Business Process Reporting Using Process Mining, Analytic Workflows and Process Cubes: A Case Study in Education

Alfredo Bolt, Massimiliano De Leoni, Wil Aalst, Pierre Gorissen

HAL Id: hal-01651891

https://inria.hal.science/hal-01651891

Submitted on 29 Nov 2017

Business Process Reporting Using Process Mining, Analytic Workflows and Process Cubes: A Case Study in Education

Alfredo Bolt[1], Massimiliano de Leoni[1], Wil M. P. van der Aalst[1], and Pierre Gorissen[2]

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
`a.bolt, m.d.leoni, w.m.p.v.d.aalst@tue.nl`
[2] Hogeschool van Arnhem en Nijmegen, Nijmegen, The Netherlands
`pierre.gorissen@han.nl`

Abstract. Business Process Intelligence (BPI) is an emerging topic that has gained popularity in the last decade. It is driven by the need for analysis techniques that allow businesses to understand and improve their processes. One of the most common applications of BPI is *reporting*, which consists on the structured generation of information (i.e., reports) from raw data. In this article, state-of-the-art process mining techniques are used to periodically produce automated reports that relate the actual performance of students of a Dutch University to their studying behavior. To avoid the tedious manual repetition of the same process mining procedure for each course, we have designed a *workflow* calling various process mining techniques using RapidProM. To ensure that the actual students' behavior is related to their actual performance (i.e., grades for courses), our analytic workflows approach leverages on *process cubes*, which enable the dataset to be *sliced* and *diced* based on courses and grades. The article discusses how the approach has been operationalized and what is the structure and concrete results of the reports that have been automatically generated. Two evaluations were performed with lecturers using the real reports. During the second evaluation round, the reports were restructured based on the feedback from the first evaluation round. Also, we analyzed an example report to show the range of insights that they provide.

Keywords: Business Process Reporting, Analytic Workflows, Process Mining, Process Cubes, Education.

1 Introduction

Business Process Reporting (BPR) refers to the provision of structured information about processes in a regular basis, and its purpose is to support decision makers. Reports can be used to analyze and compare processes from many perspectives (e.g., behavior, performance, costs, time). In order to be effective, BPR presents some challenges:

1. It should provide insights using metric-based characteristics (e.g., bottle-
   necks, throughput time, resource utilization) and behavioral characteristics
   (e.g., deviations, frequent patterns) of processes.
2. It should be repeatable (i.e., not require great efforts to repeat the analysis).
3. It should be able to analyze the data with different granularity levels (i.e.,
   analyze an organization as a whole or analyze its branches individually).

This paper shows through a case study how *process mining*, *analytic work-flows* and *process cubes* can be concretely used for business process reporting, addressing the three challenges mentioned above.

The case study presented in this paper refers to a business-process reporting service at Eindhoven University of Technology. The service produces a report each quartile (i.e., two-month academic period) for each course that is provided with video lectures. The report is sent to the responsible lecturer and provides insights about the relations between the students' usage of video lectures and their final grades on the course, among other educational data analysis results.

Process mining is a relatively young research discipline that is concerned with discovering, monitoring and improving real processes by extracting knowledge from event logs readily available in today's systems [1]. Hundreds of different process mining techniques have been proposed in literature. These are not limited to process-model discovery and the checking of conformance. Also, other perspectives (e.g., data) and operational support (e.g., predictions) are included. Process mining is supported by commercial software (e.g., Disco[3], Celonis[4]) and academic software (e.g., ProM[5] [2]) tools.

Process mining allows the extraction of insights about the overall and inner behavior contained in any given process (e.g., a student taking a course). These insights can be collected and processed into reports. When thousands of different reports need to be produced (e.g., one for each course), it can be tedious and error-prone to manually repeat all the process-mining analyses to be incorporated in the reports. Analytic workflows can be used to fully automate analytic experiments such as the generation of an arbitrary number of reports. Process cubes can be used to scope and split the overall process data into the granularity level expected by the analytic workflow. These scoped subsets of event data can be distributed into cube *cells*. Then, the event data contained in each cell can be used as input for the analytic workflow (e.g., all the students that took a given course on a given quartile).

The usefulness of the reports is evaluated with dozens of lecturers throughout two evaluation rounds in different academic periods. During the first evaluation round, an initial set of reports was sent to lecturers and feedback was collected through an evaluation form. The feedback was used to restructure the report. Then, a set of restructured reports was sent to lecturers and a group of them were interviewed to asses if the insights contained in the report were better perceived. The results show that, indeed, this is the case.

---

[3] http://fluxicon.com/disco

[4] http://www.celonis.de/en

[5] ProM tools is free to download from http://www.promtools.org

The remainder of this paper is organized as follows. Section 2 discusses related work about educational data analysis and business process reporting. Section 3 provides an overview of the case study and discusses the structure of the reports. Sections 4, and 5 summarize the related work and main concepts related to analytic workflows and process cubes and illustrate how they are concretely applied in this case study. Sections 6 and 7 discusses the reports sent and the results of the two evaluation rounds with the lecturers. Finally, Section 8 concludes the paper.

2   Related Work

This section discusses the related work done around business process reporting and educational data analysis. Related work about analytic workflows and process cubes is discussed in Sections 4 and 5 respectively.

2.1   Business Process Reporting

Business Process Intelligence (BPI) is defined by [3] as the application of Business Intelligence (BI) techniques to business processes. However, behavioral properties of processes (e.g., control-flow) cannot be represented using traditional BI tools. Alternatively, Castellanos et al. [4] provides a broader definition: BPI exploits process information by providing the means for analyzing it to give companies a better understanding of how their business processes are actually executed. It incorporates not only metric-based process analysis, but also process discovery, monitoring and conformance checking techniques as possible ways to understand a process.

Business Process Reporting can be defined as the structured and periodical production of reports containing analysis of process data obtained through BPI techniques.

