IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ENERGY EFFICIENT WIRELESS COMMUNICATIONS WITH ENERGY HARVESTING AND WIRELESS POWER TRANSFER

Received June 8, 2017, accepted June 16, 2017, date of publication June 21, 2017, date of current version July 17, 2017.

*Digital Object Identifier 10.1109/ACCESS.2017.2717858*

# Modeling and Predicting the Active Video-Viewing Time in a Large-Scale E-Learning System

**TAO XIE[1], QINGHUA ZHENG[1], (Member, IEEE), WEIZHAN ZHANG[1], (Member, IEEE), AND HUAMIN QU[2], (Member, IEEE)**

[1] 1Department of Computer Science and Technology, Xi'an Jiaotong University, Xian 710049, China
[2] 2Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Corresponding author: Tao Xie (247103210@qq.com)

This work was supported in part by the National Science Foundation of China under Grant 61472317, Grant 61428206, Grant 61221063, Grant 61472315, and Grant 91218301, in part by the MOE Innovation Research Team IRT13035, in part by the Coordinator Innovation Project for the Key Lab of Shaanxi Province under Grant 2013SZS05-Z01, in part by the Online Education Research Foundation of the MOE Research Center for Online Education under Grant 2016YB165 and Grant 2016YB169, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2016JM6027 and Grant 2016JM6080, and in part by the Project of China Knowledge Centre for Engineering Science and Technology.

**ABSTRACT** Many studies of the mining of big learning data focus on user access patterns and video-viewing behaviors, while less attention is paid to the active video-viewing time. This paper pinpoints this completely different analysis unit, models the extent to which factors influence it and further predicts when a user permanently leaves a course. The goal is to provide new insights and tutorials regarding data analytics and feature subspace construction to learning analysts, researchers of artificial intelligence in education and data mining communities. To this end, we collect video-viewing data from a large-scale e-learning system and use the Cox proportional hazard function to model the leaving time. The models mainly include the interactions between variables, non-linearity assumption and age segmentation. Finally, we use the collected hazard ratios of model covariates as the learning features and predict which users tend to prematurely and permanently leave a course using efficient machine learning algorithms. The results show that, first the modeling can be used as an efficient feature extraction and selection technology for classification problems and that, second the prediction can effectively identify users' leaving time using only a few variables. Our method is efficient and useful for analyzing massive open online courses.

**INDEX TERMS** Active video-viewing time, modeling and predicting, leaving time, leaving risk.

## I. INTRODUCTION

With the rapid development of the Internet and streaming media technology, the tools used for information dissemination have gradually been expanded from static texts and images to animations, audio and videos. In the last few decades, the research on video viewing has attracted much attention in academia. Most of the research has focused on user access patterns with respect to non-educational videos as well as integration with courses as a kind of educational technology, among others. Nearly no one has focused on the active video-viewing time, which refers to the total time that passes while a user watches a video and only for when the video player is in the playing state, when mining big learning data.

Currently, the video on demand (VoD) system is capable of accommodating video courses and is used in many learning platforms to provide convenient learning services. Video-based learning supports repeated practice and makes thousands of hours of content available on demand [1], which enables students to learn as required. Recent work on the rising Massive Open Online Courses (MOOCs) has offered the possibility of understanding the video-viewing behaviors and cognitive meanings of a user per operation. These works claim that video-viewing behaviors such as play and pause reflect learning states [2]. However, many of these works primarily associate such behaviors with learning performance, and less attention is paid to modeling and predicting the active video-viewing time of users. This topic is very important in the video-based learning. The active video-viewing time is the sum of time when the human brain interacts with the video content. Therefore, it is an effective indicator that characterizes the amount of cognitive engagement [3]. Educators may

gain insights into whether a student learns something and how much he/she has learned. For example, a duration of active watching of ten minutes indicates that a student has consumed more video content than a student who actively watched for only five minutes. Moreover, modeling and predicting the active video-viewing time has significant educational applications. On the one hand, modeling enables us to study when the number of leavers (those who leave a course before the time of interest) or non-leavers (those who leave a course at or after the time of interest) is maximum while taking those who are at risk into account. One could learn which students leave before the time of interest, what causes them to leave and how to prevent it. Moreover, modeling allows interpreting the change in leaving risk due to video viewing as the unit covariate changes, which can help educators to control some conditions to reduce the risk of students leaving a course. On the other hand, the active video-viewing time is also a predictor of academic achievement [4] because a learner is more likely to achieve a higher performance if he/she spends more time viewing the necessary videos [5]. Therefore, lengthening the active video-viewing time is liable to maximize the learning output and thus is a constantly pursued objective for many educators, such as [6] and [7]. To do this, the first and most important step is to properly identify which students are high risk based on their predicted leaving periods. If, during course registration, a student is predicted to prematurely and permanently leave a course, educators can plan to provide him/her with the proper metacognitive tools [6] to help him/her devote more time to that course. Also, an intelligent learning system could assess the teaching quality according to the possibility that a student would leave a course and recommend materials and learning companions [8], as well as offering pop-up, personalized hints [9], [10] to make the system more effective. Besides, the prediction can identify the learning time of students using carefully selected features which are used to analyze the viewing activity of video-based learning system such as MOOCs. Through this research, we hope to provide new insights and tutorials regarding data analytics and feature subspace construction to learning analysts, researchers of artificial intelligence in education and data mining communities.

Thus, this study deals with modeling and predicting the active video-viewing time in an e-learning system. The main difference from other studies is that, on the one hand, we use a completely diverse analysis unit, i.e., the active video-viewing time, while on the other hand, we adopt novel models for analyzing the effects of related variables that are mainly used for efficient feature extraction and selection. This work confused us at the outset, however, until we were inspired by a series of statistical tests. Specifically, this task can be progressively accomplished by answering the following three questions:

1) How can we understand the distribution characteristics of the active video-viewing time? This question requires us to explore the statistical properties and provides a direction for modeling techniques.
2) Which factors significantly influence a user's active video-viewing time? This question requires us to investigate the effects of multivariables on the active video-viewing time using time-to-event models and to develop a process of feature extraction and selection for prediction. The models quantify the change in leaving risk as the unit factor changes.
3) Is it possible to predict the approximate time at which a user will prematurely and permanently leave a course? This question requires us to predict the time at which a user will leave a course, which has important implications for individualized learning since one could provide strategies to increase a user's video-viewing time.

To answer these questions, we use data from a large-scale e-learning system called Skyclass and especially focus on the time during which a player is always in the playing state when a user interacts with it. We primarily use the active video-viewing time to analyze the characteristics of distribution and further use it as a dependent variable in modeling and predicting. In the modeling phase, the event of interest is defined as a user permanently leaving a course without the intention to return, and several time-to-event models are proposed with the method *enter*. Among others, interactions between variables, the non-linearity assumption and age segmentation are included. In the prediction phase, we use the mean as the numerical boundary to distinguish early leaving from late leaving and utilize several efficient machine learning algorithms to determine the hazard ratios of the covariates of statistical significance, which are considered as the learning features. The main contributions are as follows: 1) we present multiple models to show the change in leaving risk due to video viewing as the unit covariate changes; 2) the modeling can be used as a cost-effective feature extraction and selection technology for classification problems; and 3) this prediction mechanism can be used to effectively identify users' leaving times by considering only 4 factors, i.e., age, admission score, video length, and whether users are from the field of medical science.

This paper is organized as follows. Section 2 reviews the related work. Section 3 describes how data is collected and basic statistics. Section 4 discusses the distribution characteristics of the video-viewing time. Section 5 models the leaving time using the method *enter* and analyzes the factors that influence the active video-viewing time. The learning features are established in Section 6, and we predict the time at which a user prematurely and permanently leaves a course. Finally, we draw conclusions and discuss implications in Section 7.

## II. RELATED WORK

Much attention has been paid to video viewing in recent years, with part of the relevant work being presented in the

following studies [11]–[14]. However, most of them mainly considered the user access patterns [15], [16]. For example, the authors in [16] studied YouTube videos and found two types of video access: rarely accessed and frequently accessed. They attempted to describe the daily access patterns of each type and predict the number of accesses in the future. There are also some studies, such as [15], [17], that inferred the quality of a video service by modeling user behavior. However, there are few studies that measured the active video-viewing time. The authors in [18] claimed that video-viewing time is important for system planning, user engagement understanding and system quality evaluation. They measured a popular commercial VoD system called PPLive, characterized the distribution of watching time, and inferred the relation between video-viewing time and video features. Unfortunately, their analytics, which were oriented toward a business system, are probably not suitable for educational purposes.

Video-related research in the field of education is not scarce. Generally, videos are usually used as an educational technology and integrated with courses to improve students' engagement, motivation, and efficacy, among others. In comparison, studies on video-viewing behavior from the data mining point of view constitute a small proportion of the research. With the advancement of streaming technologies in recent years, the video viewing in MOOCs has caught the attention of researchers; some relevant work can be found in [19]–[22]. These studies mainly consider the clickstream behaviors of video viewing and pay little attention to the active video-viewing time, although the time component is sometimes utilized. For example, the authors in [21] defined several clickstream events, such as play, pause, and rate change, and considered the time lengths of events. Their objective was to identify the recurring behaviors of users and evaluate the impact of behaviors on performance. Alternatively, other research has aimed to associate video viewing with a user's engagement and dropout [23], [24]. Nevertheless, none of these studies attempted to model, analyze and interpret the active video-viewing time in educational systems and to effectively predict the early leavers presented in this paper. The main reason may be that the data analysis is usually based on the normality assumption and that the meaning of active video-viewing time is overlooked. Different from that of others, this work uses a completely diverse analysis unit, i.e., the active video-viewing time, which does not support the normality assumption and, in the meantime, has rich pedagogical meanings according to [3] and [4]. Another difference is that the output of our models is a set of risk values of leaving regarding the target variable with reference to the baseline, which are used as the features in machine learning. As far as we know, this is a novel feature extraction (the original variables are projected to new ones) and selection (only variables of statistical significance are selected) technology.

