IEEE TALE 2016 CONFERENCE

Analyzing Academic Discussion Forum Data with Topic Detection and Data Visualization

By Gary K. W. Wong, Simon Y. K. Li, Elby W. Y. Wong

Image Source: EDUCAUSE

香港教育大學
The Education University of Hong Kong

Document version 1.0

Agenda

1. Project Background
2. Problem of Analyzing Discussion Forum
3. Prior Works for Experiment Design
4. Experiment Design and Results
   - 4.1 Improved model of text mining
     - 4.1.1 Data Visualization – Forum Graph
   - 4.2 Data Visualization – LDAvis
     - 4.2.1 LDAvis Limitations
   - 4.3 Supplementary Visualization Means
5. Conclusion
6. Discussion and Further Development

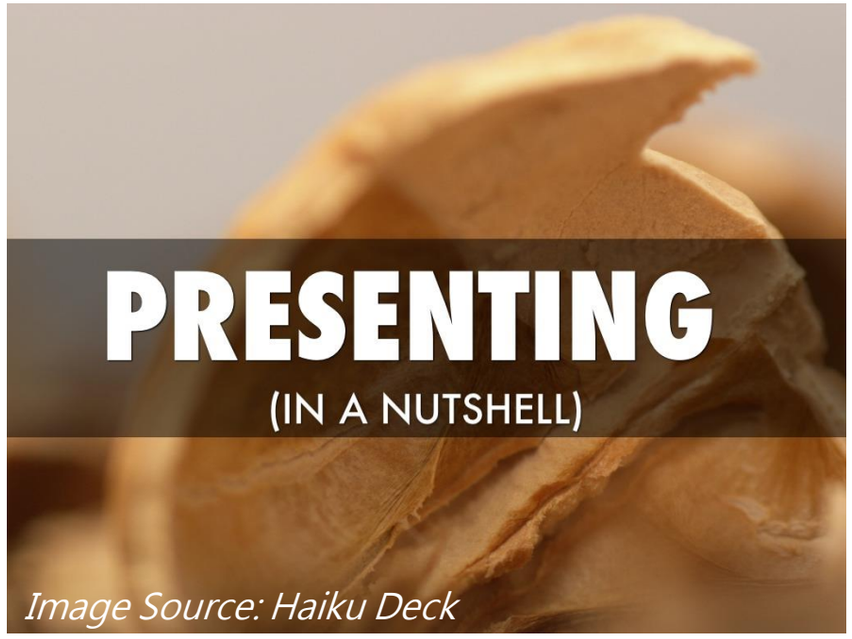
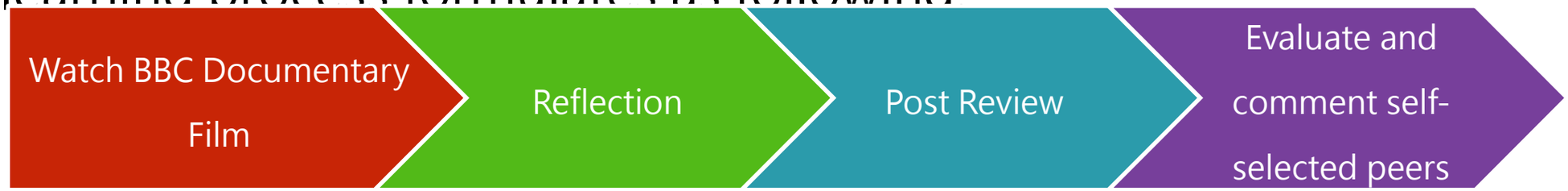


Image Source: Haiku Deck

1. Project Background

- A total of 40 undergraduate students in the Education University of Hong Kong from the General Education course called "Technology, Entertainment, and Mathematics" have been sampled for this improved experiment
- One of the course requirements was to complete at least one reflective posting on an online discussion forum in the Moodle environment of the university.
- The learning process formulates as following



- There were more than 200 posts sampled from the forum with 36 students who had completed the related study module

Image Source: University of Strathclyde

Project Background (…continued)

- At the early stages of this project, a software tools called "Polaris" from Ohsawa Laboratory was used for mining text from the following sources.

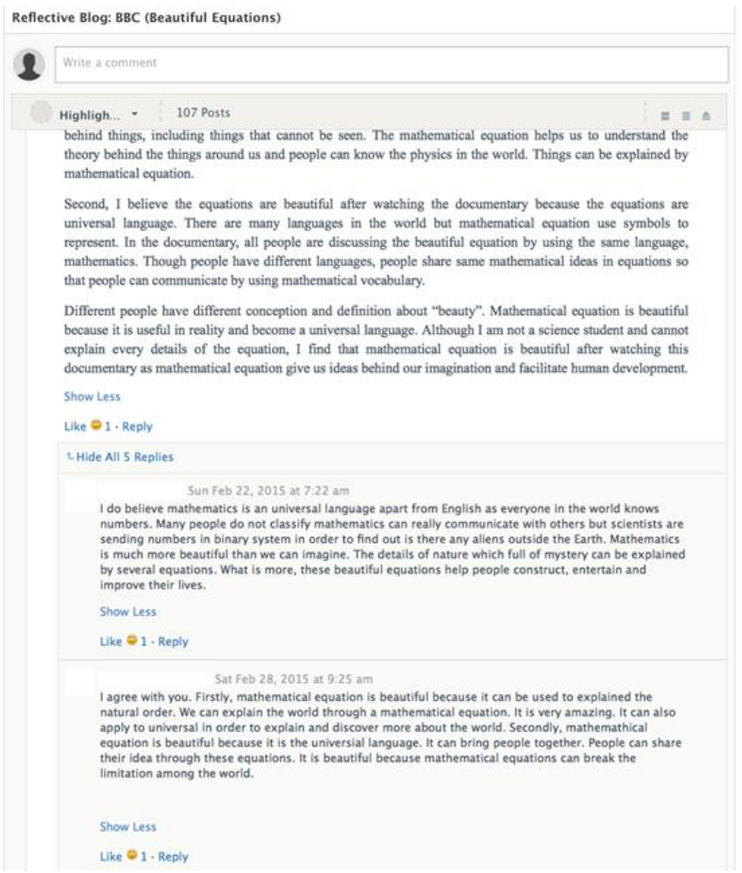
- Sources of Data (text format):

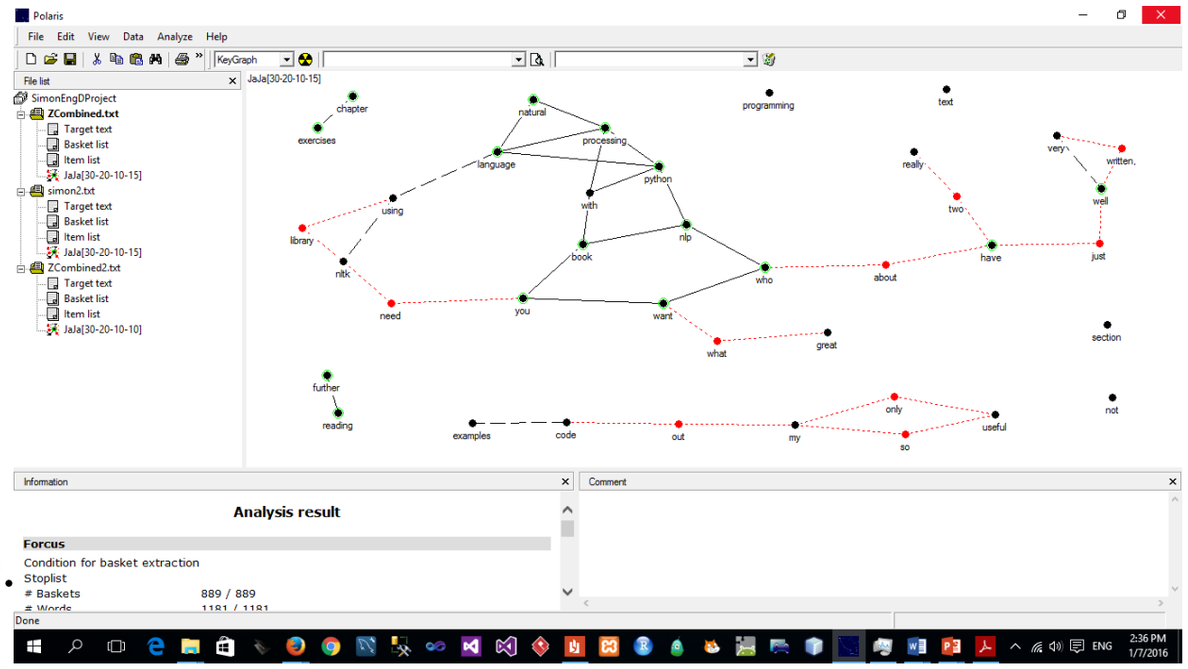
  - Online reflective discussion forum, etc.

