Visual Analysis of MOOC Forums with iForum

Siwei Fu, Jian Zhao, Weiwei Cui, and Huamin Qu

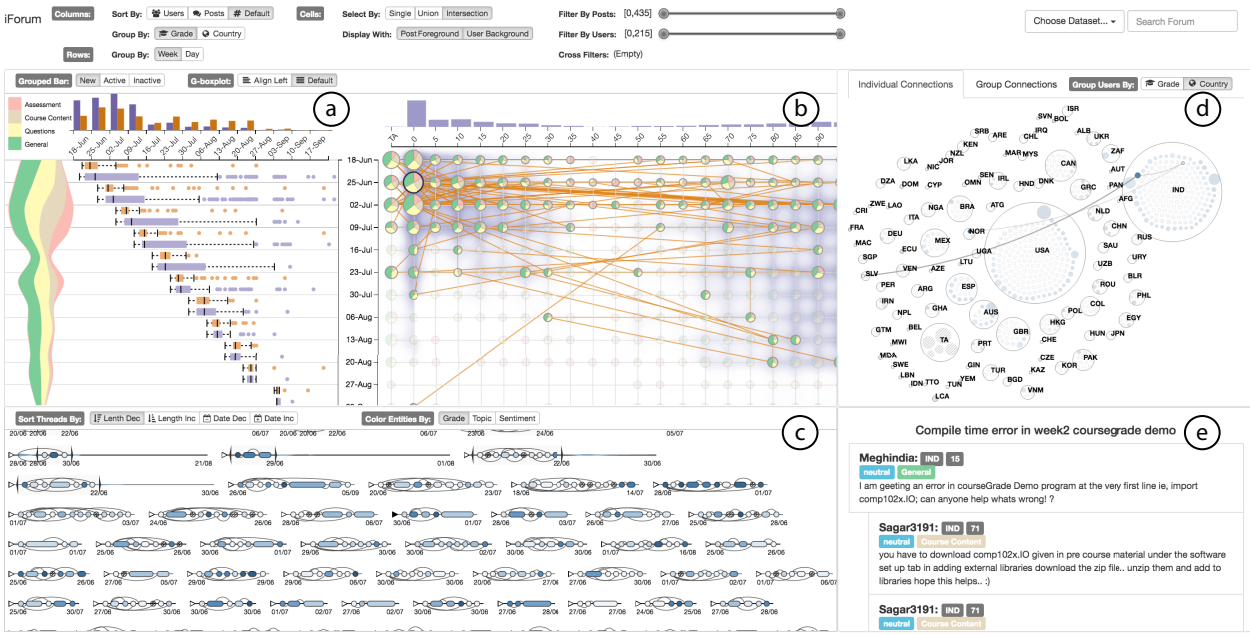

Fig. 1. Using iForum to explore the MOOC forum of a JAVA programming course that has attracted more than ten thousand students during a ten-week course period. (a) The Overview shows the overall changes of posts, threads, and users on the forum. (b) The Matrix View further enables the comparison of dynamic patterns of different user groups along time. After a cell of interest is selected, orange lines are shown on top of the matrix to indicate the threads passing through that cell. (c) Meanwhile, the Thread View presents all selected threads in a compact layout, and (d) the Social Network View reveals the interactions among corresponding users based on their replying relationships. (e) When a specific thread is selected, the Text View displays discussions in traditional indented form.

Abstract—Discussion forums of Massive Open Online Courses (MOOC) provide great opportunities for students to interact with instructional staff as well as other students. Exploration of MOOC forum data can offer valuable insights for these staff to enhance the course and prepare the next release. However, it is challenging due to the large, complicated, and heterogeneous nature of relevant datasets, which contain multiple dynamically interacting objects such as users, posts, and threads, each one including multiple attributes. In this paper, we present a design study for developing an interactive visual analytics system, called iForum, that allows for effectively discovering and understanding temporal patterns in MOOC forums. The design study was conducted with three domain experts in an iterative manner over one year, including a MOOC instructor and two official teaching assistants. iForum offers a set of novel visualization designs for presenting the three interleaving aspects of MOOC forums (i.e., posts, users, and threads) at three different scales. To demonstrate the effectiveness and usefulness of iForum, we describe a case study involving field experts, in which they use iForum to investigate real MOOC forum data for a course on JAVA programming.

Index Terms—Discussion forum, MOOC, temporal visualization, visual analytics

1 Introduction

Due to the potential for dramatic changes in higher education [24], Massive Open Online Courses (MOOC) have moved into a place of prominence in industries, in scholarly publications, and in the mind of the public [39]. Millions of students have registered for one or more MOOCs released by leading universities around the world. The MOOC forum is becoming a central hub where students are able to interact with instructional staff. According to a recent survey of 92 MOOC instructors [30], discussion forums are rated as the most useful resources in understanding class dynamics and preparing their courses for the next iteration.

Instructors of MOOCs, however, face several big challenges in analyzing the course forum. First, the forum data is complicated and heterogeneous: users initiate posts and reply to each other forming threads, which contain temporal, structural, and textual information. Second, the scale of MOOC forums is often large, typically containing thousands of users and hundreds of thousands of posts. Third, the dynamic interactions among multi-attributed MOOC forum users (including students with different grades and from different regions, as well as instructors), which partially reflect the success of a course, are difficult to examine due to the previous two challenges. Current practices of understanding MOOC forums are limited. To be specific, they only rely on reading through individual threaded discussions and

- *S. Fu and H. Qu are with Hong Kong University of Science and Technology. E-mail: {sfuaa, huamin}@ust.hk.*
- *J. Zhao is with Autodesk Research. E-mail: jian.zhao@autodesk.com.*
- *W. Cui is with Microsoft Research. E-mail: Weiwei.Cui@microsoft.com.*

Manuscript received 31 Mar. 2016; accepted 1 Aug. 2016. Date of publication 15 Aug. 2016; date of current version 23 Oct. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2598444

basic statistics analysis provided by the platform, such as the numbers of enrolled students shown in bar charts. These approaches are neither effective due to the massive number of users, threads, and posts, nor practical in solving tasks such as exploring dynamic interaction patterns among forum users.

Visual analytics techniques have been proven effective in exploring forum data in an intuitive and interactive way. For example, representing lengthy threads in hierarchical structures has helped people investigate the replying actions of users [36, 23, 33, 18]. Interactive visualizations have also facilitated the study of user interactions in online communities [17, 28, 25]. Some visualization systems have employed various analytical methods to discover insights into forums, such as automatic topic extraction and sentiment analysis [10, 35, 14]. However, none of the above techniques have sufficiently addressed all the challenges specific to analyzing MOOC forums.

To fill this gap, in this paper, we conduct a design study that involves three instructional staff in MOOCs to develop a visual analytic system, called iForum, allowing for effective discovery and understanding of dynamic patterns in course forums. We elicit domain-specific questions and tasks through multiple interviews, and design the system in a user-centered iterative approach using real datasets.

iForum offers a set of novel interactive visualizations for presenting different aspects of heterogeneous MOOC forum data, including users, posts, and threads. Analysts can explore data from multiple levels of perspectives. At the macroscopic level, the overall temporal dynamics of the entire forum are revealed, such as the topic trends of posts, the volume of the users and threads, and statistical information on lifespans of users and threads. At the mesoscopic level, a matrix-based visualization supports the comparison of different user groups in the forum with the three essential information aspects. Interactive filtering mechanisms are equipped to allow for exploring the massive amount of MOOC forum data available and diving into particular parts of interests. At the microscopic level, the subset of data of interest to analysts is detailed in three views, revealing the user replying structures of the threads, the social interactions of users, and the original text in posts. To address the problem of threads in the MOOC forum being lengthy, we propose a novel visual design that effectively summarizes the structural and temporal information of a thread with minimum screen real estate. In addition, all the visualization views in iForum are interactively coordinated to ease the exploration process.