## III. PLATFORM AND DATA

### A. SKYCLASS

Skyclass is a complete distance learning platform deployed in China. Skyclass not only transmits the classroom to remote sites but also records classes, produces courseware, manages educational resources and accommodates videos. Currently, this system is accessed via personal computers and mobile devices and is being applied in 11 districts in China, including Beijing, Zhejiang, Shanghai and Shaanxi. It covers 27 thousand institutions and is utilized by approximately 5.05 million users in total.

The VoD system is the most commonly used learning module of Skyclass and can log the durations of user interaction with the video players. Each log records information about the interaction event type, the time of an event, the video materials and so on. Some detailed statistical properties about this system can be found in the literature [25]. It has been reported that many video-viewing volumes have the characteristics of a power-law distribution, such as the cumulative number of operations of video players versus the number of users and the cumulative viewing times for each course versus the number of users.

### B. ACTIVE VIDEO-VIEWING TIME

We collected 14 million viewing logs from 57,717 unique users following the method of [26]. To answer the three questions, we chose the top 7 courses with the most interactions, which are, in order, Introduction to Mao Zedong Thought (MS), Political Economy (PE), Linear Algebra (LA), Enterprise Financial Management (EF), Marketing (MM), Microcomputer Principle and Interface (MI), and Health Assessment (HA). Each course has several video clips, and each clip contains a complete presentation of knowledge points.

Since our analysis unit is the active video-viewing time, we only consider the total time during which a player is always in the playing state. Let the triple $< E, t, playing >$ be a mathematical expression of video viewing, where $E = \{play, pause, drag, leave\ with\ return, leave\ without\ return\}$ denotes the types of player events, $t$ denotes the server time during which an event $e \in E$ occurs, $playing = \{0, 1\}$ denotes the state of a video player, with 1 indicating the playing state and 0 indicating the non-playing state. An event *play*/*pause* occurs when a user clicks the play/pause button of a video player or when play or pause occurs due to other events, e.g., a *drag* event, which can be viewed as a *pause and play* event and usually occurs when a user slides the frame of a video player to search for specific video segments. A *leave and return* event refers to one in which a user stops watching a video clip for some time but then resumes viewing the video clip when he/she takes the same course in the future; a *leave without return* event refers to one in which a user leaves a course permanently, with all subsequent interactions with the video player occurring when he/she takes a different

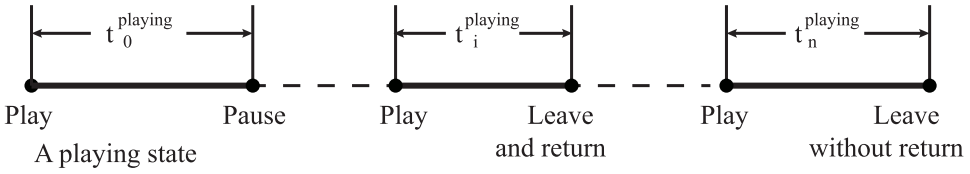



**FIGURE 1.** The illustration of video-viewing time.

course in the future. Then, the playing state is defined as the video viewing that occurs from $t_i$ to $t_{i+1}$, with the event at $t_i$ being *play* and the event at $t_{i+1}$ being $E\backslash\{play\}$, $t_i < t_{i+1}$, where $i$ is the sequence number of viewing. The active video-viewing time is the sum of all time periods during which the player state is 1, denoted as $\sum_{i=0.playing=1}^{n-1} t_i^{playing}$ where $n$ is the sum of video views. This is illustrated in Fig.1. Therefore, the active video-viewing time is an absolute measure representing the length of content consumption. In addition, we consider a relative measure called the viewing completion ratio, which is the proportion of video-viewing time with respect to the total video length and represents the progress of content consumption. We use both to analyze the characteristics of distribution and further use the active video-viewing time, because of its better properties, as the dependent variable in modeling and predicting.

### C. DATA FUSION

At the university, the features of students are usually stored in separate database systems. For example, student profiles are stored in the student management system, and the course grades are stored in the educational administration system. To study how variables affect the target variable, i.e. the active video viewing time, we need to fuse the variables of interest. Several data tables are integrated, including those related to student demographics, teacher information, course information, entrance examinations taken, area of specialty, and academic level. The student demographics comprise student ID, age, gender, region, and so on. The teacher information comprises teacher ID, age, gender, professional title and so on. To fuse the data, we use user identity fields in Skyclass to match other data tables, and then fill the missing values in the target dataset. Most of the students' gender and age fields are missing, but they can be extracted from the students' identity number; also, we obtain the missing course and teacher information through interviews. Finally, we delete the invalid data. The students represent mainly 17 areas of specialty. To achieve a greater integration, we divide the areas of specialty into five categories according to the taxonomy of the professional directory of the Ministry of Education in China [27]: Engineering, Medical Science, Management Science I, Management Science II, and Others. The difference between Management Science I and II is that the former belongs to the field of business management, with the diploma being awarded a management degree, while the latter belongs to the field of engineering management, with students being awarded either an engineering degree or a management degree. Other students not in the four categories are included in the Others category because of the small amount of data. The academic level is divided into high school to junior college (HJC) and junior college to college (JC). The HJC students refer to those who studied in high school before watching the course, while the JC students are those who have statuses as university students and want to improve the type of diploma they receive through a special examination. Correspondingly, for both levels, course authorizations are obtained via an examination. The scores of HJC students represent the sums of several courses, while the scores of JC students represent the results for a single subject. Ultimately, all the scores are standardized to a scale of 100, with 60 being the admission threshold. Additionally, we also consider the video length and video-viewing time. Thus, the age, admission score, video-viewing time, and video length are continuous variables, while the others are nominal variables.

### D. BASIC STATISTICS

The input variables are further determined in Section 5. The final data consists of the information of 7,341 users, and each user is described by nine variables.[1] The minimum age is 16, while the maximum is 58 ($mean = 29.54, SD = 6.59$). The minimum admission score is 60, while the maximum is 100 ($mean = 81.55, SD = 10.17$). The minimum video length is 1.5 minutes, while the maximum is 78 minutes ($mean = 35.59, SD = 12.29$). The minimum viewing time is 0.03 minutes, while the maximum is 60.07 minutes ($mean = 14.10, SD = 10.39$). One-way ANOVA shows that the differences between groups in terms of the continuous variables are statistically significant at the 0.05 level ($df = 6, p = 0.00$).

The proportion of male students is 44.8% and of female students is 56.2%. Four out of seven courses are taught by male teachers. 46.7% of the population are HJC students. The basic statistics across specialties are shown in Table 1. Students of Management Science comprise the greatest proportion, with the number of students of Management I being nearly twice that of Management II. The number of students in the Others category is only 6.63%. Those who have the minimum mean and standard deviation based on gender are from the Medical Science category, while those who have the maximum mean are from the Others category and those

[1] The data is available at https://knoema.com/yyzgruf/video-viewing-time.

**TABLE 1.** Statistics across areas of specialty.

| Specialty | Per.(%) | Age | | Score | | Video-viewing time | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | Mean | SD |
| Engineering | 27.11 | 30.36 | 7.05 | 81.26 | 10.60 | 14.69 | 10.70 |
| Medical science | 15.13 | 26.82 | 5.02 | 81.70 | 9.54 | 12.79 | 9.97 |
| Management (I) | 33.07 | 29.55 | 6.56 | 81.56 | 10.14 | 13.94 | 10.24 |
| Management (II) | 18.35 | 29.81 | 5.58 | 82.52 | 10.31 | 13.70 | 9.70 |
| Others | 6.33 | 32.18 | 6.93 | 81.03 | 9.79 | 16.27 | 11.37 |

**TABLE 2.** Kolmogorov-Smirnov test.

| | Viewing time | | | | Viewing completion ratio | | | |
|---|---|---|---|---|---|---|---|---|
| | skewness | kurtosis | p-value | D | skewness | kurtosis | p-value | D |
| MS | 1.062 | 1.253 | 0.000 | 0.090 | 0.569 | -0.752 | 0.000 | 0.111 |
| PE | 0.791 | 0.541 | 0.000 | 0.080 | 0.472 | -0.871 | 0.000 | 0.089 |
| LA | 0.772 | 0.438 | 0.000 | 0.075 | 0.460 | -0.956 | 0.000 | 0.103 |
| EF | 0.935 | 1.043 | 0.001 | 0.099 | 0.200 | -1.376 | 0.000 | 0.109 |
| MM | 1.039 | 1.441 | 0.000 | 0.093 | -0.056 | -1.445 | 0.000 | 0.101 |
| MI | 1.151 | 2.005 | 0.005 | 0.098 | 1.009 | 0.674 | 0.001 | 0.111 |
| HA | 1.524 | 3.624 | 0.000 | 0.106 | 0.663 | -0.545 | 0.000 | 0.138 |

who have the maximum standard deviation are from the Engineering category. The mean admission score for each specialty is approximately 82 points, with the standard deviation being approximately 10. The maximum mean and deviation of the video-viewing time occur for the Others category, while the minimum mean occurs for the Medical Science category and the minimum standard deviation occurs for the Management II category. One-way ANOVA shows that the differences between groups in terms of the age and video-viewing time have statistical significance at the 0.05 level ($df = 4, p = 0.00$), while the differences between groups in terms of the admission score have no statistical significance ($df = 4, p = 0.08$).