- Performed analysis using KeyGraph to generate the visual patterns to identify:

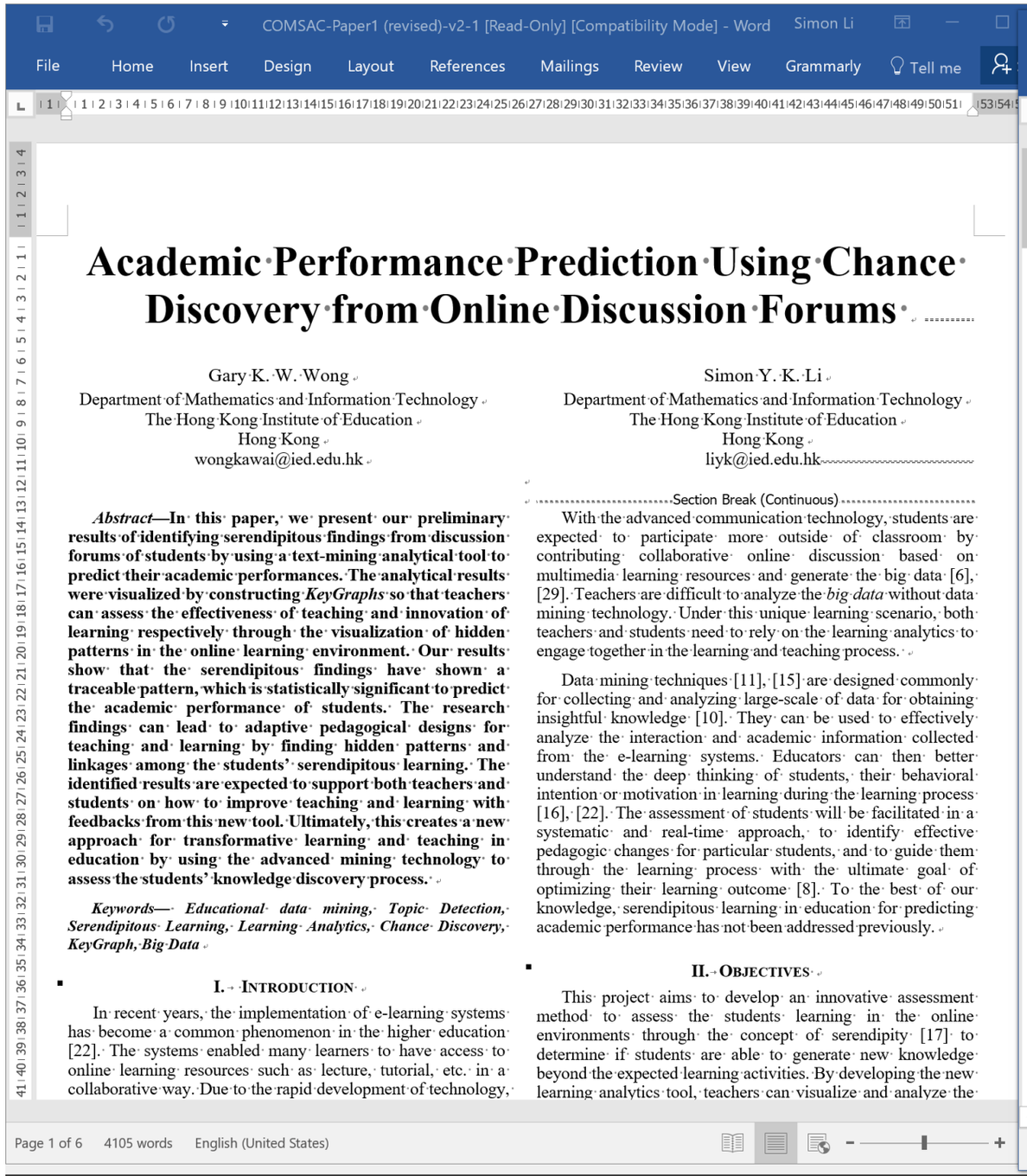
  - The formulation of key concepts from black nodes and links

  - chances (red nodes and links) for the purposes of decision making and planning in the associated areas above.





Project Background (…continued)



COMSAC-Paper1 (revised)-v2-1 [Read-Only] [Compatibility Mode] - Word — Simon Li

File Home Insert Design Layout References Mailings Review View Grammarly Tell me

Academic Performance Prediction Using Chance Discovery from Online Discussion Forums

Gary K. W. Wong
Department of Mathematics and Information Technology
The Hong Kong Institute of Education
Hong Kong
wongkawai@ied.edu.hk

Simon Y. K. Li
Department of Mathematics and Information Technology
The Hong Kong Institute of Education
Hong Kong
liyk@ied.edu.hk

***Abstract*—In this paper, we present our preliminary results of identifying serendipitous findings from discussion forums of students by using a text-mining analytical tool to predict their academic performances. The analytical results were visualized by constructing *KeyGraphs* so that teachers can assess the effectiveness of teaching and innovation of learning respectively through the visualization of hidden patterns in the online learning environment. Our results show that the serendipitous findings have shown a traceable pattern, which is statistically significant to predict the academic performance of students. The research findings can lead to adaptive pedagogical designs for teaching and learning by finding hidden patterns and linkages among the students' serendipitous learning. The identified results are expected to support both teachers and students on how to improve teaching and learning with feedbacks from this new tool. Ultimately, this creates a new approach for transformative learning and teaching in education by using the advanced mining technology to assess the students' knowledge discovery process.**

***Keywords*— Educational data mining, Topic Detection, Serendipitous Learning, Learning Analytics, Chance Discovery, KeyGraph, Big Data**

I. Introduction

In recent years, the implementation of e-learning systems has become a common phenomenon in the higher education [22]. The systems enabled many learners to have access to online learning resources such as lecture, tutorial, etc. in a collaborative way. Due to the rapid development of technology,

Section Break (Continuous)

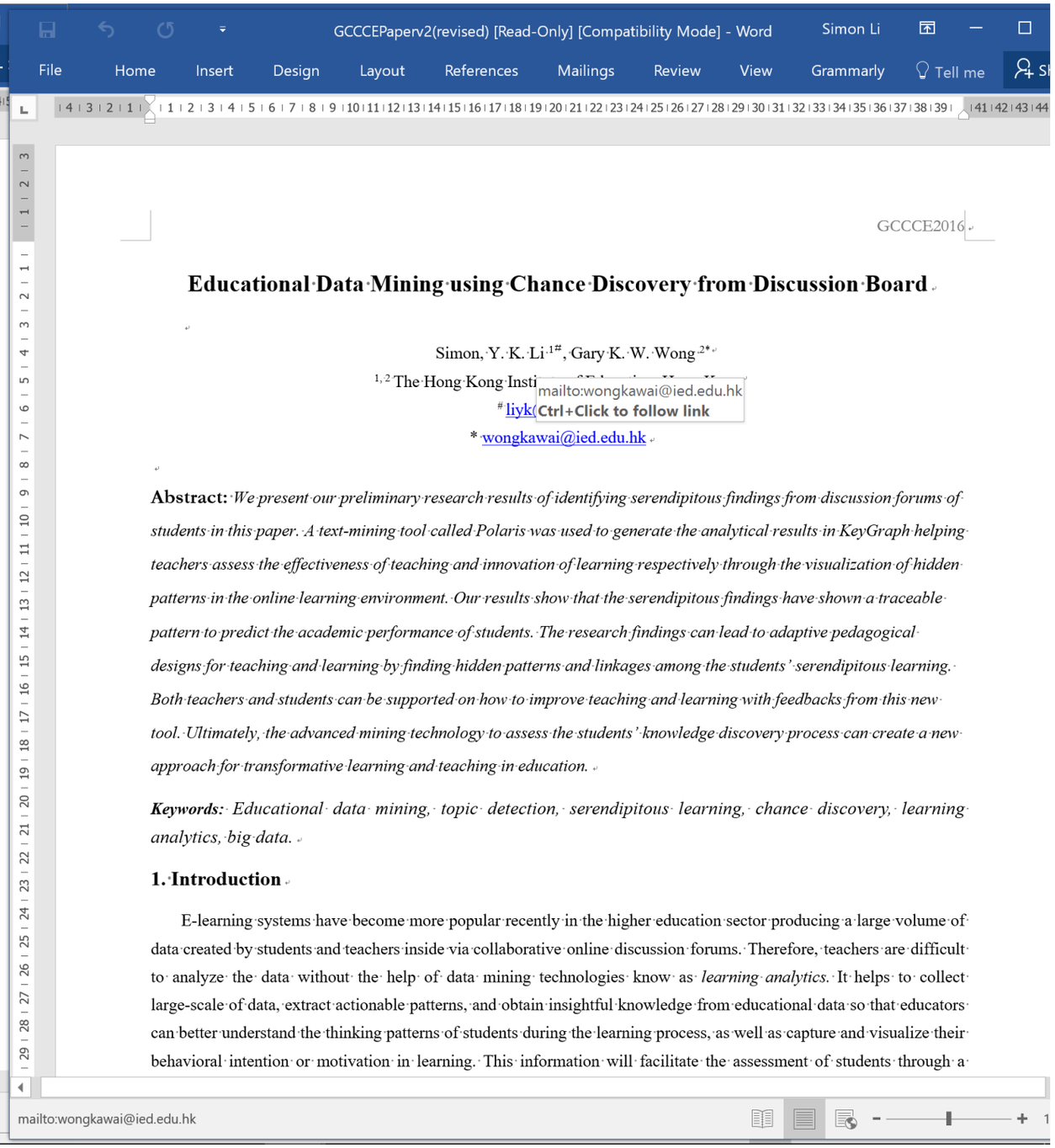
With the advanced communication technology, students are expected to participate more outside of classroom by contributing collaborative online discussion based on multimedia learning resources and generate the big data [6], [29]. Teachers are difficult to analyze the *big data* without data mining technology. Under this unique learning scenario, both teachers and students need to rely on the learning analytics to engage together in the learning and teaching process.