Our contributions in this paper include: 1) a set of domain-specific goals and design rationales derived through our interviews with instructional staff, 2) a novel visual analytics system, iForum, for the interactive exploration of dynamic patterns in massive and heterogeneous MOOC forum data, 3) a scalable and generalizable visualization of lengthy threaded discussions called Thread River, and 4) an in-depth case study to evaluate the effectiveness and usefulness of iForum on real-world datasets.

2 Related Work

In this section, we review related techniques from three main perspectives: 1) demonstrating conversational threads in forums, 2) extracting high-level information from forums, and 3) understanding social interactions among forum users. In addition, we summarize the main algorithmic approaches for analyzing MOOC forum data.

2.1 Demonstrating Conversational Threads

Due to the complex hierarchical replying structures in lengthy forum threads, researchers have leveraged visualization techniques to help analysts understand phenomena and intrinsic structure of threads. For example, Wattenberg and Millen presented the Conversation Thumbnail, which employs a focus+context visualization technique to exploit text-level metadata and navigate the overview of large-scale online conversations [36]. To better portray the hierarchical structure, Newman proposed TreeTable to obtain thread overviews and mechanisms to assist with the coherent reading of threads [23]. Yee and Hearst employed a similar visualization technique and present content-centered discussion map to depict a specific thread within a larger set of threaded conversations [40]. More recently, the tldr system has been developed to lessen the problems of information overload and help users navigate large-scale discussions efficiently [22].

The above techniques preserve the hierarchical structure of forum threads, but do not keep the temporal information of every entity in the thread, e.g., posts. To depict both the sequential models and tree models of email conversations, Venolia et al. proposed a compact chronological tree table (CCTT) [33]. On the other side, Kerr developed Thread Arcs, combining the chronology of messages with the branching tree structure of a conversational thread in a mixed-model visualization [18]. Compared with CCTT, Thread Arcs is stabler and more space efficient. Although the visual design of threads in iForum is similar to Thread Arcs, our visualization reduces visual clutter by aggregating responses that reply to the same post. Moreover, we provide a focus+context approach for visualizing significantly lengthy threads to further extend the scalability of the design.

2.2 Extracting High-Level Information

Apart from presenting the low-level structures of forum threads, many works have employed text mining techniques to detect interesting patterns from data in a dynamic and scalable way. ForumReader is a tool combining visualization techniques with automatic topic extraction algorithms to help users explore flash forums [10]. ForAVis, which integrates sentiment analysis, provides greater flexibility to the analyst in different search and exploration tasks [35].

Recently, Hoque et al. developed ConVis to support multi-faceted exploration of blog conversations [14], which contains multiple views to provide thread information at different granularities. Bum et al. presented a design study in which they characterized the domain goals of online health communities (OHCs) and derived analytical tasks to achieve these goals [20]. They proposed VisOHC, a visual analytics system that captures hidden dimensions of threads in OHCs.

The design of iForum has been inspired by many of the above visual analytics systems. However, most previous works concentrate on general analytical tasks of forum data, which is not adequate for a deeper understanding of domain-specific problems in MOOC forums. Further, none of the above systems have thoroughly considered the three dynamically interleaving aspects, i.e., users, threads, and posts, together at an entire forum level, which is essential for discovering insights in MOOC forums.

2.3 Understanding Social Interactions of Users

Many systems have been proposed to understand user social interactions in forums. For example, VISM visualizes the sequence of user interactions and subgroup formulation in a radial tree layout [17]. Pascual et al. developed a similar interface to VISM, which has a concentric and nested radial-tree layout [25]. Conversation Map computes a set of social networks in a forum based on post-reply interactions [28]. It employs users' centrality degrees in a social network to differentiate the importance of users in a community and visualizes social and semantic networks with a node-link diagram.

Although the aforementioned approaches can demonstrate social interaction among individuals in forum conversations, they are unable to reveal user connections at a higher level, i.e., connections among user groups. Differently, iForum identifies social dynamics of users at both the individual level and the group level. Moreover, our design is able to depict between- and within-group user interactions with different grouping criteria.

2.4 Analyzing MOOC Forums

Identifying content-related threads among noisy discussions is one of the most popular research topics in MOOC forum analysis. Brinton et al. proposed a unified generative model for discussion threads that allows for effectively choosing thread classifiers and ranking thread relevance [7]. Ramesh et al. characterized forum posts into three categories, course content, meta-data level discussions, and general discussions, and employed seeded topic models to classify posts into each category [26]. Cui et al. extracted linguistic features and built a classification model to identify content-related threads [9]. On the contrary, Rossi investigated several language independent features to

classify threaded discussions, such as structure, popularity, temporal dynamics of threads and diversity of the ids of the users [27].

Some researchers have focused on analyzing and examining user posting behaviors in MOOC forums. For example, the prolific posting behavior of super posters and whether they have positive contributions on the overall forum activities has previously been studied [15, 38]. In contrast, Mustafaraj and Bu explored the behavior of "passive" users who regularly read posts in a forum without posting [21].

Another research direction lies in discovering correlations between learning/teaching activities and posting behavior in MOOC forums. Yang et al. identified factors related to student behavior and social positioning in forums to explore student dropouts in MOOCs [39]. Similarly, Wen et al. observed a correlation between the sentiment ratio of daily forum posts and the number of students who drop out each day [37]. Several other works also focused on locating threads that may require an instructor's attention and students that may need assistance in MOOC forums [4, 8].

While these analytical methods have provided useful patterns in MOOC forum data with multiple perspectives, the lack of visualization components limits the ability to interactively explore data and opportunities to discover richer and deeper insights.

3 Task Characterization

In this section, we describe the method and procedure of extracting user requirements, and present the derived design rationales.

3.1 Working with Domain Experts

In this study, we worked closely with three domain experts. One of them is a MOOC instructor who has designed and released courses about JAVA programming on edX [1] and received over ten thousand students. The other two are the official teaching assistants (TAs) of this course whose responsibilities are to assist with curriculum design, interact with students in the course forum, and collect feedback from students to give to the instructor. We collected the MOOC forum data after the course completed. The dataset contains raw posts created by forum users including students, TAs, and instructors. Each data entity is a post that contains the text content, created date, authorship and to which post it replies. In addition, we gathered and processed profile data of each forum user, such as his/her grade, nationality, etc.

Over the course of this design study, we organized three formal interview sessions with all experts, during which we presented the latest prototypes to them to get feedback. Also, we scheduled frequent informal communications with the TAs. Particularly, in the first interview session, we had in-depth conversations with the experts to identify their current challenges and problems of analyzing MOOC forum data. We also provided some initial sketches of our ideas and a dashboard showing course statistics and our analytics results, in order to gather a list of concrete design requirements. In the second session, experts were presented with a working prototype, an earlier version of iForum, to test whether the major design requirements were met and collect further feedback for refining the prototype. Finally, in the third interview, a full version of iForum with a refined interface and complete set of features was demonstrated. We aimed to derive a concrete use-case scenario of analyzing the forum data of the JAVA programming course with our experts, as well as identify strengths and weaknesses of iForum and potential areas to extend (see Section 6).

3.2 Extracting User Requirements

The high-level goal of the domain experts is to understand the behavioral patterns of their students in the course forum, and to revise the courseware design for the next run. However, the three domain experts have various difficulties in analyzing the dynamic patterns of the forum data. A current MOOC platform, taking edX as an example, provides overall statistics for courses, such as the number of enrolled students, demographic distribution, etc. However, our domain experts propose analyzing the dynamics of the forum rather than static patterns. They argue that the MOOC forum is dynamic in nature because the discussions in the forum vary according to the life cycle of the course. For example, courseware-based events, such as the release of course videos, assignments, and exams, may lead to topic changes in the forum. Though statistical analysis of the MOOC forum is valuable and necessary, it is far from enough.

The instructor was busy and had no time to examine students' discussions piece by piece. He wanted to get a general idea of the course forum and to know how user groups differ in terms of learning behaviors. He mentioned, *"It would be interesting to know the learning patterns of students with high achievement over the course"*. He advocated that different learning patterns between high achieving students and low-achievement students might reveal the reasons for varying levels of performance among students.