Business process management suites (e.g., SAP, Oracle) usually provide process reporting capabilities. Often, these process reporting capabilites are an adaptation of general-purpose reporting tools (e.g., Crystal Reports, Oracle Discoverer) to process data [3]. These general-purpose reporting tools are unable to analyze the data from a process perspective (e.g., discover a model).

Most process mining tools (e.g., ProM, Disco) are able to analyze from a process perspective, but they lack reporting and capabilities Others, such as Celonis, offer business process reporting capabilities. However, they are limited to only a few process-perspective analysis components, and each report instance has to be created manually. Furthermore, the event data used as input for the report can only be filtered from the original event data; the granularity level cannot be changed. Also, most of these tools do not allow the comparison of process variants (e.g., students with different grades).

Given the limitations described above, in this paper we used a combination of process mining, analytic workflows and process cubes to provide fully automated process-oriented reports.

2.2 Educational Data Analysis

The analysis of educational data and the extraction of insights from it is related to two research communities: *Educational Data Mining* and *Learning Analytics*.

Educational data mining (EDM) is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyze educational data in order to study educational questions [5,6]. For example, knowledge discovered by EDM algorithms can be used not only to help teachers to manage their classes, understand their students' learning processes and reflect on their own teaching methods, but also to support a learner's reflections on the situation and provide feedback to learners. An extensive survey on the state of the art of EDM is presented in [5]

Learning analytics (LA) is defined by [7] as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs. According to [8], the difference between EDM and LA is that they aproach different challenges driving analytics research. EDM focuses on the technical challenge (e.g., How can we extract value from these big sets of learning-related data?), while LA focuses on the educational challenge (e.g., How can we optimise opportunities for online learning?). A discussion on the differences, similarities and collaboration opportunities between these two research communities is presented in [9].

Several process mining techniques (e.g., Fuzzy Miner [10]) have been applied successfully in the context of EDM [11] for analyzing study curriculums followed by students. Notably, the work introduced by [12] aims to obtaining better models (i.e., in terms of model *quality*) for higher education processes by performing data preprocessing and semantic log purging steps. However, most of these techniques are not suitable for analyzing video lecture usage by students, given the inherent lack of structure of such processes.

In this paper, we will use existing and new process mining techniques to obtain insights of student behavior from an educational point of view.

3 A Case Study in Education

Eindhoven University of Technology provides video lectures for many courses for study purposes and to support students who are unable to attend face-to-face lectures for various reasons. Student usage of video lectures and their course grades are logged by the University's IT systems. The purpose of this case study is to show how raw data extracted from the University's IT systems can be transformed into reports that show insights about students' video lecture usage and its relation with course grades by using process mining, process cubes and analytic workflows. Figure 1 describes the overview of this case study.

The data used in this case study contains *video lecture views*, *course grades* and *personal information* of students of the University. Each student and course
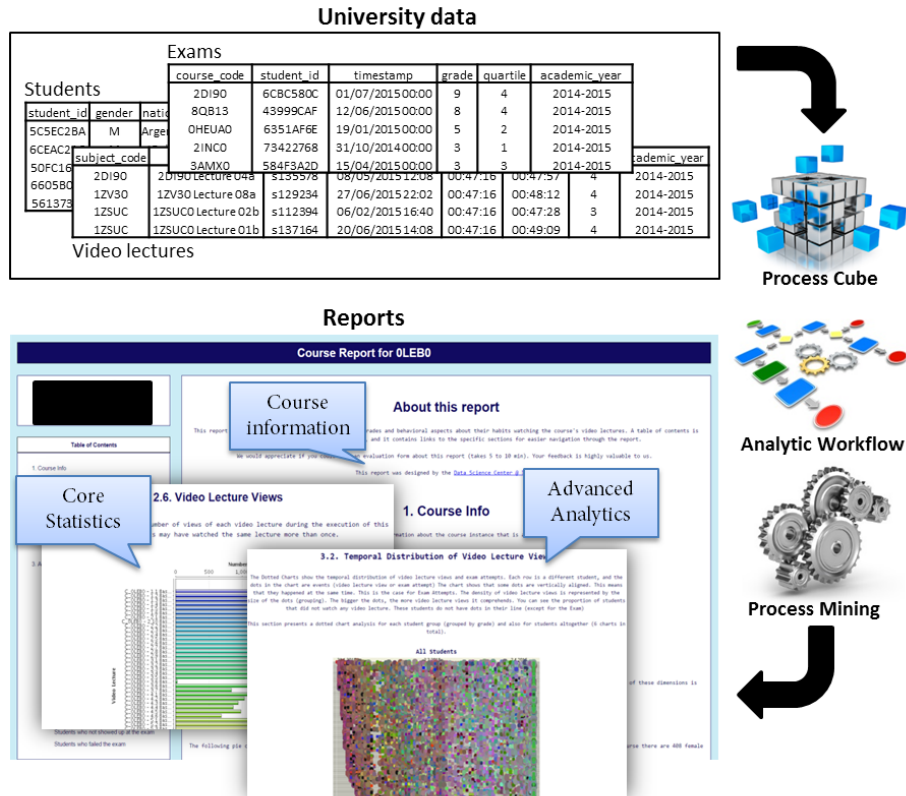
Fig. 1: Overview of the case study: University data is transformed into reports by using process mining, process cubes and analytic workflows.

has a unique identifier code (i.e., *student id, course code*). The data reveals enormous variability; e.g., thousands of students watch video lectures for thousands of courses and every course has a different set of video lectures, and they have different cultural and study backgrounds, which leads to different behavior. Therefore, we need to provide different reports and, within a report, we need to perform a comparative analysis of the students when varying the grade.

Before describing our approach and the ingredients used, we sketch the report we aim for. The report is composed of three sections: *course information*, *core statistics* and *advanced analytics*, as shown in Figure 1.[6] The analysis results refer to all students who registered for the course exam, independently whether or not they participated in it.