Data mining techniques [11], [15] are designed commonly for collecting and analyzing large-scale of data for obtaining insightful knowledge [10]. They can be used to effectively analyze the interaction and academic information collected from the e-learning systems. Educators can then better understand the deep thinking of students, their behavioral intention or motivation in learning during the learning process [16], [22]. The assessment of students will be facilitated in a systematic and real-time approach, to identify effective pedagogic changes for particular students, and to guide them through the learning process with the ultimate goal of optimizing their learning outcome [8]. To the best of our knowledge, serendipitous learning in education for predicting academic performance has not been addressed previously.

II. Objectives

This project aims to develop an innovative assessment method to assess the students learning in the online environments through the concept of serendipity [17] to determine if students are able to generate new knowledge beyond the expected learning activities. By developing the new learning analytics tool, teachers can visualize and analyze the

Page 1 of 6 4105 words English (United States)



GCCCEPaperv2(revised) [Read-Only] [Compatibility Mode] - Word — Simon Li

File Home Insert Design Layout References Mailings Review View Grammarly Tell me

GCCCE2016

Educational Data Mining using Chance Discovery from Discussion Board

Simon, Y. K. Li [1#], Gary K. W. Wong [2*]
[1, 2] The Hong Kong Insti…
mailto:wongkawai@ied.edu.hk
Ctrl+Click to follow link
[#] liyk(
* wongkawai@ied.edu.hk

Abstract: *We present our preliminary research results of identifying serendipitous findings from discussion forums of students in this paper. A text-mining tool called Polaris was used to generate the analytical results in KeyGraph helping teachers assess the effectiveness of teaching and innovation of learning respectively through the visualization of hidden patterns in the online learning environment. Our results show that the serendipitous findings have shown a traceable pattern to predict the academic performance of students. The research findings can lead to adaptive pedagogical designs for teaching and learning by finding hidden patterns and linkages among the students' serendipitous learning. Both teachers and students can be supported on how to improve teaching and learning with feedbacks from this new tool. Ultimately, the advanced mining technology to assess the students' knowledge discovery process can create a new approach for transformative learning and teaching in education.*

Keywords: *Educational data mining, topic detection, serendipitous learning, chance discovery, learning analytics, big data.*

1. Introduction

E-learning systems have become more popular recently in the higher education sector producing a large volume of data created by students and teachers inside via collaborative online discussion forums. Therefore, teachers are difficult to analyze the data without the help of data mining technologies know as *learning analytics*. It helps to collect large-scale of data, extract actionable patterns, and obtain insightful knowledge from educational data so that educators can better understand the thinking patterns of students during the learning process, as well as capture and visualize their behavioral intention or motivation in learning. This information will facilitate the assessment of students through a

mailto:wongkawai@ied.edu.hk

Two conferences papers were published before with the results analyzed using Keygraph

香港教育大學 The Education University of Hong Kong

2. Problem of Analyzing Discussion Forum

- **Problem 1**
  - Teachers usually want to know how their students perform or what the students are thinking
  - However, it is **difficult and time-consuming** to read all online discussion forum threads in details to comprehend the information inside manually

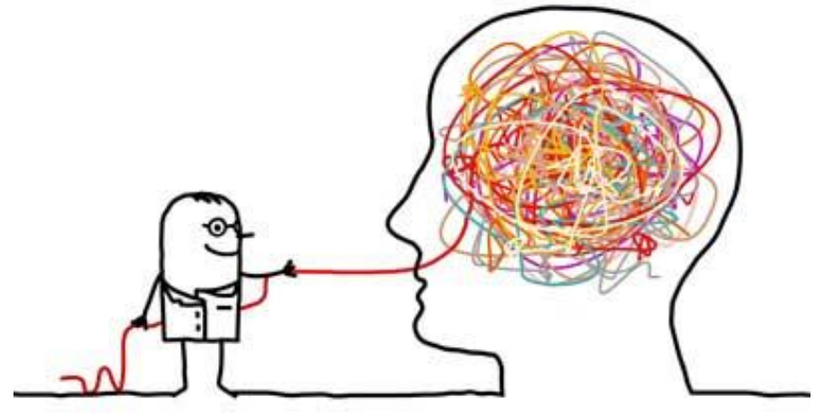
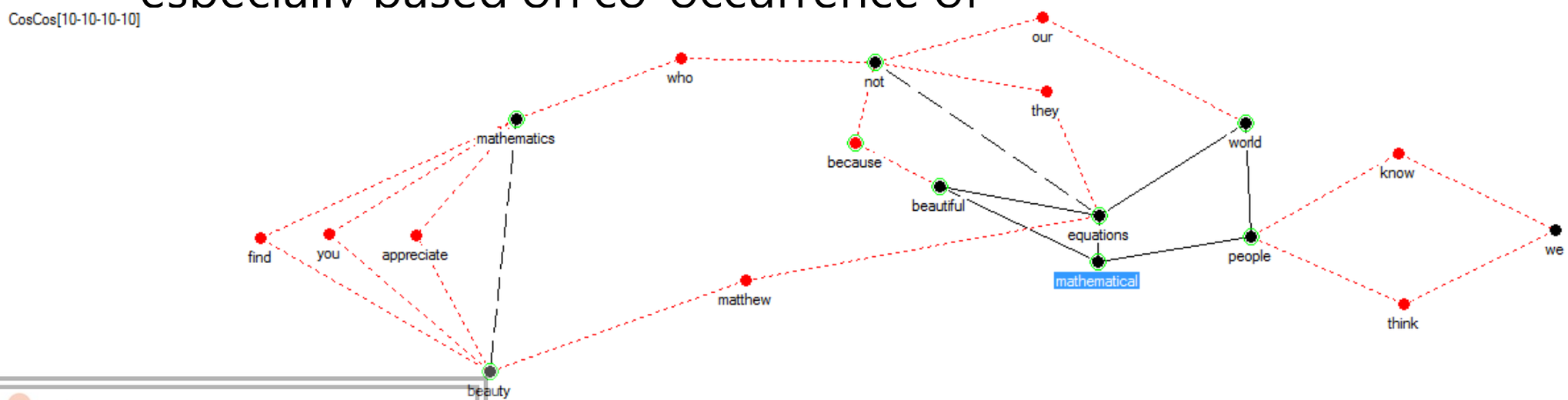


Image Source: Health Rising

- **Problem 2**
  - Keygraphs could be **difficult to interpret**, especially based on co-occurrence of



Preliminary Results – BBC (Beautiful Equations)

3. Prior Works for New Experiment Design

1. **Why using social interaction analysis for our Moodle discussion forum?**
   - Coffrin, Corrin et al. (2014) proposed visualization methods to **realize the student engagement and performance** in massive open online course (MOOC) environment.
2. **Why using a probabilistic topic model with clustering visualization approach?**

Image Source: HSE Science Olympiad

2011

Duval (2011) learning analytics could facilitate by **collecting, analyzing, and displaying the traces** that learners left behind to improve learning.

2012

One of the well-developed learning analytics systems is called Gradient's Learning Analytics System **(GLASS)** according to Leony, Pardo, et al. (2012).

This system captures and visualizes the events of learning with a **dashboard** serving as a presentation layer to display important analytics figures.

2015 - 2016

•Ezen-Can, Boyer et al. (2015) used the ideas of **clustering** to group discussion topics.

•Atapattu, Falkner et al. (2016) raised the ideas using **topic-wise classification** of discussion threads on MOOC.

Inspired by all these ideas, we selected **using probabilistic topic model together with a clustering visualization approach** in this project to visualize the student performance so that the teachers can better understand the performance-related data by using a visual mean.

4. New Experiment Design and Results

- In the latest experiment environment, we deployed a **Moodle environment to host discussion forum** of reflective postings from students.
- The contents of discussion forum of the general education course extracted from this Moodle environment for social interaction analysis which performed by a tool called **Forum Graph** (Chan, 2013)
- The data was then exported to a collection of programs written in R with packages implementing **Latent Dirichlet Allocation (LDA)** (Blei, 2012).
- The analysis results would then be visualized to

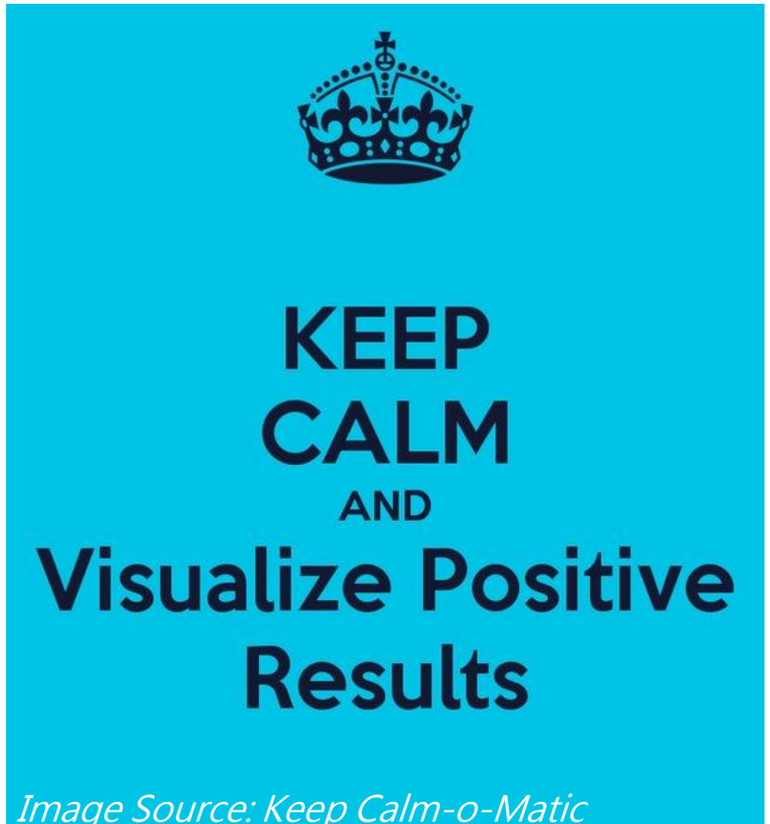


Image Source: Keep Calm-o-Matic

Image Source: Lambda Solutions

4.1 Improved model of text mining

- LDA was used as a text mining model in our latest experiment for topic discovery based on **generative statistical/probabilistic model**. It assumes that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.
- LDA aims at **uncovering the hidden thematic structure** in a collection of some document to help identify interesting and useful patterns. A topic is a multinomial distribution over many different ranked keywords of the corpus of some document. The levels of details provided for analysis can be made deeper.
- This approach is better than just relying on keygraphs, by **using clustering and sequencing co-occurrence of keywords to determine concepts as the** ords alone, to form topics.