The TAs, on the other hand, were dedicated to exploring and analyzing the large number of users and posts in the forum. Traditionally they were only able to use the web-based interface provided by the MOOC platform to perform analytical tasks. They revealed problems related to their ability to identify valuable threads in large amounts of data. As one TA said, *"Our course is quite popular, and there are hundreds of threads in the forum. We cannot read them all"*. The TA further added, *"We can only sort threads by recent activity [the last active date of the thread] or by the most activity [the post volume of the thread], the options [in the MOOC platform] are quite limited"*. Little information of individual threads is provided in the MOOC platform to help our experts locate threads of interest.

Moreover, one TA was keen on the connections among students. He noted, *"It is impossible for us to answer all questions raised by thousands of students. So we have to identify community TAs from the discussions and examine their performances"*. However, a traditional forum system is quite limited in its ability to perform this task. The TA added, *"It [the MOOC platform] does not support user-based interface to help us understand interactions among forum users"*. Besides individual level connections, the TA wanted to know stories like whether high-grade students are more likely to interact with high-grade students.

In summary, the experts demanded a visual analytics system which is specifically designed for MOOC forums to achieve the following high-level goals:

- **G1:** To explore forum data from multiple levels of scales, from the entire forum to an individual thread or user.
- **G2:** To examine the overall changes of the forum from different perspectives, and to compare behavioral patterns among different user groups.
- **G3:** To reveal various characteristics about each thread, and to assist with the exploration of a number of related threads.
- **G4:** To browse the replying actions among forum users, and to inspect the interaction patterns between specific users and between user groups.

3.3 Distilling Design Rationales

Based on the interviews with the experts, we derive the following design rationales to guide our design of the iForum system.

- **R1:** *Reveal overall temporal dynamics of the whole forum.* A MOOC forum usually contains three main interleaving elements, i.e., users, posts, and threads, that dynamically evolve over time. At each time step, some threads are newly created and become actively discussed, while others receive their last posts and will be inactive thereafter. Patterns in the lifespans of threads can reveal the overall performance of the forum from the thread perspective. Similarly, from the user perspective, depicting the active time periods of users is also essential to accurately reflecting forum dynamics. Moreover, the performance of the forum is related to the number of posts created at each time step as well as the changes of their contents such as major topics discussed over time (**G1**, **G2**).
- **R2:** *Compare behavioral patterns of user groups with different characteristics.* Forum users can be grouped by various characteristics, such as grade, nationality, activeness during the course,

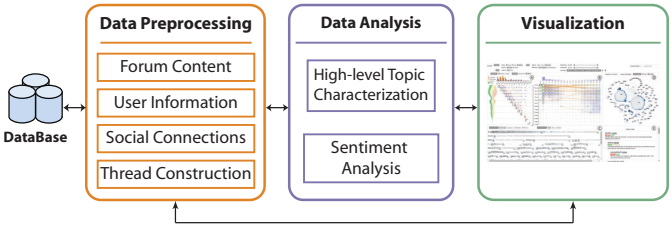

Fig. 2. Overview of the iForum system architecture.

and first/last active date in the forum. Our experts were particularly interested in two attributes, namely grade and nationality, because they may have correlations with user posting behavior in the forum. For example, students with different achievements may vary in posting tendency; and nationality that reveals the culture variation of the students may affect the topic preferences. Examining patterns in different user groups helps experts explore and gain insights into specific sets of users, thus designing more effective courseware that balance different needs (**G2**).

R3: *Discover temporal and structural patterns of individual threads*. A thread initializes from a root post followed by a collection of responses. Each response may receive several further responses. Hence, a thread is defined as a tree of posts that is structured by such responding or replying actions. In the MOOC platform, this tree structure of a thread is preserved with indentations to present the user replying relationships. However, temporal information of each post, which is crucial for browsing the dynamics of a thread, is partially discarded in the indentation presentation. Therefore, demonstrating both the temporal information and structural patterns is essential for experts to understand the evolution of a thread (**G1**, **G3**).

R4: *Identify social dynamics of users reflected by their activities in the forum*. Replying relationships among users represent their social interactions in the forum. Also, at the user group level, social connections between different groups and within the same groups are important for learning about the performance of the forum. Our experts advocated that the discussions in MOOC forums change dynamically with the release of the courseware, so as social interactions among users. Hence, identifying patterns of social dynamics between users or user groups can reflect the wellness of the forum (**G1**, **G4**).

R5: *Display original text*. All experts have suggested that showing the original text of posts in a thread similar to the traditional MOOC platform is essential for them to verify findings. Moreover, to confirm hypotheses or gain deeper understandings of forum discussions, some other information should be added on top of the textual content, such as the nationality and achievement of the posting user.

4 System Overview

The architecture of the iForum system consists of three major components: a data preprocessing module, a data analysis module, and a visualization module (Figure 2).

The data preprocessing module extracts posts from the MOOC forum data, and reconstructs threads from the replying relationships among posts. It further builds the social connections between users based on their reply relationships in threads. Moreover, we distill detailed user information from other data sources. For example, user achievement records (such as grades) are extracted from the courseware progress data, and user geographic information is derived by mapping the IP addresses in the clickstream data to countries. The most time-consuming part is the calculation of user geographic information. To accelerate preprocessing the clickstream data, we employ Map-Reduce technique [11] to process the data in parallel. In practice, for original data of about 20GB, it takes about half an hour preprocessing the data on a workstation with Intel Core i7-4930K 3.40GHz processor, equipped with 32GB of RAM and Debian 8.

In the analysis module, we employ SeededLDA [16] to identify topics of forum posts. Inspired by [31] and our interviews with domain experts, we categorize the posts into four high-level topics by utilizing a lexical seed set, including course logistics, course content, questions, and general discussions (unseeded topic). For example, Table 1 shows the seed words for each topic in a JAVA programming course. More specifically, seed words in "Course logistics" and "Course content" are employed from [26] and course syllabus respectively. Our experts suggest adding a "Question" category to collect structured questions which contain specific keywords, such as "operation system" description and problematic "code", etc. All key words are refined during the interviews with the experts. In addition, we use Natural Language Processing APIs [6] to analyze the sentiment of each post.

Topics	Seed Words
Course logistics:	thank, professor, lectures, assignments, concept, love, thanks, learned, enjoyed, forums, subject, question, hard, time, grading, peer, course, university, classroom, teaching
Course Content:	identifier, variable, expressions, memory, io, constructor, method, bool, array, string, scanner, printwriter, interface, gui, binary, stack, eclipse, java
Questions:	problem, error, bug, report, software, description, operation, system, browser, architecture, code, help, unable, suggestion, trouble, try

Table 1. Seed words for the three topics in a JAVA programming course.

The visualization module allows the end user to explore the dynamics of the MOOC forum with the three interleaving aspects, i.e., posts, users, and threads, at three different scales (Figure 1). At the macroscopic level, the overall temporal dynamics of the whole MOOC forum are shown in an *Overview* (Figure 1(a)); at the mesoscopic level, a *Matrix View* (Figure 1(b)) demonstrates the complex relationships among the three aspects of the forum; and in the microscopic level, the detailed information of each aspect is displayed with a *Thread View* (Figure 1(c)), a *Social Network View* (Figure 1(d)), and a *Text View* (Figure 1(e)) respectively. We develop the visual interface of iForum by following the aforementioned design rationales. All views in iForum are dynamically coordinated via interactive linking, allowing for a seamless exploration of forum data in different perspectives.

5 iForum Design

In this section, we describe the visual design of the iForum interface in detail, which contains five interactively coordinated views to assist the exploration of the MOOC forum data at three different scales.

5.1 Overview: Getting the Gist

The Overview demonstrates the overall temporal dynamics of the MOOC forum at the macroscopic level (**R1**). As discussed earlier, we use the SeededLDA model [16] to categorize MOOC forum posts into four topics. As shown in Figure 1(a), the topics are color-coded and shown in a flow chart along the y-axis (i.e., the time axis), where the post volume of each topic is mapped to the width of the flow.