---

[6] An example report, where student information has been anonymized, can be downloaded from `https://www.dropbox.com/s/565zz94rdo6gg2r/report.zip?dl=0`

The course information section provides general information, such as the course name, the academic year, the number of students, etc. The core statistics section provides aggregate information about the students, such as their gender, nationality, enrolled bachelor or master program, along with course grades distribution and video lecture views. The advanced analytics section contains process-oriented diagnostics obtained through process mining techniques.

The next two sections show how the desired reports can be generated.

4   Analytic Workflows as a Means to Automate Analysis

Process mining experiments usually require analysts to perform many analysis steps in a specific order. As mentioned in Section 1, it is not unusual that the same experiment has to be carried out multiple times. This is normally handled by manually executing the analysis steps of the experiment, thus requiring large periods of time and resources and introducing the risk of human-induced errors.

Current process mining tools are not designed to automatically repeat the application of the same process-mining analyses on an arbitrary number of (sub sets of) event logs. Therefore, it is not possible to automatically generate any arbitrary number of reports.
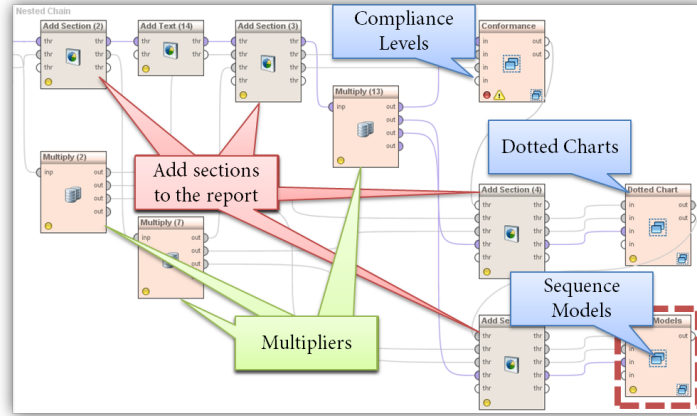
Analytic workflows can be used to address this problem. They are defined by chaining analysis and data-processing steps, each of which consumes input produced by previous steps and generates output for the next steps. Analytic workflows are a specialization of *scientific workflows* tailored towards analytic purposes. Scientific workflows have successfully been applied in many settings [13,14]. The work presented in [15] illustrate the formalization and operationalization of a framework to support process-mining analytic workflows where the steps are linked to the application of process-mining techniques.

In this paper, we combine process mining with analytic workflow systems, which allow one to design, compose, execute, archive and share workflows that represent some type of analysis or experiment. Each activity/step of an analytic workflow is one of the steps to conduct a non-trivial process-mining analysis, which can range from data filtering and transformation to process discovery or conformance checking. Once an analytic workflow is configured, it can be executed with different process data as many times as needed without reconfiguration.

In our case study, for automatically generating the course reports we used RapidProM [16,15], which extends the RapidMiner analytic workflow tool with process mining techniques.[7]

Figure 2a illustrates the analytic workflow that is used to generate each report. Figure 2b shows the explosion of the "Sequence Models" section of the analytic workflow.

---

[7] Free version and installation instructions for RapidMiner and the RapidProM extension are available at `http://www.rapidprom.org` or at the RapidMiner *Marketplace*.

(a) Advanced Analytics section sub-workflow

(b) Explosion of the "Sequence Models" sub-workflow

Fig. 2: Implemented analytic workflow used to generate the reports. Each instance of a course can be autimatically analyzed in this way resulting in the report described.

The operators shown in Figure 2 are used for different purposes: *Multipliers* allow one to use the output of an operator as input for many operators. *Filter* operators select a subset of events based on defined criteria. *Process mining operators* are used to produce analysis results. For example, the operators high-

lighted in blue in Figure 2b produce a sequence model from each filtered event data.

The complete RapidProM implementation of the analytic workflow used in this case study is available at `https://www.dropbox.com/s/g9spsziyv55vsro/single.zip?dl=0`. Readers can execute this workflow in RapidMiner to generate a report using the sample event log available at `https://www.dropbox.com/s/r3gczshxqxh6a6d/Sample.xes?dl=0`.[8]

5 Process Cubes as a Means to Select and Scope Event Data

Processes are not static within moderns organizations but their instances continuously adapt to the dynamic context requirements of modern organizations. Therefore, an event log records executions of several process variants, whose behavior depends on context information (e.g., different courses that may contain different behavior). As a consequence, the event log needs to be split into sublogs (i.e., one for each variant), each containing all the events that belong to that variant. The naive approach consists on manually filtering the event data. Naturally, this approach is unpractical in scenarios where many different process variants exist.

Process cubes [17] are used to overcome this issue: in a process cube, events are organized into cells using different dimensions. The idea is related to the well-known notion of OLAP (Online Analytical Processing) data cubes and the associated operations, such as slice, dice, roll-up, and drill-down. By applying the correct operations, each cell of the cube contains a sub-set of the event log that complies with the homogeneity assumption mentioned above. This allows one to isolate and analyze the different variants of a process.

Several approaches provide these capabilities, such as [18], which presents an exploratory view on the application of OLAP operations over events. Other process-cube approaches have been applied in specific contexts [19,20]. The term *process cube* was introduced and formalized in [17] with a working prototype presented in [21], and later improved and implemented in [22].