## IV. CHARACTERISTICS OF DISTRIBUTION

Most analysts are much too dependent on the assumption of normal distribution, which leads to the low reliability of their interpretations [28]. To overcome this problem, we first check the normality by performing a Kolmogorov-Smirnov(K-S) test. The null hypothesis is that the video-viewing time and the viewing completion ratio are subject to a specific normal distribution shape. The calculated two-tailed p = 0.000 for all courses, with 95% confidence at the 0.05 significance level. This result contradicts the null hypothesis and indicates that neither video-viewing time nor the viewing completion ratio supports the normal assumption. This result is not consistent with that of a separate study of non-educational videos, where the literature [18] reported that the viewing time of the PPLive system has an approximately normal distribution. In addition, we observe the D statistics of the K-S test. This statistic reflects the maximum distance of the cumulative distribution functions between the empirical and fitted normal distributions. We find that the D statistics are large for each course, as shown in Table 2. The small p-values in Table 2 indicate the distributions have a significant difference with normality,

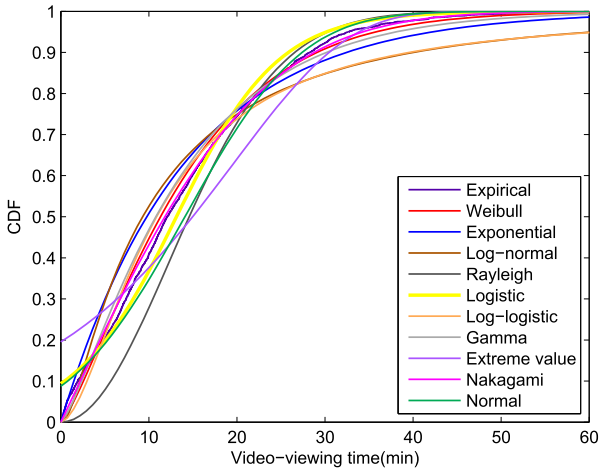



**FIGURE 2.** The CDFs of various distributions.

and most of the p-values end up with 0 means the significance is especially prominent even at the 0.001 level.

Additionally, we observe the kurtosis and skewness. They are showing that the probability density curves of the courses have sharp peaks and right skewness. Also, the absolute value of skewness is larger than 1.96 times the standard error, indicating that there is statistical significance between the skewness and normality. Similarly, the viewing completion ratio is right-skewed for all courses except MM. However, the kurtosis of the viewing completion ratio is flatter than that of video-viewing time. In summary, the normal distribution is not an ideal distribution assumption with respect to the video-viewing time and viewing completion ratio.

Now that there is no evidence for normality, is there a known distribution function that fits the data? After trying a variety of distribution functions shown in Fig. 2, we find that the Weibull distribution function achieves better goodness of fit than others. Take MS for instance; the cumulative

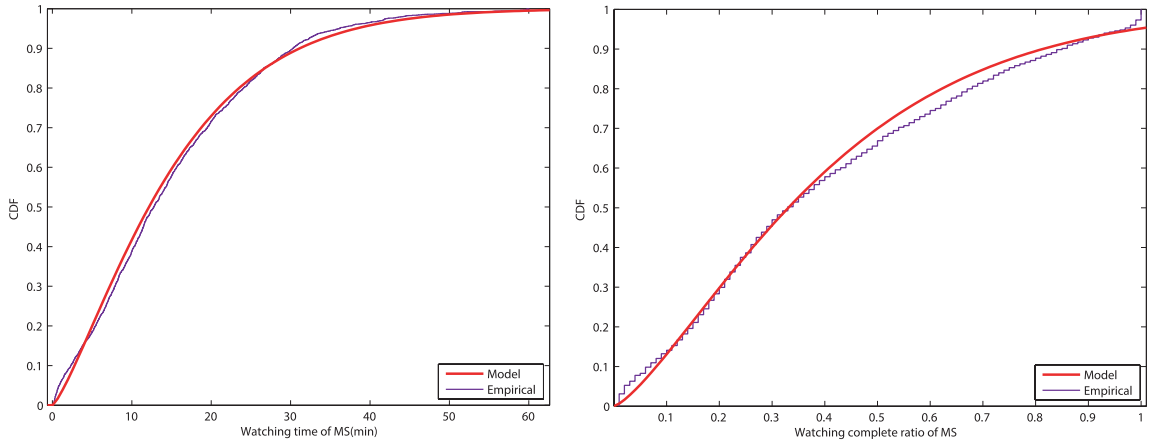



**FIGURE 3.** Cumulative probability density plots of MS with respect to the video-viewing time and viewing complete ratio.

probability plots are shown in Fig. 3. Fig. 3 shows that Weibull fits the video-viewing time better than the viewing completion ratio, as its empirical curve is shown to be closer to the model curve and the D statistic is 0.041. Although the fit of the Weibull distribution function is poor regarding the data related to students that abandoned a video after a few seconds of watching, it is very good regarding the data related to relatively long durations of viewing. This motivates us to model the leaving time with the video-viewing time in Section 5.

## V. MODELING THE LEAVING TIME

In this section, we further examine which factors significantly affect the active video-viewing time to answer the following questions: which students leave before the time of interest, what causes leaving, and how can leaving be prevented? These questions involve the amount of time that passes until an event occurs; thus, it is possible to build a survival model of video viewing. The survival model is also called the time-to-event model and is a widely used statistical technique in many fields. It is used not only to estimate the elapsed time until an event of interest occurs but also to evaluate the relationship between the elapsed time and the explanatory variables [29]. Currently, people are using the model to analyze the occurrence of dropouts [30], participant attrition [28], [31] and other problems. However, this technique has not been applied to video viewing to study the leaving time regarding users' interactions with video players. The model has several advantages with respect to our questions. First, it enables us to study when the number of leavers or non-leavers is maximum while taking those who are at risk into account. Second, it allows handling censored data to make full use of information, while the traditional methods would underestimate the mean value, considering that many users don't leave before a video ends. Last, the Weibull distribution can be used to depict the heavy tail and is one of the most commonly used assumptions in survival analysis.

The event of interest is defined as that in which a user permanently leaves a course without the intention to return. The dependent variable for this study is the elapsed time until a user permanently leaves a course, before it ends, or until a video ends, with the user having not yet left the course. We use a limited set of covariates that include both video-irrelevant (VIR) and video-relevant (VR) parts. The VIR covariates are student age, gender, area of specialty, academic level, admission score, and teacher gender. Although the teacher age, academic degree, and professional title were initially considered, their variations with age, degree, and professorship are small. Therefore, these covariates were not considered in the final set. Student gender is included because it was reported that male students perform better than female students in most cases [32] and that female attrition is larger, especially when measured based on video viewing [31]. Therefore, we expect females to be more likely to leave than males. Admission scores are included because a study found that admission scores are positively correlated with academic performance at the university level [33], [34], while low scores are directly related to student dropout [35]. Therefore, we assume that students with high admission scores are less likely to leave than those with low admission scores. Teacher gender is included because different genders may have different influences on student engagement. Additionally, we also assume that student age, area of specialty, and academic level have impacts on the video-viewing time.

The VR covariates are video category and video length. Note that students watch different videos from different categories for different amounts of time and with classic statistics, indicating that the video category is likely to affect the active video-viewing time. We then commence examining how video length affects video viewing with respect to both the video-viewing time and viewing completion ratio. To do this, we put together all the video segments, arrange them in terms of video length, and divide them into periods of 5 minutes in length. Then, we compute the mean video-viewing time and mean viewing completion ratio for each period. We learn that the relation between the mean video-viewing time and video length can be described by a quadratic polynomial $f(x) = ax^2 + bx + c$, where $a = -0.009$, $b = 0.999$,

**TABLE 3.** Results of applying the fractional polynomial approach to the continuous variables.

| | Video length | | | Age | | | Score | | |
|---|---|---|---|---|---|---|---|---|---|
| | linear | one-term | two-term | linear | one-term | two-term | linear | one-term | two-term |
| -2 log-likelihood | 111572 | 111471 | 111460 | 111572 | 111570 | 111561 | 111572 | 111572 | 111572 |
| p-value | 0.000 | 0.000 | 0.005 | 0.017 | 0.012 | 0.016 | 0.000 | 0.985 | 0.991 |
| powers | 1 | -0.5 | -2, 0 | 1 | -2 | -2, -2 | 1 | 2 | -2, -0.5 |

and $c = -3.635$. Because the coefficient $a$ is small, the parabola rises gently and the mean video-viewing time is linearly correlated with the video length when the length is small. However, when the video length surpasses a certain value, the growth rate of the mean video-viewing time becomes smaller and smaller until finally achieving negative growth. This indicates that increasing the recorded video length appropriately at the beginning can effectively enhance the mean video-viewing time of the population, while the effect gradually disappears and even becomes negative when a video is too long. The relation between the mean viewing completion ratio and video length can be depicted by a linear function $f(x) = bx + c$, where $b = -0.006$ and $c = 0.766$. This indicates that the viewing completion ratio declines linearly as the video length increases when the length exceeds 10 minutes, and every 10-minute increase in video length leads to a decrease in viewing completion ratio by a factor of 0.06. The viewing completion ratio is proportional to the video length once the length is less than 10 minutes. This probably occurs because short videos mainly include the course abstract, assignments and tests. Most students either are willing to spend more time to learn about the course outlines before formal contact is made or pay more attention to the segments that would bring them positive results. This regression analysis shows that the video length does affect the video-viewing time. However, we still do not know the significance of taking multivariable effects into account.

