Education Tech Research Dev (2021) 69:2813–2843
https://doi.org/10.1007/s11423-021-10040-5

DEVELOPMENT ARTICLE

Visualizing the learning patterns of topic-based social interaction in online discussion forums: an exploratory study

Gary K. W. Wong[1] · **Yiu Keung Li**[1] · **Xiaoyan Lai**[1]

Accepted: 7 August 2021 / Published online: 23 August 2021

Abstract

Online discussion forums are common features of learning management systems; they allow teachers to engage students in topical discussions in environments beyond physical spaces. This study presents a novel approach to operationalizing the connections between social interaction and contextual topics by visualizing posts in an online discussion forum. Using the weak ties theory, we developed a prototype of a tool that helps visualize the text-based content in online discussion forums, specifically in terms of topic relationships and student interactions. This research unveils a nuanced picture of social and topic connectivity, the nature of social interactions, and the changes in the topics being discussed when serendipity occurs. Our implementation of the tool and the results from testing show that the visualization method was able to determine that the strongly connected major topics in the discussion were related to the intended course learning outcomes, whereas the weakly connected topics could yield insights into students' unexpected learning. The proposed method of visualization may benefit both teachers and students by helping them to efficiently the learning and teaching process and thus may contribute to formative assessment design, a collaborative learning process, and unexpected learning.

Keywords Social network analysis · Topic modeling · Visualization · Weak ties · Text mining

Abbreviations

IAM	Interaction analysis model
ILO	Intended learning outcome
LDA	Latent Dirichlet allocation
LDAvis	LDA visualization
LMS	Learning management system

✉ Gary K. W. Wong
wongkwg@hku.hk

Yiu Keung Li
simonykli2004@gmail.com

Xiaoyan Lai
laixy@connect.hku.hk

[1] Faculty of Education, The University of Hong Kong, Pokfulam Road, Pok Fu Lam, Hong Kong

Introduction

The learning management system (LMS) has become a commonly used platform in higher education. From any location and at any time, teachers can use the online tools on LMSs to distribute learning resources, whereas students can use them to access those resources and interact with and learn from their peers. A common feature of the LMS is the asynchronous online discussion forum, which catalyzes learners' active learning and higher-order thinking in a text-based communication environment (Jonassen et al., 1995; McLoughlin & Mynard, 2009). The benefits of using the forum for teaching and learning purposes include overcoming spatial and temporal communications barriers with non-stop operations and providing a relatively relaxed environment where students can express their ideas freely (Chen et al., 2018). Through the forums, students are able to contribute a massive amount of text-based information and interact socially; as a result, data from their posts can be used to understand their learning patterns and topic discussions within or beyond the given topics through learning analytics.

Currently, learning analytics through data mining techniques (e.g., Clustering, association, and classification) provides a means to obtain useful information about students' learning, social patterns, and topic discussions that is not retrievable without these techniques (Han et al., 2011; Slade & Galpin, 2012). Data mining techniques are commonly designed for collecting large-scale data, extracting actionable patterns, and obtaining insightful knowledge (Gundecha & Liu, 2012; Manning et al., 2008). By applying these data mining techniques to effectively analyze the interaction data, personal data, systems information, and academic information collected from LMSs, educators can better understand the thinking patterns of students during the learning process, even running data mining techniques with small-scale data on a desktop computer (Hand et al., 2001; Ray & Saeed, 2018; Wu et al., 2013). The growth of learning analytics provides an opportunity to discover new visualization tools. Students' behavioral intentions or motivations for learning (Mazza & Milani, 2004; Romero et al., 2008) can also be captured and visualized. The data mining techniques can assist in student assessment by using a systematic real-time approach to identify pedagogic changes that may be effective for particular students and to guide students through the learning process with the ultimate goal of optimizing their learning outcomes (Foster & Ford, 2003). Under this unique learning scenario, both teachers and students may benefit from learning analytics, which allows them to engage with each other in the learning and teaching process.

Despite the consensus reached on the benefits of social interactions in learning, little attention has been paid to understanding the content of these interactions, partly because it is time-consuming to deal with the enormous amount of text data (Slade & Galpin, 2012) and partly because of the lack of an analytical framework (Tawfik et al., 2017), even when data mining provides technological support for analysis and visualization. This study combines social network analysis and the examination of text-based learning content, which have primarily been studied separately, and develops a visualization method to capture the contextualized social interactions to understand students' topic discussions and unexpected learning (Clouder & Deepwell, 2004; Havnes & Prøitz, 2016) in online discussion forums. The combination can yield insights that may provide an integrated view of the formation of networks (Aggarwal & Wang, 2011). The behavioral (You, 2016) and semantic aspects (Dicheva & Dichev, 2006) of forum discussions by students may inform our understanding of collaborative learning processes and knowledge construction on online learning forums (De Laat & Lally, 2003; Kitto et al., 2016).

The objectives of this study are to examine social interactions (networks) and their changes in topics across a discussion in an asynchronous online forum and to develop a technique to visualize these interactions. Specifically, this paper explores the differences in networking patterns for major and minor topics to reveal whether online discussion forum content might be aligned with intended learning outcomes (ILOs) or contribute to unexpected learning. The findings may help both teachers and students improve their teaching and learning, respectively, with feedback from this new visualization tool. Ultimately, this study transforms teaching and learning in higher education by creating a new approach that uses advanced data mining technologies to assess students' knowledge discovery process and provide students with innovative feedback on and formative assessment of their learning. These changes can enhance capacity for knowledge discovery in collaborative online learning.

Literature review

Social constructivism

Social constructivism attends to the sociocultural aspects of learning. The social aspects of the learning process are emphasized in contemporary learning theories (Johnson & Johnson, 2008; Stahl et al., 2006). Vygotsky (1978) posited that "human learning presupposes a specific social nature and process" (p. 78). In this regard, social interactions mediate knowledge acquisition (Johnson & Johnson, 2008). Social connections exist in the learning process, which is situated between individual cognition and sociocultural contexts (Vygotsky, 1978), as well as group psychological functioning (Stahl, 2006). Learning occurs by recognizing and interpreting patterns within a network, which is social, and is technologically enhanced when dialogues take place among learners (Siemens, 2005). People form connections with one another and obtain access to different resources, which enables learning to take place and empowers future learning (Siemens, 2005). In particular, online discussion forums are widely used in virtual learning environments to mediate social interactions of various kinds through which learners can form large or small groups, use online or hybrid methods, and learn by scripted or self-organized means (Chen et al., 2018). Our evolving understanding of learning, which is culturally and socially situated, shifts our attention from individualistic traits to social aspects to take group dynamics into account.

The weak ties theory

Granovetter (1973, 1983a, 1983b) proposed the weak ties theory to study the interpersonal relationship and emphasized the "strength of weak ties," by which he meant that the weak ties of interpersonal relationships can sociologically bridge different communities and bring them into a broader context by creating micro- and macrolinks. The theory has been applied to different contexts, such as technical science (Constant et al., 1996), social media (Haythornthwaite, 2002), information diffusion (Bakshy et al., 2012), and innovation (Ruef, 2002). In online learning, the weak ties theory has been used in the field of networked learning to study learners' relationships in creating knowledge (Jones et al., 2008), in networked learning systems (Ryberg & Larsen, 2008), and in shaping personal networks online (Haythornthwaite, 2000). The weak ties that students build in online learning environments (e.g., through an online course blog) can be transferred as social capital

(Kandakatla et al., 2020). Strong or weak relational ties and social network measures can validly predict students' learning outcomes (Wu & Nian, 2021). In addition to investigating the relations between humans, many studies have examined the strength of ties between concepts or topics in information retrieval and knowledge diffusion in the interdisciplinary sciences (Wei et al., 2016). Unlike strong ties, which imply a more significant amount of the shared network, weak ties can link diverse clusters and broadly lead to more potential opportunities (Granovetter, 1973, 1983a, 1983b). Weak ties thus have merits in novel information flows because they link the sparse information of different groups (Baer, 2010; Burt, 2004). Siemens (2005) stated that how well a concept is currently linked determines the likelihood that it will be linked in the context of learning. Learning can be regarded as a process that connects specialized nodes or information sources. The connections between those nodes or sources can be strong or weak. Weak ties can be exciting and strategic to study because they are links or bridges that allow brief connections between units of information. Furthermore, the weak ties theory may be very useful in relation to the notions of serendipity, innovation, and creativity (Siemens, 2005). Connections between disparate ideas and fields can lead to innovations.

Therefore, there is great potential in using the weak ties theory to investigate the interplay between new information and weak connections, which might create innovative ideas and social capital among learners in an online discussion forum.

Visualization

Visualization tools make it easier to obtain an overview of forum messages with visualized patterns and offer pedagogical insights (Gibbs et al., 2006); however, finding a proper tool to analyze online interactions in depth remains challenging. Visual patterns contain rich information for generating analytical pictures of multiple discussion threads and investigating different dialogues. Proper visualization tools usually indicate participation and interaction patterns and gauges of potential learning (Jyothi et al., 2012).

Various studies of visualization tools using learning analytics focus on encoding social interactions (Chen et al., 2018), modeling knowledge construction (Hou et al., 2015), examining discussion content (Lin et al., 2009), understanding the cognitive development of students (Schrire, 2004), and conducting various types of correlation analyses to predict students' learning outcomes (He, 2013; Macfadyen & Dawson, 2010; Wu & Nian, 2021). However, the number of studies that have attempted to examine the influence of social interactions about text-based content in discussion forums remain limited, due to the considerable cost of analyzing large amounts of text-based data, even with data mining technology (Hara et al., 2000; Jyothi et al., 2012), and the lack of a validated analytical framework (Goodyear, 2002; Jyothi et al., 2012).

An emerging discipline that integrates visualization, data analysis, and user interaction is visual analytics (VA) (Keim et al., 2008; Thomas & Cook, 2006), which leverages the human ability to access and evaluate information effectively and efficiently with the use of an interactive design of an interface between users and data (Fekete et al., 2008). There has been a call to integrate VA into education data sense-making for the benefit of teachers, students, and school administrators (Vieira et al., 2018). The systematic review conducted by Vieira et al. (2018) pointed out that the future development of visual learning analytics should simultaneously address three dimensions: (1) VA should be connected to educational theory; (2) VA should be connected to the visualization background; and (3) VA should apply sophisticated visualization that is interactive, novel and multilevel.

