detection and prediction of student dropout [4], low performance on quizzes or exams [5], and learner models for personalized content delivery and recommendations [6], [7]. Explaining intricate learning behaviors and the effects of learning sequence and performance are still difficult even though the aforementioned works have helped discover a few learning patterns, such as identifying the type of learners who are more likely to drop out [8]. However, most of these models are strictly based on statistical grounds with a few dominant features while some specific aspects which are meaningful to instructors might be ignored. Therefore, it could hinder the instructors' investigation of those non-dominant features which are practical to their specific contexts, let alone make possible adjustments purely based on these statistical models.

To address this issue, data visualization and visual analytics have been introduced to provide visual representations of the data and multiple interaction techniques to help instructors observe, explore, compare, and understand learning sequences. Recently, much effort has been devoted to visually analyzing student behavior, for example, Shi et al. [9] and Chen et al. [10] have used both basic and complex visualizations so instructors can explore the video clickstreams and the peaks in clickstream data. However, these works have mainly focused on an aggregated level, which eliminates sequential information. Meanwhile, some studies have concentrated on learning sequence analysis and discovered some insights. For example Guo & Reinecke [2] found most learners

- *Qing Chen, Xuanwu Yue, Xavier Plantaz, Yuanzhe Chen, Ting-Chuen Pong and Huamin Qu are with the Hong Kong University of Science and Technology.*
  E-mail: {qchenah,xyueaa,xpfplantaz,ychench,tcpong,huamin}@cse.ust.hk
- *Conglei Shi is with Airbnb Inc.*
  E-mail:shiconglei@gmail.com

exhibited a non-linear exploratory activity sequence, especially those from low student-teacher ratio countries.

However, previous works have not focused on visually analyzing learning sequences from multiple levels of granularity, which hindered instructors from gaining insights into high-level behavioral patterns. For example, a student may check a problem, review previous videos, then attend a forum discussion for assistance, and finally complete the assignment. In this example, a series of actions are performed before completing the assignment. The whole sequence could help instructors understand how explicitly learners work toward a certain problem. Sometimes, instructors are interested in a particular student or a representative of a certain learner group, thereby, providing an individual-level sequential visual exploration that aids them in further analyzing learner experience accurately. A recent review [11] by Vieira et al. reviewed that only 4 papers proposed visual analytics systems which included all five characteristics of "sophistication" including multiple visualizations, connected visualizations, visualizing data at multiple levels, interactive visualizations, novel visualizations.

To address the issues mentioned above, an interactive visual analytics system called ViSeq is introduced in this paper, to detect sequential patterns with sequential pattern mining methods, to visualize the learning sequences of different types of students from the multiple levels of granularity, and analyze the underlying stories or reasons behind their learning behaviors. ViSeq consists of four views, namely, the projection view to identify learner groups based on their learning sequence similarities, the pattern view to identify overall sequential patterns within a selected learner group, the sequence view to illustrate the transitions of learner flows among different videos and consecutive events, and the individual view to compare individual learning sequences selected from other views. Furthermore, case studies based on real datasets of three different MOOCs are conducted and several domain experts are interviewed to evaluate the system. The results and feedback received demonstrated the value of ViSeq.

The key contributions of this paper are as follows:

- An interactive multi-level visual analytics system that helps instructors explore various types of learning sequences to detect different learner groups and understand the potential correlations between learning sequences and performances.
- A novel augmented sequence design to discover the individual learning activity sequences.
- Insights gained from case studies based on real datasets and expert interviews to guide an effective course design.

2 Related Work

In this section, the most relevant studies are surveyed, covering recent research on MOOC analytics, learning sequence analytics, and event sequence visualization.

2.1 MOOC Analytics

Learner engagement and outcomes. With high enrollment and low completion rate, considerable effort has been made to study how to engage learners in the online environment, as well as which factors influence learning outcomes and their correlation with student profiles (e.g., age, gender, and country) and activity patterns (e.g., watching a video, submitting an assignment) [12].

MOOC design and curriculum. Novel methods and tools have been studied to better design a MOOC, for example, fun activities have been introduced into the system by Maalej et al. [13] to launch the course in a serious game scenario. Some research helps design automatic feedback tools for essay writing and programming; Balfour and Stephen evaluate how peer evaluation can be effectively allocated to different people besides instructors and teaching assistants [14]. Content analysis is also addressed with automated and manual approaches in the education data mining domain by Peña-Ayala [15].

Self-regulated learning (SRL). SRL, which provides personalized learning, has been studied to adapt the teaching experiences and meet specific learning requirements of individuals. Student modeling has also been used in the MOOC environment to shape different domains that characterize the learner, such as prior knowledge, learning strategies, and skills, offering opportunities for dropout prediction by Yang et al. [16] and retention enhancement along with personalized education.

Social network analysis and networked learning. Social connections in MOOCs have recently received much attention. For example, Dowell et al. used language and discourse as tools to explore their connections with performance and social centrality [17]. They found that learners tend to obtain a significant and central position in their social network when they engage with a narrative-style discourse. Moreover, Lee et al. [18] examined the transitions of student flows over a series of courses to determine the social connections among students.

Learner motivations and success criteria. Learner motivations are more diverse in online courses than in traditional education. For example, the survey by Willkowski et al. showed that only 10 % of learners reported that they aimed to earn a certificate [19]. Although numerous research has focused on analyzing the final performance of learners, the success criteria in the MOOC scenario are still questioned by researchers due to the varied motivations towards MOOC learning by Dowell et al. [17].

The aforementioned categories are not mutually exclusive because some works exert efforts in multiple directions. Several works have employed visualization techniques to identify patterns and to analyze learner group behaviors by Coffrin et al. [20]. However, most of these works still focused on aggregated behavior rather than from the sequential perspective, which hinder instructors from gaining insights into high-level behavioral patterns.

2.2 Learning Sequence Analytics

Abundant research efforts, mainly by building learner models with automatic or semi-automatic methods, have been dedicated to learning sequence detection and classification to aggregate student activities. Many powerful personalized mechanisms have been designed, such as adaptive navigation support and curriculum sequencing by Chen et al. [21], considering the interests and browsing habits of students to promote effective and efficient learning. Some works have explored the exhibited behavior of a learner around a single activity, particularly in problem solving. Shanabrook et. al proposed a semi-automatic approach to identify the state of a student while working toward a single assignment [22]. First, high-level student behavior is identified using sequence-based motif discovery, and then seven distinct meaningful groups are summarized from the 30 motifs grouped. More recently, Sinha et al. [23] tried to discover hidden structural configurations in MOOC learning sequences. Their focus is on the extraction of active and passive participants and finding their learning sequence patterns so that they can help predict dropouts in the future. However, the learning sequences in this paper

are aggregated based on their action types, such as from video interaction to forum activities, regardless of the standard sequence provided by the instructor. In contrast, our work not only considers the sequential patterns before each assignment and aggregated action types, but also takes into account the general learning sequence of all types of activities toward knowledge comprehension. In this manner, instructors can elaborate the hidden connection among different teaching units, fill in any missing parts in the designed knowledge graph, and in turn adjust course materials to ease the learning burden of students.

Therefore, although learning sequence has been exploited and several applications with personalized learning sequence recommendation have been developed, most works either fall outside the MOOC context or lack the visualization techniques to support the exploration of the sequential learning patterns.

2.3 Event Sequence Visualization

Much work has been accomplished to visually analyze the event sequence data. The most prevalent visualization used is the alignment of events along a horizontal timeline, such as in Lifelines2 by Plaisant et al. [24] and CloudLines by Krstajic et al. [25] because the sequential data contain temporal information. This encoding works well for individual level exploration on presenting the exact sequence of events, however, obstacles are encountered when dealing with large numbers of individual sequences, which increases the cognitive load for users to find valuable insights.

Recent work on complex datasets has taken advantage of the automatic computation and the flexibility of interaction exploration, such as the implementation of a self-organizing map to cluster and visualize clickstream data by Wei et. al [26]. Wang et al. [27] also discussed various unsupervised clustering methods in their clickstream analysis by partitioning a similarity graph. Coco [28] explored the results of SPM models to help identify cohorts compared with extracting subsequences for the medial data with curated sorting and filtering interactions by Malik et al. [29].

Frequence by Perer et al. [30] enhances the sequential pattern mining (SPAM) bitmap representation to handle real-world constraints and to use flow visualizations showing transitions. Liu et al. [31] selected the vertical mining of maximal sequential patterns (VMSP) algorithm [32] instead of SPAM to obtain a more compact set of patterns, provided various interactions to support easy navigation and exploration between abstraction levels. Some applications employed query interfaces, such as DecisionFlow by Gotz et al. [33] and COQUITO by Krause et al. [34], to facilitate interaction and help users with distinct goals to easily find their target patterns. In another work, Du et al. provided techniques for users to select a sequence and identify those that are similar to the selected sequence based on certain distance metrics [35].

We are inspired by all these works and implement sequential pattern mining methods and various interactions for data exploration at different levels of granularity. In particular, our work targets the learning sequence analysis in MOOCs, which contributes to visual complexity and data diversity. Meanwhile, for users exploring the data, we not only offer exploration from the event sequence aspect, but also allow them to filter students based on various attributes, such as grade and time. Therefore, the interactions require more careful design and feedback from end users with rounds of iterative development.