Apart from post topics, knowing the dynamics of user activeness is another important macroscopic task. For example, our experts want to know the following: for each week, how many new users emerge, how many old users are left, how long are the new users staying in the forum? Moreover, discovering the dynamics of thread lifespans is essential to studying a MOOC forum's wellness. Similar questions can be asked, such as how many threads are created or closed in each week. As shown in Figure 1(a), to address these questions, we leverage a visualization based on the box-and-whisker diagrams [32], named *G-boxplot*, to indicate the statistical information of user activeness and thread lifespans; we further employ the grouped bar charts to summarize the volumes of different types users and threads (i.e., new, active and inactive items) along time. In the G-boxplot visualization, both the x and y-axis represent time, where the x-axis is shared with the grouped bar charts and the y-axis is shared with the topic flow chart. For each time step (one week in this example) along the y-axis, two box-and-whisker diagrams, where orange indicates threads and purple indicates users, depict the temporal distributions of activities of the new users and threads in that time step. Taking the first purple

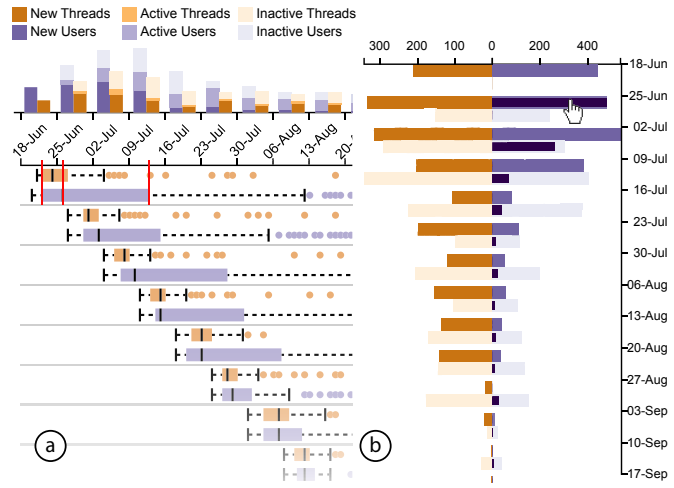

Fig. 3. The comparison between (a) the G-boxplot and (b) the grouped bar charts.

Fig. 4. An example of using crossing filtering in the Matrix View. (a) Users are grouped by grade in the matrix diagram. (b) Analysts create filters by clicking the column labels. (c) Users are grouped by country. (d) Analysts apply the filters by clicking the "✓" on the toolbar. The top five countries with the highest achieving students are revealed: India, the USA, Spain, Great Britain, and Canada.

boxplot in Figure 3(a) as an example, the boxplot shows when 25%, 50%, and 75% of the new users appearing in the first week leave the forum, indicated by the red lines.

To facilitate the exploration, two referencing lines are displayed vertically and horizontally at the mouse cursor position when the analyst moves the mouse in the Overview. In addition, the end users can choose to show either new, active or inactive items in the grouped bar charts. They can align the boxplots to the left to compare them in a relative time manner.

Design Considerations. We develop the Overview through an iterative process by working with our domain experts. As shown in Figure 3(b), in the second interview, we demonstrated a design based on grouped bar charts only to indicate the volumes of different types of users and threads along the vertical time axis, which allowed the experts to explore user activeness and thread lifespans using interactions. For example, when hovering over the bar representing the new users in the second week, these users are highlighted inside the bars with dark purple to indicate the number of users leaving in different weeks. However, our experts found that this interaction is a little cumbersome for getting a big picture of the distribution of activities across time. Therefore, we proposed the above visualization, combining the G-boxplot and the grouped bar charts on top. For example, in Figure 3(a), we can observe that nearly 75% of newly created threads at each time step close within one week in the MOOC forum. In the third interview, our experts appreciated the G-boxplot design, and would like to take actions to improve the course design, as described in Section 6.1.1.

5.2 Matrix View: Investigating User Groups

To unfold the temporal dynamics of the whole forum in different user groups (**R2**), we develop the Matrix View that includes a matrix diagram and a bar chart on top to reveal information at the mesoscopic level (Figure 1(b)).

The matrix diagram is a major analytical component in this view, sharing the same vertical time axis with the Overview. The x-axis of the matrix diagram represents user groups, where the analyst can choose to group users by different meta-data attributes, such as grade and country. Each cell in the matrix diagram indicates the information of one user group (column) at one time step (row) from two aspects: users and posts. First, the number of posts is mapped to the size of a pie chart, where each portion corresponds to a topic using the same color-coding in the flow chart of the Overview. Second, the number of users is encoded with a density background using white-purple gradient, where the darker the color the larger the user volume. A summary of user distribution is further indicated as a bar chart on top of the matrix, where the height of the bar maps to the total user volume in each group. To avoid visual clutter and information overloading in the matrix diagram, the analyst can also choose to display only the pie charts or the density background. The third aspect of the forum data, threads, is revealed on demand as the analyst clicks the pie charts in the matrix cells. The threads passing through *all* the selected cells (i.e., a group of user at a particular time step) are drawn as orange lines across the matrix diagram, indicating how different user groups interact with each other in those threads along time. Meanwhile, all the unrelated cells are fade-out to ease the exploration. Alternatively, the analyst can set the combining method of threads to union, i.e., threads passing *any* of the selected cells would be displayed.

Three different interaction techniques, sorting, aggregation, and filtering, are integrated with the matrix diagram to help the analyst transit from the mesoscopic exploration to the microscopic level. First, the matrix columns can be sorted according to the user volume or the post volume of each user group, which is useful when the analyst would like to investigate the most or least popular user groups. By default, columns are sorted numerically or alphabetically according to user grouping attributes. Second, multi-scale exploration is supported in the matrix diagram by allowing for dynamic aggregation of the time axis, e.g., by day or by week. When the data is available, the analyst can also choose to operate the column of the matrix in a similar multi-scale manner, e.g., grouping users with different grade chunk sizes. Third, iForum integrates two kinds of filtering mechanisms, traditional filtering and cross filtering, to locate points of interests in the forum data. First, traditional filtering can be used to filter matrix cells based on the volumes of users and posts. Second, the analyst are equipped with cross filtering to create a series of dynamic, removable filters along multiple user meta-data attributes by simply clicking the column header (Figure 4).

In addition, interactive linkings are supported to relate the matrix diagram with the bar chart on top. For example, hovering over a cell in the matrix highlights its portion of the whole user volume on the corresponding bar in the bar chart.

Design considerations. We choose the matrix metaphor as the base of the visualization design because it is versatile in showing both discrete and continuous attributes in a multi-scale manner and allowing for a quick overview and comparison of different parts in the massive forum data with a semantic subdivision, i.e., time and user groups. The iForum prototype presented in the second interview was without crossing filtering function. However, our experts sometimes asked questions like: "what is the grade distribution of American students", "what is the geographic distribution of high achieving students", and "what is the activities of American high achieving students in the forum"? Traditional filtering with a multi-attributed column in the matrix diagram does not provide sufficient answers. We thus implement cross filtering to empower the matrix diagram, which is even more powerful when more user attributes are included. For example, in Figure 4(d), after applying filters, we are able to see the top five countries with the highest achieving students. They are India, the USA, Spain, Great Britain and Canada. In the third interview, our experts mentioned that the interaction was more intuitive than tables with lists of numbers for obtaining similar patterns (Section 6.1.2).

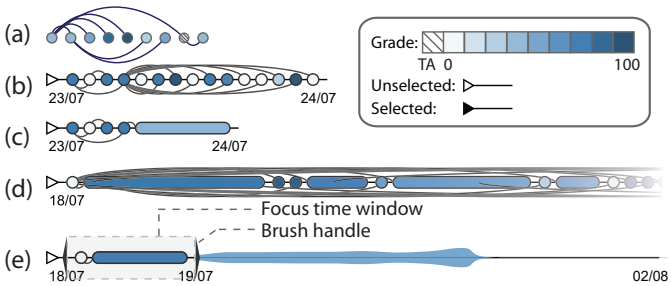

Fig. 5. The design process of Thread River. (a) The basic Thread Arcs. (b) Thread Arcs not scaling well for a number of posts. (c) The revised Thread Arcs with aggregation in subsequent posts replying to the same post, showing the same thread as that in (b). (d) The design of revised Thread Arcs showing lengthy thread with complex reply relationships, which is too long to display. (e) The Thread River design presenting the same thread as that in (d).