### A. MODELS

We prefer to avoid any distribution assumption in the process of model adoption. Therefore, we use the Cox proportional hazard function to develop $M_1 \sim M_4$ utilizing the *enter* method. It uses the mathematical model to fit the relationship between the survival distribution and the influencing factors, and evaluate the impact of the influencing factors on the distribution of the survival function. The Cox proportional hazard function is a semiparametric regression model with the form

$$h(t, \mathbf{x}, \beta) = h_0(t)e^{\mathbf{x}\beta}. \tag{1}$$

The left-hand side is called the hazard function, which is equivalent to the product of two functions; the baseline function $h_0(t)$ characterizes how the hazard function changes as a function of survival time, and $e^{\mathbf{x}\beta}$ denotes how the hazard function changes as a function of the subject covariates $\mathbf{x}$ [36], [37]. The coefficient $\beta$ is the natural logarithm of the hazard ratio when the corresponding covariate is increased by one unit. Whenever the value is larger than 1, the covariate is associated with an increasing hazard of leaving; when the value is less than 1, the covariate is associated with a decreasing hazard of leaving. The estimated hazard ratio of leaving can be calculated by exponentiating the coefficient.

The developed models are as follows:

1) $M_1$: A model that includes the VIR covariates only.
2) $M_2$: A model that includes both the VR and VIR covariates.
3) $M_3$: A model that adds interaction terms between covariates based on $M_2$.
4) $M_3{}^*$: A variant of $M_3$ that considers the age segmentation.
5) $M_4$: A model that takes the non-linearity assumption of continuous variables into account. This is done because the log hazard ratios of continuous variables may not vary linearly and the test is usually neglected in practice. To do this, we use the fractional polynomial approach [38], which decomposes a continuous covariate into a combination of several power functions. When the linearity assumption is not satisfied, we expect to achieve a better fitting through transformation. Assume that the log hazard function is

$$ln\left(h\left(t\right)\right) = ln\left(h_0\left(t\right)\right) + \sum_{i=1}^{J} F_j(x)\beta_j, \tag{2}$$

where

$$F_j(x) = \begin{cases} x^p, if\, p_j \neq p_{j-1} \\ F_{j-1}(x)\ln(x), otherwise \end{cases} \tag{3}$$

and $p \in \wp$, $\wp$ is a possible exponent collection.

We apply the fractional polynomial approach to the three continuous variables and obtain the statistics shown in Table 3. Table 3 shows that the video length and age are statistically significant in the case of linear, one-term, and two-term power functions. Therefore, we choose the model with the largest log partial likelihood. We use the generated powers to reconstruct the variables and use the new variables as the inputs of models. The admission score is significant only for the linear function, indicating that this covariate satisfies the linearity assumption. According to the formula, $M_4$ decomposes the original *video length* item into an item $F_1\left(videolength\right) = \left(videolength\right)^{-2}$ and an item $F_2\left(videolength\right) = \ln\left(videolength\right)$ and decomposes the original *age* item into an item $F_1(age) = (age)^{-2}$ and an item $F_2(age) = (age)^{-2}ln(age)$. To avoid the incidental numerical disaster [37], one may scale down the numerical variable of interest.

**TABLE 4.** Statistics of interest.

| Variable | $M_1$ | | $M_2$ | | $M_3$ | | $M_4$ | |
|---|---|---|---|---|---|---|---|---|
| | $\beta$ | p-value | $\beta$ | p-value | $\beta$ | p-value | $\beta$ | p-value |
| Age | | | | | -0.023 | 0.000 | | |
| Student(male) | 0.096 | 0.001 | 0.098 | 0.001 | 0.290 | 0.000 | | |
| Teacher(male) | -0.056 | 0.041 | | | | | | |
| Medical Science | 0.263 | 0.000 | 0.212 | 0.000 | 0.183 | 0.000 | 0.097 | 0.001 |
| Management(I) | 0.089 | 0.006 | | | | | | |
| Management (II) | 0.104 | 0.029 | | | | | | |
| JC | | | | | -0.250 | 0.057 | | |
| Admission score | -0.003 | 0.014 | -0.005 | 0.000 | -0.005 | 0.000 | -0.004 | 0.000 |
| Video length | | | -0.025 | 0.000 | -0.025 | 0.000 | | |
| Age*level | | | | | 0.015 | 0.000 | | |
| Gender*level | | | | | -0.131 | 0.010 | | |
| Age_2 | | | | | | | -0.828 | 0.006 |
| Video length_1 | | | | | | | 0.018 | 0.000 |
| Video length_2 | | | | | | | -0.792 | 0.000 |

Notes: the units that are not statistically significant are left blank or omitted; the combination of an underline and a number in the variable column indicates the part of the transformation due to application of the fractional polynomial approach.

### B. RESULTS

The results are shown in Table 4. The interpretation of the fitted Cox's proportional hazard models requires drawing inferences from the estimated coefficients in the corresponding models. The estimated coefficient for a covariate represents the rate of change of the dependent variable as the unit covariate changes. The extent to which the covariates influence the leaving risk can be calculated by exponentiating the estimated coefficients. When only the VIR covariates are considered, i.e., $M_1$, student age is not significant, while teacher gender is significant. When the VR covariates are added, as in $M_2$, the impact of teacher gender becomes insignificant. When the interactions between variables are considered, as in $M_3$, student age starts to become a predictor of leaving, and the academic level is significant at 0.1. The course categories are not significant in the models, and only the video-viewing time of the Medical Science students has statistical significance compared to that of the Engineering students. Take $M_2$ for example; the estimated hazard ratio of Medical Science students is $e^{0.212} = 1.236$, indicating that they have a 23.6% higher rate of leaving a video compared with Engineering students. When the interaction is considered, the growth rate reduces by 20.1%. Furthermore, the running result with respect to student gender is inconsistent with the initial expectation, i.e., the leaving risk of males is higher than that of females in the models. The admission scores always significantly influence the video-viewing time of a population. As is expected, a higher score leads to a lower leaving risk. For clarity, we explain each of the models as follows.

In $M_1$, the statistically significant covariates are student gender, teacher gender, admission score and whether students are from Medical Science or Management Science. Using these covariates as the predictors, male students have a 10.08% higher leaving risk than their female counterparts. A male instructor yields a leaving risk that is 5.45% lower than that of a female instructor. Students from Medical Science, Management I, and Management II have a 30.8%, 9.31%, and 10.96% higher leaving risk, respectively, than Engineering students. Every 10-point increase in admission score tends to reduce the leaving risk by 2.96%.

In $M_2$, the number of predictors decreases relative to $M_1$. The main reason is that increasing the video length dilutes the influence of some VIR covariates. Male students have a leaving risk of more than 10.3% compared with female students. Only students from Medical Science have statistical significance, with 23.6% higher leaving risk compared to Engineering students. Every 10-point increase in admission score tends to reduce the leaving risk by 4.88%. Every 1-minute increase in video length reduces the leaving risk by 2.47%, while every 5-minute increase reduces the leaving risk by 11.75%.