3 Problem Characterization

In this section, we describe the data preprocessing procedure, summarize the analytics tasks, and identify four major design rationales for designing the system.

3.1 Data Preprocessing

We extracted learner profiles and learner web log data from the databases of three edX courses. The preprocessed learner profile mainly includes learners' performance, since the lack of information about the country, age, and educational background information of learners does not allow making sound assumptions. The web log data included various types of events (i.e., video watching, problem access, and discussion posting). Table. 1 presents the general statistics of the three courses we exploited later for the case studies.

Every event performed is recorded in a chronological order, particularly for the web log. For each learner, we first extracted all the events conducted by him or her throughout the whole course period with the exact timestamp. Then, we formalized each learner's events as a sequence S = [E_1, E_2,...,E_i,...,E_n], where i denotes the index of the event and n is the total number of events in the sequence. In the interests of grouping learners based on their learning sequences, we needed to first build a similarity matrix based on their learning sequences. However, since the sequences are too long and are of different lengths, we tried to extract subsequences from the original sequences as features to compare the similarity. Referring to Wang et al.'s work [27], T_k was defined as the set of all possible k-grams (k consecutive elements) in sequence S: $T_k(S) = \{k-gram | k-gram = (s_j s_{j+1} ... s_{j+k-1}), j \in [1, n+1-k]\}$. To make two different sequences the same length for computing the distance, we calculated both the common k-grams in the two sequences and their count. Suppose the two sequences are S_1 and S_2, with a given k, we first computed the set of all possible k-grams from both as $T = T_k(S_1) \cup T_k(S_2)$. Thus we transformed the two sequences into the same length. Then, we calculated the normalized frequency of each k-grams within each sequence $l (l = 1, 2)$ as array $[c_{l_1}, c_{l_2}, ..., c_{l_n}] (n = |T|)$ where c_{l_j} indicates the normalized frequency for the jth k-gram in sequence l. Finally, we computed the normalized polar distance between the two arrays $D(S_1, S_2) = \frac{1}{\pi} \cos^{-1} \frac{\Sigma_{j=1}^{n} c_{1j} \times c_{2j}}{\sqrt{\Sigma_{j=1}^{n} (c_{1j})^2} \times \sqrt{\Sigma_{j=1}^{n} (c_{2j})^2}}$. This normalized distance ranges from 0 to 1, with a small number indicating a small distance (i.e., high similarity) between the two sequences. In our dataset, we tested different ks from 1 to 10 and finally selected $k = 5$, which presents the most distinct patterns. As to the computation complexity, this method has $O(n)$ complexity where n is the number of sequences. This method is scalable in terms of speed when computing a large number of sequences. Besides, the speed is also affected by the average length of sequences and the total number of all possible k-grams, but in a real MOOC dataset, these two terms often have a specific range. Therefore, the speed is mainly affected by the number of sequences. In our case, with each chosen k, it took us 3 to 4 hours to compute all the sequences. After building this similarity matrix, we projected learners into 2D space based on their similarity.

For the projection method, we utilize t-SNE because it is superior to the existing techniques in creating a single map that reveals structures of various scales by Maaten et al. [36]. This technique visualizes high-dimensional data by converting similarities between data points to joint probabilities and providing

Course	Platform	#Event Logs	#Learners	#Weeks	#Videos	#Problems	#Forums	#Distinct Events
J1	Edx	20558353	18832	11	122	89	8	219
J2	Edx	11933314	33402	7	65	67	6	138
E1	Edx	5879876	19073	6	46	102	6	154

TABLE 1: Overview of the courses information

each data point a location in a 2D map. By applying t-SNE to our dataset, we can see groups more clearly compared with Multidimensional Scaling (MDS) projection which is also piloted in the design process. Referring to the well-established open-source sequential data mining library called SPMF by Fournier-Viger et al. [37], there are several possible choices regarding the type of data we have, maximal sequential pattern mining (MaxSP, VMSP) without loss of information and sequential generator patterns (VGEN, FEAT, FSGP) with loss of information. We tested each result with the minimum support 0.1 and finally chose the VMSP algorithm because of its conciseness. Maximal sequential patterns are appropriate in the learning sequence context since a more compact set of patterns are often required, and it also works faster than other methods. Each time the user selects a group of learners, the sequential mining algorithm will run again with the average time cost within one minute. Therefore, we decided to use VMSP so as to efficiently mine maximal sequential patterns in a compact representation. Moreover, the customer visit path data from Liu et al. [31] are similar to ours, so their choice of the VMSP also gave us some support on the decision made.

3.2 Task Analysis

Since the beginning of 2014, we have started our collaboration with several MOOC instructors, with previous explorations and visual analytics on the aggregated behavior of video clickstreams and forum interactions. Apart from the previous interviews with domain experts, we conducted a literature review of previous works on MOOC sequence analytics (including [20], [38], [39], [40], [41], [42], [43]), and further extended the analytics tasks at different levels of detail.

T1. Can we identify any learner groups based on learning sequences, and how are they generally distributed? Identifying a learner group is a crucial research task in learning sequence analysis. Instructors want to determine the grouping requirements for students based on their learning sequences so as to have a rough idea about the general distribution of the learners in each group. This overview helps instructors to rapidly drill down their interest to a specific group.

T2. What are the typical learning sequence patterns for different learner groups? With prior knowledge of the log data, instructors fully understand that seeing all the learning sequences of all the students is impossible. They are more interested in understanding the typical patterns of various learner groups. Moreover, the instructors sometimes have already had specific sequences in mind to explore. Therefore, pattern query functionality is required in the system to offer customized exploration.

T3. What is the overview of non-linear transitions between consecutive events for different learner groups? Although most MOOCs have a designed learning sequence, students are not required to follow it. They can adjust the designed sequence to facilitate personalized learning, which greatly differs from traditional classroom education. A common question raised by the instructors is that, "what happens before or after students conduct event X?" In this paper, these preceding or succeeding events are considered to be consecutive events, and the designed transitions between consecutive events are considered as linear transitions. A situation may occur in which students jump to the materials of a different week while studying for the current week's materials. Hence, a visualization for exploring the overall transitions between weeks for different learner groups are required.

T4. What are the consecutive events for a designated event of interest? Given a rough understanding of the overview transitions between consecutive events, sometimes more detailed information of the exact event sequence is required. One scenario has been raised several times by instructors during our previous interviews: some students go directly to assignments and later go back to watch some corresponding videos, while others tend to complete the assignment after watching the video. For a designated event, such as a specific assignment, knowing how users conduct their activities before and after this particular event is interesting and valuable. In particular, if the overall student performance of a designated event is out of the expectations of the instructors (either too high or too low), then the consecutive events could offer hints to explain why this occurred.

T5. What is the temporal difference for learning sequences? Temporal information affects how students learn, especially when they learn for the first time and review for quizzes or exams. Therefore, the learning sequences might exhibit different patterns for various time periods. A typical situation happens when students prepare for the final exam. Some may go through the materials of different weeks in the same setting, which manifests itself as jumping back and forth between various learning events. Therefore, users need the temporal filtering function to help them investigate such patterns.

T6. What is an individual learning sequence? During our previous interviews, instructors mentioned the individual level analysis several times. One instructor remarked that, "The previous visualizations are quite useful for me to discover learner groups, however, if I want to understand exactly how a student has been doing over different course periods, there is currently no tool to help me achieve this goal." In [44], Graesser also envisioned the research on analyzing unique learning sequences of individuals. However, how to drill down to a specific learner also poses challenges for the design of our system. A system that could support different levels of granularity is helpful.

T7. How is each individual sequence different from others and how to find similar or dissimilar individuals? After investigating an individual sequence, another question is raised as end users want to know whether other learners share similar sequences with the selected individual. Users could rapidly identify students of interest and compare different individual sequences straightforwardly with a similarity sorting function. It might enlighten instructors on giving advice to students, for example, "based on the performance of the students who share a similar learning sequence with you, you would have a high probability of catching up with the course by revisiting these learning units."

3.3 Design Rationales

Since the system is designed especially for course instructors and education analysts, several major design rationales are identified from interviews and feedback from end users.

R1. Facilitate quick identification of learner groups and most frequent patterns. One of our primary tasks for learning sequence analysis is to identify learner groups with similar patterns. It is critical for users to identify unknown learner groups based on sequential behaviors, as well as finding the learner groups of interest with determined directions for grouping. Therefore, some automatic methods should be provided to help present visualizations to facilitate quick identification of learner groups. Meanwhile, some query functions could further support users' determined groups to balance the fully automatic methods for pattern discovery and users' specific needs according to their interests and domain knowledge.

R2. Display an adequate amount of information step by step. It is often impossible to present all the information at once. Our ultimate goal is to design a visualization interface that is easy to use while conveying essential information for users to determine their interaction operations for the next step. Therefore, we adopt the "overview first, details on command" [45] mantra and display the visualizations from simple to complex, and from familiar to novel. Our system provides four different visualizations including scatterplot, chord diagram, arc diagram, and augmented chain chart. We introduce these visualizations from simple to complex while following the analytics tasks from general to detailed.