Moore, in his seminal work on online learning, categorized interaction into three types: learner–learner, learner–teacher, and learner–content (Moore, 1989). Learner–learner and learner–teacher interactions refer to interpersonal relations in online discussion forums, whereas learner–content interaction refers to how learners develop the ideas and whether they stay on topic (Moore, 1989). However, analyses of the learner–content level of interaction in forums are still lacking (Jarvela & Hakkinen, 2003). Therefore, we categorized the interaction levels with the online discussion forum into three types (1) learner–forum, (2) learner–learner/teacher, and (3) learner–content—and mapped examples of the visualization tools. We use the phrase "learner–forum interaction" to refer to learners' engagement and participation in the online discussion, which can be indicated by the degree of (1) cognitive and social "presence" in a forum (Garrison et al., 1999), (2) mandatory or non-mandatory participation (Caspi et al., 2003), and (3) lurking (Beaudoin, 2002). The analysis of visualized learner–content connections is underdeveloped, partly because text data are involved. Emerging research has developed prototypes of visualization tools to combine text mining and network analysis to deepen our understanding of learner–content connections in the context of online interactions (Musabirov & Bulygin, 2020). The application of visualization tools and data mining techniques can address teachers' need for assessment tools to reduce the cumbersome manual assessment process (Garrison et al., 2001), and students' need for effective discussions that help them understand public opinions, which can increase their socio-scientific reasoning performance (Chen et al., 2020) and social–cognitive engagement (Ouyang et al., 2021). Previous studies using data mining as a strategy for assessing synchronous discussion forums have focused largely on investigating participation, interaction, and topical focus, which were studied separately (see Table 1). However, they have neglected to answer certain questions, such as what the topical focus of a specific group of interacting students is and how the strong or weak interactions between students influence topic changes or vice versa. In particular, the above methods did not focus on visualizing the weak ties of social interactions and topics, which can yield insights into unexpected learning; therefore, it is necessary to develop methods to visualize them. These previous studies are categorized and summarized in Table 1, revealing the enormous potential for visualization tools to comprehensively visualize the data mining results of different levels of connection.

To address the above research gaps, the present study posited the following two consolidated research questions: (1) what is an effective method for analyzing the learner–forum, learner–learner, and learner–content interactions on course-based discussion forums? and (2) how do we visualize and interpret the interactions to support the assessment of student learning?

Related works

Some research has centered on learner–learner interactions; however, the content level of such interactions does not explain the nature and dynamics of learning networks (Tawfik et al., 2017). In the study conducted by Tawfik et al. (2017), learner interaction was conceptualized under the interaction analysis model (IAM) (Gunawardena et al., 1997). This model shows the progression of student interaction from sharing information (phase 1) to constructing knowledge (phase 5), allowing for texts to be classified into different phases to reveal the nature of the interaction. Although the study used both SNA and content analysis for a holistic understanding of the dynamics of student interaction, it discussed the student interaction and themes separately. Few applications have addressed multiple

Table 1 Mapped visualization tools and data mining techniques in asynchronous online discussion forums

Level of interaction	Indicative examples (Dringus & Ellis, 2005)	Examples of visualization tools examples (Jyothi et al., 2012)
Learner–forum	Navigational patterns in learning management systems (Poon et al., 2017) Degree of "presence" in a forum; cognitive and social presence (Garrison et al., 1999) Mandatory/nonmandatory participation (Caspi et al., 2003) Lurking (Beaudoin, 2002)	Sequential pattern mining (Poon et al., 2017) CourseViz (Mazza & Dimitrova, 2007) TrAVis (May et al., 2007, 2008)
Learner–learner/teacher	Social learning analytics (Chen et al., 2018) Level of interaction in the forum (Jarvela & Hakkinen, 2003) Learner–learner, teacher–learner interaction activity (Moore, 1989) Teaching presence (Garrison et al., 1999)	CanvasNet (Chen et al., 2018) CourseViz (Mazza & Dimitrova, 2007) igraph (Figueira & Laranjeiro, 2007)
Learner–content	Computational thinking topical features (Cutumisu & Guo, 2019) Accuracy of message content (Jeong, 2003) Shifts in topical focus: initiating versus responding to topics (Williams & Murphy, 2002) Staying on topic (Moore, 1989)	Prototyped text mining and network analysis tools (Musabirov & Bulygin, 2020) Ligilo (Kent et al., 2019) Meerkat-ED (Rabbany et al., 2014)

aspects of learning, such as behavioral, social, and semantic aspects (Kitto et al., 2016). Despite an appeal for more holistic views of learning, meaningfully integrating information from multiple analytical aspects remains challenging (Chen et al., 2018). Additionally, He (2013) used educational data mining to investigate the significant patterns of participation and interaction (social presence, teaching presence, and cognitive presence) framed by the community of inquiry framework by analyzing online questions and chat messages in the context of a live video streaming course. The prominent themes and interaction patterns were outlined in the study, but the holistic interaction network and dynamics were not explored. Ouyang et al. (2021) visualized three kinds of network analysis—social network, topic network, and cognitive network—to understand students social-cognitive engagement in online discussion. They argued that multiple visual analytics should be integrated to understand learning contexts from different angles and empower both active and inactive students.

Methods

This exploratory study, which used social network analysis, examined the discussion forum data generated in a general studies module and investigated precisely how the strong or weak ties between topics influenced learning differently. Two kinds of network were constructed. One type of network was the topic relationship (Wei et al., 2016), in which the nodes were the topics in the discussion forum posts on the LMS, indicating the prevalence of topics in the whole discussion forum and per student as well as the relations between topics and terms. The other type of network was the social interaction network (Dawson, 2010), in which the nodes were the students, and the edges presented the relationships or flows between the nodes, reflecting the patterns of densely knit clusters and the extent of connectivity based on students' interaction, thus causing a spectrum of differences within the online learning network for a comparative study of the strong and weak ties. With the use of text mining, the study intended to combine topic modeling and social network analysis, representing the visual patterns of an online asynchronous discussion forum to construct a learning ecosystem (Aggarwal & Wang, 2011).

Methods

A total of 35 undergraduate students in a general education course called "Technology, Entertainment, and Mathematics" at a publicly funded university in Hong Kong in 2016 formed the convenience census for this improved experiment. The students who took this free elective course were in Years 2 to 4 of their 4-year undergraduate programs. Informed consent was acquired from the students before data collection and analysis. One of the course requirements was to contribute at least one reflective post to an online discussion forum in the university Moodle LMS environment. The students were asked to watch a BBC documentary called "Beautiful Equations" or other relevant movies related to mathematics, and then post in the forum their reflections on why and how the movies they watched were related to mathematics. Students were required to write their self-reflections, and were also encouraged to comment on one to three posts of others (self-selected peers) with critiques and suggestions; commenting was a way that they could obtain bonus credit for the course requirement. Thus, the intensity of learner–forum participation was expected and pragmatic. All of these posts were extracted for analysis in our experiment.

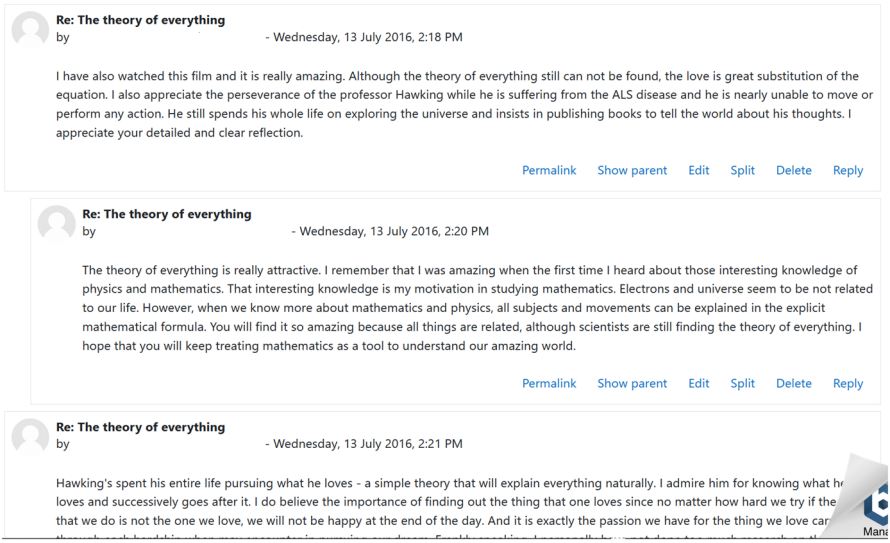


Fig. 1 Sample reflective posts from students

The selected discussion forum used in our experiment is described in Fig. 1. Sixty-eight posts from 35 students who had completed the study module were analyzed. The suggested learning analytics approach built upon previous work on capturing the interaction between students in an e-learning discussion forum to investigate the degree of students' skills and perspectives within the course (Li & Wong, 2016a, 2016b; Wong & Li, 2016).

We followed the analytic approaches of using both topic modeling (TM) by Latent Dirichlet Allocation (LDA) (Ponweiser, 2012) and social network analysis to examine the research questions concerning learner–forum, learner–learner, and learner–content relationships in the online discussion forum. We defined learner–forum interaction as students' levels of participation, which was measured by the overall number of posts that the students contributed to the forum discussion. Learner–learner interaction was the interpersonal interaction between students on the online discussion forum, which was calculated using social network analysis to identify the authorities and hubs. Learner–content interaction referred to the relationship between students and topics, which was indicated by the major and minor topics that students discussed, as well as the major and minor contributors of a specific topic.

All of the data were anonymized to ensure confidentiality. For ease of discussion, we also set random aliases for the students. We complied with text cleaning procedures to process the texts (Weiss et al., 2015). At the beginning of the text processing, all of the words were tokenized and stemmed, and the punctuation marks were removed. The stop words, the existence of which made no difference in meaning, were filtered out because the number of these words would have had adverse effects on the calculations by taking up space in the corpus. First, we conducted topic modeling of the discussion posts and generated significant topics and terms with probabilities in the whole forum and per individual post. In doing so, we constructed a document-term matrix, which is a matrix structure describing the frequency of terms in a collection of documents (Weiss et al., 2015). Four tests of measurement were conducted to select the optimal number of topics (Arun et al., 2010; Cao et al., 2009; Deveaud et al., 2014; Griffiths & Steyvers, 2004; Ponweiser, 2012). We

visualized the topic prevalence and calculated the distance between topics to investigate the variability of topic–topic relationships (Sievert & Shirley, 2014).

LDA is a generative probabilistic algorithm that samples terms from a corpus of documents randomly using Monte Carlo and presents latent/probabilistic connections of the terms by linking them together to form topics. LDA can also be regarded as a form of probabilistic topic modeling of a corpus of documents as a topic is a probability distribution of those terms (Ponweiser, 2012). The analysis for the corpus of documents was performed based on the discussion forum with students' postings, which were written in human languages. Therefore, it was desirable to use a natural language processing (NLP) approach to analysis to understand the contents, and LDA is one effective implementation of analytical algorithms that use NLP concepts (Sun et al., 2017). Therefore, LDA was selected to analyze the posts of the students in this study. LDA was designed as a data/text-mining algorithm which can be used for big data sets (Tirunillai & Tellis, 2014; Zhang et al., 2007) to derive intelligence, but LDA can also be used for small data sets (Krestel et al., 2009; Lu et al., 2016). LDA is thus a scalable data/text mining algorithm which can be used for both large and small data sets for topic modeling to derive intelligence. LDAvis (Sievert & Shirley, 2014) is an implementation of the LDA algorithm that adds web-based visualization facilities so that visualization of the user interactions can be possible. LDAvis is an open-sourced tool that can be deployed on an ordinary and affordable hardware platform. Therefore, most educational institutes can afford to use it without investing in expensive hardware and software.