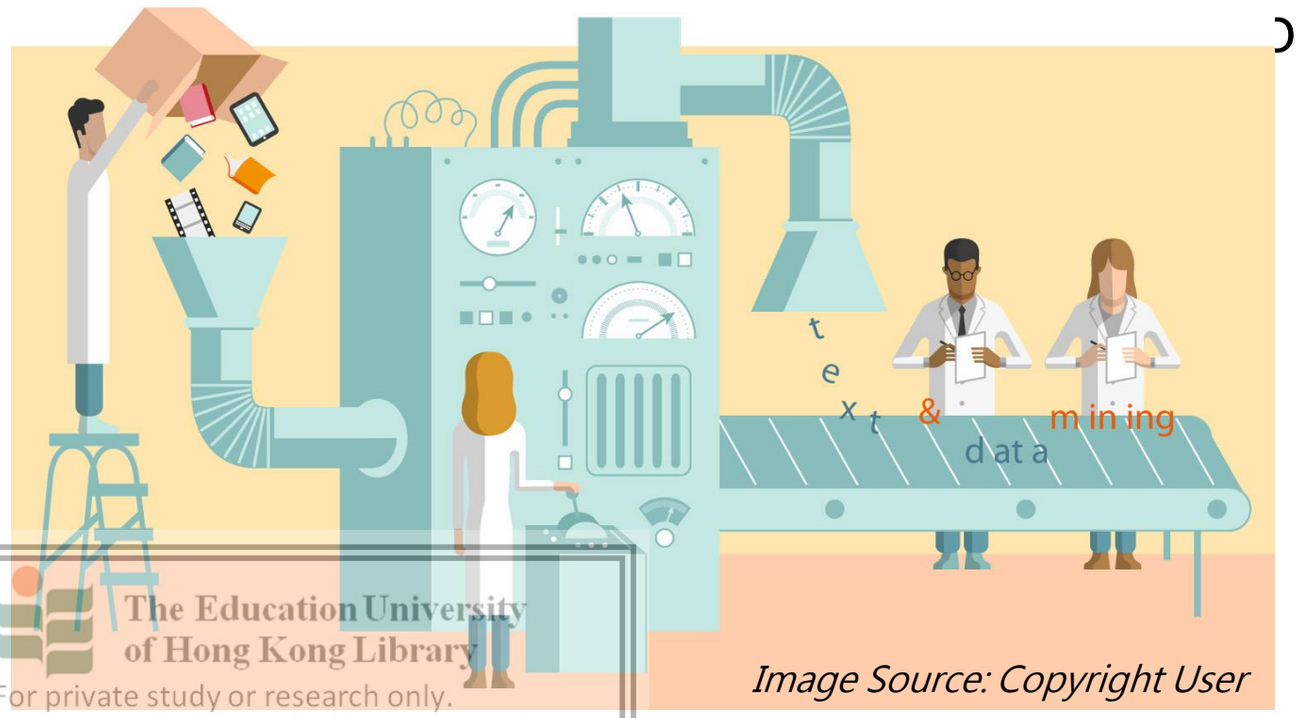
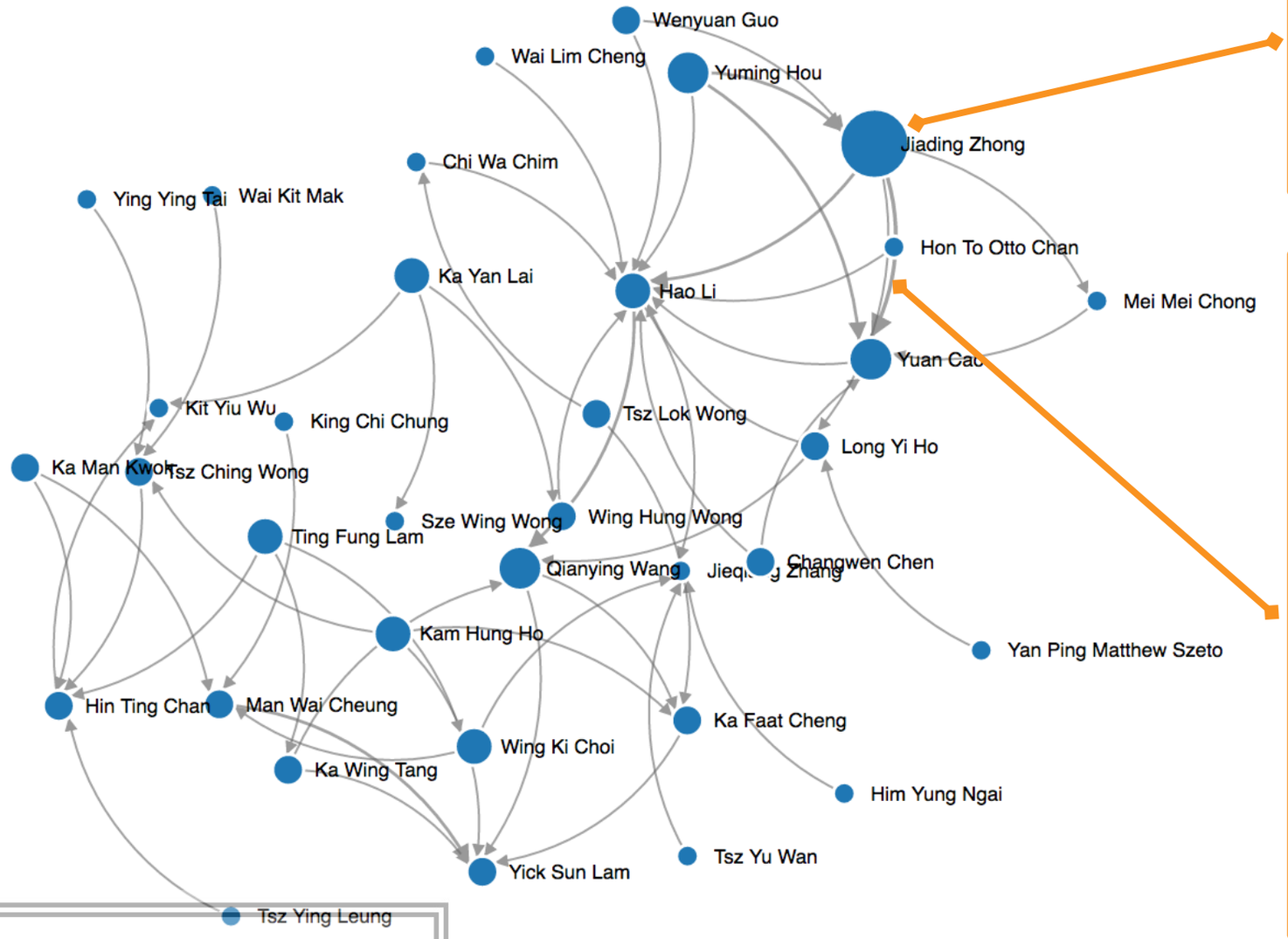


Image Source: Copyright User

4.1.1 Data Visualization – Forum Graph

- The social interaction inside one of the course discussion forums was visualized using "node" and "edges" in which "nodes" refer to the student while "edges" refer to interaction performance. The general rules to comprehend the Forum Graph are stated as follows.