5.3 Thread View: Diving into Individual Threads

After the analyst identifies points of interests by interacting with the Overview and the Matrix View, she can shift the exploration focus to the microscopic level with the Thread View (Figure 1(c)). This view displays detailed information of all selected threads, initiated by the focusing cells in the Matrix View.

To unfold the temporal and structural characteristics of individual threads (**R3**), we design a novel visualization, named *Thread River*, to present lengthy threads with complex replying relationships. We are first inspired by Thread Arcs [18] which depicts both the hierarchical structure of a conversational thread and its chronological sequence of messages in a compact layout. As Figure 5(a) shows, each post in the thread is represented as a node located on the time axis, and a replying relationship is indicated as a curve connecting the two nodes. Although intuitive, Thread Arcs is not scalable for threads with large numbers of posts due to the visual clutter of curves (Figure 5(b)). We then simplify the hierarchical structure by grouping all subsequent posts of a post if they all reply to that post and are adjacent in time. As shown in Figure 5(c), the thread is the same as that depicted in (b). The first rectangle on the right represents all posts replying to the forth post. Such design eliminates many nodes and curves in the original Thread Arcs while preserving the tree structures. However, it still has limitations in visualizing significant long threads with very complex replying relationships (Figure 5(d)). To further extend scalability, the final Thread River design employs the focus+context technique to show lengthy threads with a focus window of the detailed thread structures and a background context of the post volumes across time (Figure 5(e)). The analyst can freely adjust the focus window by dragging the black handles and a staged animation is played after to ease the transition between different visual states of the thread.

Rich interactions are also supported in the Thread View. First, the analyst can choose to map different attributes, such as the grade of the post owner, the topic, or the sentiment of the post, to the color of the nodes. For aggregated nodes, i.e., the rectangles in Thread River, our current implementation shows the average value for continuous attributes (e.g., grades) or the majority value for categorical attributes (e.g., topics). Second, various sorting strategies are embedded to enable the analyst to organize threads with specific criteria, including thread length, starting date, closing date, and so forth. Interactive linkings between this view to other views in iForum are also embedded. For example, hovering over a thread highlights the corresponding cells in the Matrix View and the visual elements in the Social Network View that will be introduced later.

Design Considerations. With the help of node aggregation and the focus+context technique, Thread River is able to present threads containing hundreds of posts while keeping the temporal and hierarchical structure of a thread. During the design process, our experts can easily explore the longest thread (628 posts) in their MOOC forum data in a fluid manner, which might be impossible with the traditional Thread Arcs visualization. However, certain aspects of the Thread River can still be enhanced, such as using color gradients to encode post attributes in the aggregated nodes in order to reveal more detail. All experts liked the design of Thread River in the second interview, when it was first presented. To accelerate the exploration of threads of interest, they further suggested sorting all threads according to different criteria, such as thread length, starting date, etc.

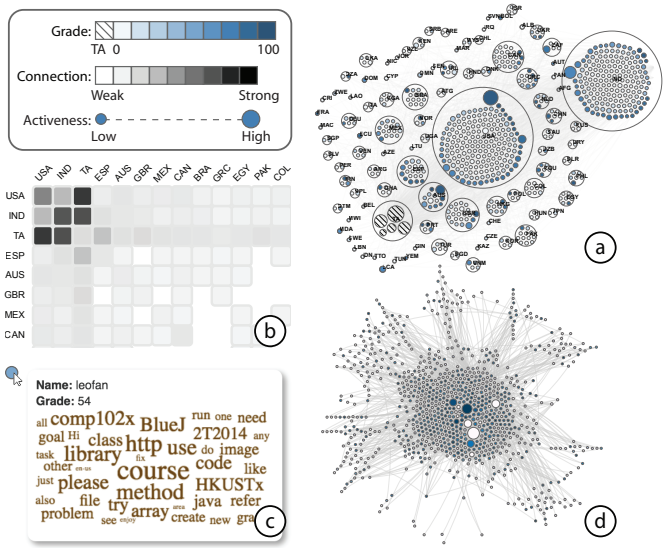

Fig. 6. (a) The grouped node-link diagram. (b) The heatmap matrix showing group level connections. (c) Informative tooltip pops out when a user hovers over a node. (d) The basic node-link diagram based on MDS and force directed layout.

5.4 Social Network View: Examining User Connections

The Social Network View serves the microscopic exploration of data from the user perspective (Figure 1(d)), revealing the social dynamics of users in MOOC forums (**R4**). This view shows all users appearing in the Thread View, in correspondence to the focusing cells in the Matrix View, or the entire social connection in forum data.

Two types of visualizations are designed in the Social Network View, a grouped node-link diagram and a heatmap matrix, to represent user social connections in different forms. In the grouped node-link diagram (Figure 6(a)), users are represented as circles where the size indicates user activeness level and the color indicates the user attribute value (e.g., grade). Users within the same attribute group are drawn together using the circle packing algorithm [34]. Social connections are shown as gray links between users. In the heatmap matrix (Figure 6(b)), both rows and columns depict user groups, and each cell is color-coded with the connection strength (the darker the stronger) between the corresponding user groups. Thus, diagonal cells present within-group connections and other cells show between-group connections. Synchronized with the Matrix View, users can be grouped by grade or by country in the Social Network View. Further, when the analyst hovers over a node in the grouped node-link diagram, a tooltip window pops up with some basic information of that user (e.g., grade) and a word cloud capturing the most frequent words in the user's posts (Figure 6(c)).

Design Considerations. We iteratively refine our design of the Social Network View with the help of our experts. In the second interview, we illustrated the most straightforward node-link diagram to show the social network of users (Figure 6(d)), and we experimented on the graph layout with the multidimensional scaling (MDS) [19] and the force directed method [5]. However, it failed to show social patterns at a higher level, such as connections among user groups, in an intuitive way, which was more important to our experts. Further, our experts found it hard to locate students of interest. We thus develop the grouped node-link diagram; and to overcome the visual clutter of links, we further augment it with a heatmap matrix after another discussion with the TAs. Each of the two visualizations have their own benefits, where the analyst can leverage both to explore data more effectively.

More specifically, the grouped node-link diagram can reveal information at the individual level or user group level, and the heatmap matrix is connection oriented and presents clearer relationships. In the third interview, our experts were excited about the heatmap matrix because they were able to discover within- and between- group interactions with little effort. Moreover, the grouped node-link diagram enabled the experts to conveniently explore a lot of students/TAs of interests and examine their forum activities (Section 6.1.4).

5.5 Text View: Revealing Original Posts

To enable the analyst to examine the raw data at the lowest level (**R5**), we design the Text View that provides a conventional presentation of forum threads (Figure 1(e)). Posts of the thread are arranged vertically with various levels of indentations to reveal the hierarchical structure. Beside the textual content of posts, we selectively show some important attributes with minimum visualizations, including the name, grade, and country of the user as well as the topic category and the sentiment of the post.

The Text View are linked with the Thread View and the Social Network View through interactions. On the one hand, after selecting a thread in the Thread View, or clicking a user in the Social Network View, the Text View updates with the content of the thread, or all posts created by the user. On the other hand, when clicking an entry in the Text View, the relevant post and user are highlighted in the Thread View and the Social Network View respectively.

6 Case Study

We assess the effectiveness and usefulness of iForum with an in-depth case study. We conducted a semi-structured interview with the same three domain experts, one course instructor and two official TAs, whom we worked with to iteratively develop iForum. The interview took about 60 minutes. We demonstrated the system in the first ten minutes. The following 40 minutes were used for free exploration. We instructed them to do the exploration in a think-aloud protocol, encouraging experts to speak out whatever they were looking at, thinking, doing and feeling in the exploration. We took the notes about their feedback. Finally a post-interview discussion was conducted to further discuss the strengths and weaknesses of iForum. We recorded the whole session for later analyses.