At the same time, we employed social network analysis to understand learner–learner interactions. We examined centrality to identify the authorities and hubs (Kleinberg, 1999) and to discover the most active and responsive students within the learning network. Additionally, we detected communities based on the feature of betweenness (Newman & Girvan, 2004), by which students were divided into several clusters that had engaged in further discussion. In addition, to measure the strongly connected and weakly connected students, the strength of ties between students was calculated based on the number of adjacent edges.

The learner–content interaction was examined from two angles. One was the topic–topic network of students, which indicated what a particular student had discussed, including major and minor topics. The other was the learner–learner network by topic, which reflected that within a particular topic, some students were main contributors and some were not. This approach allowed us to examine the hidden patterns for unexpected learning that might have taken place among the weakly connected students in relation to a particular topic. These steps are illustrated in Fig. 2 as a pathway of data analysis of the online discussion forum posts.

Data analysis

Learner–forum interaction

We calculated the students' participation in the forum based on the number of posts that they had submitted. Based on the instructions of the course requirement, students were required to post a self-reflection, and it was suggested that they also comment on one to three posts of others. We thus categorized the participation levels of learner–forum interaction into (1) full participation (submitted a reflection and commented on three posts); (2) partial participation (submitted a reflection and commented on fewer than three posts);

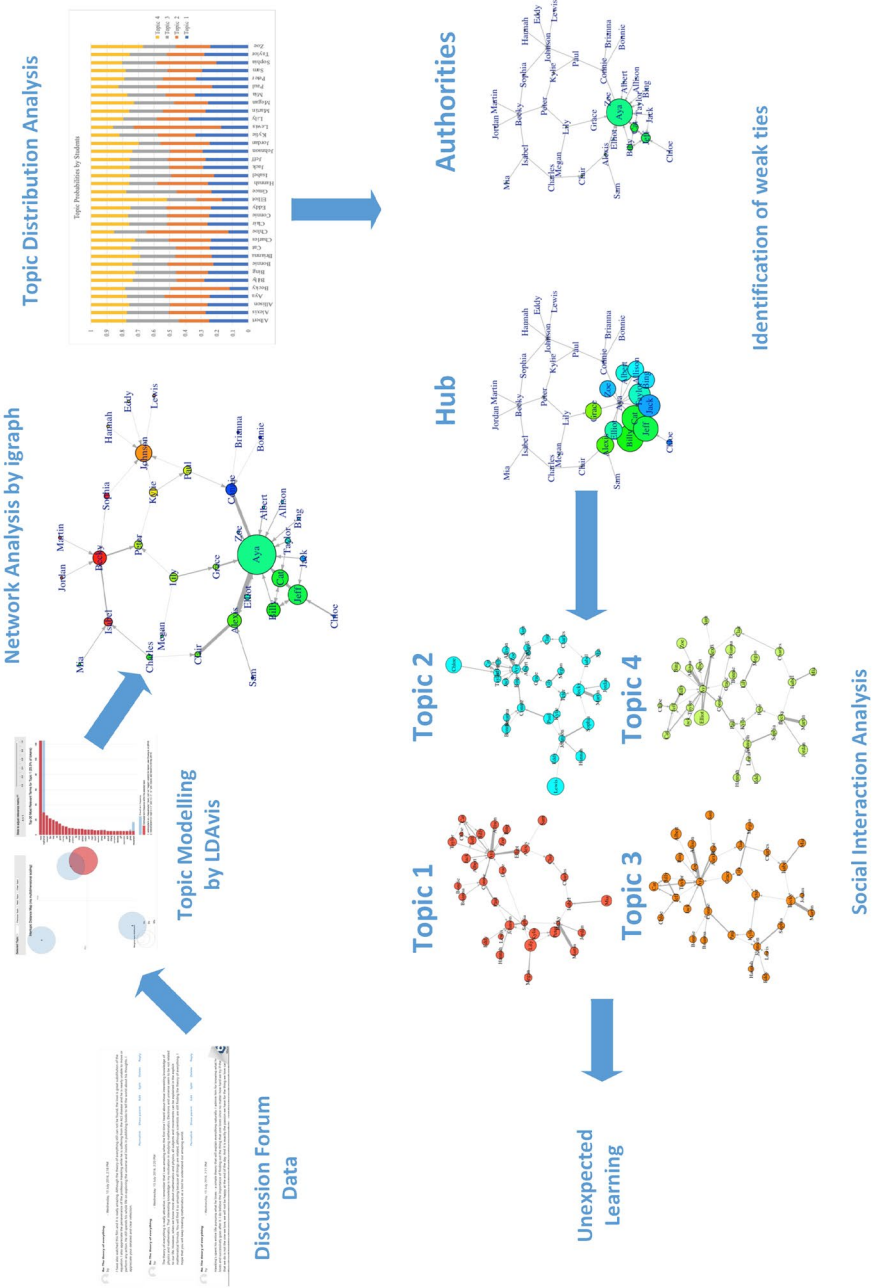


Fig. 2 Pathway of data analysis for the online discussion forum posts

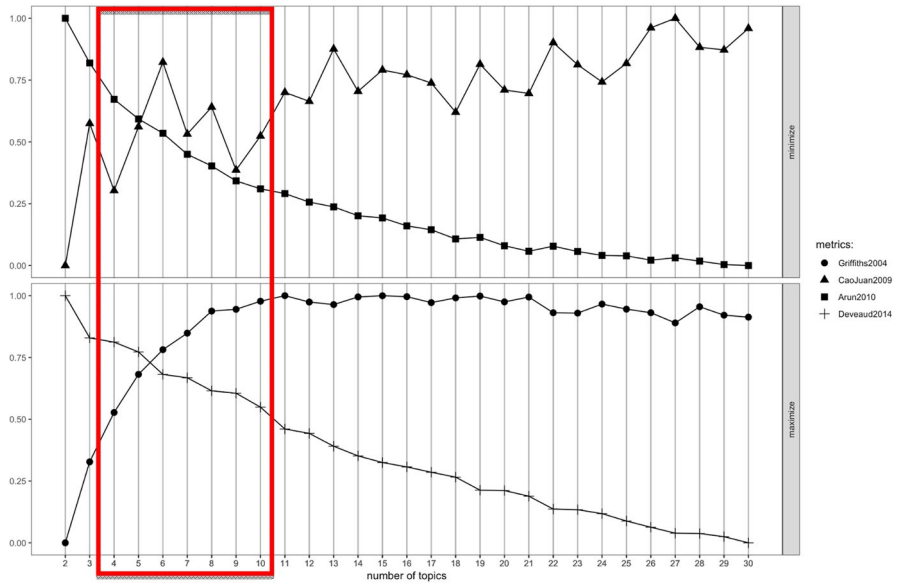


Fig. 3 Metrics of the optimal number of topics

(3) limited participation (did not submit a reflection but commented on three posts) and (4) minimal participation (did not submit a reflection and commented on fewer than three posts). The results showed that three students (Billy, Cat, Jeff) fully participated in the online discussion and 11 students partially participated. Two students fell into the category of limited participation and 19 students had minimal participation.

Topic selection

We used the metrics from Arun et al. (2010), Cao et al. (2009), Deveaud et al. (2014), and Griffiths and Steyvers (2004) to identify the range for selecting the optimal number of topics. These metrics are based on statistical inference algorithms for LDA (Ponweiser, 2012), which is a generative model for documents. Each document contains a mixture of some topics, and each topic contains a number of terms. While such an algorithm can be used to gain insight into the content of documents made up of complex texts, these metrics have been shown to select the optimal number of topics that can reflect the meaningful aspects of the documents. An optimal number of topics can be decided when the values of the metrics based on Arun et al. (2010) and Cao et al. (2009) are minimized and the values of the metrics based on Griffiths and Steyvers (2004) and Deveaud et al. (2014) are maximized. Figure 3 summarizes the values of the metrics with regard to the number of topics, from 2 to 30, after the 68 posts were analyzed using R programming language. On this basis, we decided that the optimal number of topics fell in a range between four and ten. After trials of all of the possible numbers of topics, when the topic number was set to four, we obtained four sets of topics that were relatively distinct from each other for analysis.

Topics and terms distribution

In this case, each document was considered a mixture of four topics. The top six terms for each topic are listed in Table 2. Topics 1 and 2 were more closely related to the course content, which was mathematics and calculation, whereas topics 3 and 4 were more related to the movie content, and each topic emphasized different themes. From this point of view, Topics 3 and 4 were beyond the intended learning outcomes. Document term frequency (Ponweiser, 2012) was used to describe the statistical distribution of the appearance of a term over a corpus of documents. In our case, this method was used to produce a cross reference that depicted the distribution of a topic over the corpus of each discussion post. Table 3 shows the results of the probabilities with which each topic was assigned to a post. In addition, the algorithm assigned each post to the primary topic, which had the highest probabilities, and the distribution of assignments is summarized in Table 4 and Fig. 4. Posts that had the same probability of 0.25 among the four topics were assigned to each topic, and duplicated assignments caused the total number of posts in Table 4 to be over 68. The results showed that more posts were assigned to topics 1 and 2, which means that more posts had higher probabilities of discussing the major topics of mathematics and calculation.

Table 2 Top six terms of the four topics

	Topic 1	Topic 2	Topic 3	Topic 4
1	Mathemat	Calcul	Watch	Movi
2	Life	Anim	Gambl	Dream
3	Math	Charact	Realli	Time
4	Mani	Lot	Film	Find
5	Interest	Make	Good	Earth
6	Student	Movement	Game	Watney

Table 3 Topic probabilities by post

Post	Topic 1	Topic 2	Topic 3	Topic 4
1	0.1266	0.5601	0.1137	0.1996
2	0.2669	0.3941	0.1992	0.1398
3	0.2101	0.3116	0.2246	0.2536
4	0.2721	0.2721	0.2132	0.2426
5	0.2441	0.3147	0.1382	0.3029
6	0.2905	0.2095	0.2365	0.2635
……				
64	0.25	0.2661	0.25	0.2339
65	0.2333	0.2333	0.22	0.3133
66	0.2202	0.3274	0.2679	0.1845
67	0.2230	0.2905	0.2230	0.2635
68	0.2256	0.3841	0.2012	0.1890

Table 4 Topic assignment of posts

	Topic 1	Topic 2	Topic 3	Topic 4
No. of posts	25	22	19	13
Proportion of all posts	59.49%		40.51%	

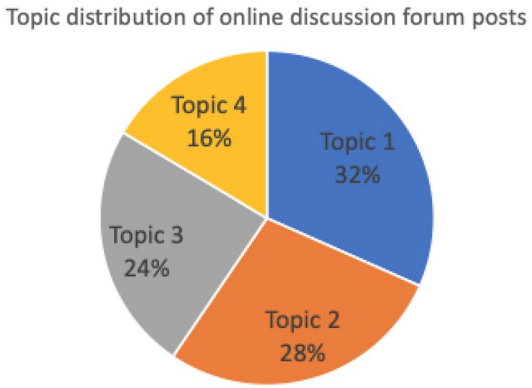


Fig. 4 Topic distribution of online discussion forum posts

Visualized topic relationship

A holistic view of topic prevalence

We used LDAvis to analyze the posts and topics by calculating the distance and prevalence of the four topics. LDA, an algorithm for topic modeling, had been used in our previous experiments (Wong & Li, 2016; Wong et al., 2016) as a text mining model for topic discovery based on the generative probabilistic model, which regards each document as a mixture of some topics, with each topic containing certain terms. LDA can uncover the hidden thematic structure from a collection of documents and help identify interesting and useful patterns. A topic is viewed as a multinomial distribution over many different ranked keywords of the corpus of some documents to reveal more profound levels of detail to be provided for analysis. This approach is superior to using only the co-occurrence of keywords to determine concepts as relevant, rather than frequent, keywords to form topics.