R3. Support interactive filtering with immediate feedback. The system provides both direct selection on the visualizations and the separate control panel for exact filtering. For example, users indicate a keen interest in **T5** to identify temporal differences for learning sequences. Therefore, the temporal range is selected as a brushing criterion in the sequence view. With a greater need to study learner groups with different grades, interactive filtering for grades is available for each view so that the visualized patterns of the selected group in different views could appear simultaneously for correlation analysis. Moreover, when users select a particular transition in the sequence view, a corresponding grade distribution histogram is shown to indicate how the selected group of learners perform. This gives users an immediate visual feedback, which is important for designing fluid interactions by Elmqvist et al. [46]. Moreover, the query in the pattern view also make the system by Yi et al. [47] more responsive .

R4. Design consistently and link appropriately for multiple views. As indicated by Wang et al. [48], several rules should be employed when using multiple views. The rule of consistency is a must as we keep the color and shape encoding consistent for all the weeks and types of events. Moreover, we consider linking between different views as one of our most important interactions for exploring patterns of different learner groups. The attention management corresponds to the tasks users perform. Each view serves for one or several specific tasks and complement each other regarding various data aspects or different levels of detail.

4 Visualization Design

4.1 System Overview

Upon entering the system, users can select a course to explore from the menu at the top. In the main interface, the system presents the projection view and the pattern view on the left, the sequence view on the right, and the individual view in the pop-up window on the menu bar. The projection view is presented with a scatterplot for users to identify learner groups based on their learning sequence similarity. In the pattern view, all sequential patterns detected by the VMSP algorithm are listed with their bar charts to the left showing the frequency of occurrence.The Sequence View is constituted by an interactive three-level chord diagram on consecutive sequence transitions. The pattern view and the sequence view are updated by selecting a group of learners from the projection view, and the individual view appears with the individual learning sequence of learners throughout the entire course period from the pop-up window. In addition, each view has an independent control panel to filter learners grades, as well as the active time period over the course timeline. The system also provides a history callback function that allows users to record the current explored data before investigating it at the next level. Users can detect the patterns for the specific learner groups or individuals of their interests by exploring multiple views iteratively.

4.2 Visual Encoding

Based on the analytic tasks identified from the literature review and expert interviews (Section 3.2), several visual designs have been proposed. Furthermore, multi-level views are aligned in accordance with the design rationales in Section 3.3. In particular, to maintain consistent encoding, we consider the major visual channels for our visual designs of the learning event in different views. Color, shape, and texture are the top three possible channels for categorical data according to Munzner's book [49]. However, each event can only take a small space given the excessive number of events. Therefore, texture encoding is excluded, whereas color and shape channels are retained. Three shapes, namely, rectangles, circles, and triangles are used to encode three types of learning events, namely, video watching, assignment solving, and forum discussion, respectively. Diverging colors are selected to represent different weeks. The exact encoding is labeled on the ring of the chord in the sequence view. Twelve is set as the upper bound as most MOOCs are designed with less than 12 weeks of content. Users can then perceive the subsequent weeks easily because they are marked with similar colors.

4.2.1 Projection View

To facilitate investigation into meaningful patterns, we sorted the mined sequential patterns after applying the VMSP algorithm in the pattern view, which further supports the search function to allow users to filter interested patterns. We employ the projection view based on learners' sequence similarity across the entire course period to discover learner groups and determine the overall learner distribution across different groups (**T1**). The projection view is presented as a scatterplot of all the students, in which each student is represented by a node. The algorithms for calculating sequence similarity and projection are discussed in Section 3.1. The predefined time period is the entire semester. Users can also use the control panel to filter a more specific duration. The control panel also facilitates filtering based on learner grades. For example, in Fig. 1 learners with grades over 60 are shown by default. We find two evident groups with a large number of learners and the rest scattered around the border. This kind of projection is helpful to narrow down a specific learner group and eventually drill down to individual sequences.

Interactions. Linking and Brushing are implemented in the projection view. Users can either brush on the control panel for

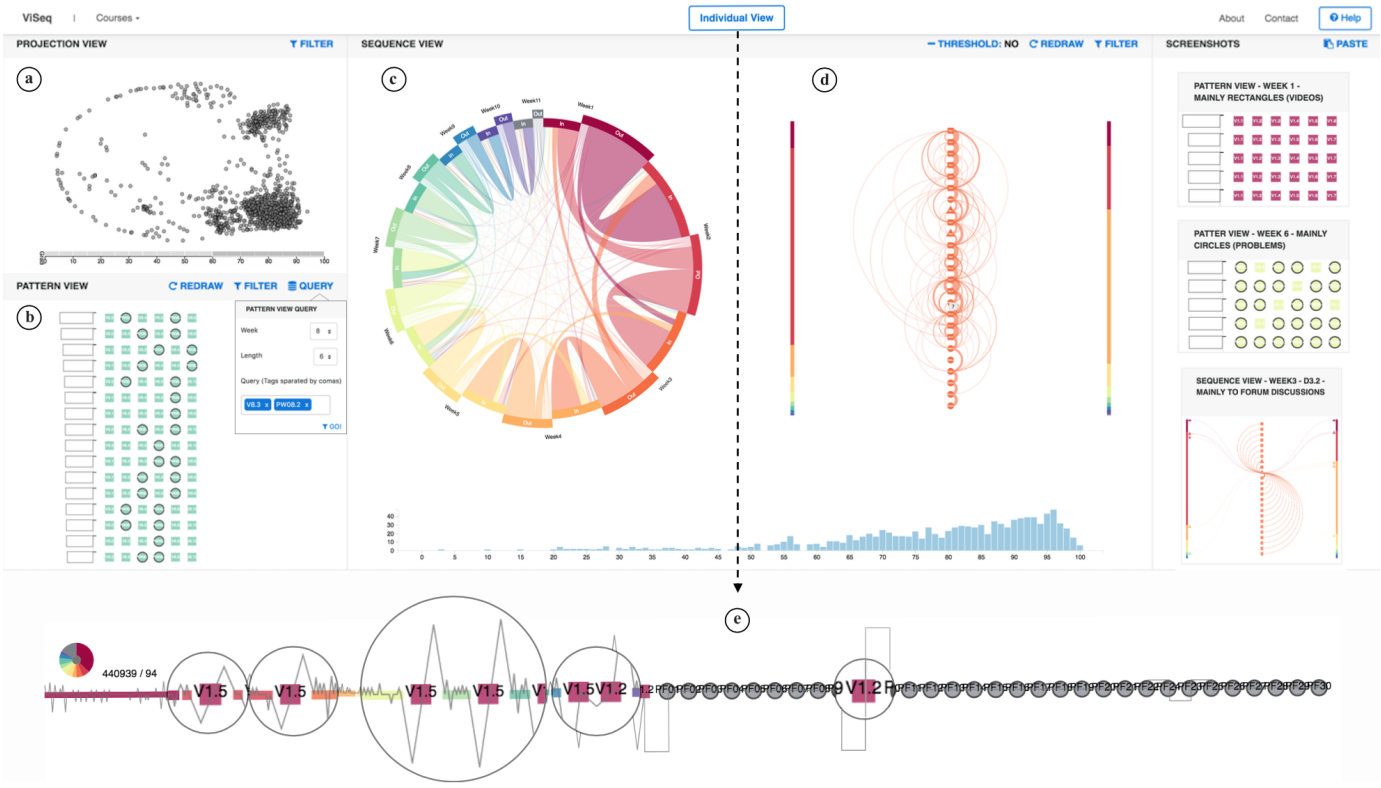


Fig. 1: ViSeq system consists of four major views. The projection view (a) identifies learner groups based on their sequence similarity throughout the whole course period; the pattern view (b) presents the mined sequential patterns within a selected learner group; the sequence view shows the non-linear consecutive events between weeks (c) and inside a selected week (d); the individual view (e) helps explore individual learning sequence and find similar individuals. The screenshot function aims to record stories found and to return to previous explorations.

learner grades and time duration or select directly by dragging a rectangle onto the scatterplot. To facilitate the flexibility of choices of the latter type, multiple selections of areas on the scatterplot are provided. Thus, users can make small modifications on the items selected for more accurate selection (**R1**). While brushing alone is of limited use according to Kosara et al. [50], we also linked this view with others so that users can see the visualizations of the same group of learners in different views (**R4**).

4.2.2 *Pattern View*

The pattern view presents users with typical learning sequence patterns for different groups of learners (**T2**).

After selecting a learner group from the projection view, the pattern view presents all the sequential patterns mined from VMSP with their bar charts to the left, showing the occurrence frequency. The length of each bar chart shows the number of learners with the sequential bar chart. However, the number of patterns might still be large even after the first-round of filtering learners from the projection view.Therefore, various attributes allow users to examine patterns by either sorting according to the pattern length, or classifying them based on different time durations (i.e., weeks). Each row presents a mined sequential pattern with each event marked in colored rectangles, circles, and triangles. Moreover, as mentioned in **T2**, instructors sometimes have clear ideas about the learning sequences where they want to search by inputting some constraints. Hence, we offer the pattern query function to filter the designated events, as displayed in Fig. 1 (b). Both the inclusion and exclusion of a certain event or a set of events are provided to facilitate user exploration.