Size of Node

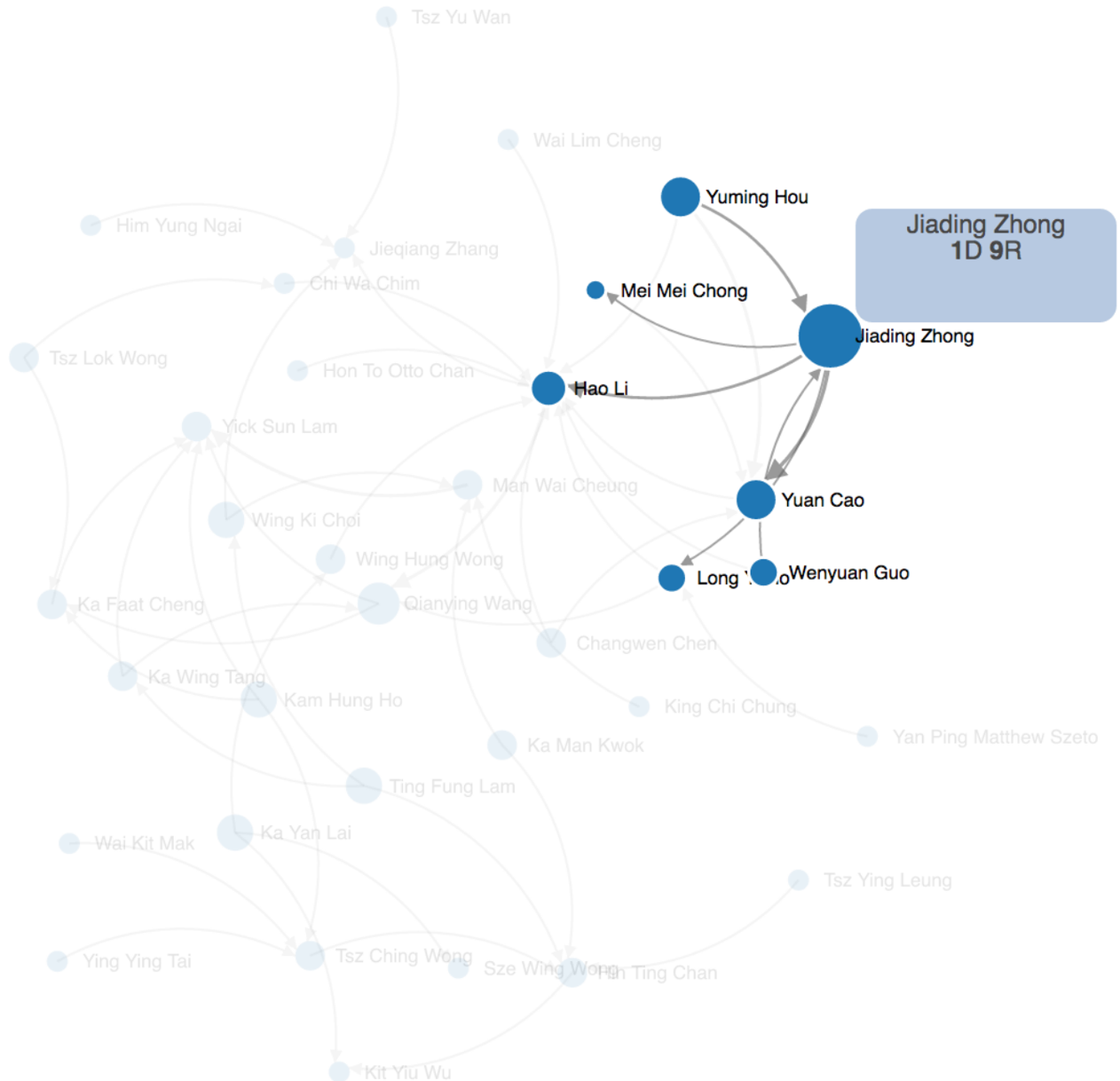
Biggest size of a node identifies most responsive students.

Size and direction of Edge

Thick edge can be understood as a strong relationship between nodes. Also, the arrow of edge demonstrates which node are passively receiving some messages or who actively replying other's discussion.

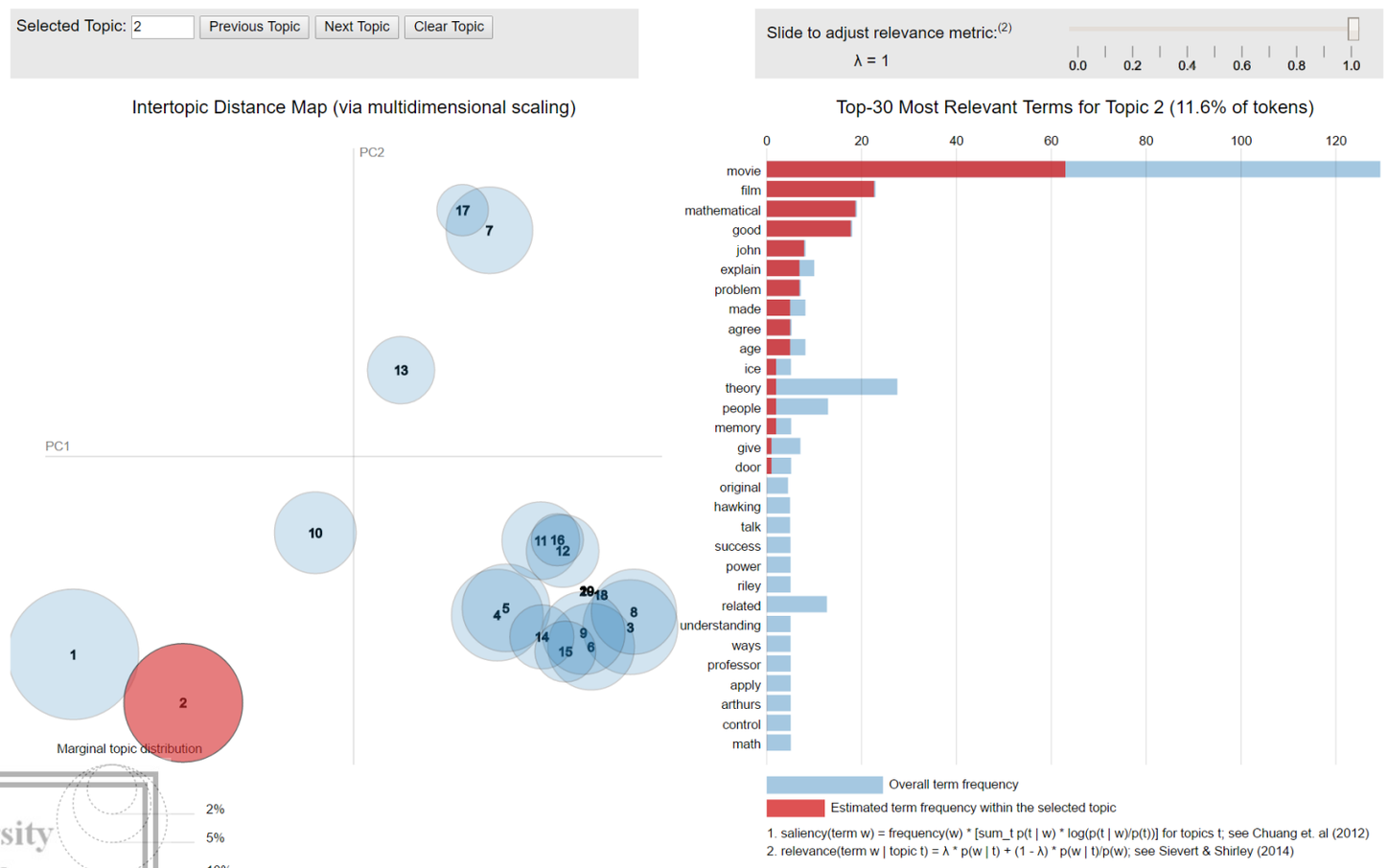
4.1.1 Data Visualization – Forum Graph (Cont'd)

- Scenario 1: teachers want to identify how student perform in discussion and does the students act as a source or receiver in discussion
- The forum graph provides a "hover" function for users
- We can hover on one node and the graph indicates the response of the specific students
- It can illustrate how they think of their study, as **the graph can identify the student's interaction** with different discussion threads and **investigate the performance group of students** by using forum graph.



4.2 Data Visualization – LDAvis

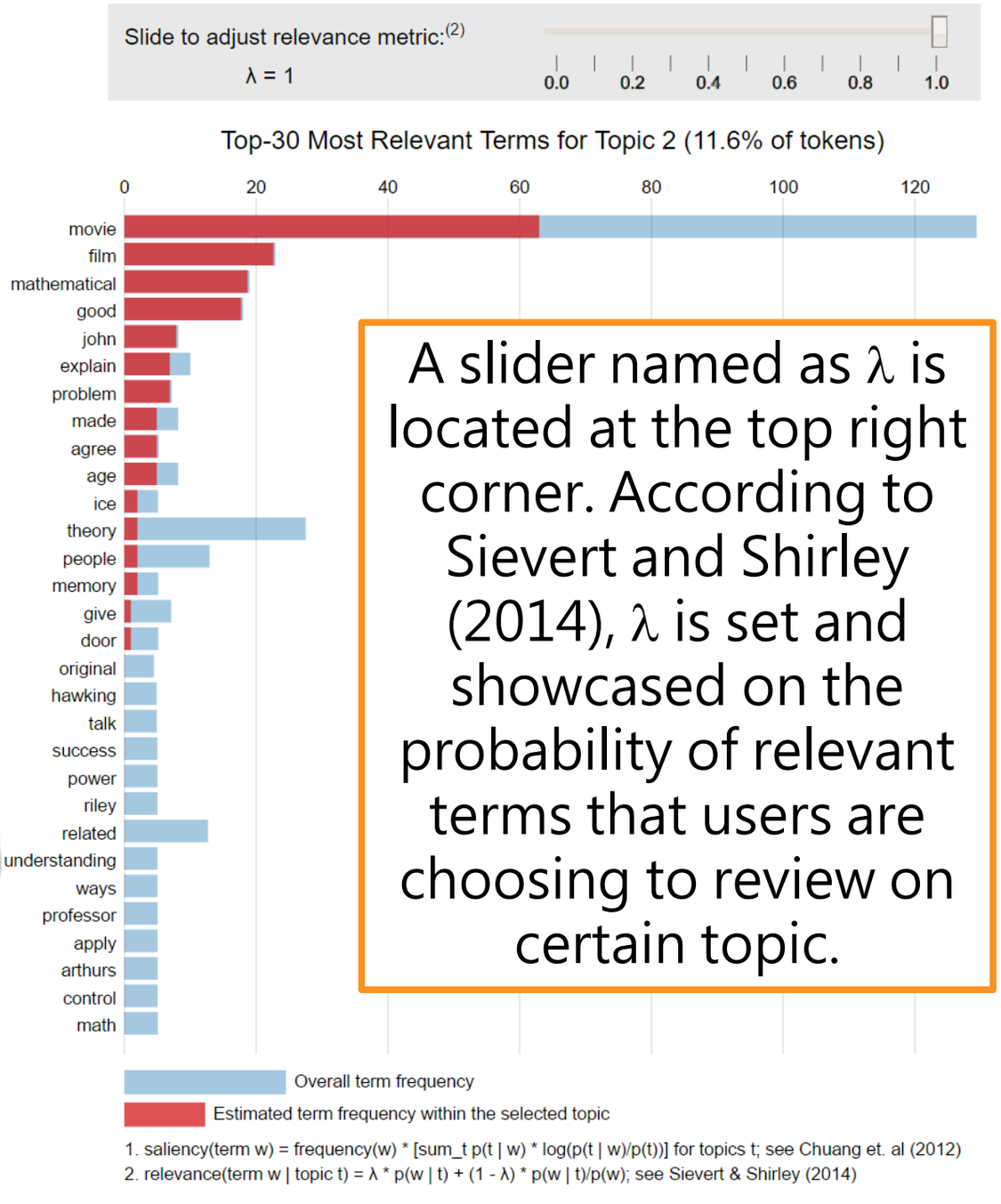
- LDAvis is capable of providing the **key term relevance in fixed size of topic models (TMs)** in which LDAvis sufficiently visualizes the correlation of term among TMs and **provides an interactive platform** for users to select specific terms to reveal its related distribution of TMs
- In our experiment, the topic model parameter (k) has been initiated to 20 while setting up 5,000 iterations of (G) to **execute the likelihood of MCMC (Markov Chain Monte Carlo) algorithm in LDAvis**. The LDAvis graph, which contained the analysis results, was generated as following:



4.2 Data Visualization – LDAvis (Cont'd)

- Basic Concepts of LDAvis:

we have set parameter k to 20, the topic model will contain 20 topics/clusters.



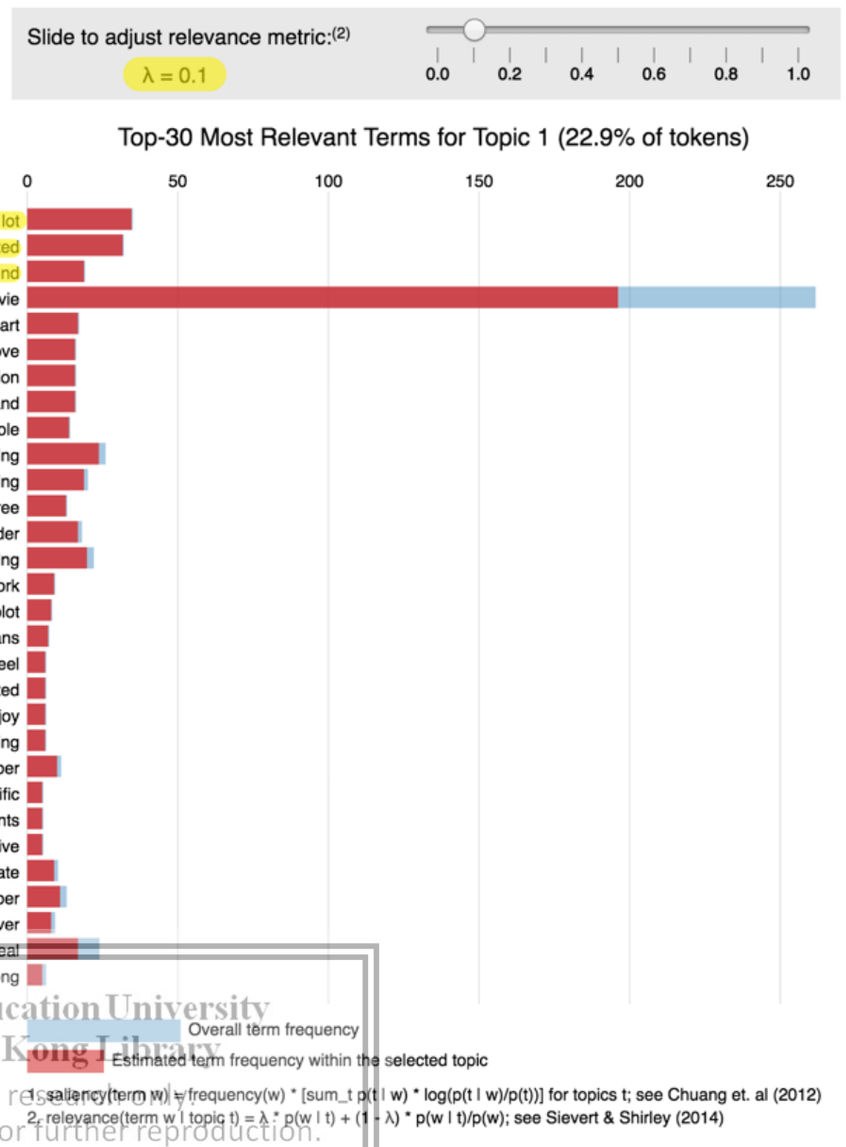
A slider named as λ is located at the top right corner. According to Sievert and Shirley (2014), λ is set and showcased on the probability of relevant terms that users are choosing to review on certain topic.

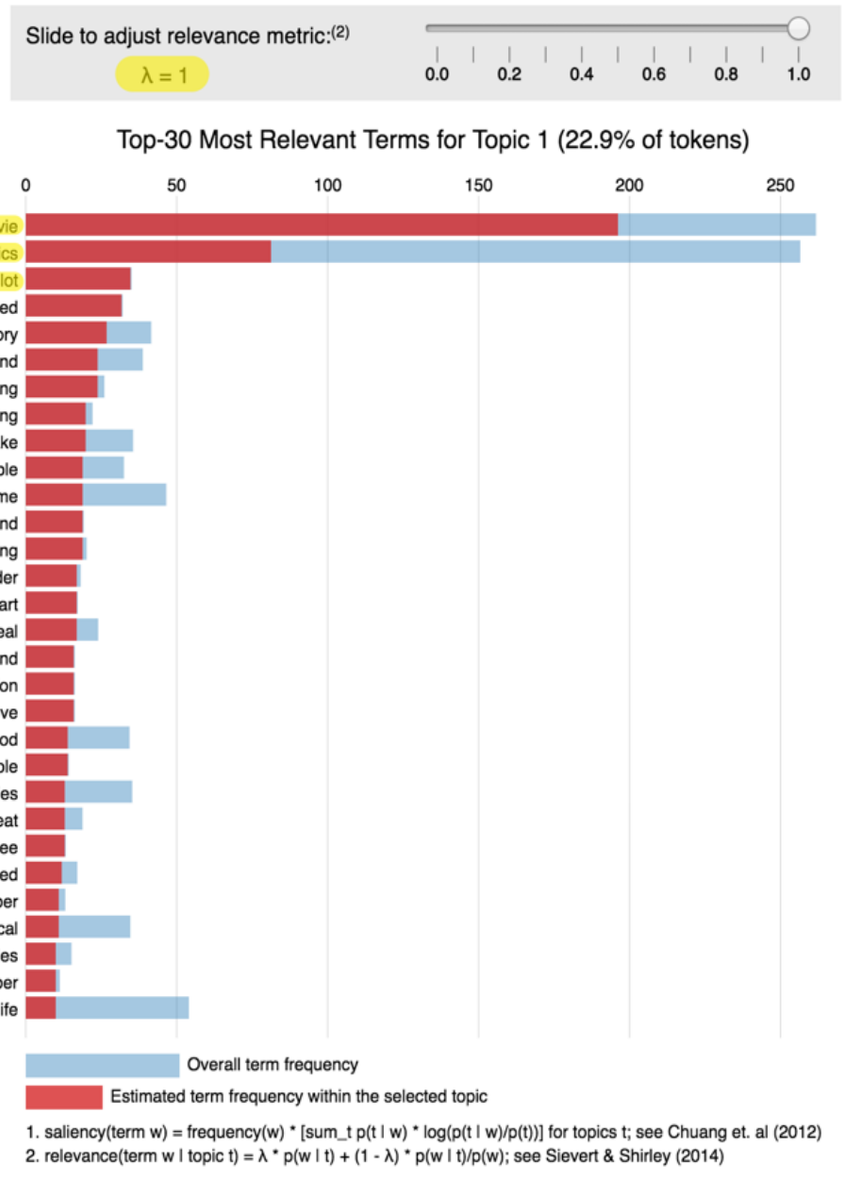
4.2 Data Visualization – LDAvis (Cont'd)

- Scenario 2: teachers want to evaluate the relevant terms in Topic 2
- At the top left corner, we can select "Topic" for reviewing the corresponding topic in detail.
- Also, some specific topics were selected to review by pressing "Previous topic" or "Next topic" button.