Furthermore, LDAvis (Sievert & Shirley, 2014), an extension of LDA, was used as a data visualization tool in our experiment. LDAvis is a web-based interactive visualization tool. It provides a high-level overview of the topics identified from the corpus of the document to show their similarities and differences by calculating the distances among them. This allows the viewer to consider the meaning and prevalence of these topics by inspecting the relevant terms (i.e., keywords) within each found topic. The LDAvis display panel enables the viewer to better understand how a particular topic can be formed by those relevant terms (Fig. 5). The left panel indicates the distance between the topics, whereas the right panel indicates the composition of the keywords within a highlighted topic. The size of the bubble indicates the prevalence of the topic. The two panels are linked so that viewers can browse all of the different found topics together with their components to understand the correlations. This critical feature allows viewers to interactively explore the themes of the corpus of the document and the associated keywords constituting the themes with relevance figures. LDAvis can provide the key term relevance for a topic model (TM) because

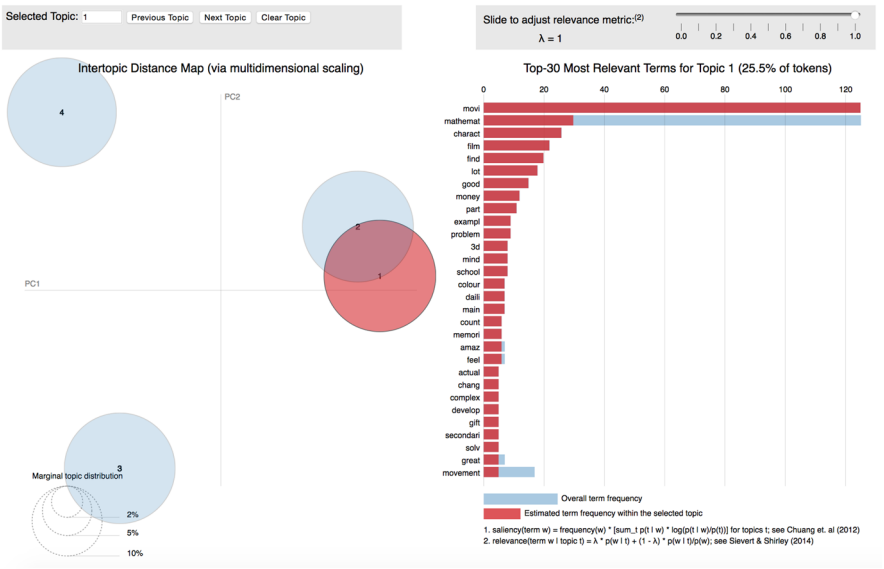


Fig. 5 Visual overview of LDAvis analysis

it sufficiently visualizes the correlation of terms among TMs and provides an interactive platform for users to select specific terms to reveal the related TM distribution (Sievert & Shirley, 2014).

In our experiment, the topic model parameter (k) was initiated at 4 for 5000 iterations of (G) to execute the likelihood of the MCMC algorithm in LDAvis. The LDAvis graph, which contained the visualization results, was generated as shown in Fig. 5. The results show that Topic 1 and 2 were closely related, whereas topic 3 and topic 4 were distant among the topics.

Topic probabilities by posts

Overall, we can see that the topic probabilities of the posts were relatively evenly distributed between 0.2 and 0.3 (see Fig. 6). Topic probabilities refer to the proportion of words within the corpus for the post that were represented by elements categorized in a particular topic. However, some posts had higher probabilities for a particular topic than for other topics. For example, posts 1, 7, and 56 had the highest probabilities in topic 2, at more than 0.5. Similarly, posts 20 and 27 had the highest probabilities in topic 1. Post 57 had the highest probabilities in topic 3. Post 61 had the highest probabilities in topic 4. Partial examples of posts are given in Figs. 7, 8, 9, 10, 11. As shown in Fig. 7, post 1 emphasized the calculations of movement, and such terms occurred frequently in topic 2 in a way that enhanced the student's understanding of mathematics. Meanwhile, the student who posted it also talked about the movie production, referring to "scenes," "character," and "colour code of the cartoon." These terms were related to topic 3 and 4, which indicates that this was an unexpected response, which may have some learning effects (i.e., unexpected connections of those topics). However, future research is needed to determine whether these responses actually led to learning opportunities. Our calculation of topic distribution by

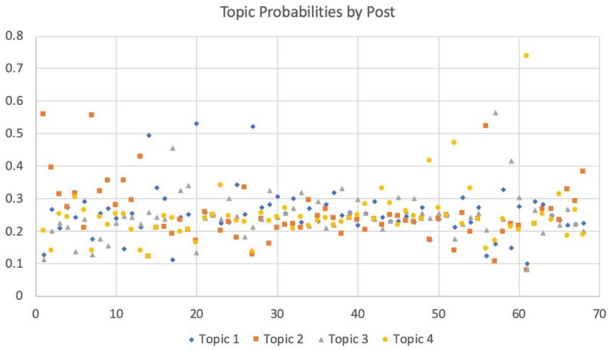


Fig. 6 Topic probabilities by post

reflectiveBlog1.txt — Edited

"I have selected a movie called 'Inside Out', a 3D animated cartoon presented by Pixar Animation Studios released in 2015. The setting of the story is the mind of an eleven-year-old girl named Riley Andersen. The story starts with introducing the five emotions, Joy, Sadness, Fear, Anger and Disgust. Riley has to move away from her hometown as her father has to work in another city. The movie is different from the other cartoon. It emphasizes on the internal feelings of Riley. It shows how the mind of Riley which is controlled by the internal panel of the five feelings reacts. The movie has shown different scenes that Riley has to face. It is funny to understand how our mind reacts when we face different sceneries. It also gives the message that different feelings are essential and special as they play different roles in one's life.

Different concepts of mathematics calculation can be found in the movie 'Inside Out'. Firstly, it is related to the scenes in the cartoon. It does not take much effort to notice that there are different movements of the characters. This requires a wide variety of calculation related to movements and motions. The production team has to pay attention to the every movement of the characters. Besides the movements, the selection of the colour of the characters required a lot of mathematics concepts too. The production team of the cartoon has to select the specific colour code from a tide of colour codes. The colour code is mostly consisted of a 5-digit number. It requires the production team to key the suitable code for the selected colour. It seems a minor part of the production, the colour of the characters play an important role in the movie!

Other mathematics concepts are geometric knowledge in 3-dimensional space using trigonometric functions and the concept of loci including maintain a fixed, equal distance from different conditions of points. Take the starting scene as an example. While Joy and Sadness are arguing for the control panel, this scene requires the producers to have a calculation in the loci in order to make the scene more realistic.

The below link is the trailer of the movie 'Inside Out'. https://www.youtube.com/watch?v=seMwpP0yeu4

Fig. 7 Forum post example 1

reflectiveBlog7.txt

"
Toy story 3 is a touching story, I love it very much. In this 3D animation, calculation really helps a lot in creating the details features and also the movements of the character, especially the mouth shape when they are talking and also the movement of their eye blow and hair strains. Maths really brings us a lot of amusement although in daily life we may not be able to notices their relation to mathematics. "

Fig. 8 Forum post example 2

probabilities (topic 1, 0.1266; topic 2, 0.5601; topic 3, 0.1137 and topic 4, 0.1996) also validated the topics of post 1.

Topic probabilities by student

Some students heavily discussed one specific topic, whereas others evenly discussed the four topics in the online discussion forum (see Fig. 12). For example, Lily was the main

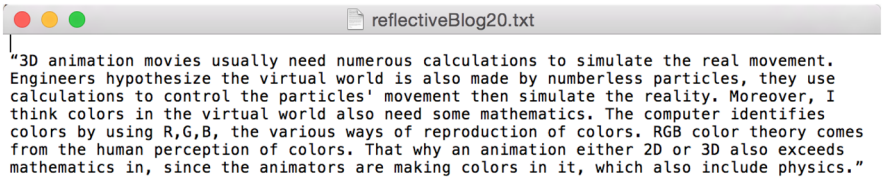


Fig. 9 Forum post example 3

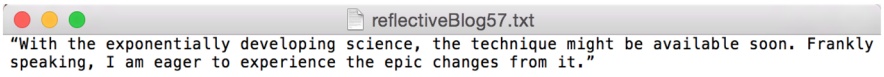


Fig. 10 Forum post example 4

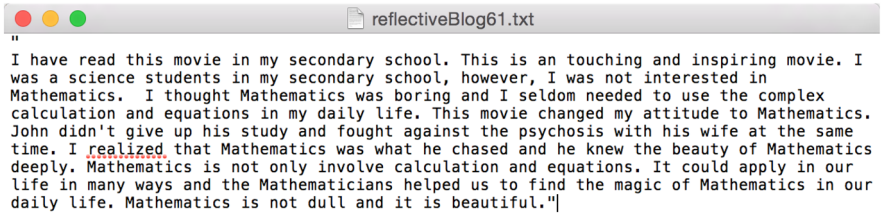


Fig. 11 Forum post example 5

contributor to topic 1. Chloe and Lewis were the active discussants of topic 2. The identification of a discussant as a main contributor to a topic meant that their statements (or statement) were made up primarily of that topic. For topic 3, the variation was not significant, but Albert contributed relatively more than did the other students. Elliot had the highest probability among all the students of discussing topic 4.

Visualized social interaction

We used igraph (https://igraph.org/), an open-source and free network analysis package for R programming language, to conduct the network analysis for the data from the discussion forum. The data was organized as shown in Table 5, with the visualized results shown in Figs. 13, 14 in a directed network.