4.2.3 *Sequence View*

Sequence View consists of a three-level interactive graph to visualize the non-linear transitions between consecutive events from different levels of detail (**T3, T4**) and various perspectives (**T5**). At the first level as shown in Figure 1 (c), we use a two-way chord diagram to demonstrate the aggregated transitions between different weeks (**T3**). We separate the incoming and outgoing flows between different weeks. The width of each flow shows the number of learners with that transition. An evident observation is that the most outgoing flow goes to the next week, and the most incoming flow comes from the previous week. The number of people who experienced this path is displayed on the tooltip when hovering over the flow between two weeks.

One of the tasks we aim to solve with this design is to compare the learning sequences among different learner groups. To fulfill this task, the filtering function is provided for users to compare different learner groups with various ranks of overall performances throughout a selected time range. As an example illustrated in Fig 2, the time range is set from the beginning of the course to the week before the exam, so that we can examine how students with various grades behave differently in the regular course period. Suppose we rank the students' grades from A to E (high to low), O stands for the overall student body. With the radial layout in the chord, we can better compare the distribution of different corresponding chords. In this figure, we can clearly compare the differences in the non-linear sequence transitions among various weeks for learners with different grades.

Two other design alternatives have also been taken into con-

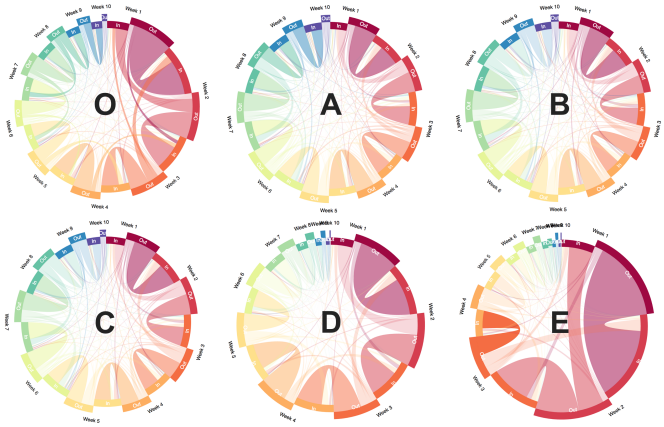


Fig. 2: The first-level sequence view shows the non-linear sequence transition flows through different weeks. O stands for the overall student body, and the other five groups of students with different ranks of overall performance (from A to E indicating high to low grades) are compared.

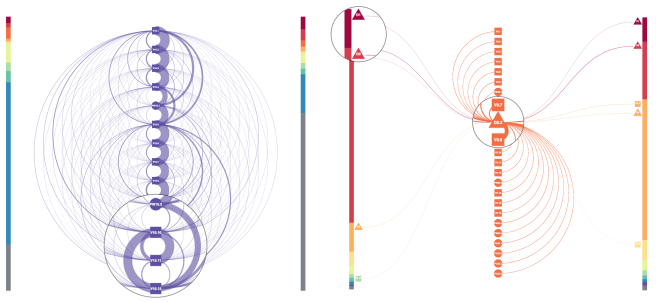


Fig. 3: The second and third level of the sequence view. (a) highlights a majority of sequence transitions between consecutive events within the selected week; (b) illustrates the preceding and succeeding events of a selected forum discussion event.

sideration. The first one is using a matrix to represent the two-way transition relationship, where the opacity of each element in the matrix shows the number of incoming and outgoing transitions. However, since the grid in the matrix is of the same size, the number of transitions can only be discerned by color, which is not easy for comparison, especially when they do not align. Then we turn to another common choice, the sankey diagram. However, the more levels there are in the sankey diagram, the more complicated and space-consuming it is to draw on the screen. The chord diagram is more space-efficient and can show the transitions more clearly. Moreover, it is more aesthetically pleasing according to the end users and fits our current layout better. Concerning the scalability issue, the matrix design is definitely preferable. However, when many transition flows occur with a small number of learners, a threshold can be set by the users to hide these flows in the chord diagram to easily observe the majority flows. Meanwhile, if the edges are too dense, we can also choose to do edge bundling, however, at the current stage, it works fine so we do not perform the edge bundling algorithm.

A histogram with the grade distribution of these learners appears at the bottom by clicking on the transition flow, which is illustrated at the bottom of the sequence view in Figure 1. Accordingly, the individual view is updated with each individual sequence of these students. This histogram was implemented after gathering feedback from the prototyping interviews, as one instructor mentioned the need for an overview of the filtered results on students' grade distributions, which could provide hints for deciding whether it would be worth further investigating more details in the individual view.

The second level opens when users click on the edge of the circle to a designated week to explore event sequence transitions in the same week. An arc diagram is then shown with the structured events aligned vertically in the selected week. The arcs on the right indicate forward transitions, while the arcs on the left show backward jumps. This arc diagram provides much information on how students jump between different learning events normally, given that most sequence transitions occur within each week are often related to the same key topic. However, even if we identify some sequence paths such as in Fig. 3 (a), many students might jump from PW10.2 to V10.12. Then, maybe these students cannot easily follow the course; hence, they go back to V10.10 and V10.11 to catch up. We are uncertain whether these students are the same group given that the transitions between two events are merely the aggregated sum. To confirm these assumptions, users can click on the sequence to filter the learning sequences of the same group. A more concrete example is illustrated in Fig.7 with descriptions in Section 5.2.

After exploring the overall sequence flows between different weeks and in the same week, as mentioned in **T4**, the instructors are also concerned about the incoming and outgoing flows of a specific learning event. Therefore, the next step from the second level is to select on the designated event, where the third level expands with two color bars on both sides of the arc diagram. The small multiples aligned on the two vertical color bars indicate the top destinations of the flows coming into and going out from the selected event. Both the relative scale and the absolute scale have been implemented for different weeks at the second level. For example, Fig. 3 (b) illustrates the preceding and succeeding events of a selected forum discussion event. We later find that the antecedents and sequelae of a discussion event are more likely to be another discussion event.

Interactions. Users can interact with the sequence view in several ways so that information could be presented step-by-step during exploration (**R2**). First, users can hover over transition flows, and a tooltip is shown with the precise number of the transition (**T3**). Second, they can click on week edges to perform week selection. When a week is selected, the next-level sequence view is expanded to display the in-week transitions (**T4**). If users click on the sequence path, the sequence view supports an additional visualization with a histogram at the bottom showing the grade distribution of these learners. In this way, users can get instant feedback on the filtering result and decide whether to explore next (**R3**). Third, the time range and learner grades are also available for filtering learners in the control panel (**T5**).

4.2.4 *Individual View*

After drilling down from the entire student body to a specific group, the next level of exploration is the individual view, where we present each individual learner sequence horizontally (**T6**). The individual view is in a separated pop-up window rather than as part of the main interface. This is to relieve the cognitive burden to switch between the context as all the other views aim at group analysis (**R2**). We created an augmented sequence chain design for individual-level exploration Fig. 4. The events are aligned in a

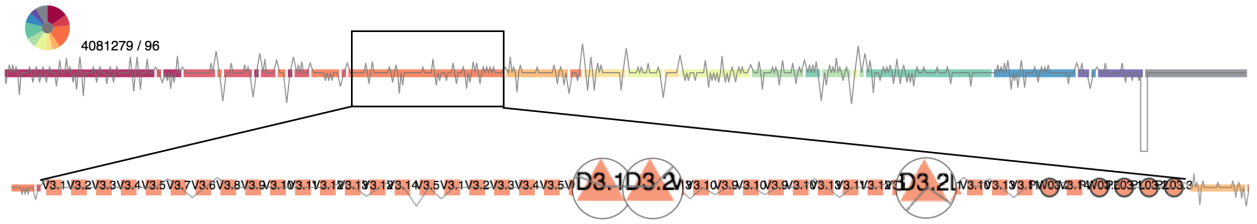


Fig. 4: An augmented chain sequence design is displayed with two levels of detail, the aggregated level and the expanded detail level.

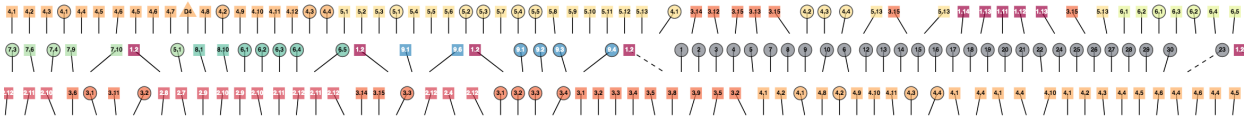


Fig. 5: The alternative balloon design corresponds the slope of lines to the forward or backward distance of two consecutive events.

chronological order. Each small multiple represents an event, with the shape indicating the event type and color representing the week that the event belongs to. All the events performed sequentially by the same student are aligned horizontally. However, we use an overlaid line chart to encode the distance between consecutive events. If the student follows the designed course sequence strictly, we can draw a straight line to connect the consecutive events. However, if the student skips several learning units, an upper triangular line is used to connect the consecutive events. Similarly, if the student goes back to review the previous learning units, a lower triangular line is drawn. When one of the events comes from final exam problems, a rectangular line is applied to show transitions more explicitly. The height of the triangular or rectangular line indicates the relative distance between two consecutive events. Therefore, a more fluctuated line indicates that the learner often skips between different learning units; whereas a smooth line indicates that the student strictly follows the designed learning sequence (as illustrated in the expanded sequence segment from Fig. 4) . We also considered an alternative design using a gestalt line to indicate the transition distance in Fig. 5. In this "balloon" glyph, the slope of the line below each event corresponds to the forward or backward distance of two consecutive events, and the dashed line indicates jumps from the final exam. Although it saves the vertical space to some extent and helps detect forward or backward movements, the instructors found it difficult to perceive the trend compared with the augmented sequence chain design.