4.2 Data Visualization – LDAvis (Cont'd)

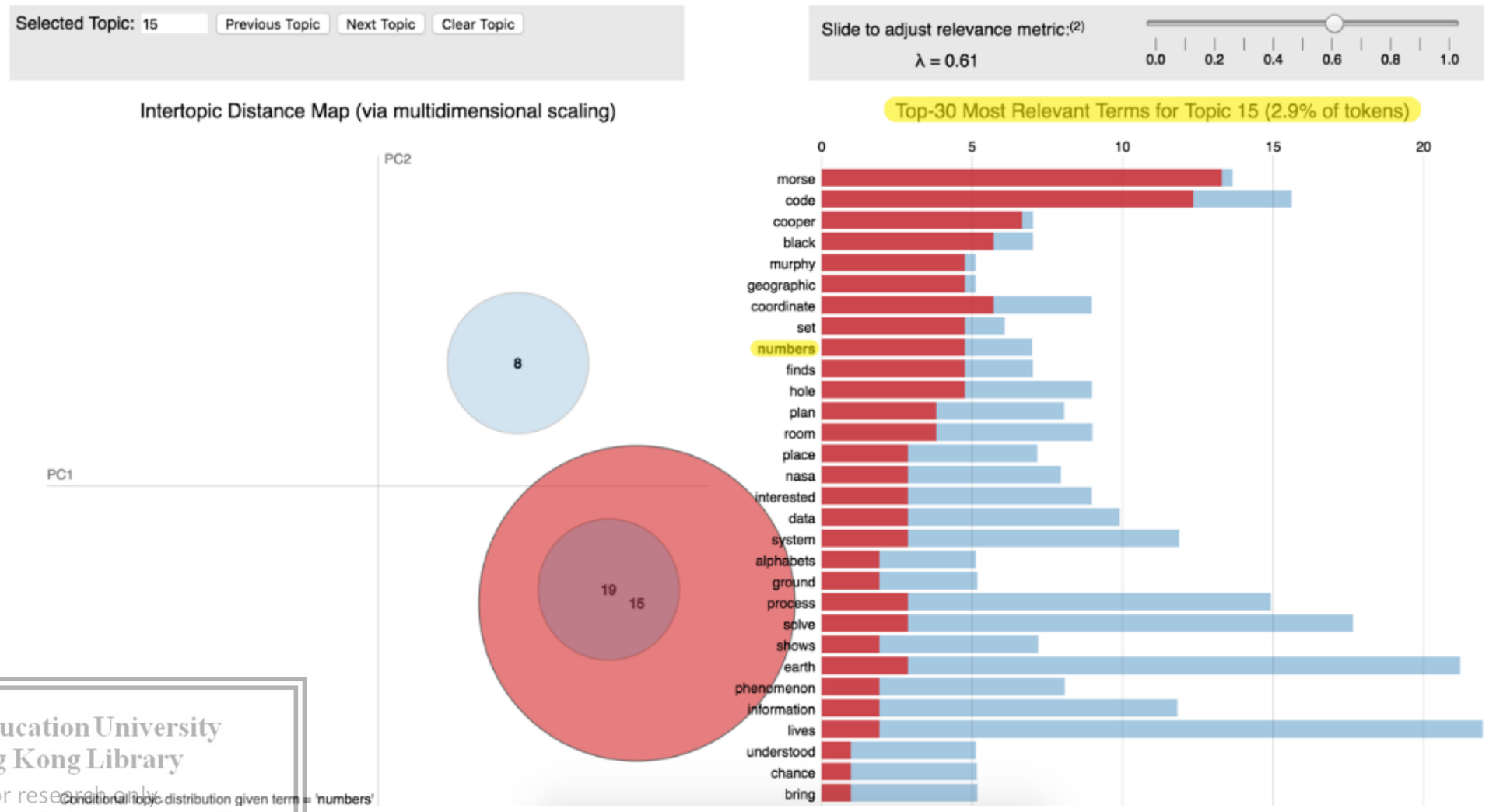
- Scenario 3: teachers want to evaluate the relevance terms in Topic 2, and investigate how students elaborate and think of Topic 2 with correlated terms.
- LDAvis allows setting λ into 0.1 and 1 to review. Simply, λ is acting as a probability of terms relevance. Once, λ is larger and the terms relevance of certain TMs will become more abstract (Sievert and Shirley, 2014).





4.2 Data Visualization – LDAvis (Cont'd)

- Scenario 3: teachers want to know the terms of "numbers" occurred in which topic.
- Hover on the "numbers" and the related topic groups will highlight as red circle
- "Topic 15" and "Topic 19" are not mutually exclusive but they are highly correlated, in which, the terms – "numbers" is distributed in both "Topic 15" and "Topic 19". Besides, "Topic 19" performs as the subset of "To



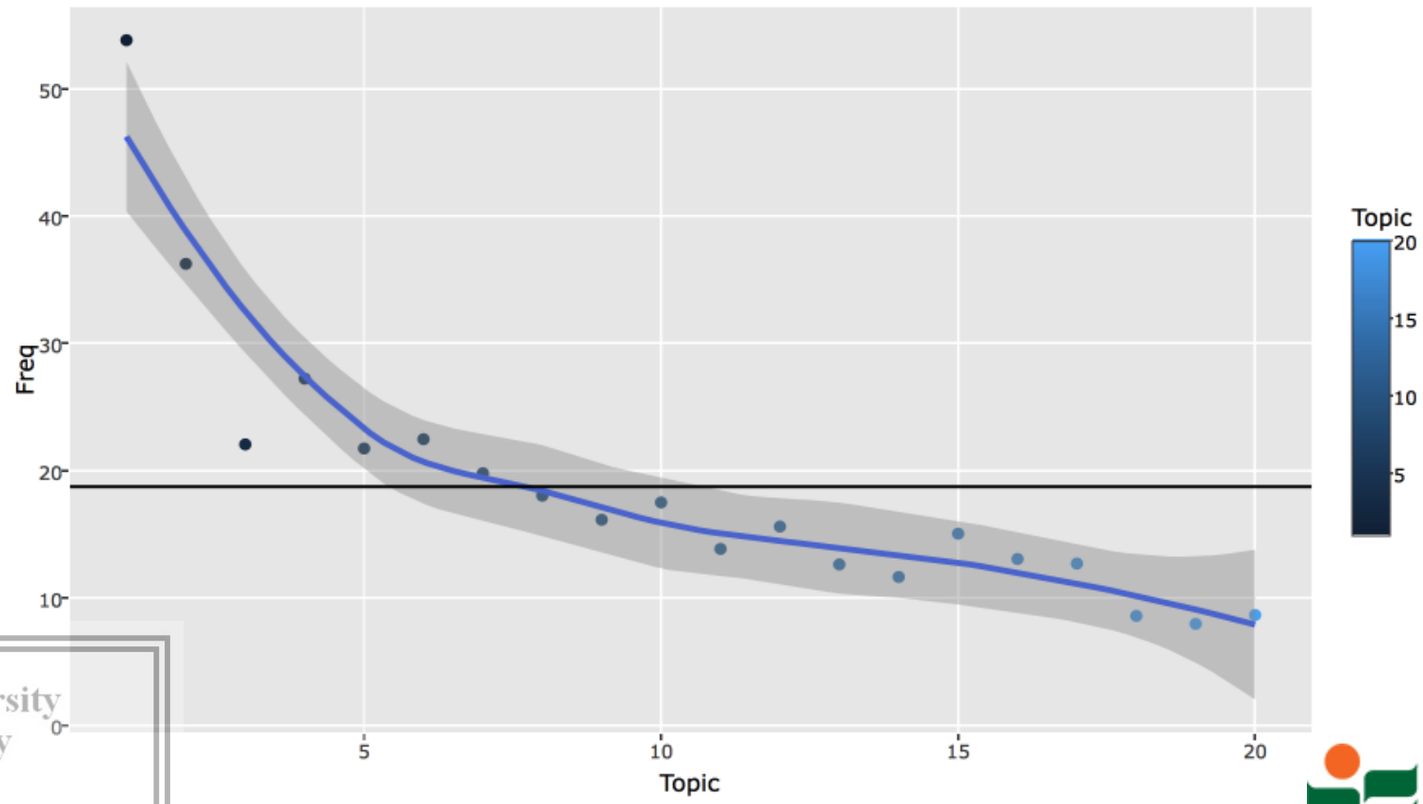
4.2.1 LDAvis Limitations

- LDAvis can visualize the generic relationship of topic groups and their relevant terms. However, the **topic model visualization is an abstract and board concept**, and users might not be able to read a concise analysis report comparing all of them easily all at once.
- In our experiment, 20 clusters of topic groups have been compiled, and users **have to investigate the topic group one by one on LDAvis panels which are not an efficient action to review and compare the topic from one to another**. Also, the relevant terms are determined by a wide variety of keywords. The user can hardly spot the difference in between the topic group at a glance.
- Apart from the interactive topic selection, **LDAvis is difficult to identify the essential topic usage**, for example, we can only notify that there are 20 topic groups without any frequency ranking illustrated in LDAvis. Hence, it is not easy to estimate the prevalence of topics within 20 topic groups.