$M_3$ reveals significant interactions between variables. The number of predictors increases to 8. Student age becomes a significant variable influencing the leaving risk for the first time, and the leaving risk of a student reduces by 2.27% for every 1 year older that student is. From a larger age span perspective, for every 5 years older a student is, the leaving risk reduces by as much as 10.86%. Male students have a 33.64% higher leaving risk compared with female students. The leaving risk of students from Medical Science is 20.08% higher than that of those from Engineering. The leaving risk of JC students is 22.12% lower than that of HJC students at the 0.1 significance level. The influences of admission score and video length are invariant compared to those in $M_2$. Student age and gender interact with the academic level. The plots of the estimated log hazard are shown in Fig. 4. Both groups of lines show a departure from being parallel, and each of them is statistically significant. The JC students leave at a higher rate than the HJC students, and the difference in leaving risk between them becomes larger as age increases. When considering the interaction between gender and

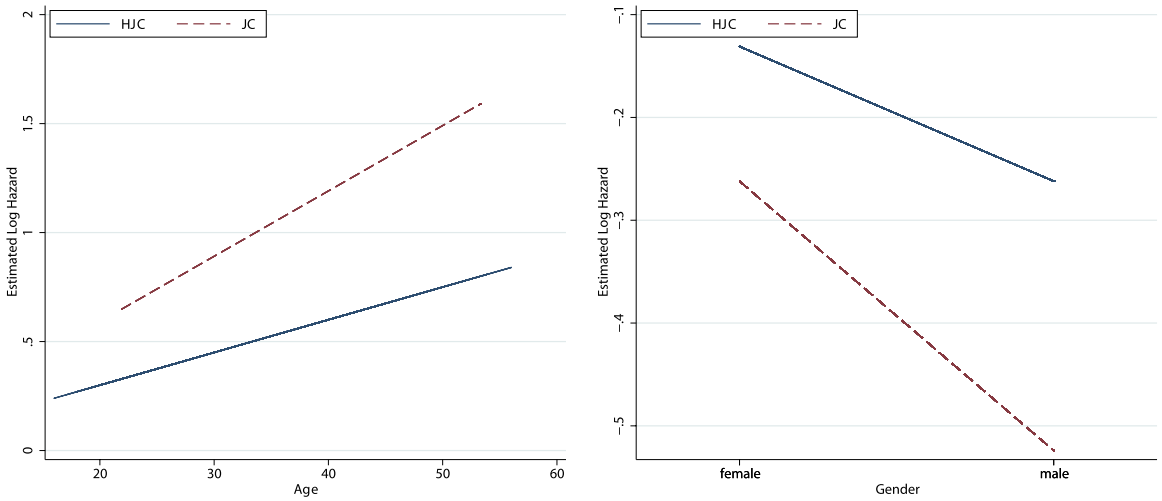



**FIGURE 4.** Plots of considering interaction items. They are showing a departure from being parallel.

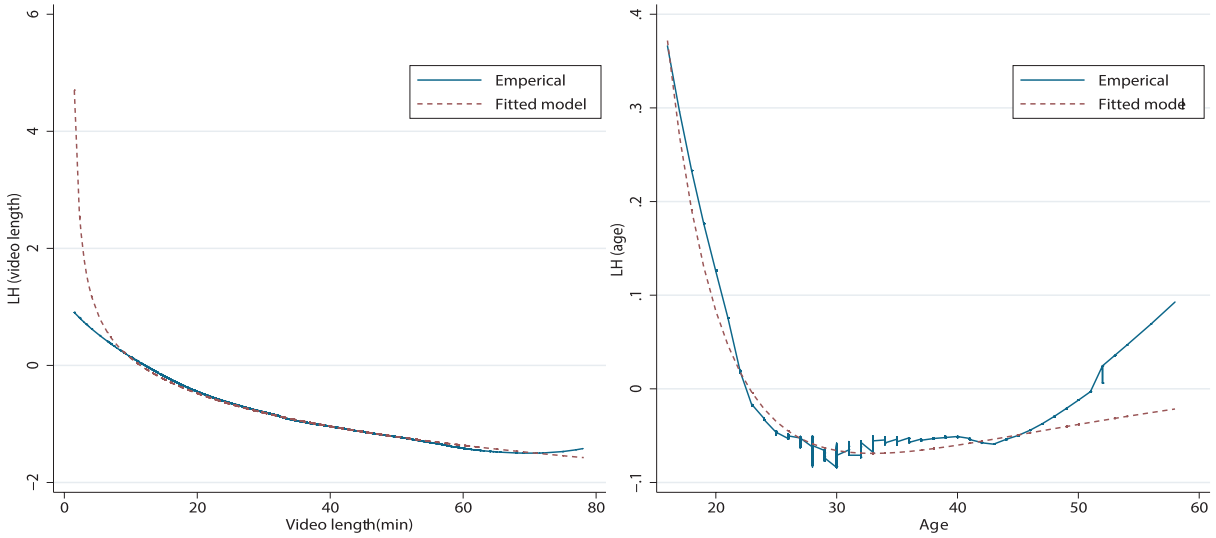



**FIGURE 5.** The log hazard ratio of video length and age.

academic level, the leaving risk of HJC students is higher than that of JC students. At either academic level, male students have a lower risk of leaving compared to female students. For both levels, the difference in male leaving risk is larger than that of female leaving risk. In other words, JC male students have the lowest leaving risk, while HJC female students have the highest.

$M_4$ generates new variables by transforming the original non-linear variables. For the video length variable $x$, the new variable is computed by $f(x) = 0.018x^{-2} - 0.792\ln(x)$; for the age variable $y$, the new variable is computed by $f(y) = -0.828y^{-2}\ln(y)$. To determine how the new and old variables affect the leaving risk, we use a ratio estimator [39] to infer the log hazard functions of the original variables. The plots of estimated log hazard are shown in Fig. 5. The fractional polynomial model fits the log hazard function of variables well, except for video lengths less than 10 minutes and ages greater than 45 years. The reason why this phenomenon occurs may be the sparse sample size at both ends. In the main body of the curve, the log hazard of video length monotonically decreases as the length increases. However, the estimated log hazard declines rapidly when the length is less than 18 minutes, while it undergoes a gentle decline when the length is larger than 18 minutes. The log hazard of age presents a left-inclined and fat L-shaped curve as the age increases, and the minimum risk is obtained for an age of approximately 33 years. The log hazard ratio decreases rapidly before the age of 33 and increases slowly after the age of 33. This indicates that the influence of video length on the leaving risk in $M_4$ is basically consistent with that of other models; the difference is the magnitude of risk reduction, which becomes smaller rather than remains constant per unit increase in length. The interpretation of the

**TABLE 5. Hazard ratios of age groups.**

| Variable | age $\leq$ 26 ($n = 3065$) | 27 $\leq$ age $\leq$ 33 ($n = 2432$) | age $\geq$ 34 ($n = 1845$) |
|---|---|---|---|
| Age | $0.907^{*}$ | 1.018 | 0.988 |
| Student(male) | $1.514^{**}$ | $1.761^{***}$ | 0.779 |
| Medical science | $1.271^{***}$ | 1.067 | $1.446^{**}$ |
| Management(II) | 1.014 | $0.836^{*}$ | 1.128 |
| JC | $0.391^{\diamond}$ | 1.617 | $0.427^{\diamond}$ |
| Admission score | $0.995^{*}$ | $0.996^{\diamond}$ | $0.995^{\diamond}$ |
| Video length | $0.971^{***}$ | $0.976^{***}$ | $0.981^{***}$ |
| age*level | $1.049^{*}$ | 1.001 | 1.014 |
| gender*level | $0.837^{*}$ | $0.723^{***}$ | $1.258^{*}$ |

Notes: $^{\diamond}p < 0.1$, $^{*}p < 0.05$, $^{***}p < 0.001$.

age effect must be conducted in segments: from the minimum age to approximately 26 years old, the hazard ratio decreases exponentially as the age increases; from 26 to 33 years old, the hazard ratio decreases slowly to a minimum; and above 33 years old, the value increases slightly. Considering that the newly established variables in $M_4$ change the meanings of the original variables, we apply the age-segmented hazard model to the covariates of $M_3$; the pros and cons of using the age-segmented model and the non-age-segmented model are presented in the next section. The estimated hazard values are shown in Table 5.

The effect of age is statistically significant only below 26 years old, and the leaving risk decreases by 9.3% for every 1-year increase in age, which is much bigger than the group of mixed ages (2.27%). The leaving risk of male students in the medium age-group is higher than that of those in the low age-group. However, the leaving risk of male students in the high age-group is not statistically significant among groups. The leaving risk of Management II students in the medium age-group is lower than that of Engineering students, while other age groups have significantly increased risk compared to Engineering students. The admission score is statistically significant in the low age-group and not very significant in the high age-groups. However, the effect of video length on the leaving risk is significantly different for all age groups, where the leaving risk declines as the video length increases. The interaction between age and academic level is only significant in the low age-group. The interaction between gender and academic level demonstrates two trends: the estimated log hazard plot for the medium-low age-group is a straight line with a negative slope; while that for the high age-group has a positive slope.

### C. DISCUSSION

With the addition and decomposition of variables, the models present different explanatory abilities. From the application perspective, $M_3$ and $M_3{}^{*}$ are preferable because the former exhibits interactions between variables, while the latter takes the statistical differences in age segmentation into account; from the perspective of analysis, $M_4$ removes unimportant components of variables and preserves the important components through fractional polynomial transformations, which causes the information to not be completely discarded. Regarding the experimental results, some violate our initial assumptions. For example, we would have expected a significant interaction between student gender and teacher gender, but teacher gender has no significant effect on the dependent variable for $M_2$. This phenomenon occurs because an important variable, i.e., video length is added. Another violation is that male students are more likely to leave a video course than female students. The idea is that personalized assistance of the system needs to create a balance between gender differences and other influencing variables. The main limitation of modeling is that we do not directly consider the effect of difficulty of knowledge on the active video-viewing time. Instead, we use the categories of courses and assume that the students learning each category will not experience significant differences in difficulty. This is a simplification of the problem because the degree of difficulty of knowledge varies across individuals, and experts' definitions of difficulty do not always cover all students.