Interactions. Since the learner sequence is sometimes extremely long (some exceeding 500 events), it poses new challenges for exploration and comparison interactively. To allow users to see longer sequences at once, the system offers an aggregation mode to condense the events in the same week (visually appearing in the same color) together as a color bar, and the users can expand any segment by clicking on it (**R2**). The pie chart in front of each sequence shows the aggregated percentage of events from various weeks. Learners with low grades are much more likely to contribute more in the initial one or two weeks, whereas learners with high grades tend to perform more uniformly in each week. If a learner with high grades presents a pie chart with a high proportion in the first several weeks, users might be interested in investigating such individuals. To facilitate searching for similar learners, the previously calculated arbitrary similarity distance between every two learners is used in the individual view (**T7**). Moreover, we provide users with a brushing functionality to allow them to identify students who share similar learning sequences given that the similarity judgments are subjective. When users select an interested individual, the other individual sequences are sorted according to the similarity with the selected individual.

5 Case Studies

We applied ViSeq to three MOOCs offered by our university. The first two are the same course with different offering rounds and modified course structures as one consisted of 10 weeks' materials (referred to as J1), the other (J2) comprises only 5. The third course (E1) is a language course with 6 weeks' materials.

5.1 Learner grouping and typical sequential patterns

Upon entering the system, we first observe the projection view and filter learners with different grade ranges. When we filter those with passing grades (i.e., score above 60 points out of 100) for the three courses (Figure 6), two evident groups are identified for the first two courses (J1 and J2, marked as 1 and 2), whereas only one obvious group appears for the third course (E1 as 3). We select these groups to examine their learning sequences in the individual view. By comparing the individuals' learning sequences from these two groups, we find that the major difference lies in the way they work in the final exams. For all the three courses, learners are allowed to access any course material during the final exam; hence, the final exam sequence patterns can vary for different individuals. The groups marked in orange rectangles are constantly referring to previous weeks' materials, as a large number of rectangular lines are shown in (A) which indicates the frequent jumps from the exam problems to previous materials. The other groups marked in blue rectangles, as opposed to the orange ones, presents a linear exploration when completing the final exam as there are very few rectangular jumps in (B). Since the first two courses are in nature the same one, learner behaviors tend to be similar, where the two groups described could both be recognized. On the contrary, the third course has only one group marked in orange. The sequences of this group (C) exhibit similar patterns as in (A), which implies there is no group of learners who manage to finish the final exam without referring to previous materials in the third course.

From the pattern view, we then explore all the mined sequences from different weeks of various lengths. Almost all the most frequent patterns follow the course structure except the last exam week, wherein some learning units may be skipped but no reverse patterns are found, or at least not those frequent enough to be detected as a common pattern. All the forum discussion units in the course structure except the first week did not appear in the most frequent sequential patterns either. Apart from the discussion

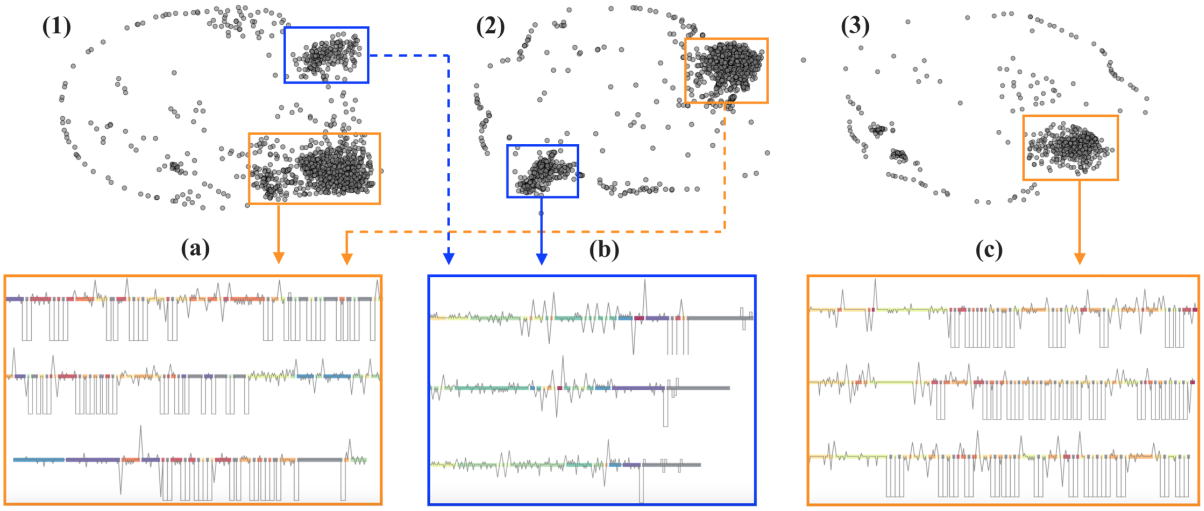


Fig. 6: Three projection views (1) ∼ (3) from the three MOOCs are presented with different groups. (A) ∼ (C) shows the typical individual sequences for the identified groups, where (A) and (C) indicates similar patterns with frequent jumps (rectangular lines) during the final exam week.

units, for the first few weeks, the skipped learning units are mainly assignment problems. However, in later weeks, most problems are included in the most frequent patterns. This condition is probably caused by the "diligent" students who worked hard on assignments and remained until the later half of the course. Another assumption is that those who persisted in the course have determined the "right" approach to learn the course by practicing the assignments. In particular, week 6 includes the largest number of assignment problems, and most of the frequent sequential patterns are focused on the problems, such as "PW06.1-PW06.2-PW06.3-PW06.4-V6.8-PW06.5-PW06.6" (referring to the second screenshot in Fig. 1(b)). Even in week 7, many learners still worked on week 6 problems. The reasons behind such patterns were later discussed with course instructors in the interviews. The instructors reported that, in the first five weeks, all the basic components are introduced, and starting from week 6, more practical or advanced problems are proposed, which might have caused the students to spend more time and effort on week 6 problems. Students are required to apply the basic concepts learned from previous weeks to the week 6 problem solving. This is also one of the reasons why instructors decided to split the course into two halves for the next-round offering, so that students can determine whether they are prepared for the next level of study.

5.2 Non-linear transitions from different levels of detail

Next, we explore the sequence view to discover the non-linear transitions between consecutive events. From the first level of transitions between different weeks illustrated in Fig. 1 (c), the most frequent outgoing flow is from the current week (that is, week N) to the next week (week N+1) because the inner week transitions are ignored here. Similarly, the most frequent incoming flow is from the current week (week N) to the previous week (week N-1). Week 6 has the largest percentage of incoming flows from various weeks other than the previous week, which is almost 50%. This result verifies our observation from the pattern view.

In Section 5.1, it is discovered that there are two typical ways learners to work on the final exam, so we wonder whether the way various learners prepare for the final exam also differs. Therefore, we filter the time period before the final exam in the sequence view and find that learners with varied grades tend to revise differently. For example, for the J1 course, before the final exam, given that we filter the sequence view in week 10 (Fig. 7), students with high-scores already started to review videos from the first two weeks, whereas few students with low-scores jumped back to previous weeks to review for the final exam. Then, during the final exam, the students with low-scores still focused on the materials from the last two weeks, whereas students with high-scores still navigated through previous weeks other than the last week, and extremely few jumped back to week 9. Students with low-scores could probably not catch up to the last part of the course or they do not exert sufficient effort to review previous components.

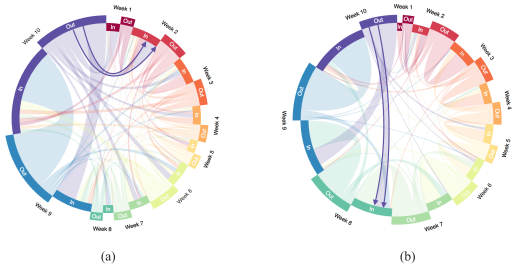


Fig. 7: In the week before the final exam, learners with high scores started to review videos from previous weeks, shown in (a) with a large proportion of sequence transitions flowing from week 10 to week 1 and 2; whereas most learners with low grades were still struggling with the materials from the previous week, displayed in (b) with the large proportion of sequence transitions flowing from week 10 to week9.

After seeking patterns from the first level, we open the second level by clicking on the border of a random week to allow instructors to explore the flow of student transitions through the learning units in each week, like week 8 for an example. The flows with fewer than 100 learners are hidden by setting the threshold to 100. Fig. 8 shows that several evident non-linear sequence paths exist given that some students jumps from V8.6 to V8.10 (marked with path 1), go back to V8.7 (path 2), continue with V8.8 and V8.9 (path 3), and finally skip the previously viewed V8.10 to get

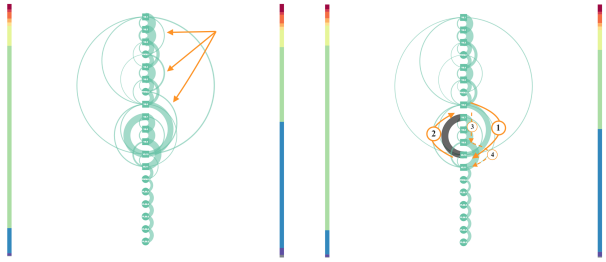


Fig. 8: After setting the threshold to display the learning sequence transitions (a) within week 8, users select one specific sequence path to filter the learning sequences from the same group of learners (b) to validate the assumption of the case in Section 5.2.