5.1 Basic Concepts

A process cube is characterized by a set of *dimensions*, each of which is associated with one or a group of event's data properties. For each combination of values for the different dimensions, a cell exists in the process cube. Hence, each process-cube cell contains the events that assign certain values to the data properties. Each cell of the process cube contains event data that can be used by process mining techniques. Please note that certain dimensions may be considered as

---

[8] When running the workflow, make sure that the *Read File* operator points to the sample event log and the "HTML output directory" parameter of the *Generate Report* operator points to the desired output folder.
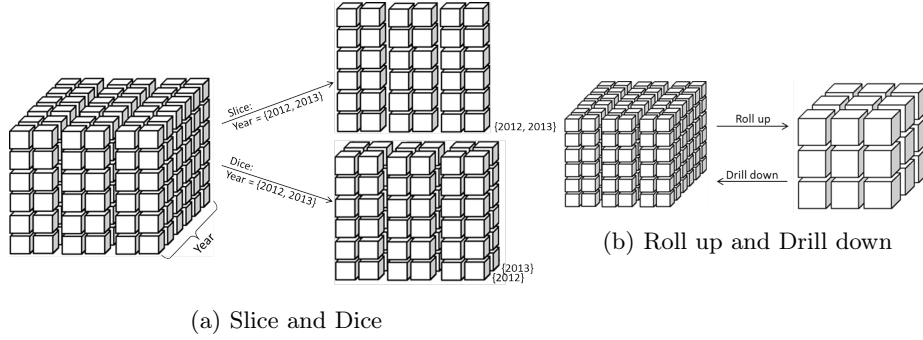
(a) Slice and Dice

(b) Roll up and Drill down

Fig. 3: Schematic examples of cube operations

irrelevant and, therefore, they are ignored and are not *visible* in the cube. Also, some dimensions may be not readily available in the event data; however, they can be derived from the existing dimensions. For example, the "Year" and "Day" dimensions can be derived from the "Timestamp" dimension.

The *slice* operation selects a subset of values of a dimension while removing that dimension from the analysis. For example, if the "Year" dimension is sliced for Year = {2012, 2013}, only the events in those years are retained. Also, the "Year" dimension is removed from the cube as shown in Figure 3a. The latter implies that cells with different values for the "Year" dimension and the same values for the other dimensions are merged.

The *dice* operation is similar to the *slice* operation, with the difference that the dicing dimension is retained. So, the dice operation is only removing cells without merging any cells: the dicing dimension can still be used for further exploration of the event data, as shown in Figure 3a.

The *roll up* and *drill down* operations change the granularity level of a dimension. As shown in Figure 3b, if a dimension is rolled up, an attribute with a more coarse-grained granularity will be used to create the cells of the cube, and if a dimension is drilled down, an attribute with a more fine-grained granularity will be conversely used. For example, the "Day" dimension can be rolled up to "Month", and the "Month" dimension can be drilled down to "Day".

5.2 Application to the Case Study

For performing process cube operations over the University data we used the *Process Mining Cube* (PMC) tool introduced in [22]. As mentioned before, the starting point is an event data set. This event data set has been obtained by defining and running opportune joins of tables of the database underlying the video-lecture system of the University (see Section 3). A fragment of the event data set is shown in Table 1.

Using the event data, we created a process cube with the following dimensions: Student Id, Student Gender, Student Nationality, Student Education

Table 1: A fragment of event data generated from the University's system: each row corresponds to an event.

Ev. Id	Student Id	Nat.	Ed. Code	Course Code	Activity	Quartile	Acad. Year	Timestamp	Grade	...
1	1025	Dutch	BIS	2II05	Lecture 1	1	2014-2015	03/09/2012 12:05	6	...
2	1025	Dutch	BIS	2II05	Lecture 2	1	2014-2015	10/09/2012 23:15	6	...
3	1025	Dutch	BIS	1CV00	Lecture 10	3	2014-2015	02/03/2012 15:36	7	...
4	2220	Spanish	INF	1CV00	Lecture 1	3	2014-2015	20/03/2013 16:24	8	...
5	1025	Dutch	BIS	2II05	Exam	2	2014-2015	13/12/2012 12:00	6	...
6	2220	Spanish	INF	1CV00	Lecture 4	3	2014-2015	25/03/2013 11:12	8	...
7	2220	Spanish	INF	1CV00	Exam	3	2014-2015	04/04/2013 12:00	8	...
...	...	...	...	...	...	...	...	...	...	...

Code, Student Education Phase, Course Code, Course Department, Activity, Activity Type, Grade, Timestamp, Quartile and Academic Year.

After *slicing* and *dicing* the cube, thousands of cells are produced: one for each combination of values of the "Course Code", "Quartile" and "Course Grade" dimensions. Each cell corresponds to an event log that can be analyzed using process mining techniques.

We applied our approach that combines process mining, analytic workflows and process cubes to the case study presented in Section 3 in two evaluation rounds. The following sections describe the work, reports, results and the feedback obtained on each round.

6 Initial Report

The first evaluation round was conducted in August 2015 and it used the event data corresponding to the academic year 2014-2015. The data used in this round contains **246.526** *video lecture views* and **110.056** *course grades* of **8.122** students, **8.437** video lectures and **1.750** courses. Concretely, we automatically generated a total of **8.750** course reports for 1750 courses given at the University in each of the 5 quartiles (i.e., 4 normal quartiles + *interim* quartile) of the academic year 2014-2015. For reliability of our analysis, we only selected the reports of courses where, on average, each student watched at least 3 video lectures. In total, 89 courses were selected and their reports were sent to the corresponding lecturers.

Section 6.1 shows the first report structure through an example of the reports sent to lecturers in this evaluation round. It also provides a detailed analysis of the findings that we could extract from the report for a particular course. Along with the report, we also sent an evaluation form to the lecturers. The purpose of the evaluation forms is to verify whether lecturers were able to correctly interpret the analysis contained in the report. The results obtained in the first evaluation round are discussed in Section 6.2.

6.1 Structure of the Report

To illustrate the structure, contents and value of the reports, we selected an example course: "Introduction to modeling - from problems to numbers and back"

given in the third quartile of the academic year 2014-2015 by the Innovation Sciences department at the University. This course is compulsory for all first-year students from all programs at the University. In total, 1621 students attended this course in the period considered. This course is developed in a "flipped classroom" setting, where students watch online lectures containing the course topics and related contents, and in the classroom, they engage these topics in practical settings with the guidance of the instructor.