## VI. PREDICTING THE LEAVING TIME

Note that a user is more likely to leave a course prematurely if he/she has a higher leaving risk. We have obtained a subspace of the original feature set by statistically testing the models and calculated the hazard ratios of covariates by exponentiating their coefficients. We consider two cases: one involves measuring how many unit differences exist between the variable value of interest and the reference value in the case of continuous variables; the other involves examining the statistical differences of the category of interest with respect to the reference category in the case of nominal variables. To this end, different students have different combinations of hazard ratios. The goal of this section is to predict, at the time of registration, whether a student would prematurely and permanently leave a course by using the hazard ratios of covariates of statistical significance as the learning features. The extracted features are a set of leaving risks which have values around 1. Whenever the value is larger than 1, it is more likely to be classified into the early-leaving category compared to the reference level; and vice versa. Doing this has very important implications. First, it separates students into groups and provides criteria for who should be

particularly considered based on a student's active video-viewing time. This occurs because the active video-viewing time is an indicator of academic achievement [4], and one is more likely to achieve higher performance if he/she spends more time viewing the necessary videos [5]. Second, once we achieve this target, we can provide students predicted to leave at different times with different metacognitive tools to help them devote more time to courses [6] or make the system more effective by having it recommend materials and companions and send pop-up, personalized hints. Furthermore, this problem is also important because if the prediction performance of a model (especially $M_3$ or $M_4$) is superior to that of others, it presents a more extensive value than the other models. However, predicting the exact time at which a student will leave a course is very difficult or even impossible, although it is possible to predict the approximate time of leaving. To do this, we first need to determine the numerical boundary to distinguish early leaving from late leaving. In this study, we use the mean as the partition criteria and define a person as an early leaver if his/her active video-viewing time does not reach the mean value and as a late leaver if the opposite is true. Because of this, the research becomes a classification problem of binary data in machine learning.

Given a set $\Omega$ of $n$ training cases $(x, y) \in X^l \times Y$, where $x$ is a vector denoting the leaving risks of covariates and $y \in \{0, 1\}$ represents the category value, the goal is to predict the value of a class label for a test case $q$ described by the same variable space $X$ as that of the training data. The training set $\Omega_E$ and test set $\Omega_T$ are always chosen such that $\Omega_E \cap \Omega_T = \emptyset$ and $\Omega_E \cup \Omega_T = \Omega$. According to the definition, we have

$$y_i = \begin{cases} 0, \, if \sum_{j=0}^{m} t_{ij} \leq mean \\ 1, \, otherwise, \end{cases} \tag{4}$$

where $y_i$ represents the $i'th$ student's class label in $\Omega_E$, $t_{ij}$ is the time of the $j'th$ viewing of the $i'th$ student, and $m$ is the number of video views after registering for a course.

### A. ALGORITHMS

SVM classifier. We choose this because it creates a complex decision boundary by using different kernel functions. This algorithm was adopted by Brinton et al. to predict MOOC performance, and they acquired good results when they used the probabilities as the learning features [22]. Therefore, we study whether this algorithm could also produce a good prediction effect when we use the hazard ratios of model covariates as the corresponding features. Then, we solve the following optimization:

$$\underset{\mathbf{w},\varepsilon}{minimize}\left(\frac{1}{2}||\mathbf{w}|| + \frac{C}{l}\sum_{i=1}^{l}\varepsilon_i\right), \tag{5}$$

*s.t.* $y_i(x_i \cdot w + b) \geq 1 - \varepsilon_i$, where $\varepsilon_i \geq 0, \forall i \in \Omega_T$. We use the Gaussian kernel because of its high popularity and flexibility in mapping within a multidimensional space [40].

$k*$ classifier. This is an instance-based algorithm that classifies an instance by comparing it to a set of pre-classified examples. A main task of instance-based learning is to determine the distance between two instances. We choose $k*$ because it uses the information entropy as a distance measure and always performs well against a range of both rule-based and instance-based learners [41]. Assume that $a$ and $b$ are two instances; the program transforms $a$ into $b$ and forms a finite sequence of transformation starting at $a$ and terminating at $b$. $k*$ computes the distance sums over all possible transformations between them and estimates the probability function $P*$ that traverses all paths from $a$ to $b$. This can be formulized as

$$\begin{cases} k*(a|b) = -\log_2 P*(b|a) \\ P*(b|a) = \sum_{\bar{t} \in P:\bar{t}(a)=b} p(\bar{t}) \end{cases}, \tag{6}$$

where $P$ is the set of all prefix codes from transformations and $\bar{t}$ denotes the members of transformations.

Bootstrap aggregating (Bagging). This generates multiple versions of a classifier and uses them to get an aggregated predictor that reduces the variance associated with the prediction [42]. Given a classifier $\varphi(x, \wp)$ and a learning set $\{\wp_k\}$ each consisting of $N$ independent observations, the goal is to replace $\varphi(x, \wp)$ with the average of $\varphi(x, \wp_k)$ by taking repeated bootstrap samples [43]. Here, we use a randomly generated decision tree as the classifier.

Random forest (RF). A random forest is a classifier consisting of a collection of tree-structured classifiers $\varphi(x, \wp_k)$, where $\{\wp_k\}$ are independent and identically-distributed random vectors and each tree casts a unit vote for the most popular class at input $\boldsymbol{x}$. We train multiple decision trees by randomly sampling the training data and randomly selecting features and the combine results of models by taking a majority vote.

The algorithms Bagging and RF are both ensemble learners. We choose them because each works by running a base learning algorithm multiple times and forming a vote based on the resulting hypotheses [44]. They have been reported to improve the prediction ability well, e.g., [45].

### B. PROCEDURE

Method. Our goal is to determine, relative to the original feature set, which classifiers' prediction performances are improved and which models can be further used for prediction when we use the hazard ratios of model covariates as the learning features. Therefore, we choose the original dataset $M_0$ as the benchmark and apply algorithms to the determined feature sets of models 1-4 for comparison. In addition, we apply algorithms to the feature set of the age-segmented model and show the pros and cons with respect to prediction. Relative to the alias $M_3$, this dataset is named after $M_3{}^*$.

Metrics. Let $TP$, $FP$, $TN$, and $FN$ be the true positive, false positive, true negative and false negative, respectively, that are obtained from a classifier. The first metric we choose is the accuracy, which represents the percentage of correctly classified samples with respect to the total, computed as

**TABLE 6.** Evaluating based on accuracy.

| | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_3{}^*$ | $M_4$ |
|---|---|---|---|---|---|---|
| SVM | **0.758** | 0.52 | 0.553 | 0.542 | 0.55 | 0.525 |
| k* | 0.736 | 0.565 | 0.617 | 0.766 | **0.771** | **0.782** |
| Bagging | 0.747 | 0.563 | **0.825** | 0.815 | 0.815 | **0.841** |
| RF | 0.796 | 0.559 | **0.841** | 0.829 | 0.834 | **0.854** |

**TABLE 7.** Evaluating based on F-score.

| | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_3{}^*$ | $M_4$ |
|---|---|---|---|---|---|---|
| SVM | **0.792** | 0.578 | 0.587 | 0.603 | 0.605 | 0.573 |
| k* | 0.776 | 0.706 | 0.698 | 0.799 | **0.803** | **0.813** |
| Bagging | 0.786 | 0.644 | **0.851** | 0.842 | 0.842 | **0.864** |
| RF | 0.825 | 0.661 | **0.861** | 0.852 | 0.856 | **0.873** |

**TABLE 8.** Evaluating based on ROC.

| | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_3{}^*$ | $M_4$ |
|---|---|---|---|---|---|---|
| SVM | **0.75** | 0.511 | 0.552 | 0.53 | 0.54 | 0.519 |
| k* | 0.809 | 0.559 | 0.665 | 0.848 | **0.853** | **0.858** |
| Bagging | 0.824 | 0.559 | 0.878 | 0.881 | **0.882** | **0.906** |
| RF | 0.885 | 0.558 | 0.894 | 0.901 | **0.903** | **0.93** |

$(TP + TN)/(TP + TN + FP + FN)$. Considering that the distribution of class labels of the data is a little skewed, with 56.9% positive and 43.1% negative, the accuracy may not be the best evaluation indicator. Thus, we also use the F-score which is the harmonic mean of precision and recall, denoted as $2TP/(2TP+FN+FP)$. In addition, observing the imbalance of positive and negative samples in the test set, the receiver operating characteristic (ROC) curve is chosen as the third metric. It remains stable as the distribution of samples changes and characterizes the ability to distinguish positive from negative samples for different thresholds.

k-fold cross-validation (CV). CV is a method of model selection that uses some samples for training and the others for testing. To do this, we divide $\Omega$ into $k$ subsets $\Omega_1$, $\Omega_2$,..., $\Omega_k$ such that $\Omega_1 \cup ... \cup \Omega_k = \Omega$, $\Omega_1 \cap ... \cap \Omega_k = \emptyset$, and $|\Omega_j| = |\Omega|/k$, $1 \leq j \leq k$, where $|\cdot|$ denotes the size of samples in a set. For each $j$, $1 \leq j \leq k$ , CV uses $\Omega \backslash \Omega_j$ as $\Omega_E$ and $\Omega_j$ as $\Omega_T$ and computes the generalization error of the model. When all the models are obtained, CV computes the mean generalization errors and selects the model of minimum error. Here, we set $k = 10$.