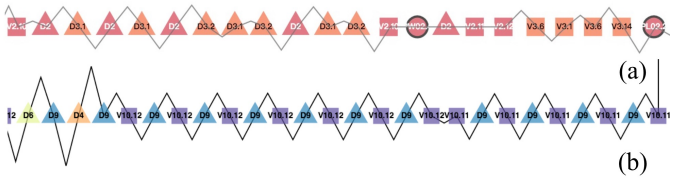


Fig. 9: (a) A typical sequence segment for those students jump to discussions regularly. (b) An outlier who frequently jumps from learning unit D9 to V10.11. constantly.

back on the right track with the course designed order (path 3). This series of behavior probably indicates that the students who skipped V8.7 to V8.9 found them difficult to follow; thus, they jumped backward to make up for the skipped content in order to catch up. However, we are still uncertain whether they are the same group of students because the transitions between two events are merely the aggregated sum. To confirm these assumptions, we click on the backward jump sequence from V8.10 to V8.7 in order to filter the learning sequences of the same group of learners. The filtered results show that the majority of the original flows remained, and only a few forward jumping sequences disappeared (as the orange arrows pointed out), such as from V8.1 to PW8.1. The number of students who jumped directly from the first video to the first assignment in this week and from the PW8.1 and PW8.2 to forward events, decreased. Therefore, the students who skipped V8.7 to V8.9 and those who went back from 8.10 to V8.7 are not the same students who regularly jump through learning units.

5.3 Individual sequence exploration and comparison

When the students are filtered from the pattern view and the sequence view, we could then explore individual sequences. From the third level of the sequence view, we select the students who jump from a forum discussion to watch a previous week's video and then finish an assignment (from D3.2 to V2.10 to be exact). A common question raised by end users is that whether these students tend to refer to the forum a lot during their whole learning period. Since the individual view is updated accordingly after the selection in the sequence view, we find many of these individual sequences include a few short colored segments with a non-linear order. By clicking on these segments to expand the events, most of such events correspond to the forum discussion activities, which validates the assumption (Fig. 9 (b)). For another example, the selected learner in Fig. 1(e) often the revisited video from the first week when learning latter weeks materials, as we can observe abrupt jump lines and the appearance of the red color bar among later sequential color bars. By clicking on it and expanding the folded color bar, exact activities show that video 1.5 and video 1.2 are the most reviewed videos. There are also some outliers who constantly transit between certain learning units, in most cases, the transition is between a specific video and a discussion unit. For example, in Figure 9 (b), the learner frequently jumps from D9 to V10.11 and V10.12. Such frequent jumps are usually small jumps between consecutive weeks, which might give instructors hints on reinforcing the connection between these learning units.

6 Expert Interview

We conducted in-depth interviews with four domain experts to evaluate how our system works for real datasets. The first two experts are MOOC instructors (MI1, MI2) who have had the experience of working with us before on MOOC visual analytics. The other two are university lecturers (UL1, UL2) who had never heard of our system before and only viewed some basic data statistics provided on e-learning platforms. In each interview, we first asked them about their background knowledge on MOOC related analytics, then introduced them to the purposes and features of the system with a live demonstration tutorial. Following that, they could ask questions and explore ViSeq on their own to understand the capabilities of the system. To ensure they understand how the system works, we asked them to perform some small tasks which are listed as follows.

- Identify the number of learner groups after filtering a certain grade range in the projection view. (T1) Example: how many learner groups can you find with grades greater than 60?
- Find the most frequent pattern in one week with a given length in the pattern view. (T2) Example: in week 8, what are the top three most frequent patterns with a length of 6?
- Explore the three-level sequence view step by step to identify a required transition between consecutive events. (T3∼T5) Example: how many students have conducted consecutive events from week 3 to week 5 and what is their grade distribution?
- Observe the updated individual view and seek out the most similar individuals of the given sequence. (T6∼T7) Example: what is the longest jump backwards that this individual student has taken?

After that, we conducted a semi-structured interview regarding system usability, visual designs, and interaction. Their suggestions for the system and other potential uses are discussed below. Each interview lasted for around 90 minutes.

System usability. All the experts were very impressed by the system, especially the two ULs who had never used such analytics tools before. We asked the two MIs who have experience working with us not to be constrained by the guided tasks, and to explore scenarios of interest to them and from their own experience. MI1 commented, *"although previous video-based analytics are useful, this one gives me more direct feedback to make adjustments for course design"*. For a direct query function from the mined patterns in the pattern view, MI2 appreciated both the inclusion and exclusion of a certain event or a set of events, *"it's now easier for me to filter events of interest quickly"*. However, MI2 also mentioned that he sometimes found some potential patterns later on when exploring individual level sequences, then when he wanted to verify his assumption, it would become a little difficult to formulate all the possibilities of event combinations for that potential pattern. If the system could allow users to include

some descriptive words in their query, it would be more user-friendly. Apart from the pattern view, the two ULs are interested in the sequence view as well. UL1 suggested we could add each assignment problem's average score in the small multiple design, *"it would be great if the system could give me direct hints on which assignment might be problematic in the sense of being either too difficult or too easy for the students"*.

Visual design and interactions. The visual designs were appreciated by the experts. They found the circular chord diagram *"visually appealing"* and *"would like to explore it more"*. UI2 commented that the transition comparison among learner groups with different grades was helpful, where he was constantly using the filtering function in the sequence view. MI2 particularly liked the individual view and found the similarity sorting functionality useful to find students with similar learning sequences, *"this could help me give direct feedback to my students in the real classrooms when they encounter similar problems"*. Since he also asked his students to watch the lecture videos before coming to the classroom, he said this function could help him identify those who need specific guidance. Moreover, he suggested sorting learners based on the similarities of a selected segment of sequences. UL1 also found the augmented chain sequence design interesting and easy to understand, *"individual patterns found in this view can work as a good example to explain what students actually do in their learning process."*. He commented that the augmented chain design could be easily extended to other time series data with a predefined sequence order, since *"The wave shape is catchy and the two levels of detail is quite flexible to show both the aggregated and detailed information"*. MI1 mentioned it would be better to show some content of each learning unit, such as the title and keywords for the video to fully understand why students perform some non-linear learning behavior. The interactions were considered helpful in general. For the interactions between different views, UL1 and UL2 felt some degree of confusion in the trial. However, MI1 and MI2 did not encounter any difficulty when exploring the system. In this light, the training process of using such analytics systems still requires improvement.

7 Discussion

In this section, we discuss several system limitations, design reflections learnt from the process, potential generalizations for such applications, and the implications for the education domain.

7.1 System Limitations

Loss of temporal information for the similarity metric. We choose the current similarity metric because it can better reflect the similarity of sequences with varying lengths, especially when compared with a traditional similarity metric such as editing distance. However, this metric lost temporal information. To reduce such information loss, we plan to explore hashing based techniques to speed up the computation. In this way, we can update the projection view with filtered sequences when users select a specific time period. Therefore, different behaviors are less likely to be computed as similar.

Concerns about the scalability issue. In the projection view, we display the learner group with grades higher than 60 as the default setting. If we show all the nodes, they will gather together so that users could barely identify any particular group. Even though scatterplot is one of the visualizations that have good scalability, with 20k learners, the overlapping problem still appears. One way to deal with it is using visual aggregation such as glyphs. The other approach is to employ interactive techniques such as focus+context visualization.

Lack of content-aware analysis of learning sequences. The pre-defined learning sequence in this paper follows a linear structure derived from the course database. However, the structure arrangement of learning elements is sometimes considered by the course instructors which could not be retrieved directly from the current data. In this case, the learning elements aligned might not be in a straight line. Thus, it requires some layout algorithms to draw the arcs between different components. Concept-based maps like booc.io by Schwab et al. [51] and KTGraph by Zhao et al. [52] for collaborative sensemaking are inspirations for such improvements. Regarding content analysis, the forum text analytics with the video content is also an exciting topic. An example task is to examine whether students who previously failed at a specific assignment turned out to get the correct answer after discussing it in the forum or reviewing a particular video. Instructors can then evaluate the setting of the assignments and whether students can seek help from the platform.

Lack of support for side-by-side comparison during the exploration process. When designing the system interface, we considered having a separate comparison view. On the one hand, a side-by-side comparison function could minimize some cognitive load for the comparison task. On the other hand, it might cause the interface and interactions to become more complicated, which in turn increase the cognitive overhead (e.g., context switching). From our interviews, two users supported the idea of adding a comparison view while the other two said they did not care. However, one did mention that he wanted to have a tool to help organize what he had explored. For future work, narrative techniques could help convey analytics stories and communicate among instructors, education experts, administrators, and students.