6.1 Use-Case of iForum

During the interview, we asked the experts to use the system to examine the MOOC forum data generated by the course they had taught. This JAVA programming course was released on edX during Jun.-Aug. 2014. The forum data contains a total of 1,976 threads, involving 2,221 users and 11,915 posts. In addition to the high-level goals outlined in Section 3, in this case study, our experts were particularly interested in identifying the difference of behaviors between high- and low-achievement students and examining social interactions among different user groups.

6.1.1 Digesting the Forum

First of all, the course instructor wanted to examine the overall patterns during the course period (**R1**). After the data were loaded into iForum, the Overview provided a big picture of the entire forum (Figure 1(a)). From the flow chart showing the overall volume of posts, he noticed that the activities on the forum kept increasing during the first three weeks of the course and then decreased thereafter: *"students might be just signing up and deciding to stay or leave"*. The instructor also observed a big gap at the 5th week. One TA explained: *"We released the course project at the 6th week, and the project attracted much attention from our students."* The instructor knew from past experience that forum discussions changed periodically due to the release of new course videos every week. So he shifted his eyes to the G-boxplots on the right of the flow chart summarizing the patterns of users and threads weekly (Figure 3(a)). He observed that for each week, about 50% of newly active users became inactive on the forum after the week, showing a turnover rate in discussion participation of approximately 50%. In addition, over 50% of newly created threads

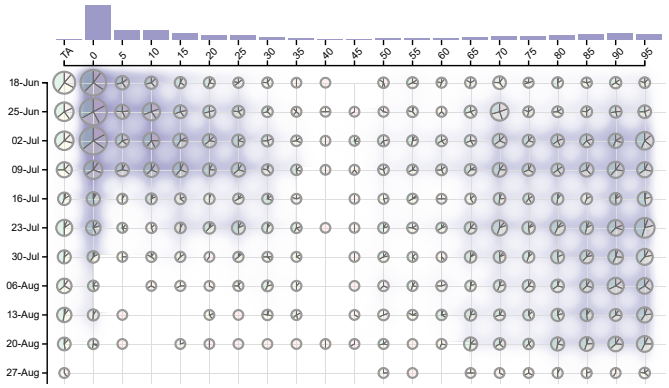

Fig. 7. The temporal dynamics of the whole forum in different user groups. Users are grouped by grade.

closed within one week. The instructor said to one TA, *"maybe we can summarize these threads weekly, and provide weekly FAQs in the next run [next release of the course]"*. Moreover, he mentioned that the G-boxplots were effective because it summarized the important statistics and dynamics of the whole forum in one view and allowed for comparisons across the course period.

6.1.2 Examining User Groups

Next, the instructor wondered what the behaviors of students with different grades might differ (**R2**). In the Matrix View, he chose to show the density background in the matrix diagram only, since he wanted to get an overall idea of user distribution by time and grades (Figure 7(a)). The instructor found that it was darker at the upper left corner, indicating that most low-achievement students were active in the first few weeks only. Moreover, the right part of the matrix was also in dark purple, showing that most high achieving students were continuously active across the course period. *"It is reasonable, but we haven't expected to examine this pattern before"*, the instructor added, *"Maybe the ten-week course lasted too long for many students studying online. We can adjust our course syllabus for the next run"*. Further, the instructor observed that in the first column of the matrix diagram, where laid the TA user group, circles were large with a nearly white background, indicating that few TAs contributed a large number of posts. *"Our TAs worked hard and community TAs had high motivation during the course period"*, the instructor mentioned.

To further investigate the topics of the discussions, the instructor chose to show the foreground pie charts of the matrix diagram. He found that pie charts in the columns of grades 40 and 60 had higher proportions of red, indicating students with grades between 40 and 50 tended to discuss more about course logistics. *"Maybe people on the edge of passing the course care more about course syllabus and how they are evaluated, or they are likely to judge the course"*, the instructor murmured. To examine what they were talking about, he then double-clicked one pie chart full of red proportion in the matrix diagram, which was located in the column "55" and row "27-Aug". A long orange line passing through multiple pie charts in various columns (as illustrated in Figure 8) attracted his attention. In particular, some posts were created after the course ends, indicated by last three pie charts in the matrix diagram. In the Thread View, only one thread appeared with starting date 27th June. and ending date 20th September. *"This thread was active across the whole course period"*, the instructor exclaimed. He selected this thread to see content details. The Social Network View was updated to show all social connections among individual users. As shown in Figure 9, users from different countries were involved in this thread. In addition, a thread titled "VOTE for the professor and crew to next do the Android App course! (click =>)" was displayed in the Text View. The instructor laughed, *"Aha! We plan to release the JAVA programming course for the second run. Perhaps Android application development is a good topic"*.

Then, the instructor adjusted the columns of the matrix to represent students' nationalities (Figure 4(c)), and found that most of the students were from India and America. He further wondered if there

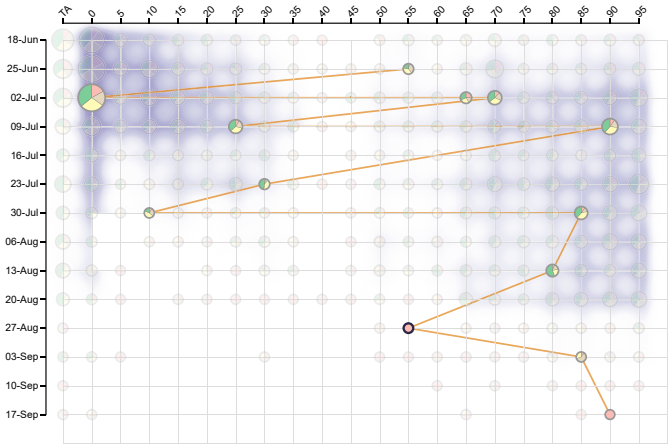

Fig. 8. An orange line passes through multiple pie charts starting from the second row till the last row in the matrix diagram, showing that the thread was active across a long time period. In addition, these pie charts locate in different columns, indicating users with different performance were involved into the thread.

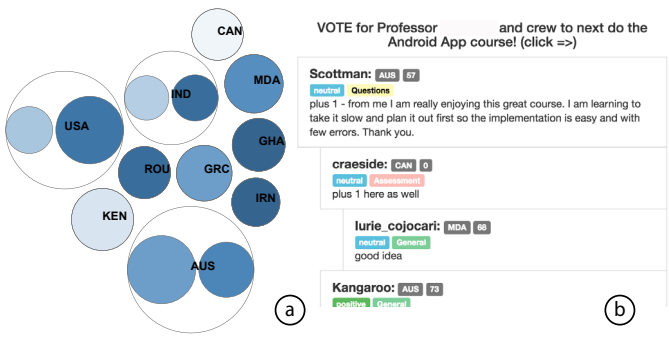

Fig. 9. (a) In the Social Network View, users are grouped by nationality. Users from various countries participate into the thread. (b) The thread content is shown in the Text View.

were any differences in the achievements of these two big groups of students. Thus, the instructor performed a series of cross filtering operations as shown in Figure 4, which resulted in a matrix diagram of the distribution of high grade students in different countries (Figure 4(d)). Through the percentage information on the bar chart above the matrix, the instructor found that 21% of Indian students got grades higher than 80 while high achieving students accounted for 17% of American students. *"Hmm, it seems Indian students performed better [compared with American students]"*, the instructor added, *"Our TAs have obtained similar results using R, but the results are shown in tables. I like this visual way because it is handy, intuitive, and interactive. We don't need to write any program."*