The video lectures provided for this course are mapped onto weeks (1 to 7). Within each week, video lectures are numbered to indicate the order in which students should watch them (i.e., 1.1 correspond to the first video lecture of the first week). As indicated by the course's lecturer, the first video lectures of each week contain the course topics for that week, and the last video lectures of each week contain complementary material (e.g., workshops, tutorials). The number of video lectures provided for each week depends on the week's topics and related activities, hence, it varies.

Students's behavior can be analyzed from many perspectives. As mentioned in Section 2.2, several process mining techniques have been applied in the context of educational data analysis [11].

Initially, we applied traditional process model discovery techniques (e.g., Fuzzy Miner [10], ILP Miner [23], Inductive Visual Miner [24]) to the educational data. However, given the unstructured nature of this data (i.e., students watching video lectures), the produced models were very complex (i.e., *spaghetti* or *flower* models) and did not provide clear insights. Therefore, we opted for other process mining techniques that could help us understand the behavior of students:

Figure 4.a shows for each video lecture the number of students that watched it. We can observe that the number of students that watch the video lectures decreases as the course develops: most students watched the video lectures corresponding to the first week (i.e., 1.X) but less than half of them watched the video lectures corresponding to the last week (i.e., 7.X). Note that within each week, students tend to watch the first video lectures (i.e., X.1, X.2) more than the last ones (i.e., X.5, X.6). This was discussed with the course's lecturer. It is explained by the fact that, as mentioned before, the first video lectures of each week contain the topics, and the last ones contain complementary material.

Figure 4.b shows for each student group (i.e., grouped by their grade) the level of conformance, averaged over all students in that group, of the real order in which students watch video lectures, compared with the "natural" or logical order, namely with watching them in sequence (i.e., from 1.1 to 7.4). The conformance level of each student is measured as the *replay fitness* of the data over a process model that contains only the "natural" sequential order. The replay fitness was calculated using conformance checking techniques [25]. We can observe that students with higher grades have higher levels of conformance than students with lower grades.

Figure 4.c shows the grade distribution for this course where each bar is composed by two parts corresponding to the number of students who watched
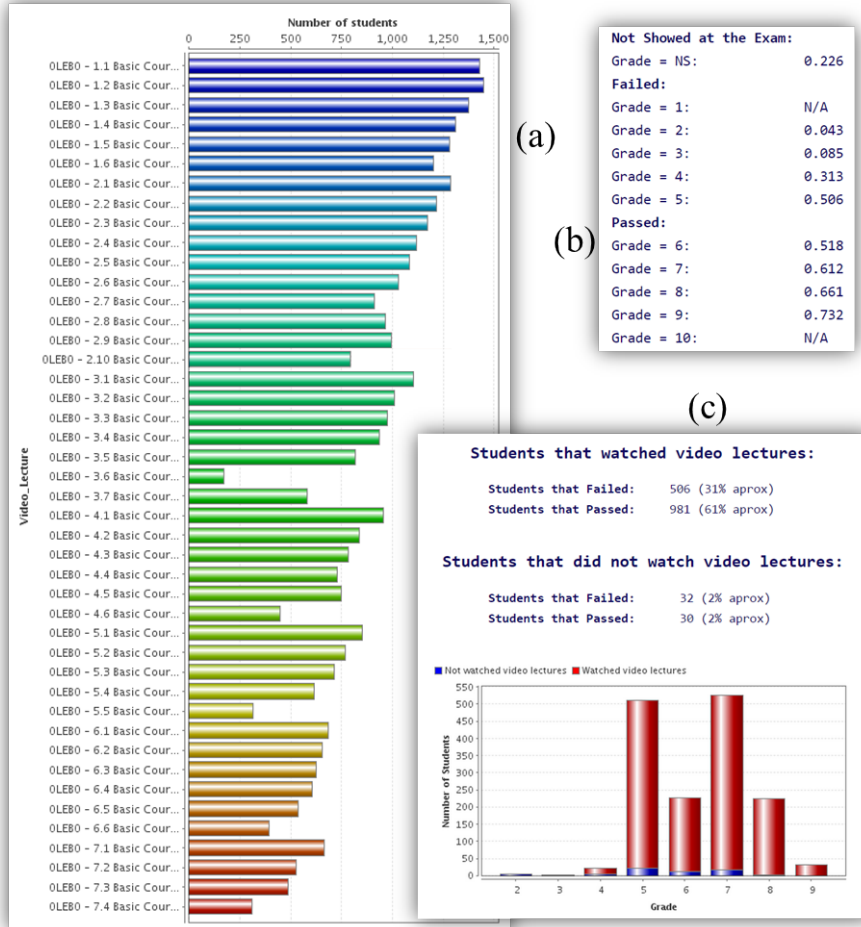
Fig. 4: Analysis results contained in the report of the course 0LEB0:
(a) Number of students that watched each video lecture
(b) Conformance with the natural viewing order by course grade
(c) Grades distribution for students who watched video lectures (in red)
or did not (in blue)

at least one (red part) video lecture and the number of students who did not (blue part). We can observe that the best students (i.e., with a grade of 8 or above) use video lectures. On the other hand, we observe that watching video lectures does not guarantee that the student will pass the course, as shown in the columns of students that failed the course (i.e. grade ≤ 5).