### C. RESULTS

For observational purposes, we highlight the values of interest in bold. The results of evaluating based on accuracy are shown in Table 6. For almost all models, the SVM yields very poor results, except $M_0$. Because the feature space of $M_1 \sim M_4$ is a set of leaving risk values which are closely distributed around 1, most of them are not linearly separable by super-planes using the kernel function; while the original features of $M_0$ are more easily separated by super-planes. For $M_4$, the $k*$, bagging, and RF algorithms yield satisfactory classification accuracy, with RF producing the best accuracy. This indicates that the prediction ability of $M_4$ outperforms that of the others in terms of accuracy when we use the hazard ratios of covariates as the learning features. Furthermore, the algorithms applied to the hazard ratio set of covariates of $M_3{}^*$ yield slightly higher accuracies compared with the algorithms applied to $M_3$, which reveals the advantage of the age-segmented model in prediction. The performances of the bagging and RF algorithms for $M_2$ are second only to those for $M_4$ but superior to those for the other models. The results of evaluation based on the F-score are very similar to those of the evaluation based on accuracy, as shown in Table 7. Using both metrics, we draw the same conclusions: 1) classification is much better when we use the hazard ratio set of covariates of $M_4$ as the learning features; and 2) RF is a satisfying classifier when it is used with the hazard ratio set of covariates.

The results of evaluating based on the ROC curve are shown in Table 8. It has both similarities and dissimilarity relative to the accuracy and F-score measures. The similarities are as follows: 1) the SVM can only be used for prediction for the original feature set but is not appropriate for the hazard ratio set; 2) the $k*$, bagging, and RF algorithms have the best prediction performances when they are used for $M_4$; 3) the prediction of the RF algorithm is the best for all models; and 4) the prediction can be improved to some extent when the age-segmented case is considered for $M_3$. However, the difference is that the $k*$, bagging, and RF algorithms yield good prediction performances when applied to $M_3{}^*$, compared with when they are applied to $M_2$, that are inferior only to those achieved when applied to $M_4$ and superior to those achieved when applied to all other models.

Now, we discuss the ecological costs of the models, which reveals how many variables are required for prediction.

This could help us understand the inputs of economy and technology. Note that the original space has 9 variables, with $M_1$, $M_2$ and $M_4$ each using 4, while $M_3$ uses 6. The costs of $M_1 \sim M_4$ are, respectively, 4/9, 4/9, 6/9, and 4/9. Apparently, $M_4$ has the highest performance with the lowest cost. Specifically, one could easily achieve a prediction performance of more than 85% when using the RF classifier with 4 factors, i.e., student age, admission score, video length, and whether the student is from the Medical Science category.

In summary, the prediction performances achieved using the feature sets of $M_2$, $M_3$ and $M_4$ are much better than that of the original one. $M_4$ is optimal and the most economical. The improved percentage (IP) with respect to the accuracy is as little as 6.25% and as much as 12.58%. The IP with respect to the F-score is as little as 4.77% and as much as 9.92%. The IP with respect to the ROC curve is as little as 5.08% and as much as 9.95%. We obtain the worst prediction ability using $M_1$, which considers only the VIR covariates.

### D. DISCUSSION

The prediction results are encouraging and suggest that $M_3$ and $M_4$ are not only able to interpret the impact of specific conditions or their combinations on the leaving risk but also have the potential to predict the leaving time accurately. However, we did not additionally tune the model parameters because the experiments were of a preliminary nature and considered fairness in the performance comparison. The prediction error mainly results from the crisp boundary problem caused by the explicit numerical boundary, which leads to an arbitrary segmentation of the data. One promising optimization is to define a fuzzy interval around the boundary so that users in this interval can be included in both categories. This approach is more reasonable considering that users who fall within the interval have relatively small differences in active video-viewing time. However, the implementation of this approach is beyond the scope of this article and is left for future work. Additionally, using the median, mode or other quantiles as the numerical boundary is also possible despite the fact that each would bring about different classification results.

The prediction provides valuable implications for designing an intelligent e-learning system. A successful intelligent e-learning system should consider both usefulness and effectiveness. Usefulness requires that the system automatically identifies the student groups and provides correct feedback instantly, while effectiveness means that the system is simple enough to satisfy specific calculation and application contexts. This work presents a novel method for significantly improving the prediction accuracy and, at the same time, avoids intensive calculations to enhance the performances of algorithms. These improvements can be attributed to two aspects of technical optimization. The first one is the transformation of the feature space, which transforms the raw variables into values that contribute to the leaving risk. The second one is the condensed feature space, which merely selects the components that are statistically significant for the dependent variable so that the feature space is reduced to the maximum extent. Additionally, the system can achieve useful and effective identification by inputting as little as four variables instead of complex data. For example, when a 22-year-old Computer Science student who achieves a score of 90 points on his/her entrance examination watches a 20-minute video, the system would estimate whether he/she has a greater probability of permanently leaving the course before the active video-viewing time reaches 8 minutes (the imaginary mean value) using the random forest classifier. This indicates that the system can successfully achieve the prediction function when developers provide a small number of database interfaces. Moreover, the effectiveness makes it possible to extend the system to mobile learning scenarios that are based on smart terminals.

## VII. CONCLUSIONS AND IMPLICATIONS

In this paper, we studied the active video-viewing time of an educational VoD system and dealt with the task of modeling students' leaving times and predicting the times at which they will leave a course. The study was conducted in three phases. First, we determined the distribution patterns of active video-viewing time and learned the shapes and characteristics. Second, we developed several statistical models with the *enter* method, which quantified the extent to which covariates in a model affect the active video-viewing time. We were especially concerned with the models with interaction terms, a non-linearity assumption of continuous variables and age segmentation. Finally, we used the hazard ratio of model covariates of statistical significance as the learning features and used efficient classification algorithms in machine learning to predict the time at which a user left a video without the intention to return.

The main findings are two-fold. On the one hand, the modeling results show the usefulness of the models in interpreting the influence of variables on the active video-viewing time. In general, 1) the amount and category of covariates as well as the extent of influence on the video-viewing time vary across models; 2) the increasing of the video length in $M_2$ dilutes the influence of some VIR covariates; 3) student age and gender have significant interactions with academic level; and 4) the fractional polynomial approach in $M_4$ displays significant non-linear characteristics of continuous covariates with respect to video length and age. It is worth noting that the non-linearity of age suggests that one could segment age to improve the explanatory ability of a model. On the other hand, the prediction results show that it is possible to estimate whether a student would leave permanently by considering only a few factors. Moreover, the outstanding prediction ability further illustrates a more extensive value than that of the models. Putting both the prediction ability and the ecological cost together, we find that using the fractional polynomial model, which considers the non-linearity of continuous variables, achieved the highest performance with the lowest cost. In comparison, the algorithms applied to the original features not only needed more variables but also

yielded poor prediction performances. Last but not least, the prediction ability of the age-segmented model was superior to those of the non-age-segmented models.

The findings of this paper have important implications. First, one could choose a proper model to explain the leaving risk of users so that teachers or policymakers could gain deep insights into which users will leaving before the time of interest, what causes leaving and how to prevent it. Additionally, the prediction ability implies that one could estimate whether a user would prematurely and permanently withdraw from a course by using only a small number of variables so that they can intervene with metacognitive strategies to lengthen the user's video-viewing time or so that the system can recommend materials and learning companions and offer pop-up, personalized hints to make the system more effective based on the extent to which he/she would leave. Moreover, this work presents valuable guidelines for designing an intelligent e-learning system and provides new insights and tutorials regarding the data analytics and feature subspace construction to learning analysts, researchers of artificial intelligence in education and data mining communities.

## REFERENCES


[1] J. Kamahara, T. Nagamatsu, M. Tada, Y. Kaieda, and Y. Ishii, "Instructional video content employing user behavior analysis: Time dependent annotation with levels of detail," in *User Modeling, Adaptation, and Personalization*. Berlin, Germany: Springer, 2010, pp. 87–98.

[2] T. Sinha. (Jul. 2014). "'Your click decides your fate': Leveraging clickstream patterns from MOOC videos to infer students' information processing and attrition behavior." [Online]. Available: https://arxiv.org/abs/1407.7143

[3] R. Moreno, "Constructing knowledge with an agent-based instructional program: A comparison of cooperative and individual meaning making," *Learn. Instruct.*, vol. 19, no. 5, pp. 433–444, 2009.

[4] M. L. G. Gonzalez and J. C. Arroyo, "Relationship between study time, self-regulation of learning and academic achievement in University students," *CPU-E Revista Invest. Edu.*, vol. 23, pp. 142–166, Jul. 2016.

[5] B. Xu and D. Yang, "Motivation classification and grade prediction for MOOCs learners," *Comput. Intell. Neurosci.*, vol. 2016, Jan. 2016, Art. no. 2174613.

[6] E. M. Brodhagen and M. Gettinger, "Academic learning time," in *Encyclopedia of the Sciences of Learning*. Boston, MA, USA: Springer, 2012, pp. 33–36.

[7] I.-H. Jo, D. Kim, and M. Yoon, "Analyzing the log patterns of adult learners in LMS using learning analytics," in *Proc. 4th Int. Conf. Learn. Anal. Knowl. (LAK)*, New York, NY, USA, 2014, pp. 183–187.

[8] C.-Y. Chou, T.-W. Chan, and C.-J. Lin, "An approach of implementing general learning companions for problem solving," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 6, pp. 1376–1386, Nov. 2002.

[9] P. J. Muñoz-Merino and C. D. Kloos, "An architecture for combining semantic Web techniques with intelligent tutoring systems," in *Proc. 9th Int. Conf. Intell. Tutoring Syst. (ITS)*, Montreal, QC, Canada, Jun. 2008, pp. 540–550.