6.1.3 Diving into Threads

To look into the discussion threads in the forum, the instructor selected the three biggest cells from the Matrix View, since he wanted to examine which threads were related to these cells (**R3**). This resulted in seven threads in the Thread View, and then the instructor examined them one by one together with the Text View. He found that five of these threads were social-based, such as "Hi friends, Anyone from India - Tamilnadu!!!" and "Introduce Yourself!". Moreover, he noticed that these threads were composed with nodes in white or light blue, indicating that they were actively supported by a number of low achieving students. *"It is not surprising"*, the instructor added, *"But I never thought this pattern [social-based threads attracting low-grade students] is obvious like that"*. Then, the instructor decided to identify which threads attracted high achieving students. He cleared the previous selections and clicked the biggest cell in the last column (grade 95-100). As shown in Figure 10(a), a thread depicted with the Thread River caught his eye. From the time line in the diagram, the instructor found that this thread was active for one month, which

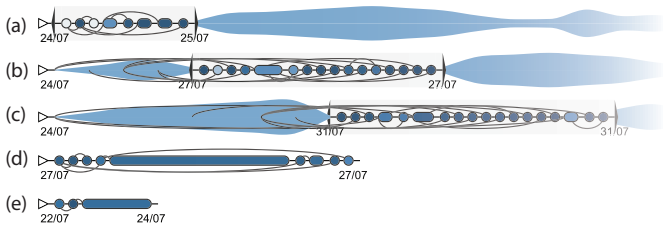

Fig. 10. (a), (b) and (c) demonstrate the results of dragging the time window on the Thread River. (d) and (e) display two threads that attract high achieving students.

was quite unusual. He dragged the focus time window of the thread and saw that most nodes were in dark blue (Figure 10(b) and (c)), indicating that high achieving students actively participated in this thread. After clicking it, from the Text View the instructor found it was "[OFFICIAL] Q&A - Project", which was created by one TA to allow students to post questions related to the course project. He wondered, *"Students caring about the project and technical details are likely to get higher grades?"* To confirm this hypothesis, the instructor further selected some other threads that were full of dark blue nodes, such as in Figure 10(d) and (e). In the Text View, he observed that these threads were technically oriented and targeted on specific problems, such as "LAB 4 tast 4 missing return statement?" and "Lab 04 task 1 Incompatible types error" (**R5**). One TA pointed to the screen and said that he like the flexible color encoding because it allowed him to inspect threads from different perspectives. He suggested adding more user attributes to the color mapping for further insights.

6.1.4 Exploring Social Connection

After the above exploration, the instructor shifted his focus to the social dynamics between different groups of students (**R4**). From the heatmap matrix showing the group-level connections of all the students by their nationalities (Figure 11(a)), he observed that most interactions happened in the groups "USA", "IND"and "TA", indicated by a few dark gray cells. Particularly, darker cells were along the diagonal of the matrix, indicating students tended to discuss with their peers in the same country. The instructor said that this pattern made sense but was not expected, and he suggested that MOOC forums should be designed for breaking these boundaries to encourage more cross-group discussions. However, the TAs were the opposite, showing significantly more connections with other groups of users than their within-group interactions, *"reflecting TAs' special roles in the forum."* Next, the instructor chose to group users by grade in the Social Network View to identify social connections between students with different achievements (Figure 11(b)). He noticed that the diagonal corners of the matrix were much darker than other parts, indicating that besides interacting with TAs, high achieving students were more likely to discuss with each other and so do low achieving students. Moreover, from the dark cells at the top left corner of the matrix, the instructor found that TAs interacted with low-grade students most often. *"They [students with grades between zero and five] gave up themselves although we tried our best to help them"*, added one TA. Also, the instructor appreciated the Social Network View, *"Social connections among TAs and students are important for us to evaluate the performance of community TAs during the course period. However, edX doesn't provide this insight. iForum does a good job by allowing us to analyze such information."*

6.2 General Feedback

The instructor and TAs showed great interest in using iForum, and they appreciated the insights found in the current prototype. The instructor mentioned: *"Your system can help us generate hypotheses. Some results are useful and inspiring and we have not thought of computing them before. For example, [the matrix diagram indicates that] the ten-week course may be too long for many students, and we should refine the course design for the next run."* Therefore, the instructor decided to try the "self-paced" mode of presenting the course materials, thus providing more flexibility to students, in the

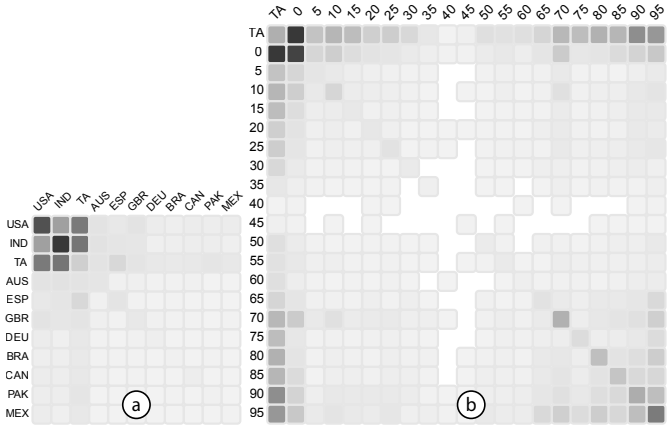

Fig. 11. The heatmap matrices with users grouped by country (a) and by grade (b).

the second release later in the same year. In addition, the instructor proposed two other directions of refining their course design. First, the TAs should compile a weekly FAQ during the course period. Second, the instructor considered to include Android application development in the next run.

From the post-interview discussion, the experts also suggested two general directions for a deeper understanding of students' learning behavior in MOOCs. First, passive users, who surf forums without posting, account for the majority of forum users [21]. For these users who do not necessarily create posts, they may search and acquire enough information in the forum. *"Users creating posts in the forum are limited"*, therefore the instructor recommended integrating the analysis of forum browsing and posting to obtain a comprehensive picture of user activities in MOOC forums. Second, various learning activities are recorded in heterogeneous MOOC data sources. For example, discussions among students are presented in the forum data and students' video watching behaviors are stored in the clickstream data. *"I believe you can get deeper behavioral patterns by combining the clickstream data and the forum data"*, the expert suggested to conduct a joint analysis among different data sources to gain a better understanding of students' learning behavior in MOOCs.

More encouragingly, the expert would like to use iForum for presenting insights when delivering talks outside the campus. To fit to screens with different resolutions, he suggested showing detailed views, i.e., the Thread View, the Social Network View, and the Text View, in a tabbed panel.

7 Discussions

Although our iterative user-centered design and the resulting case study reveal the effectiveness of iForum in exploring the temporal dynamics of MOOC forum data from multiple perspectives and multiple levels, some limitations still exist in the current prototype.

First, some parts in the analytical component of iForum need further improvement. For example, in the sentiment analysis, one comment is marked as negative with the content "*I have sent the lab 01 and there were no problems at all*". It is challenging due to the complexity and variation of human languages. For instance, the state-of-the-art work in predicting fine-grained sentiment labels for all phrases has only achieved a precision of 80.7% [29]. Moreover, to apply seededLDA to other courses, we need to manually generate a set of seed words for each topic, which lacks the flexibility of vast deployment of the system. However, when more advanced analytical techniques are developed in the future, such limitations may be resolved.

Second, the G-boxplot is designed based on the box-and-whisker diagram in statistics, which is not a commonly-seen visualization. This design requires bit learning for users who are not familiar with the box-and-whisker diagram. Although we do not encounter any difficulty of understanding the G-boxplot during our design study with the experts, designing a more intuitive visualization is necessary to widen the audience for using iForum in the future.

Third, as shown in Figure 1(b), we use orange lines to depict threads passing through all the selected cells in the matrix diagram. However, lines connecting cells in the same row are usually overlaid together. In addition, in the grouped node-link diagram, visual clutter exists due to a large number of lines linking individual users. To resolve these problems, we show entities with different opacities in the current prototype. However, detailed connections are hard to perceive by adjusting opacity only. We plan to employ the edge bundling technique [13] or other methods summarized in [12], such as sampling, to illustrate messy lines in a clean and informative way.