In our case, the node size was measured based on the degree of the node, i.e., the number of adjacent edges. Each node represented each student who participated in the online discussion forum. The larger the node, the more posts the student received or sent, showing that the student was more active. Each edge was associated with a direction and weight. Kleinberg's hub and authority centrality scores were calculated to map the hubs and authorities in the online discussion forum (Kleinberg, 1999). The term "authorities" was used to refer to a node to which many other nodes were directed, whereas "hubs" were nodes that were directed to the authorities. In this case, the authorities were Aya, Billy, Cat, and Jeff, who received the most responses, as shown in Fig. 15. Comparing the learner–forum and learner–learner interaction, students who fully participated largely overlapped with the

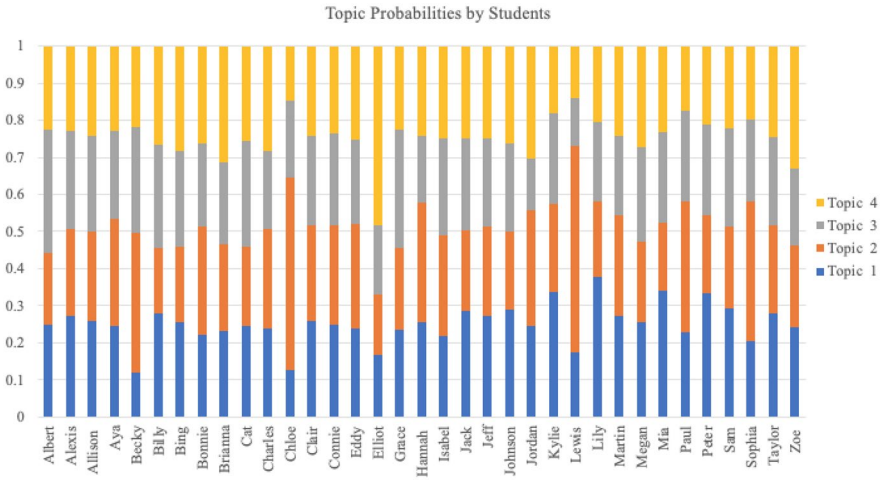


Fig. 12 Topic probabilities by student

center of the social network of learners. The slight difference was due to the measure for learner–learner interaction calculating both sending and receiving behaviors, whereas the measure for learner–forum interaction counted only the number of comments.

The size of the edge was measured by the betweenness of the edge, which indicated how many of the shortest paths passed through the edge (Newman & Girvan, 2004). The thicker the edge, the stronger the connection between the two students. Community detection was calculated based on the edge betweenness of which the scores measured the number of shortest paths (Newman & Girvan, 2004). This algorithm detected seven communities (see Fig. 16); the bridges were those edges that were the only paths linking the communities, such that their removal would lead to splitting components.

Social interaction by topic

The node size was calculated by the probability that a student would discuss a specific topic (see Figs. 17, 18, 19, 20). For example, the size of the node "Lily" in Topic 1 referred to the probability that Lily was discussing Topic 1. In contrast to the previously constructed networks, the width of the edge was calculated by the centrality of the student, indicating the activeness of the student in the online discussion forum. In this case, the width of the edge indicated how well the students were connected in terms of social interaction.

Weak ties and unexpected responses

We determined the ties between students to be weak when they had fewer shared edges, such that the thickness between the weakly connected students was thinner than that between the strongly connected students. As shown in Figs. 17, 18, 19, 20, interestingly, the strongly connected students tended to have similar probabilities in topic distribution, which meant that they had similar chances of discussing similar topics. For example, Aya and Allison were strongly connected, and their probabilities of discussing topics 1 to 4

Table 5 Three examples of data organization

From	To	Post
Sophia	Becky	Inside out is definitely a fantastic movie! I enjoy it a lot. After the introduction of the application of mathematics in Frozen by our tutor, Gary. I suddenly think that the animation of this movie is so fabulous that makes me totally paying attention to this movie because it makes people think that it's real and all this should be credits to the use of mathematics
Isabel	Becky	In this movie, I think they have also applied mathematics formula for creating the large sum of memory balls instead of drawing them one by one. Inside out is mainly focused on those memory ball containing memory of the girl. Some might be long forgotten, but when the memory flash back, no matter good or bad memories, those five emotion will affect the girl. Therefore, I think even the girl's facial expression can be generated under specific mathematical formula too!
Martin	Becky	It is surprised that many movies becoming made by mathematics. I think 'Inside Out' is a great movie and I'm getting more interesting about the equations in making this animation. It needed to be very patient during calculating complex equations. I can't imagine that how did people had the thought of using mathematics to make animation. It is a good invention in mathematics

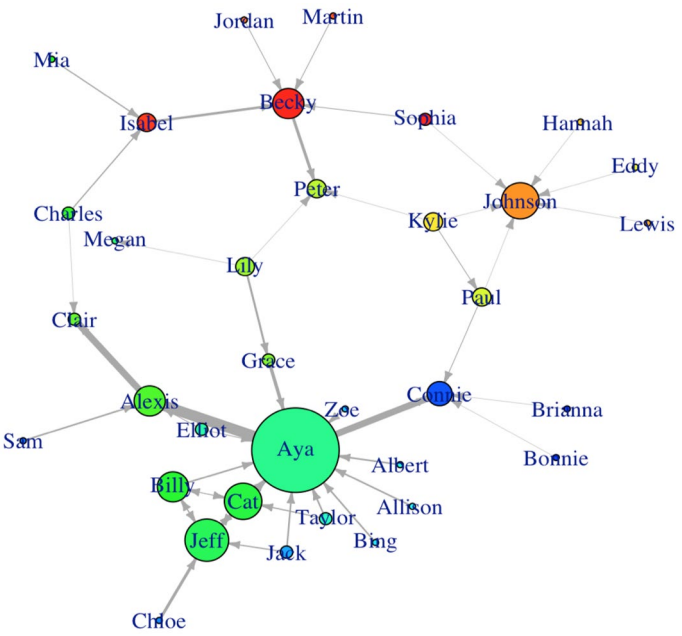


Fig. 13 Overview of the online discussion network

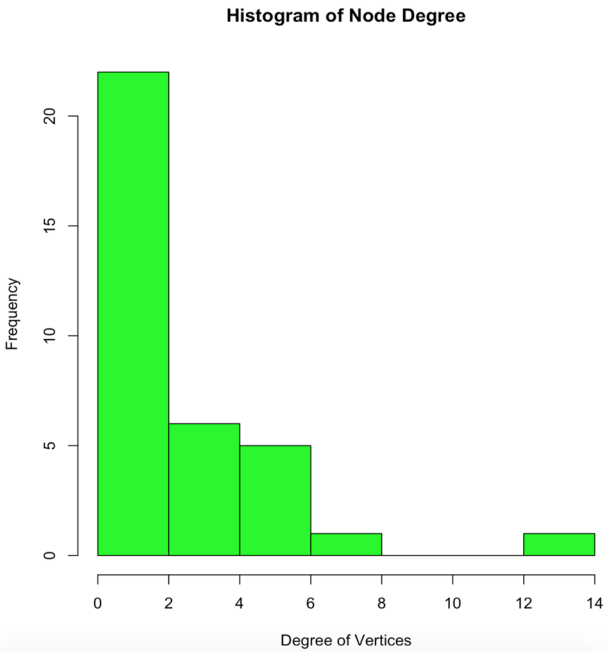


Fig. 14 Distribution of node degree and frequency

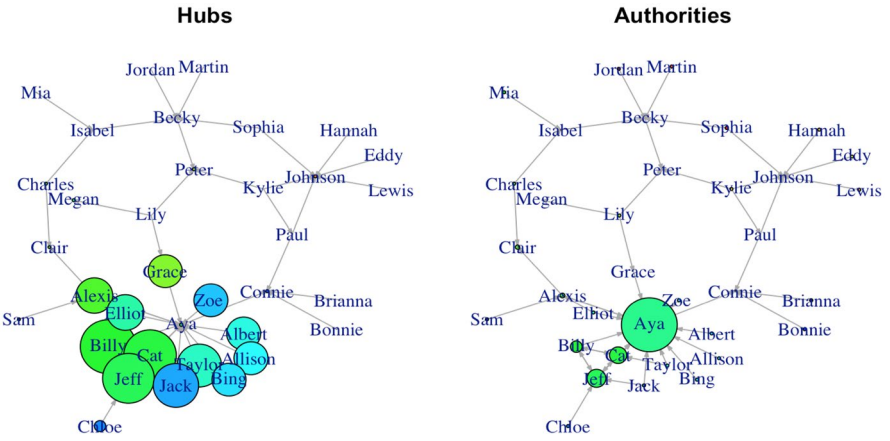


Fig. 15 Online discussion networks of hubs and authorities

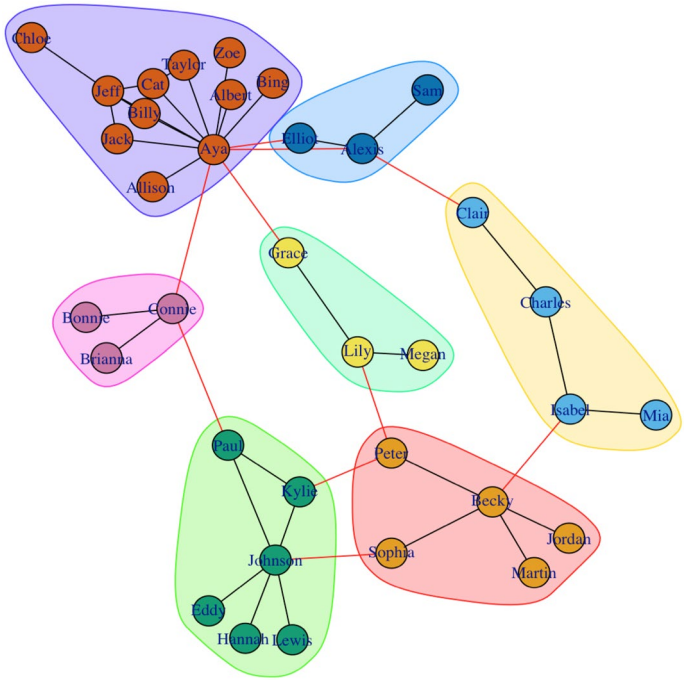


Fig. 16 Communities within online discussion networks

were nearly the same. Similar examples could be found in the case of Becky and Martin, although there were some variations. Conversely, some weakly connected students presented a different picture in which they varied in the probabilities. Lewis and Johnson, for example, were weakly connected, and the difference could be observed in their probabilities of discussing topics 1 to 4. This is a reflection that weak ties may lead to