Lack of connection between discovered patterns to the outcomes Currently, our system provides some correlated analysis between learning sequence patterns with performance. However, more visual aids could further connect the discovered patterns to the outcomes. For example, in the pattern view, the average grades of learners who perform each mined pattern can be explicitly displayed on the bar using color transparency or an overlayed line to indicate whether this sequence is conducted by the high-performance group or the low-performance group. Other learner attributes can also be added such as the drop-off rate.

7.2 Design Reflections

Interactions can be complicated with multiple linked views. Users sometimes need guidance to interact with the system more efficiently. Although our system supports learner group filtering to identify individuals with specific characteristics, it is not easy to interact between group analysis and individual analysis. For example, if users identify a specific individual, the current visualization would not show the individual's sequence in the group-level visualization. A possible solution is by adding tags to learners during the exploration process so that we could further analyze and discover some general patterns for learners with the same tags. Then, different types of learners with multiple sequencing attributes can be defined and analyzed, which is a standard way in the education domain to discover new learner groups.

Our study also implies that the transitions among different weeks' materials indeed vary and have a different reviewing

patterns for students with varying grades. The decision then leaves to the users as to how to nudge the students, provide recommendations and adequate intervention. The scatterplot is useful in identifying learner groups with similar sequences; the chord diagram explicitly shows "non-linear" transitions between different weeks and the arc diagram to show the transitions in the same week. By displaying one level after another, the interface looks less cluttered and can reduce cognitive load (**R2**). Since T4 is intrinsically a subtask of T3, the general shape of the chord diagram is preserved with an animated transition when users select a specific week and explore the arc diagram (**R4**).

7.3 Generalizations

While event sequence data is ubiquitous, such applications can not only be exploited in the online education domain but also in other event sequence data analysis. Most of the functions offered in the current system could be generalized. For example, similar systems can be implemented to analyze online shopping behavior which contains the user log data. Likewise, other online platforms which record user logs, such as video viewing websites, can perform a similar analysis. Apart from the web log data, there are other scenarios concerned with event sequence data on which to apply such applications. For example, in the medical domain, event sequence data can be essential for cohort exploration and symptom diagnosis. Another example is the career path recommendation and event sequence related decision and sensemaking problems. Different views from the system could help perform similar tasks to those mentioned above. In this light, ViSeq is quite a versatile system which can be adapted to various application domains. However, these datasets display some differences to our online learning data, the most obvious one being that we have a default sequence. While this is a major concern, we believe that in most cases, the experts can manually define a default order based on their experience. Take the online shopping clickstream as an example, analysts probably know how a typical customer would behave. Once a standard order is defined, they could integrate our system. The current limitation is that we could not allow users to adjust this standard since it will cause us a lot of time to recompute the sequence. It could be our future work to consider how to enhance the efficiency of this process and allow users to freely manipulate different baselines for a standard sequence order.

7.4 Implications for the Education Domain

After a thorough review of sequence analytics in the education field and the discussions with the education expert, we further summarize three potential directions in the education domain.

The first direction is to utilize students' actual learning sequences to examine pedagogical assumptions and inform instructors for course redesign. In a recent work, Lai et al. [53] conclude with implications for designing online learning experiences, "such as activity sequence, can have a meaningful impact without increasing learning time." One limitation mentioned is "the small number of participants, the lack of collaborative group work and the scope of the cultural backgrounds used." Our work complements this work as they could use the system with MOOC data for a more significant number of participants from various backgrounds. If we had more courses to compare the differences in learning sequences of students from various educational backgrounds, it would bring more benefit to the education domain.

The second is to facilitate individualized recommendations and to promote self-regulated learning. Maldonado-Mahauad et al. [43] identified three types of learners; comprehensive learners who follow the course structure; targeting learners who engage in some specific course materials which could help them pass assessments; and sampling learners who are more unsettled without clear goals. They also defined six types of self-regulated learning strategies: goal-setting strategies, strategic planning, self-evaluation, task strategies, elaboration, and help-seeking. We also have some similar findings. For example, the help-seeking type is shown in Figure 9(a) as a typical sequence segment for those students who regularly jump to forum discussions. However, without proper theoretical models from the education domain, it is not in our expertise to conclude different types of learning strategies. Therefore, to support self-regulated learning is another research direction for visual analytics in data-driven education. One suggestion from a recent review paper by Vieira et al. on visual learning analytics is that "future studies can explore the effect of using visual learning analytics on student metacognitive development and self-regulation, and how to provide personalized formative feedback using visual learning analytics" [11], which is in accordance with our findings.

The third research direction is to follow the traits of "successful" students when designing a student-facing visualization dashboard. We could further involve some social comparison theory to increase engagement and reduce the dropout rate by nudging students and presenting their behavior compared with slightly-better-performing peers. We can then examine how the use of social motivation and peer pressure could influence low-performing learners. The need for student-facing visualization tools are also mentioned in the review paper mentioned above.

8 Conclusion and Future Work

This paper presents ViSeq, an interactive visual analytics system that helps instructors explore learning sequences in MOOCs and understand the causes and results behind learning sequences. We described the key views of ViSeq, including: (a) the projection view to discover learner groups; (b) the pattern view to identify the most frequent sequential patterns; (c) the sequence view to explore the transitions between consecutive events; and (d) the individual view to present each learner's sequence and compare similar individuals. Various interaction techniques are employed, such as filtering, searching, highlighting, sorting, and history callback. We also introduced the methods used for similarity calculation and learner grouping, along with the analytics tasks summarized from interviews with domain experts. Case studies and expert interviews demonstrated the usefulness and effectiveness of the system.

While the initial results are promising, there is still room for improvement in the future. First, the grade information for each assignment raised from the interviews suggests directions for exploring more meaningful patterns. Second, the roles of age, language, and educational culture might also help identify different learner groups and understand how these different learner attributes affect learning sequences. In the future, we will try to retrieve such information and apply it to multi-level analysis. Moreover, visualizing longer sequential patterns for group-level analysis is also a useful but challenging task for potential future work. High-order networks (HoN) could be applied to study longer sequential patterns by creating additional nodes to represent

higher-order dependencies [54]. We would like to further explore such visualizations for HoNs to tackle this issue.

ACKNOWLEDGMENTS

The authors thank Qiaomu Shen for his help on the interactivity part and Ling Li for her feedback from the pedagogy perspective. This work is supported by the Innovation Technology Fund of Hong Kong under Grant No. ITS/306/15FP.

REFERENCES

[1] "By the numbers: Moocs in 2017," https://www.class-central.com/report/mooc-stats-2017/.

[2] P. J. Guo and K. Reinecke, "Demographic differences in how students navigate through moocs," in *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 2014, pp. 21–30.

[3] D. Gasevic, V. Kovanovic, S. Joksimovic, and G. Siemens, "Where is research on massive open online courses headed? a data analysis of the mooc research initiative," *The International Review of Research in Open and Distributed Learning*, vol. 15, no. 5, 2014.

[4] S. Halawa, D. Greene, and J. Mitchell, "Dropout prediction in moocs using learner activity features," *Experiences and best practices in and around MOOCs*, vol. 7, 2014.

[5] S. Jiang, A. Williams, K. Schenke, M. Warschauer, and D. O'dowd, "Predicting mooc performance with week 1 behavior," in *Educational Data Mining 2014*, 2014.

[6] A. Ramesh, D. Goldwasser, B. Huang, H. Daumé III, and L. Getoor, "Modeling learner engagement in moocs using probabilistic soft logic," in *NIPS Workshop on Data Driven Education*, vol. 21, 2013, p. 62.

[7] M. Wen and C. P. Rosé, "Identifying latent study habits by mining learner behavior patterns in massive open online courses," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 1983–1986.

[8] D. Davis, G. Chen, C. Hauff, and G.-J. Houben, "Gauging mooc learners' adherence to the designed learning path," in *Proceedings of the 9th International Conference on Educational Data Mining (EDM). Raleigh, NC, USA*, 2016.

[9] C. Shi, S. Fu, Q. Chen, and H. Qu, "Vismooc: Visualizing video clickstream data from massive open online courses," in *Visualization Symposium (PacificVis), 2015 IEEE Pacific*. IEEE, 2015, pp. 159–166.

[10] Q. Chen, Y. Chen, D. Liu, C. Shi, Y. Wu, and H. Qu, "Peakvizor: Visual analytics of peaks in video clickstreams from massive open online courses," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 10, pp. 2315–2330, 2016.

[11] C. Vieira, P. Parsons, and V. Byrd, "Visual learning analytics of educational data: A systematic literature review and research agenda," *Computers & Education*, vol. 122, pp. 119–135, 2018.

[12] K. F. Hew, "Promoting engagement in online courses: What strategies can we learn from three highly rated moocs," *British Journal of Educational Technology*, vol. 47, no. 2, pp. 320–341, 2016.

[13] W. Maalej, S. Msaed, P. Pernelle, and T. Carron, "Adaptive and playful approach in the mooc: thanks to serious game," in *Digital Information Management (ICDIM), 2014 Ninth International Conference on*. IEEE, 2014, pp. 201–204.

[14] S. P. Balfour, "Assessing writing in moocs: Automated essay scoring and calibrated peer review (tm)," *Research & Practice in Assessment*, vol. 8, 2013.

[15] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert systems with applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[16] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé, "Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses," in *Proceedings of the 2013 NIPS Data-driven education workshop*, vol. 11, 2013, p. 14.