Figure 5 shows dotted charts [26] highlighting the temporal distribution of video-lecture watching for two student groups: (a) students that failed the course

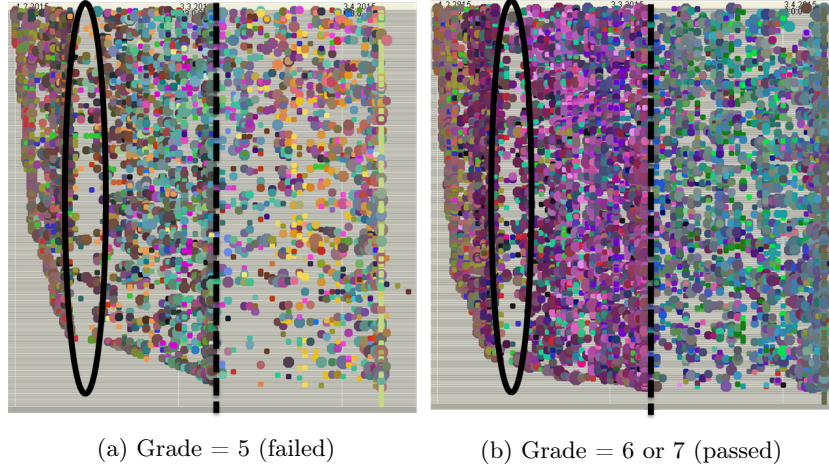(a) Grade = 5 (failed)   (b) Grade = 6 or 7 (passed)

Fig. 5: Dotted charts for students grouped by their course grades

with a grade of 5, and (b) students that passed the course with a grade of 6 or 7. Each row corresponds to a student and each dot in a row represents that student watching a video lecture or taking the exam. Note that both charts show a gap where very few video lectures were watched, which is highlighted in the pictures through an oval. This gap coincides with the *Carnaval* holidays. We can observe that, in general, students that failed watched fewer video lectures. Also note that in Fig. 5.a the density of events heavily decreases after the mid-term exam (highlighted through a vertical dashed line). This could be explained by students being discouraged after a bad mid-term result. This phenomenon is also present in (b), but not equally evident. We can also observe that most students tend to constantly use video lectures. This is confirmed by the low number of students with only a few associated events.

Figure 6 shows sequence analysis models that, given any ordered sequence of activities, reflects the frequency of directly-follows relations[9] as percentage annotations and as the thickness of edges. The highest deviations from the ordered sequence order are highlighted in colored edges (i.e., black edges correspond to the natural order). This technique was tailored for the generation of reports and it is implemented using a customized RapidProM extension. When comparing (a) students that passed the course with a grade of 6 or 7 with (b) students that had a grade of 8 or 9, we can observe that both groups tend to make roughly the same deviations. Most of these deviations correspond to specific video lectures being skipped. These skipped video lectures correspond in most cases to complementary material. In general, one can observe that the thickness (i.e., frequency) of the arcs denoting the "natural" order (i.e., black arcs) is higher for

---

[9] The frequency of directly-follows relations is defined for any pair of activities (A, B) as the ratio between the number of times that B is directly executed after A and the total number of times that A is executed.

Table 2: Summary of the classification of statement evaluations performed by
lecturers

Statement Evaluation	Core Statistics Section	Advanced Analytics Section			Sub Total	Total (%)
		Conformance	Temp. Dist.	Seq. Analysis		
Correct	261	30	67	32	**390** (89%)	**61%**
Incorrect	28	5	8	6	**47** (11%)	
Unknown	95	61	69	58	**283**	**39%**

(b), i.e., those with higher grades. Note that at the beginning of each week we
can observe a *recovery* effect (i.e., the frequencies of the natural order tend to
increase).

6.2   Lecturers Evaluation

In addition to the qualitative analysis for some courses like such as the course
analyzed in Section 6.1, we have also asked lecturers for feedback through an eval-
uation form linked to each report.[10] The evaluation form provided 30 statements
about the analysis contained in the reports (e.g., "Higher grades are associated
with a higher proportion of students watching video lectures", "Video lecture
views are evenly distributed throughout the course period"). Lecturers evalu-
ated each statement on the basis of the conclusions that they could draw from
the report. For each of the 30 statements, lecturers could decide if they agreed
or disagreed with the statement, or, alternatively, indicate that they could not
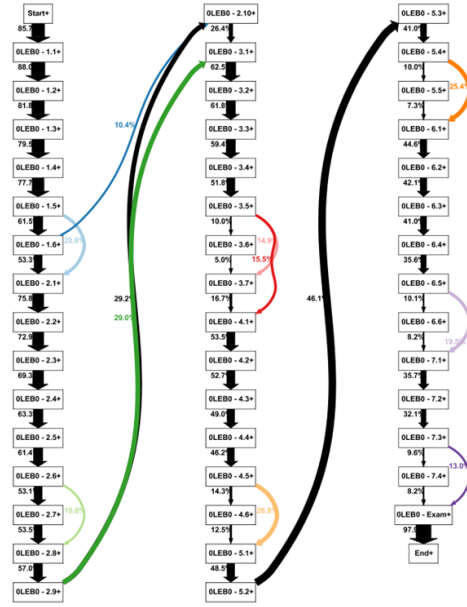evaluate the statement (i.e., "I don't know").

In total, 24 of the 89 lecturers answered the evaluation form. Out of the 720
(24 x 30) possible statement evaluations, 437 statements were answered with
"agree" or "disagree". The remaining cases in which the statement could not be
evaluated can be explained by three possible causes: the statement is unclear,
the analysis is not understandable, or the data shows no conclusive evidence.

In the case that a statement was evaluated with "agree" or "disagree", we com-
pared the provided evaluation with our own interpretation of the same statement
for that report and classified the response as *correct* or *incorrect*. In the case that
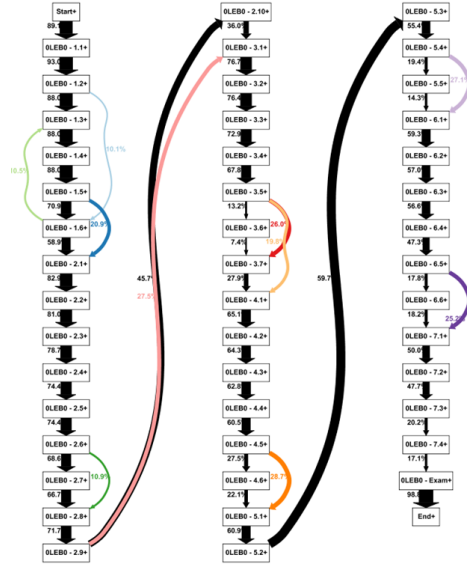a statement was not evaluated, the respose was classified as *unknown*.