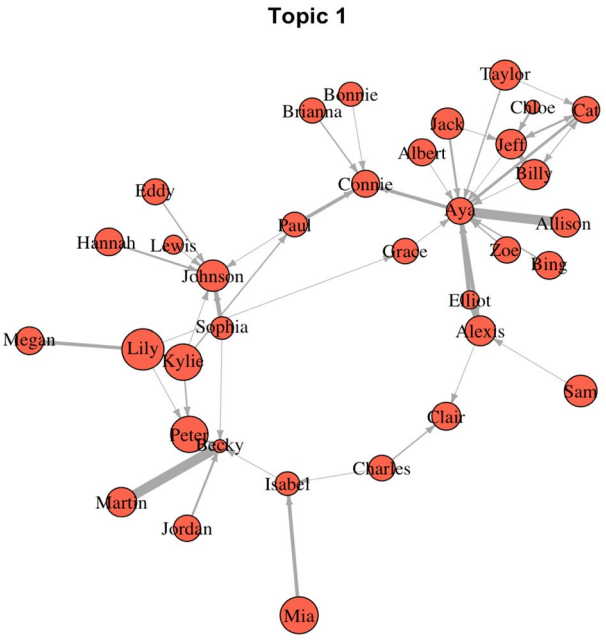


Fig. 17 Indicators of student prevalence in topic 1 network

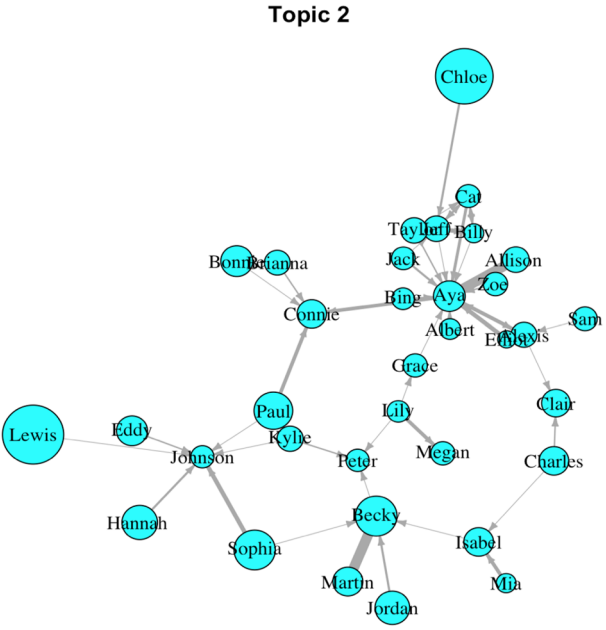


Fig. 18 Indicators of student prevalence in topic 2 network

innovative ideas (Siemens, 2005), because students who are weakly connected may discuss a wider range of topics than strongly connected students. In this sense, weakly connected students can pool together diverse and innovative ideas. However, there were indeed

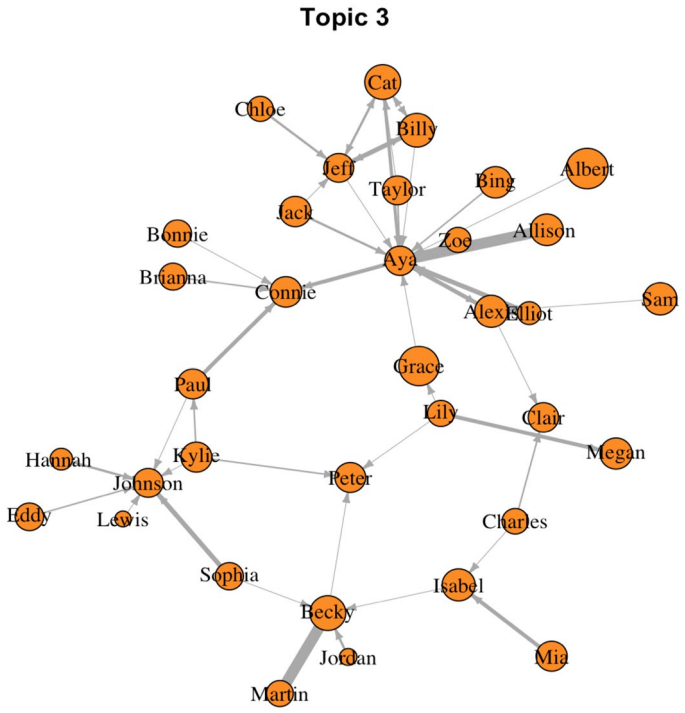
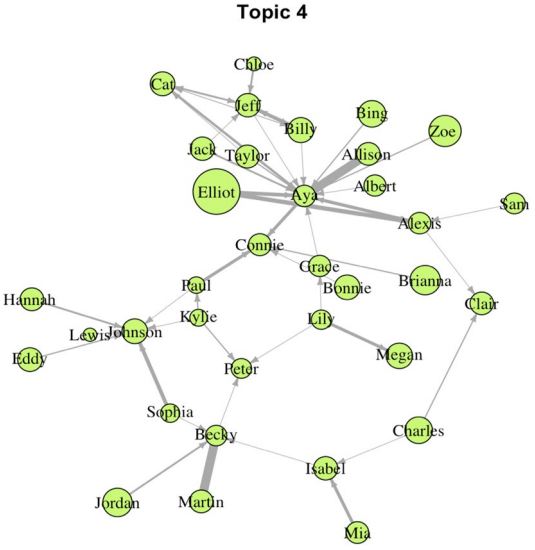


Fig. 19 Indicators of student prevalence in topic 3 network

Fig. 20 Indicators of student prevalence in topic 4 network



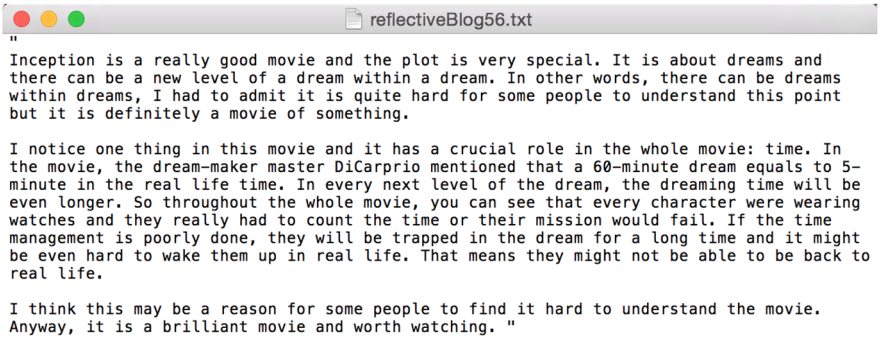


Fig. 21 Forum post example 6

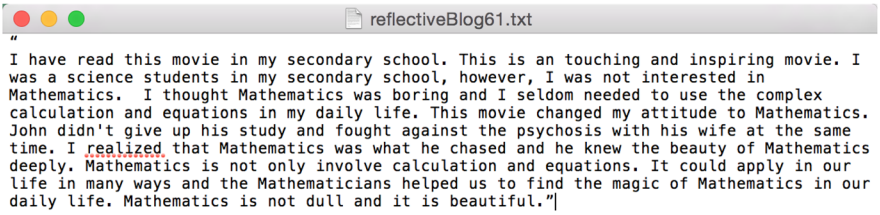


Fig. 22 Forum post example 7

some exceptions of students who were weakly connected and had similar patterns in topic distribution.

All of the students had some unexpected responses when they were posting in the discussion forum because they were discussing both mathematics and the movie, but the difference lay in the degree. In addition, as shown in Fig. 5, topics 1 and 2 were relatively more strongly connected than other relationships among topics. Chloe had fewer probabilities in topic 1, and we reviewed her post and found that it discussed issues beyond mathematics. Her post discussed the characters in the movie, so it was related to topic 2. Additionally, she discussed the concept of time and time management, which was an example of unexpected learning in the discussion (see Fig. 21 for the example).

Another example was Elliot, who appeared to be the primary content contributor to topic 4, as shown in Fig. 22, indicating a relatively higher level of unexpected responses. This example showed how Elliot developed ideas about mathematics from movies and life examples. These elements were intertwined in the reflective posts. In Fig. 23, Elliot discussed gambling, which seemed to have no connection with mathematics. However, Elliot learned the lesson, "don't be addicted to gambling," which was not the intended learning outcome of this course but was an example of serendipity, from which the student could benefit beyond the classroom.

Constructing topic-dependent social networks can shed light on the formation and dynamics of innovation, serendipity, and creativity. We can draw information from both social interaction and topic relationships. On the one hand, weaklyconnected topics (for example, topic 1 and topic 4 in our study) were largely irrelevant. However, the students were able to connect weaklyconnected topics in their discussion posts, which shows their ability to make connections to diverse topics. In this case, topic 1 and topic 2 were closely related to the course content, whereas topic 3 and topic 4 were largely unrelated,

reflectiveBlog35.txt

"
21 is an exciting movie. Ben is talented in Mathematics and he applied his talent in gambling. We always think gambling is controlled by fortune and luck. However, this movie tells us gambling is also about Mathematics although it must be very difficult to practice.

After watching this film, I recommended it to one of my relatives who love visiting the casino in Macau. I would like to let him know gambling is not only about fortune and luck but also a ruse. This film is also reflective because it also mentioned that people would miss their mind in vanity easily. Gambling is a special way to 'earn' money, however, don't be addicted to gambling."

Fig. 23 Forum post example 8

demonstrating the unexpected learning that goes beyond the intended learning outcomes that students generated innovative and creative ideas from the course-related topics. We can thus see that weak ties in social interaction and topic relationships, often considered together, can yield insights into how innovative ideas are formed and developed.

Discussion

Teachers usually want to know how their students are performing and what they are thinking. However, it is difficult and very time-consuming to read all online discussion forum threads in detail to glean this information. By using the text mining technique, teachers may come to understand how students develop their topics and more efficiently understand the dynamics of different topics. By using this visualization methodology, we can analyze (1) learner–forum, (2) learner–learner, and (3) learner–content social interactions, showing the changing dynamics among topics.

The topics in the analyzed online discussion forums had four themes. The algorithm did an excellent job of assigning the posts to different topics. This confirmed that topic models can produce two kinds of distribution: (1) the distribution of topics by their proportion in each text and (2) the distribution of words by the probability that they are related to each topic. In light of this, the results are to be interpreted in terms of the most probable words and text within the prevalent topic of interest (Musabirov, & Bulygin, 2020). Some of the topics were the hubs of the discussion (Wu & Nian, 2021). Among them, topics 1 and 2 were closely related to the course content, whereas topics 3 and 4 were comparatively irrelevant. Based on the topic assignment results, more posts were likely to be related to topics 1 and 2, as these two topics were about mathematics and calculation, which were the intended learning outcomes of the course. However, there were also some posts corresponding to topics 3 and 4. From this, we concluded that topics 1 and 2 were the major topics in the discussion forum, and topics 3 and 4 were the minor ones. In other words, topics 1 and 2 were strongly connected to each other. In addition, the combination of topics 1 and 3, topics 1 and 4, topics 2 and 3, topics 2 and 4, and topics 3 and 4 were weakly connected. When we examined the topic probabilities post by post, we found that some posts were predominantly focused on specific topics, whereas others were not. This is also referred to as the public opinions effect of online discussion forums, in which different students begin to use words similarly, reflecting the way that being exposed to public discussion gradually influences individuals' word choices (Chen et al., 2020).