[17] N. M. Dowell, O. Skrypnyk, S. Joksimovic, A. C. Graesser, S. Dawson, D. GaĹevic, T. A. Hennis, P. de Vries, and V. Kovanovic, "Modeling learners' social centrality and performance through language and discourse." *International Educational Data Mining Society*, 2015.

[18] Y.-C. Lee, W.-C. Lin, F.-Y. Cherng, H.-C. Wang, C.-Y. Sung, and J.-T. King, "Using time-anchored peer comments to enhance social interaction in online educational videos," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 689–698.

[19] J. Wilkowski, A. Deutsch, and D. M. Russell, "Student skill and goal achievement in the mapping with google mooc," in *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 2014, pp. 3–10.

[20] C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy, "Visualizing patterns of student engagement and performance in moocs," in *Proceedings of the fourth international conference on learning analytics and knowledge*. ACM, 2014, pp. 83–92.

[21] C.-M. Chen, H.-M. Lee, and Y.-H. Chen, "Personalized e-learning system using item response theory," *Computers & Education*, vol. 44, no. 3, pp. 237–255, 2005.

[22] D. H. Shanabrook, D. G. Cooper, B. P. Woolf, and I. Arroyo, "Identifying high-level student behavior using sequence-based motif discovery," in *Educational Data Mining 2010*, 2010.

[23] T. Sinha, N. Li, P. Jermann, and P. Dillenbourg, "Capturing" attrition intensifying" structural traits from didactic interaction sequences of mooc learners," *arXiv preprint arXiv:1409.5887*, 2014.

[24] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman, "Lifelines: visualizing personal histories," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1996, pp. 221–227.

[25] M. Krstajic, E. Bertini, and D. Keim, "Cloudlines: Compact display of event episodes in multiple time-series," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2432–2439, 2011.

[26] J. Wei, Z. Shen, N. Sundaresan, and K.-L. Ma, "Visual cluster exploration of web clickstream data," in *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 2012, pp. 3–12.

[27] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, "Unsupervised clickstream clustering for user behavior analysis," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 225–236.

[28] 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 *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 2015, pp. 38–49.

[29] S. Malik and E. Koh, "High-volume hypothesis testing for large-scale web log analysis," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2016, pp. 1583–1590.

[30] A. Perer and F. Wang, "Frequence: interactive mining and visualization of temporal frequent event sequences," in *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 2014, pp. 153–162.

[31] Z. Liu, Y. Wang, M. Dontcheva, M. Hoffman, S. Walker, and A. Wilson, "Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 321–330, 2017.

[32] P. Fournier-Viger, C.-W. Wu, A. Gomariz, and V. S. Tseng, "Vmsp: Efficient vertical mining of maximal sequential patterns," in *Canadian Conference on Artificial Intelligence*. Springer, 2014, pp. 83–94.

[33] D. Gotz and H. Stavropoulos, "Decisionflow: Visual analytics for high-dimensional temporal event sequence data," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1783–1792, 2014.

[34] J. Krause, A. Perer, and H. Stavropoulos, "Supporting iterative cohort construction with visual temporal queries," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 91–100, 2016.

[35] F. Du, C. Plaisant, N. Spring, and B. Shneiderman, "Eventaction: Visual analytics for temporal event sequence recommendation," *Proceedings of the IEEE Visual Analytics Science and Technology*, 2016.

[36] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[37] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C.-W. Wu, V. S. Tseng *et al.*, "Spmf: a java open-source pattern mining library." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3389–3393, 2014.

[38] K. Stephens-Martinez, M. A. Hearst, and A. Fox, "Monitoring moocs: which information sources do instructors value?" in *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 2014, pp. 79–88.

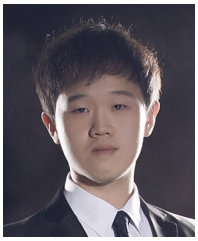
[39] C. Zinn and O. Scheuer, "Getting to know your student in distance learning contexts," in *European Conference on Technology Enhanced Learning*. Springer, 2006, pp. 437–451.

[40] R. Mazza and V. Dimitrova, "Informing the design of a course data visualisator: an empirical study," in *5th International Conference on New Educational Environments (ICNEE 2003)*, 2003, pp. 215–220.

[41] A. M. F. Yousef, M. A. Chatti, U. Schroeder, and M. Wosnitza, "What drives a successful mooc? an empirical examination of criteria to assure

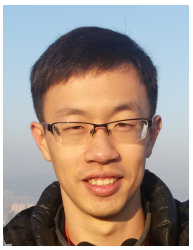
design quality of moocs," in *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on*. IEEE, 2014, pp. 44–48.
[42] H. Fournier, R. Kop, and H. Sitlia, "The value of learning analytics to networked learning on a personal learning environment," in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. ACM, 2011, pp. 104–109.
[43] J. Maldonado-Mahauad, M. Pérez-Sanagustín, R. F. Kizilcec, N. Morales, and J. Munoz-Gama, "Mining theory-based patterns from big data: Identifying self-regulated learning strategies in massive open online courses," *Computers in Human Behavior*, vol. 80, pp. 179–196, 2018.
[44] S. D. Sparks, "Data mining gains traction in education," *Education Week*, 2010.
[45] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996, pp. 336–343.
[46] N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly, "Fluid interaction for information visualization," *Information Visualization*, vol. 10, no. 4, pp. 327–340, 2011.
[47] J. S. Yi, Y. ah Kang, and J. Stasko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.
[48] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky, "Guidelines for using multiple views in information visualization," in *Proceedings of the working conference on Advanced visual interfaces*. ACM, 2000, pp. 110–119.
[49] T. Munzner, *Visualization analysis and design*. CRC Press, 2014.
[50] R. Kosara, H. Hauser, and D. L. Gresh, "An interaction view on information visualization," *State-of-the-Art Report. Proceedings of EUROGRAPHICS*, pp. 123–137, 2003.
[51] M. Schwab, H. Strobelt, J. Tompkin, C. Fredericks, C. Huff, D. Higgins, A. Strezhnev, M. Komisarchik, G. King, and H. Pfister, "booc. io: An education system with hierarchical concept maps and dynamic non-linear learning plans," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 571–580, 2017.
[52] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan, "Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 340–350, 2018.
[53] P. K. Lai, A. Portolese, and M. J. Jacobson, "Does sequence matter? productive failure and designing online authentic learning for process engineering," *British Journal of Educational Technology*, vol. 48, no. 6, pp. 1217–1227, 2017.
[54] J. Tao, J. Xu, C. Wang, and N. V. Chawla, "Honvis: Visualizing and exploring higher-order networks," in *Pacific Visualization Symposium (PacificVis), 2017 IEEE*. IEEE, 2017, pp. 1–10.

Qing Chen Qing Chen received the BEng degree in digital media technology from the Department of Computer Science, Zhejiang University and her Ph.D. degree in the Department of Computer Science and Engineering from the Hong Kong University of Science and Technology (HKUST). Her research interests include information visualization, visual analytics, human-computer interaction, and online education. For more information, please visit http://qingchen.website/



Xuanwu Yue Xuanwu Yue received the BEng degree in software engineering from the Department of Sofware, Shandong University, China, in 2016. He is currently working toward the PhD degree in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). His research interests include visual analytics of financial data, human-computer interaction and information visualization.

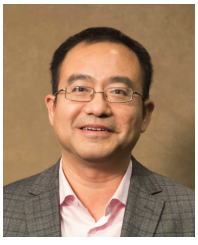
Xavier Plantaz Xavier Plantaz received his Master of Philosophy in Technology Leadership and Entrepreneurship program with a focus on hi-tech innovation and entrepreneurship, from the Hong Kong University of Science and Technology (HKUST) in 2018. He gained his Master of Engineering from Ecole Centrale Paris in 2017. His research interests include information visualization, visual design, and data analytics.



Yuanzhe Chen Yuanzhe Chen received the BE and ME degrees in electrical engineering from Shanghai Jiao Tong University, China, in 2014. He is currently working toward the PhD degree in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. His research interests include information visualization and human-computer interaction.

CongleiShi Conglei Shi is a data visualization engineer in Airbnb. Before that he was Research Staff Member in IBM T.J. Watson Research Center. He received his Ph.D. degree in Computer Science from the Hong Kong University of Science and Technology (HKUST) and his B.Sc. degree in Shanghai Jiao Tong University in major of Computer Science. His main research interests are information visualization, visual analytics , and human computer interaction. For more information, please visit http://www.conglei.org/

Ting-Chuen PONG is a professor in the Department of Computer Science and Engineering and a senior advisor to the Executive Vice-President and Provost (Teaching Innovation and E-learning) at the Hong Kong University of Science and Technology. His research interests include Multimedia Computing, E-learning, Computer Vision and Image Processing, Artificial Intelligence, and Pattern Recognition. He received his PhD in Computer Science from Virginia Polytechnic Institute and State University.



Huamin Qu is a professor in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. His main research interests are in visualization and human-computer interaction, with focuses on urban informatics, social network analysis, e-learning, text visualization, and explainable artificial intelligence. He obtained a BS in Mathematics from Xian Jiaotong University, China, an MS and a PhD in Computer Science from the Stony Brook University.