[10] P. J. Muñoz-Merino and C. D. Kloos, "A software player for providing hints in problem-based learning according to a new specification," *Comput. Appl. Eng. Edu.*, vol. 17, no. 3, pp. 272–284, 2009.

[11] L. Chen, Y. Zhou, and D. M. Chiu, "Video browsing—A study of user behavior in online VoD services," in *Proc. 22nd Int. Conf. Comput. Commun. Netw. (ICCCN)*, 2013, pp. 1–7.

[12] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of Internet short video sharing: A YouTube-based measurement study," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1184–1194, Aug. 2013.

[13] M. Z. Shafiq, J. Erman, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Understanding the impact of network dynamics on mobile video user engagement," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 42, no. 1, pp. 367–379, 2014.

[14] Z. Li *et al.*, "Watching videos from everywhere: A study of the PPTV mobile VoD system," in *Proc. ACM Conf. Internet Meas. Conf. (IMC)*, New York, NY, USA, 2012, pp. 185–198.

[15] C. Zhou, Y. Guo, Y. Chen, X. Nie, and W. Zhu, "Characterizing user watching behavior and video quality in mobile devices," in *Proc. 23rd Int. Conf. Comput. Communi. Netw. (ICCCN)*, Aug. 2014, pp. 1–6.

[16] G. Gürsun, M. Crovella, and I. Matta, "Describing and forecasting video access patterns," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 16–20.

[17] R. K. Mok, E. W. Chan, X. Luo, and R. K. Chang, "Inferring the QoE of HTTP video streaming from user-viewing activities," in *Proc. 1st ACM SIGCOMM Workshop Meas. Stack (W-MUST)*, New York, NY, USA, 2011, pp. 31–36

[18] Y. Chen, B. Zhang, Y. Liu, and W. Zhu, "Measurement and modeling of video watching time in a large-scale Internet video-on-demand system," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2087–2098, Dec. 2013.

[19] C. Shi, S. Fu, Q. Chen, and H. Qu, "VisMOOC: Visualizing video clickstream data from massive open online courses," in *Proc. IEEE Pacific Vis. Symp. (PacificVis)*, Apr. 2015, pp. 159–166.

[20] R. H. Kay, "Exploring the use of video podcasts in education: A comprehensive review of the literature," *Comput. Human Behavior*, vol. 28, no. 3, pp. 820–831, 2012.

[21] C. G. Brinton, S. Buccapatnam, M. Chiang, and H. V. Poor. (Mar. 2015). "Mining MOOC clickstreams: On the relationship between learner behavior and performance." [Online]. Available: https://arxiv.org/abs/1503.06489

[22] C. G. Brinton and M. Chiang, "MOOC performance prediction via clickstream data and social learning networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 2299–2307.

[23] T. Sinha, N. Li, P. Jermann, and P. Dillenbourg. (Sep. 2014). "Capturing 'attrition intensifying' structural traits from didactic interaction sequences of MOOC learners." [Online]. Available: https://arxiv.org/abs/1409.5887

[24] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses," in *Proc. 3rd Int. Conf. Learn. Anal. Knowl. (LAK)*, New York, NY, USA, 2013, pp. 170–179.

[25] N. Xue *et al.*, "Probabilistic modeling towards understanding the power law distribution of video viewing behavior in large-scale e-learning," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, vol. 2., Aug. 2015, pp. 136–142.

[26] T. Xie, Q. Zheng, and W. Zhang, "A behavioral sequence analyzing framework for grouping students in an e-learning system," *Knowl.-Based Syst.*, vol. 111, pp. 36–50, Nov. 2016.

[27] *Ministry of Education of the People's Republic of China*, accessed on Oct. 4, 2016. [Online]. Available: http://en.moe.gov.cn/

[28] M. Eagle and T. Barnes, "Survival analysis on duration data in intelligent tutors," in *Proc. 12th Int. Conf. Intell. Tutoring Syst. (ITS)*, Honolulu, HI, USA, Jun. 2014, pp. 178–187.

[29] M. K. Islam, "Introducing survival and event history analysis," *Can. Stud. Population*, vol. 41, nos. 1–2, pp. 223–224, 2014.

[30] S. B. Plank, S. DeLuca, and A. Estacion, "High school dropout and the role of career and technical education: A survival analysis of surviving high school," *Sociol. Edu.*, vol. 81, no. 4, pp. 345–370, 2008.

[31] G. Allione and R. M. Stein, "Mass attrition: An analysis of drop out from principles of microeconomics MOOC," *J. Econ. Edu.*, vol. 47, no. 2, pp. 174–186, 2016.

[32] H. Finch, D. K. Lapsley, and M. Baker-Boudissa, "A survival analysis of student mobility and retention in Indiana charter schools," *Edu. Policy Anal. Arch.*, vol. 17, no. 18, pp. 1–18, 2009.

[33] K. McKenzie and R. Schweitzer, "Who succeeds at University? Factors predicting academic performance in first year Australian University students," *Higher Edu. Res. Develop.*, vol. 20, no. 1, pp. 21–33, 2001.

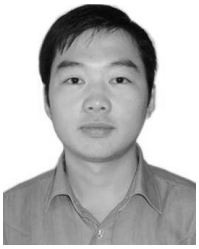
[34] G. R. Pike and J. L. Saupe, "Does high school matter? An analysis of three methods of predicting first-year grades," *Res. Higher Edu.*, vol. 43, no. 2, pp. 187–207, 2002.

[35] J. C. S. Uribe, C. M. L. Gómez, and M. C. J. Elorza, "Identification of factors that affect the loss of student status using a logit survival model for discrete time data," *Dyna*, vol. 79, no. 171, pp. 16–22, 2012.

[36] M. Bradburn, T. Clark, S. Love, and D. Altman, "Survival analysis part III: Multivariate data analysis—Choosing a model and assessing its adequacy and fit," *Brit. J. Cancer*, vol. 89, no. 4, pp. 605–611, 2003.

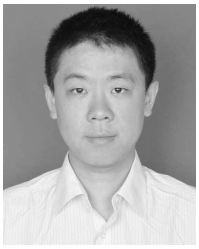
[37] A. Palloni, "Applied survival analysis: Regression modeling of time to event data," *Sociol. Methods Res.*, vol. 31, no. 4, pp. 557–559, 2003.

[38] P. Royston and D. G. Altman, "Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling," *J. Roy. Statist. Soc. Ser. C, Appl. Statist.*, vol. 43, no. 3, pp. 429–467, 1994.
[39] P. M. Grambsch, T. M. Therneau, and T. R. Fleming, "Diagnostic plots to reveal functional form for covariates in multiplicative intensity models," *Biometrics*, vol. 51, no. 4, pp. 1469–1482, 1995.
[40] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Netw.*, vol. 12, no. 6, pp. 783–789, 1999.
[41] J. G. Cleary and L. E. Trigg, "K*: An instance-based learner using an entropic distance measure," in *Proc. 12th Int. Conf. Mach. Learn.*, vol. 5, 1995, pp. 108–114.
[42] N. Dey, *Classification and Clustering in Biomedical Signal Processing*. USA: IGI Global, 2016.
[43] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[44] H. B. Mitchell, "Ensemble learning," in *Image Fusion: Theories, Techniques and Applications*. Berlin, Germany: Springer, 2010, pp. 125–142.
[45] Z.-H. Zhou, "Ensemble learning," in *Encyclopedia of Biometrics*. Boston, MA, USA: Springer, 2009, pp. 270–273.



**TAO XIE** received the B.S. and M.S. degrees in educational technology from Southwest University, China, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree in computer science with Xi'an Jiaotong University, China. His current research interests include mobile learning, location-based service, data mining, and technology-enhanced learning.

**QINGHUA ZHENG** received the B.S. degree in computer software, the M.S. degree in computer organization and architecture, and the Ph.D. degree in system engineering from Xi'an Jiaotong University, China, in 1990, 1993, and 1997, respectively. He did post-doctoral research at Harvard University in 2002 and was a Visiting Professor of Research with the Hong Kong University from 2004 to 2005. He is currently a Professor with the Department of Computer Science and Technology, Xi'an Jiaotong University, where he serves as the Vice President. His research interests include intelligent e-learning theory and algorithm, computer network, and trusted software. He received the First Prize for National Teaching Achievement, State Education Ministry in 2005 and the First Prize for Scientific and Technological Development of Shanghai City and Shaanxi Province, in 2004 and 2003, respectively.



**WEIZHAN ZHANG** received the B.S. degree in electronics engineering from Zhejiang University, China, in 1999, and the Ph.D. degree in computer science from Xi'an Jiaotong University, China, in 2010. He served as a Software Engineer with Datang Telecom Corporation from 1999 to 2002. He is currently an Associate Professor with the Department of Computer Science and Technology, Xi'an Jiaotong University. His current research interests include multimedia networking, cloud computing, and mobile computing.

**HUAMIN QU** received the B.S. degree in mathematics from Xi'an Jiaotong University, China, and the M.S. and Ph.D. degrees in computer science from the Stony Brook University. He is currently a Professor with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. His main research interests include visualization and computer graphics, with focuses on urban informatics, social network analysis, e-learning, and text visualization.