The students interacted with each other on different levels, as shown in Figs. 12, 13. This provided evidence for the social capital that students built through strong or weak social ties by working with their peers on the online course blog (Kandakatla et al., 2020). The majority of the degree of vertices were below two, with some having higher degrees between 12 and 14, meaning that the majority of students were not active in the online discussion forum. As Figs. 14, 16, 17, 18, 19 show, there was no apparent relationship between the opinion leaders in the four topics and the authorities and hubs in the overall online discussion network as were calculated by the degree of nodes, regardless of the context, meaning that the topic leaders were different from the forum authorities. This demonstrates that multiple visual analytics that combine social network analysis and text mining techniques are needed to have a comprehensive understanding of forum dynamics (Ouyang et al., 2021).

We visualized the topic-based social interaction, as shown in Figs. 16, 17, 18, 19, representing the different interactions among different topics. These visualizations established the connections between the entities in texts, demonstrating who or what was mentioned together in the discussions (Musabirov & Bulygin, 2020). Such a visualization can allow teachers to identify the significant contributors to specific topics related to the intended learning outcomes, and can help students determine whether their posts are on or off topic. In addition, we found that it was around weakly connected topics where unexpected learning usually took place. This finding echoes a previous study that found that socially active students could sustain a high level of social elicitation and responsiveness, whereas peripheral students could form self-awareness of the learning process (Ouyang et al., 2021). Students' levels of learner–forum interaction do not reflect their learner-content interaction levels, because the former type of analysis is content–independent (counting the number of posts) whereas the latter is content-dependent. An awareness of this distinction was the reason why our study focused on learner–content analysis rather than students' levels of participation, because the latter does not take the content of discussion into consideration.

By combining analyses of social network analysis and topic modeling, this study contributes to visual analytics by providing a prototype for analyzing content-dependent social networks and visualizing the content-level of social interaction (Tawfik et al., 2017). An integrated view of the formation of networks (Aggarwal & Wang, 2011) is presented in the nuanced picture of how students' interaction can change when different topics are discussed. Learner–forum interaction can yield insights into the behavioral aspects (You, 2016) of the collaborative learning processes, and learner–content interaction has implications for the semantic aspects (Dicheva & Dichev, 2006).

This study adds to the literature of social constructivism by contextualizing LMS, which are composed of student interactions and topic networks. The connectivity of these networks shapes students' learning process and how they articulate and develop topics, posts, and ideas. Siemens (2005) suggested that the weak ties theory plays a role in examinations of knowledge creation, discovery, and serendipity. The results of this study support this argument and demonstrate that weak ties, both in social interactions and topic relations, have value in relation to students' unexpected learning. In addition, this study provides empirical evidence supporting Kop's (2012) statement that a higher level of serendipity can be achieved if the information provider is somewhat distant from the information collector.

Limitations

The study had several limitations. The calculations did not take into account the real-life relationships among the students, which might have influenced their online interactions. The data were captured as a snapshot to gain a holistic view of the interactions between weak ties, topics, and serendipity in a nuanced way. However, this method might have overlooked the time series formation of the online discussion network. Further research could include more student attributes, such as grade and gender, for further analysis. The network formed at different points in time could be examined to study the evolution of networks and causal relationships.

The algorithm is not perfect, which means there may be some errors in calculating the topic probabilities. However, the outcomes that it produced in this study were accurate enough for a meaningful discussion. Therefore, this study was exploratory, and used case analyses to interpret topic probabilities by triangulating different aspects of data, reflective posts, topic probabilities, and social network attributes.

Enhancements

The tool is being developed further to allow its features to be embedded within Moodle so that the data can be extracted automatically from the LMS. This will allow the feedback and visualization to be provided instantly to both students and teachers while the online discussion is being generated. This will provide instructors with a new and systemic perspective for understanding what students discuss in online discussion forums, thus aiding in formative assessment. Instructors may be able to use the topic visualization in innovative ways during the learning process, such as by developing timely prompts to facilitate student discussion on the prescribed topics. Meanwhile, students may be able to use the visualization of how they developed the topics as a guide for learning how to develop their discussion.

The writing styles and the choices of keywords used by forum content contributors also have an impact on readability. Therefore, automated tools that can help students better understand their performance without having to read the online contributions of all of the other students may be helpful.

Existing methods of learning analytics seem to focus on analyzing data collected from learners for the purpose of understanding the degree to which expected learning outcomes have been achieved. The innovation and creative elements of students' unexpected responses have been less explored, but such explorations may become a trend in curriculum design and may be pursued by some educators. Limited studies have addressed how *serendipity* (Merton & Barber, 2006) may occur among students through their participation in LMSs as a by-product of collaborative online learning and, more importantly, how unexpected responses in a discussion forum can be identified. The concept of serendipity may be worth exploring because through serendipity, accidentally relevant and surprisingly useful links can be identified for generating innovation. Both teachers and students could further explore these links to foster innovative learning and teaching.

In light of this, this study points to possible future research avenues. First, given that this study was exploratory in nature and had a small sample size, it is suggested that the study be replicated with a larger sample size. Second, a longitudinal study could further

explore the changes in major topics over time when more students interact in a discussion, such as in the context of Massive Open Online Courses. Third, further studies could examine the extent to which the learners' familiarity with each other in real life might influence their online interactions.

Conclusion

This paper presents an exploratory study examining the interplay between topic relationships and student interactions and how the confluences can be used to visualize student learning and unexpected responses in online discussion forums. The findings of the study, based on the weak ties theory, open the door to the visualization and measurement of unexpected learning. The development of more advanced text mining techniques, such as those used in this study, can allow for topics to be assigned more accurately. The results show that within online discussions, there is a divide between major topics, which are more related to course content, and minor topics, which are related to unexpected learning. Additionally, students who are not main contributors to major topics might steer the discussion to other topics through their unexpected responses. The findings pertaining to unexpected responses and serendipitous learning in online discussion forums also highlight the need to take both weak ties in social interaction and topic relationships into consideration. Understanding the discussion content and context in which the online social interactions are situated can contribute to a more holistic view of the learning process. This can lead students to better understand their process of learning, how they form topics, and their unexpected responses. The proposed visualization tool, which depicts the dynamics of topic relationships, social interactions and topic-dependent interactions, can also help teachers effectively evaluate how their students learn and whether they have achieved the expected learning outcomes or gone beyond them. Although this exploratory study had some limitations, it was able to analyze an online discussion on an asynchronous forum in terms of its social interactions (networks) and the topic changes that occurred, developing a technique to visualize these interactions. The study sheds light on a research area that bridges social interactions and topical relationships, which has seldom been addressed before. It also highlights directions for future research, such as the investigation of forum context, discussion content, and social interaction in an integrated way, which could yield insights that allow the better visualization and interpretation of interactions, thereby providing further support for the assessment of students' learning.

Acknowledgements Not applicable

Funding Not applicable.

Data availability The datasets used and/or analyzed in the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interests.

Informed consent All participants were informed with their consent to participate in this project. All information in this paper is anonymous.

Research involving human participants and/or animals Human participants were invited.

References

Aggarwal, C. C., & Wang, H. (2011). Text mining in social networks. In C. C. Aggarwal (Ed.), *Social network data analytics* (pp. 353–378). Springer.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391–402). Springer.

Baer, M. (2010). The strength-of-weak-ties perspective on creativity: A comprehensive examination and extension. *Journal of Applied Psychology, 95*(3), 592–601.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web* (pp. 519–528). ACM Press.

Beaudoin, M. F. (2002). Learning or lurking?: Tracking the "'invisible"' online student. *The Internet and Higher Education, 5*(2), 147–155.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110*(2), 349–399.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7–9), 1775–1781.

Caspi, A., Gorsky, P., & Chajut, E. (2003). The influence of group size on nonmandatory asynchronous instructional discussion groups. *The Internet and Higher Education, 6*(3), 227–240.

Chen, B., Chang, Y.-H., Ouyang, F., & Zhou, W. (2018). Fostering student engagement in online discussion through social learning analytics. *The Internet and Higher Education, 37*, 21–30.

Chen, C. M., Li, M. C., & Huang, Y. L. (2020). Developing an instant semantic analysis and feedback system to facilitate learning performance of online discussion. *Interactive Learning Environments,*. https://doi.org/10.1080/10494820.2020.1839505

Cheng, C. K., Paré, D. E., Collimore, L. M., & Joordens, S. (2011). Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers and Education, 56*(1), 253–261.

Clouder, D. L. & Deepwell, F. (2004). Reflections on unexpected outcomes: Learning from student collaboration in an online discussion forum. In S. Banks, P. Goodyear, V. Hodgson, C. Jones, V. Lally, D. McConnell & C. Steeples (Eds.) *Proceedings of the 2004 networked learning conference* (pp. 429–435). Lancaster University

Constant, D., Sproull, L., & Kiesler, S. (1996). The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization Science, 7*(2), 119–135.

Cutumisu, M., & Guo, Q. (2019). Using topic modeling to extract pre-service teachers' understandings of computational thinking from their coding reflections. *IEEE Transactions on Education, 62*(4), 325–332. https://doi.org/10.1109/te.2019.2925253

Dawson, S. (2010). "Seeing" the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology, 41*(5), 736–752. https://doi.org/10.1111/j.1467-8535.2009.00970.x

De Laat, M., & Lally, V. (2003). Complexity, theory and praxis: Researching collaborative learning and tutoring processes in a networked learning community. *Instructional Science, 31*(1–2), 7–39.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*(1), 61–84.

Dicheva, D., & Dichev, C. (2006). TM4L: Creating and browsing educational topic maps. *British Journal of Educational Technology, 37*(3), 391–404.

Dringus, L. P., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers and Education, 45*(1), 141–160.

Fekete, J.-D., van Wijk, J. J., Stasko, J. T., & North, C. (2008). The value of information visualization. In A. Kerren, J. T. Stasko, J. D. Fekete, & C. North (Eds.), *Information visualization lecture notes in computer science* (pp. 1–18). Springer.

Figueira, Á. R., & Laranjeiro, J. B. (2007). Interaction visualization in web-based learning using igraph. In *Proceedings of the 8th ACM conference on hypertext and hypermedia* (pp. 45–46). ACM Press.

Foster, A., & Ford, N. (2003). Serendipity and information seeking: An empirical study. *Journal of Documentation, 59*(3), 321–340.

Garrison, D. R., Anderson, T., & Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education, 2*(2–3), 87–105.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education, 15*(1), 7–23.

Gibbs, W. J., Olexa, V., & Bernas, R. S. (2006). A visualization tool for managing and studying online communications. *Journal of Educational Technology and Society, 9*(3), 232–243.