There also exists limitations of our design study. For example, we only involve experts of one MOOC course in this study, which might be insufficient to generalize our results to other domains. Further, since the experts participate the iterative design process from the beginning, and are familiar with the system during the study, potential problems of iForum may not be fully revealed. Therefore we plan to evaluate iForum with more MOOC courses and carry out interviews with instructors from different domains.

Several interesting directions are promising to generalize the current iForum design. First, the proposed design of visualizing lengthy threads, Thread River, is flexible enough to be applied in exploring and analyzing other asynchronous online discussions, such as blog discussions, Email conversations and Twitter comments. Moreover, although we conduct our design study in the context of analyzing MOOC forum data, we believe most of the visual design in iForum can also be applied to the exploration of other kinds of forums, such as Stackoverflow [3], Reddit [2], etc. Because MOOC forum data is similar to other forum data in structure, except that some education-specific attributes (e.g., grade) are embeded in user profiles. Thirdly, in this design study, we only collect user information from two aspects: grade and country. It would be straightforward to extend iForum to analyze richer user attributes, such as gender, age, and education level, to gain a deeper understanding of their behaviors in MOOC forum discussions. However, such multi-attributed user information may require more advanced visualization techniques to support effective faceted browsing in addition to the cross filtering mechanisms in the Matrix View.

8 Conclusion and Future Work

In this paper, we have presented a design study for developing a visual analytics system, named iForum, which allows analysts to effectively discover and understand dynamic patterns of MOOC forums. iForum enables the interactive exploration of the complex and heterogeneous MOOC forum data that includes three interleaving aspects, i.e., posts, users, and threads, at three different scales. Moreover, through this iterative user-centered design process, we have outlined a set of domain-specific goals and design rationales that could further inform the future design of similar visualization systems. We have also proposed Thread River that can illustrate temporal and structural information of lengthy threaded discussions. The results of one case study have demonstrated the effectiveness and usefulness of iForum in exploring real-world MOOC forum data based on working with our domain experts.

In the future, we plan to enhance the reliability of the analytical component of iForum by applying more advanced natural language processing techniques. Next, since FAQ is important in MOOC courses, we would like to enhance iForum by extracting FAQs automatically to accelerate the analysis of forum data. Moreover, our experts rely on their previous knowledge to interpret the visualization. We plan to integrate course events into iForum to suggest possible reasons of noticeable patterns detected by the system. Further, we plan to support real-time browsing of MOOC forums so that instructors and TAs can dynamically adjust their strategies for delivering class lectures or react to certain behaviors from students in the forums.

Acknowledgments

We thank all the domain experts, i.e., Ting-Chuen Pong, Tony W K Fung and Amy Quon, involved in the studies. We also thank Hao Dong for the constructive feedback on the early prototype. The project is partially supported by RGC 618313 and ITF ITS/306/15FP.

REFERENCES

[1] edx, free online courses from the world's best universities. https://www.edx.org/. Accessed: 2016-1-27.

[2] Reddit: the front page of the internet. https://www.reddit.com/. Accessed: 2016-1-23.

[3] Stack overflow. http://stackoverflow.com/. Accessed: 2015-10-11.

[4] J. Arguello and K. Shaffer. Predicting speech acts in mooc forum posts. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[5] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall PTR, 1998.

[6] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[7] C. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. Wong. Learning about social learning in moocs: From statistical analysis to generative model. *Learning Technologies, IEEE Transactions on*, 7(4):346–359, 2014.

[8] S. Chaturvedi, D. Goldwasser, and H. Daumé III. Predicting instructor's intervention in mooc forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1501–1511, 2014.

[9] Y. Cui and A. F. Wise. Identifying content-related threads in mooc discussion forums. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15, pages 299–303, 2015.

[10] K. Dave, M. Wattenberg, and M. Muller. Flash forums and forumreader: Navigating a new kind of large-scale online discussion. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, pages 232–241, 2004.

[11] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.

[12] G. Ellis and A. Dix. A taxonomy of clutter reduction for information visualisation. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1216–1223, 2007.

[13] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):741–748, 2006.

[14] E. Hoque and G. Carenini. Convis: A visual text analytic system for exploring blog conversations. *Computer Graphics Forum*, 33(3):221–230, 2014.

[15] J. Huang, A. Dasgupta, A. Ghosh, J. Manning, and M. Sanders. Superposter behavior in mooc forums. In *Proceedings of the First (2014) ACM Conference on Learning @ Scale Conference*, L@S '14, pages 117–126, 2014.

[16] J. Jagarlamudi, H. Daumé, III, and R. Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 204–213, 2012.

[17] S. Jyothi, C. McAvinia, and J. Keating. A visualisation tool to aid exploration of students' interactions in asynchronous online communication. *Comput. Educ.*, 58(1):30–42, 2012.

[18] B. Kerr. Thread arcs: an email thread visualization. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 211–218, 2003.

[19] J. B. Kruskal and M. Wish. Multidimensional scaling. *Quantitative Applications in the social Sciences Series*, 1978.

[20] B. Kwon, S. Kim, S. Lee, J. Choo, J. Huh, and J. Yi. Visohc: Designing visual analytics for online health communities. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):71–80, 2016.

[21] E. Mustafaraj and J. Bu. The visible and invisible in a mooc discussion forum. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15, pages 351–354, 2015.

[22] S. Narayan and C. Cheshire. Not too long to read: The tldr interface for exploring and navigating large-scale discussion spaces. In *System Sciences, 2010. Proceedings of the 43rd Annual Hawaii International Conference on*, pages 1–10, 2010.

[23] P. S. Newman. Exploring discussion lists: Steps and directions. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '02, pages 126–134, 2002.

[24] L. Pappano. The year of the mooc. *The New York Times*, 2(12):2012, 2012.

[25] V. Pascual-Cid and A. Kaltenbrunner. Exploring asynchronous online discussions through hierarchical visualisation. In *Information Visualization, 2009 13th International Conference*, pages 191–196, 2009.

[26] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. Understanding mooc discussion forums using seeded lda. In *9th ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[27] L. Rossi and O. Gnawali. Language independent analysis and classification of discussion threads in coursera mooc forums. In *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on*, pages 654–661, 2014.

[28] W. Sack. Discourse diagrams: Interface design for very large-scale conversations. In *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*, page 10. IEEE, 2000.

[29] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642, 2013.

[30] K. Stephens-Martinez, M. A. Hearst, and A. Fox. Monitoring moocs: Which information sources do instructors value? In *Proceedings of the First (2014) ACM Conference on Learning @ Scale Conference*, L@S '14, pages 79–88, 2014.

[31] G. S. Stump, J. DeBoer, J. Whittinghill, and L. Breslow. Development of a framework to classify mooc discussion forum posts: Methodology and challenges. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, 2013.

[32] J. W. Tukey. Exploratory data analysis. *Addison–Wesley series in behavioral science: quantitative methods*, page 688, 1977.

[33] G. D. Venolia and C. Neustaedter. Understanding sequence and reply relationships within email conversations: A mixed-model visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 361–368, 2003.

[34] W. Wang, H. Wang, G. Dai, and H. Wang. Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 517–520, 2006.

[35] F. Wanner, T. Ramm, and D. A. Keim. Foravis: Explorative user forum analysis. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 1–10, 2011.

[36] M. Wattenberg and D. Millen. Conversation thumbnails for large-scale discussions. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pages 742–743, 2003.

[37] M. Wen, D. Yang, and C. Rose. Sentiment analysis in mooc discussion forums: What does it tell us? In *Educational Data Mining 2014*, 2014.

[38] J.-S. Wong, B. Pursel, A. Divinsky, and B. Jansen. An analysis of mooc discussion forum interactions from the most active users. In N. Agarwal, K. Xu, and N. Osgood, editors, *Social Computing, Behavioral-Cultural Modeling, and Prediction*, volume 9021 of *Lecture Notes in Computer Science*, pages 452–457. Springer International Publishing, 2015.

[39] D. Yang, T. Sinha, D. Adamson, and C. P. Rose. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, 2013.

[40] K.-P. Yee and M. Hearst. Content-centered discussion mapping. *Online Deliberation*, 2005.