Table 2 shows a summary of the response classification for each section of
the report. In total, **89%** of the statement evaluations were classified as *correct*.
This indicates that lecturers were capable to correctly interpret the analysis
provided in the reports. Note that the *Conformance* section had the highest rate
of *unknown* classifications (63.5%). This could be related to understandability
issues of the analysis presented in that section.

The evaluation form also contained a few general questions. One of such
questions was: "Do you think this report satisfies its purpose, which is to provide
insights about student behavior?", for which 7 lecturers answered "yes", 4 lectur-
ers answered "no" and 13 lecturers answered "partially". All the lecturers that
responded "partially" provided written feedback indicating the improvements

---

[10] The evaluation form is available at `https://www.dropbox.com/s/09y4ypklt70y6d9/form.zip?dl=0`

(a) Grade = 6 or 7

(b) Grade = 8 or 9

Fig. 6: Sequence analysis for students grouped by their course grades

they would like to see in the report. Some of the related comments received were: "It would be very interesting to know if students: a) did NOT attend the

lectures and did NOT watch the video lectures, b) did NOT attend the lectures, but DID watch the video lectures instead, c) did attend the lectures AND watch the video lectures too. This related to their grades", "The report itself gives too few insights/hides insights", "It is nice to see how many students use the video lectures. That information is fine for me and all I need to know", and "I would appreciate a written explanation together with your diagrams, next time". Another question in the evaluation form was: "Do you plan to introduce changes in the course's video lectures based on the insights provided by this report?", for which 4 lecturers answered "yes" and 20 answered "no". The results show that the analysis is generally perceived as useful, but that more actionable information is needed, such as face-to-face lecture attendance. However, this information is currently not being recorded by the TU/e. The feedback provided by lecturers was used to improve the report. These improvements are discussed in the next Section.

7 Final Report

We modified the reports based on the feedback obtained in the first evaluation round. The detail of the changes is presented in Section 7.1. To assess the quality of the improved report, we conducted a second evaluation round, which was conducted in March 2016 and it used the event data corresponding to the first two quartiles of the academic year 2015-2016. The data used in this round contains **89.936** *video lecture views* and **49.078** *course grades* of **10.152** students, **2.718** video lectures and **1.104** courses. Concretely, we automatically generated a total of **2.208** course reports for 1104 courses given at the University in each of the 2 first quartiles of the academic year 2015-2016. For reliability of our analysis, we only selected the reports of courses where, on average, each student watched at least 3 video lectures. In total, 56 courses were selected and their reports were sent to the corresponding lecturers.

Section 7.1 shows the changes introduced in the report based on the feedback obtained from lecturers in the first evaluation round. It also provides examples of the findings that several lecturers could extract from the report. Along with the report, we also sent an evaluation form to the lecturers. The purpose of the evaluation forms is to verify whether lecturers were able to correctly interpret the analysis contained in the improved report. Unfortunately, in this evaluation round no lecturer answered the evaluation form. Therefore, we held face-to-face meetings with four lecturers, where the results included in the report were discussed. The insights obtained in these meetings are discussed in Section 7.2.

7.1 Changes in the Report

According to the feedback obtained from lecturers (reported in Table 2 in Section 6.2), the most problematic sections (i.e., highest rate of *unknown* classifications) were the Conformance (63.5% *unknown*) and Sequential Analysis (60.4% *unknown*) sections of the report (described in Section 6.1).

Given this feedback and the fact that the interpretation of a specific *replay fitness* value can be misleading for non-process-mining-experts, the conformance section was replaced for a section that describes the *compliance* of students with the "natural" order of watching video lectures based on simpler calculations, defined as follows.

Definition 1 (Compliance Score). *For any given student, their compliance score (CS) w.r.t. the natural order is calculated as $CS = \sum_{i=1}^{n-1} \frac{df(a_i, a_{i+1})}{count(a_i)}$, where $df(a_i, a_{i+1})$ is the number of times that the student watched lecture a_{i+1} directly after a_i, $count(a_i)$ is the number of times that the student watched the lecture a_i and n is the number of video lectures available for the course.*
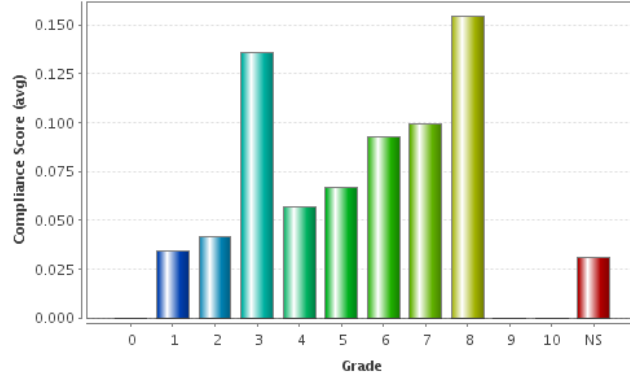
This new compliance score is easier to interpret: a value of X means that X percent of the video lectures watched by the student were watched in the natural order.