Image Source: SQLAuthority

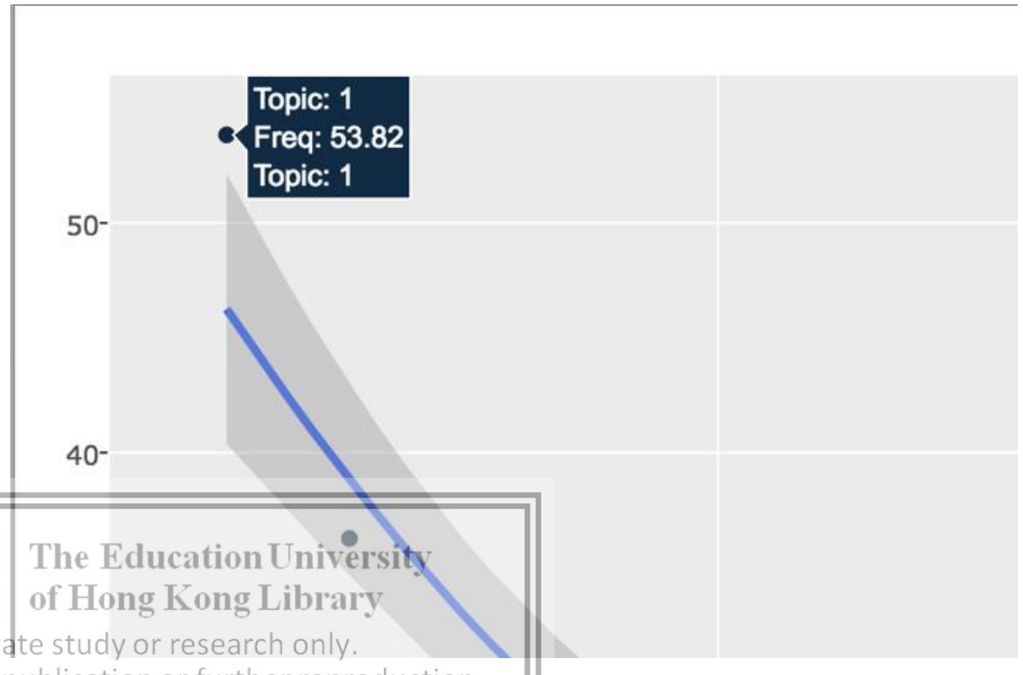
4.3 Supplementary Visualization Means

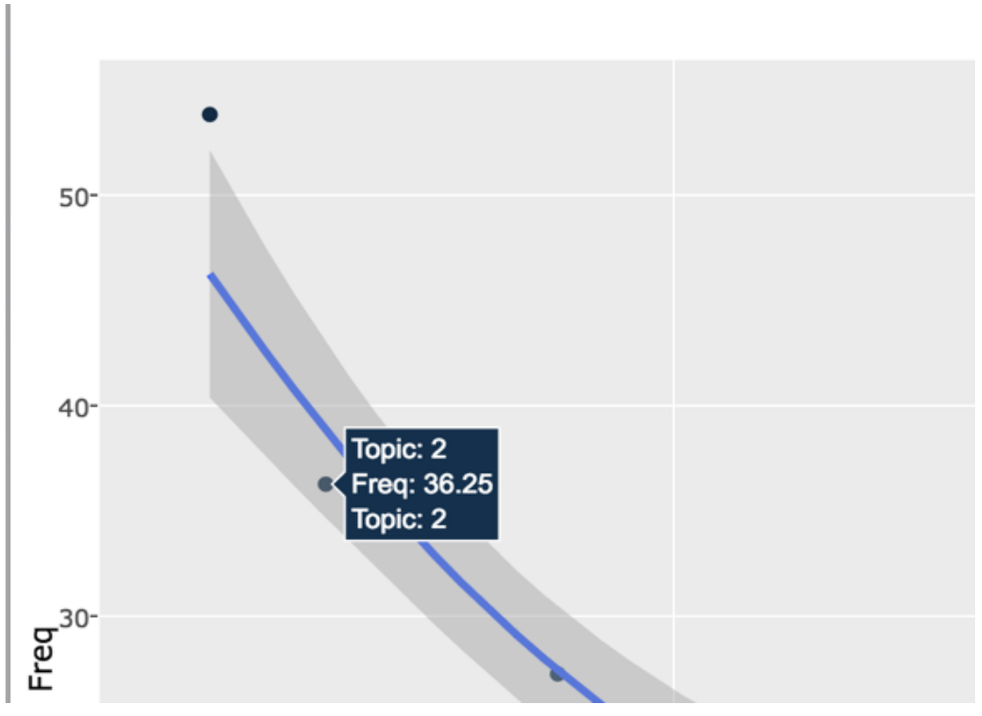
- By using "keyword of term" aggregation in the topic group, the frequency of each "keyword of term" would then be rounded-up in the "topic group" as well.
- Calculating the average of frequency can **differentiate the overall importance of the topic group** to identify which topic group is categorized as the upper level of topic frequency or below average.
- The following graph indicated the frequency of topic group from Topic 1 to Topic 20 with the corresponding accumulated frequencies. Also, topics from 1 to 7 are categorized as the top level of frequency group meanwhile Topic Group 8 to T[illegible] overall average freq[illegible]



4.3 Supplementary Visualization Means (Cont'd)

- **Scenario 4: teachers want to know the dominant topic group and identify which topic group is not popular in the discussion.**
- Line chart provides hover function for each topic group while it indicates the average term frequency in the graph
- The line chart also proves that LDAvis topic distance does not matter to the frequency which means LDAvis only distinguishes the similarity of topic keywords occur in between the topic groups.
- Moreover, the average frequency of line chart can identify which topic group often distributed by the relevant terms that help us to **investigate the importance of the topic and produce corresponding teaching strategy for each topic**.





5. Conclusion

- By using the Forum Graph, academic administrators can perform social network analysis to **understand the interaction among students and teachers** to identify the frequent contributors and passive observers. The Forum Graph can probably help **teachers better understand the participations of their students** in the forum discussion.
- The LDAvis visualization tool also provides a huge potential to **help teachers understand and research the existing and growing topics of discussion and probably discover accentual findings**.
- LDAVis would help teachers spot whether their students can meet regular learning objectives and some **unexpected learning outcomes** by spotting some themes being closely located with themselves or being separately scattered as outliners.
- These can be regarded as some major improvements to our experiment in learning analytics research.

6. Discussion and Further Development

User Acceptance Test (UAT) with Teachers

- We are going to invite serving teachers to test drive the methods we present in this paper by importing their students' works into the text mining and visualization to reveal any potential findings.

- Feedback would be collected to improve and further develop the methods based on field testing, which is the next few stages of our experiment.

Continuous Enhancement

- Another important enhancement in visualization can be network analysis using igraph and other R packages (Katya, 2014). This especially further supplement the LDAvis.

- Using network analysis for the spotted topics periodically (e.g. weekly) may further help spot those changes.

Further Development

- LDAvis do not come with a meaningful name as a topic label.

- It is difficult to understand the topic the relevant keywords are grouped to.

- Therefore, a taxonomy approach will be deployed to help resolve this particular problem so that the relevant keywords can be further classified into a meaningful topic name instead of just using topic numbers.

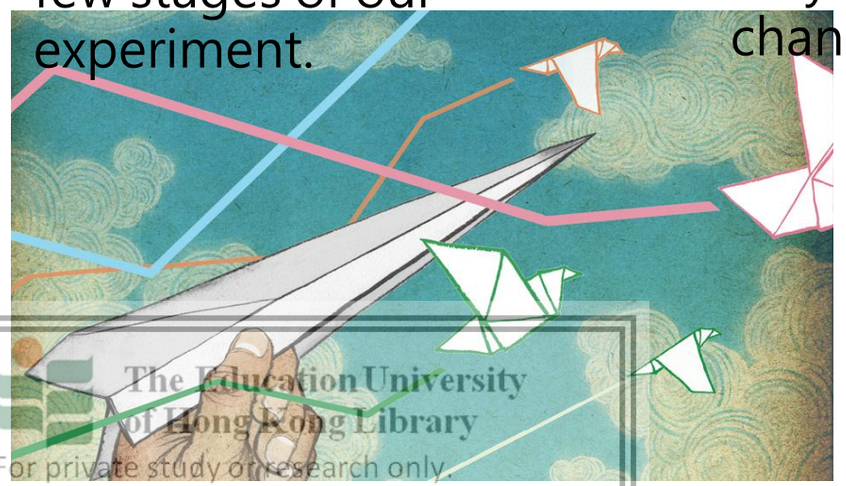


Image Source: Deloitte University Press

End of Presentation – Thank you

Reference

1. Atapattu, T., Falkner, K. and Tarmazdi, H. (2016). *Topic-wise classification of MOOC discussions: A visual analytics approach.* Proceedings of the 9th International conference on Educational Data Mining (EDM), Raleigh, NC, USA
2. Andy, C. 2013, Reports: Forum Graph., Moodle. https://moodle.org/plugins/report_forumgraph
3. Blei, D. M. 2012. Probabilistic Topic Models. Communications of the ACM 55(4):77–84
4. C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy. (2015). Visualizing patterns of student engagement and performance in MOOCs. pages 83–92, New York, New York, USA, 2014. ACM
5. Duval, E. (2011, February). Attention please!: learning analytics for visualization and recommendation. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 9-17). ACM.
6. Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 146-150).
7. KATYA, O., 2014. Network Analysis with R. http://kateto.net/network-visualization
8. Leony, D., Pardo, A., de la Fuente Valentín, L., de Castro, D. S., & Kloos, C. D. (2012, April). GLASS: a learning analytics visualization tool. in proceedings of the 2nd international conference on learning analytics and knowledge (pp. 162-163). ACM.
9. Li, S. & Wong, G. 2016, Educational Data Mining using Chance Discovery from Discussion Board. In Proceedings of GCCCE' 16, The Hong Kong University of Education (pp. 712-715).
10. Ohsawa, Y., Benson, N. E., & Yachida, M., 1998. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on (pp. 12-18). IEEE.
11. Sievert, C., and Shirley, K. E. 2014. LDAvis: A Method for Visualizing and Interpreting Topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces 63–70.
12. Wong, G & Li, S. 2016, Academic Performance Prediction Using Chance Discovery from Online Discussion Forums. In Proceedings of COMPSAC' 16, IEEE (pp. 706-711)