Goodyear, P. (2002). Psychological foundations for networked learning. In C. Steeples & C. Jones (Eds.), *Networked learning: Perspectives and issues* (pp. 49–75). Springer.
Granovetter, M. (1983a). The strength of weak ties: A network theory revisited. *Sociological Theory, 1*(1), 201–233.
Granovetter, M. (1983b). The strength of weak ties: A network theory revisited. *Sociological Theory, 1*, 201–233.
Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360–1380.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Supplement 1), 5228–5235.
Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research, 17*(4), 397–431.
Gundecha, P., & Liu, H. (2012). Mining social media: A brief introduction. In P. Mirchandani (Ed.), *Informs tutorials in operations research* (pp. 1–17). INFORMS.
Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation. *Tourism Management, 59*, 467–483.
Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Elsevier.
Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT Press.
Hara, N., Bonk, C. J., & Angeli, C. (2000). Content analysis of online discussion in an applied educational psychology course. *Instructional Science, 28*(2), 115–152.
Havnes, A., & Prøitz, T. S. (2016). Why use learning outcomes in higher education? Exploring the grounds for academic resistance and reclaiming the value of unexpected learning. *Educational Assessment, Evaluation and Accountability, 28*(3), 205–223.
Haythornthwaite, C. (2000). Online personal networks. *New Media and Society, 2*(2), 195–226.
Haythornthwaite, C. (2002). Strong, weak, and latent ties and the impact of new media. *The Information Society, 18*(5), 385–401.
He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior, 29*(1), 90–102.
Hou, H.-T., Wang, S.-M., Lin, P.-C., & Chang, K.-E. (2015). Exploring the learner's knowledge construction and cognitive patterns of different asynchronous platforms: Comparison of an online discussion forum and Facebook. *Innovations in Education and Teaching International, 52*(6), 610–620.
Jarvela, S., & Hakkinen, P. (2003). The levels of web-based discussions: Using perspective-taking theory as an analytical tool. In H. van Oostendorp (Ed.), *Cognition in a digital world* (pp. 77–95). Lawrence Erlbaum Associates.
Jeong, A. C. (2003). The sequential analysis of group interaction and critical thinking in online. *The American Journal of Distance Education, 17*(1), 25–43.
Johnson, D., & Johnson, R. (2008). Cooperation and the use of technology. In J. M. Spector, M. D. Merrill, J. van Merrienboer, & M. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 659–670). Routledge.
Jonassen, D., Davidson, M., Collins, M., Campbell, J., & Haag, B. B. (1995). Constructivism and computer-mediated communication in distance education. *American Journal of Distance Education, 9*(2), 7–26.
Jones, C. R., Ferreday, D., & Hodgson, V. (2008). Networked learning a relational approach: Weak and strong ties. *Journal of Computer Assisted Learning, 24*(2), 90–102.
Jyothi, S., McAvinia, C., & Keating, J. (2012). A visualisation tool to aid exploration of students' interactions in asynchronous online communication. *Computers and Education, 58*(1), 30–42.
Kandakatla, R., Berger, E., Rhoads, J. F., & DeBoer, J. (2020). The development of social capital in an active, blended, and collaborative engineering class. *International Journal of Engineering Education, 36*(3), 1034–1048.
Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information visualization* (pp. 154–175). Springer.
Kent, C., Rechavi, A., & Rafaeli, S. (2019). Networked learning analytics: A theoretically informed methodology for analytics of collaborative learning. *Learning in a networked society* (pp. 145–175). Springer.
Kitto, K., Bakharia, A., Lupton, M., Mallet, D., Banks, J., Bruza, P. et al. (2016). The connected learning analytics toolkit. In *Proceedings of the 6th international conference on learning analytics and knowledge* (pp. 548–549). ACM Press.
Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604–632.
Kop, R. (2012). The unexpected connection: Serendipity and human mediation in networked learning. *Educational Technology and Society, 15*(2), 2–11.

Krestel, R., Fankhauser, P., & Nejdl, W. (2009, October). Latent Dirichlet Allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems* (pp. 61–68).

Li, S. Y., & Wong, K. W. G. (2016). Educational data mining using chance discovery from discussion board. In *Proceedings of the 20th global Chinese conference on computers in education 2016* (pp. 712–715). The Hong Kong Institute of Education.

Li, Y. K., & Wong, G. K. (2016, November). Visualizing the asynchronous discussion forum data with topic detection. In *SIGGRAPH ASIA 2016 Symposium on Education: Talks* (p. 17). ACM.

Lin, F.-R., Hsieh, L.-S., & Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers and Education, 52*(2), 481–495.

Lu, H. M., Wei, C. P., & Hsiao, F. Y. (2016). Modeling healthcare data using multiple-channel latent Dirichlet allocation. *Journal of Biomedical Informatics, 60*, 210–223.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education, 54*(2), 588–599.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

May, M., George, S., & Prevot, P. (2007). Tracking, analyzing and visualizing learners' activities on discussion forums. In *Proceedings of the 6th IASTED international conference on Web-Based Education (WBE)* (pp. 649–656). WBE.

May, M., George, S., & Prevot, P. (2008). A closer look at tracking human and computer interactions in Web-based communications. *Interactive Technology and Smart Education, 5*(3), 170–188.

Mazza, R., & Dimitrova, V. (2007). Coursevis: A graphical student monitoring tool for supporting instructors in Web-based distance courses. *International Journal of Human-Computer Studies, 65*(2), 125–139.

Mazza, R. & Milani, C. (2004). 'GISMO: A graphical interactive student monitoring tool for course management systems', paper presented at *The T.E.L.'04 Technology Enhanced Learning'04 International Conference*, Milan, Italy (18–19 November).

McLoughlin, D., & Mynard, J. (2009). An analysis of higher order thinking in online discussions. *Innovations in Education and Teaching International, 46*(2), 147–160.

Merton, R. K., & Barber, E. (2006). *The travels and adventures of serendipity: A study in sociological semantics and the sociology of science*. Princeton University Press.

Moore, M. G. (1989). Editorial: Three types of interaction. *American Journal of Distance Education, 3*(2), 1–6.

Musabirov, I., & Bulygin, D. (2020). Prototyping text mining and network analysis tools to support netnographic student projects. *International Journal of Emerging Technologies in Learning, 15*(10), 223–232.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 6113.

Ouyang, F., Chen, S., & Li, X. (2021). Effect of three network visualizations on students' social-cognitive engagement in online discussions. *British Journal of Educational Technology,*. https://doi.org/10.1111/bjet.13126

Ponweiser, M. (2012). *Latent Dirichlet allocation in R* (Diploma Thesis). Vienna University of Economics and Business.

Poon, L. K. M., Kong, S.-C., Yau, T. S. H., Wong, M., & Ling, M. H. (2017). Learning analytics for monitoring students participation online: Visualizing navigational patterns on learning management system. In S. K. S. Cheung, L. Kwok, W. W. K. Ma, L.-K. Lee, & H. Yang (Eds.), *Blended learning. New challenges and innovative practices* (pp. 166–176). Springer International Publishing.

Rabbany, R., Elatia, S., Takaffoli, M., & Zaïane, O. R. (2014). Collaborative learning of students in online discussion forums: A social network analysis perspective. *Educational data mining* (pp. 441–466). Springer.

Ray, S., & Saeed, M. (2018). Applications of educational data mining and learning analytics tools in handling big data in higher education. In M. M. Alani, H. Tawfik, M. Saeed, & O. Anya (Eds.), *Applications of big data analytics* (pp. 135–160). Springer.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers and Education, 51*(1), 368–384.

Ruef, M. (2002). Strong ties, weak ties and islands: Structural and cultural predictors of organizational innovation. *Industrial and Corporate Change, 11*(3), 427–449.

Ryberg, T., & Larsen, M. C. (2008). Networked identities: Understanding relationships between strong and weak ties in networked environments. *Journal of Computer Assisted Learning, 24*(2), 103–115.

Schrire, S. (2004). Interaction and cognition in asynchronous computer conferencing. *Instructional Science, 32*(6), 475–502.

Shinde, P. P., Oza, K. S., & Kamat, R. K. (2017, February). Big data predictive analysis: Using R analytical tool. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 839–842). IEEE.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning, 2*(1), 1–8.
Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).
Slade, S., & Galpin, F. (2012). Learning analytics and higher education: Ethical perspectives. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 16–17). ACM Press.
Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. MIT Press.
Stahl, G., Koschmann, T., & Suthers, D. D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409–426). Cambridge University Press.
Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion, 36*, 10–25.
Tawfik, A. A., Reeves, T. D., Stich, A. E., Gill, A., Hong, C., McDade, J., et al. (2017). The nature and level of learner–learner interaction in a chemistry massive open online course (MOOC). *Journal of Computing in Higher Education, 29*(3), 411–431.
Thomas, J. J., & Cook, K. A. (2006). A visual analytics agenda. *IEEE Computer Graphics and Applications, 26*(1), 10–13.
Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using Latent Dirichlet Allocation. *Journal of Marketing Research, 51*(4), 463–479.
Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers and Education, 122*, 119–135.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Cole, V. John-Steiner, S. Scribner, E. Souberman, Trans.). Harvard University Press.
Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of predictive text mining*. Springer.
Wei, L., Xu, H., Wang, Z., Dong, K., Wang, C., Fang, S., et al. (2016). Topic detection based on weak tie analysis: A case study of LIS research. *Journal of Data and Information Science, 1*(4), 81–101. https://doi.org/10.20309/jdis.201626
Wong, G. K., & Li, S. Y. (2016). Academic performance prediction using chance discovery from online discussion forums. In *2016 IEEE 40th annual computer software and applications conference (COMPSAC)* (pp. 706–711). IEEE.
Wong, G. K., Li, S. Y., & Wong, E. W. (2016). Analyzing academic discussion forum data with topic detection and data visualization. In *2016 IEEE international conference on teaching, assessment, and learning for engineering (TALE)* (pp. 109–115). IEEE.
Wu, J. Y., & Nian, M. W. (2021). The dynamics of an online learning community in a hybrid statistics classroom over time: Implications for the question-oriented problem-solving course design with the social network analysis approach. *Computers and Education,*. https://doi.org/10.1016/j.compedu.2020.104120
Wu, X., Zhu, X., Wu, G., & Ding, W. (2013). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97–107.
Williams, C. B., & Murphy, T. (2002). Electronic discussion groups: How initial parameters influence classroom performance. *Educause Quarterly, 25*(4), 21–29.
You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education, 29*, 23–30.
Zhang, H., Qiu, B., Giles, C. L., Foley, H. C., & Yen, J. (2007, May). An LDA-based community structure discovery approach for large-scale social networks. In *2007 IEEE Intelligence and Security Informatics* (pp. 200–207). IEEE.

Gary K. W. Wong is an assistant professor in Faculty of Education at the University of Hong Kong. His research interests are computer science education, computer-mediated learning environment, learning design with technology and educational technology.

Yiu Keung Li is the research assistant in the Faculty of Education at the University of Hong Kong.

Xiaoyan Lai is a graduate student in information and technology studies.