Figure 7 shows an example of the new compliance section of the report. It refers to the course "5ECC0 - Electronic circuits 2" (more details will be given later). Figure 7.a shows the average compliance scores according to the student's grades, while Figure 7.b shows the distribution of students according to their compliance scores.
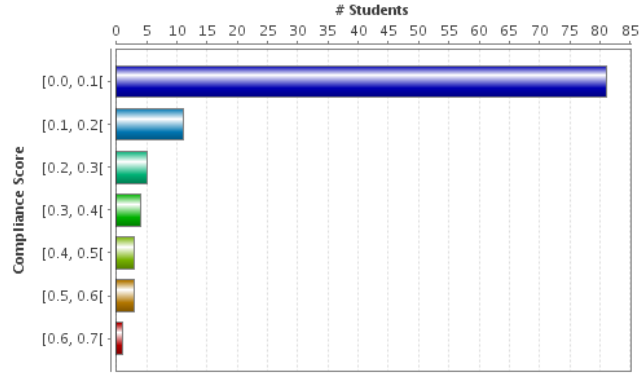
Regarding the Sequence Analysis section, we simplified the explanatory text of this section in the report. However, this section presents inherent difficulties associated to the analysis of process models: most lecturers are not familiar with process models. Previously, sequence models only referred to frequency deviations. In this round, we decided to incorporate sequence models that show performance information. In these sequence models, an arc indicates the time between the start of the source activity (i.e., video lecture or exam) and the start of the target activity. From these models, one can observe if a given lecture is being fully watched, or if students are skipping most of it after watching a few minutes. Figure 8 shows an example of a sequence model annotated with performance information. This model was obtained from one of the course reports (7U855 - Research methods for the built environment) sent in this evaluation round. From these models we can get interesting insights about the students' behavior on this course. For example, in Figure 8.a (i.e., students that obtained a 6 or a 7 in the exam) the arrow between *Lecture 01* and *Lecture 02* states that students that watched *Lecture 02* directly after *Lecture 01*, started watching *Lecture 02* 14 seconds (in average) after started watching *Lecture 01*. However, in Figure 8.b (i.e., students that obtained a 8, 9 or 10 in the exam) this specific behavior is not observed.

7.2   Lecturers Evaluation

As mentioned before, from the 56 reports sent to lecturers in this evaluation round, we obtained no responses to the corresponding evaluation forms. Therefore, we held face-to-face meetings with four lecturers from different departments of the University to discuss the report in general, and to evaluate if the changes

(a) Average student compliance with the "natural" order according to the student's grades

(b) Student distribution over compliance level by range.

Fig. 7: New compliance section of the report for an example course (5ECC0 - Electronic circuits 2)

introduced in this evaluation round did actually improve the understandability of the report.

In the remainder of this section, we summarize the insights obtained by lecturers when discussing the reports in the face-to-face meetings.

The first lecturer we met was responsible for the course 1CV00 - Deterministic Operations Management, provided by the Industrial Engineering department. In this course, lectures are grouped by topic (i.e., 2 lectures per topic) and topics are independent from each other. Figure 9.a shows the distribution of students according to their compliance scores for this course. In this chart, we can observe that students have a very low compliance score in general, and it had no correlation with grades. The lecturer defined this behavior as "expected" since the course topics are independent. Figure 9.b shows the dotted chart containing all the students of the course. Here we can observe two peaks of video lecture usage
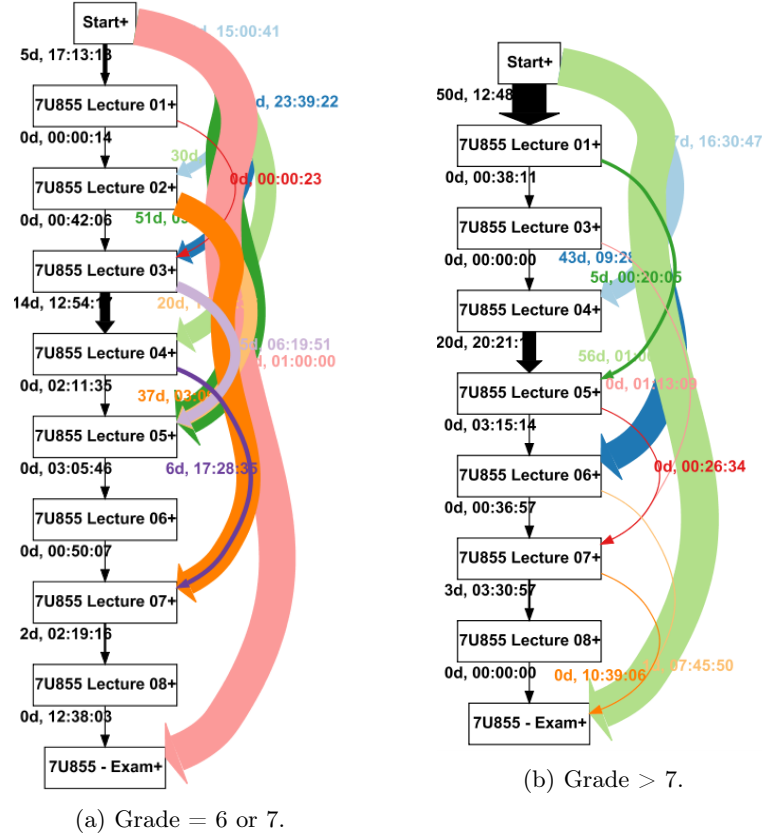
(a) Grade = 6 or 7.

(b) Grade > 7.

Fig. 8: Sequence models annotated with performance information for students grouped by their grade. The models were obtained from the report of course 7U855 - Research methods for the built environment.

in weeks 4 and 7 (highlighted with vertical yellow lines), but without context information, we cannot explain why they happened. The lecturer immediately identified these two peaks as the two *mid-term* exams that are part of the course. The interpretation given by the lecturer was that students were using the video lectures to study for these exams. This behavior was expected by the lecturer, but in the past he did not have the information to either confirm or deny it.

The second lecturer was responsible for the course 4EB00 - Thermodynamics, provided by the Mechanical Engineering department. In this course, some topics build on top of knowledge acquired in previous topics, but others are independent. Figure 10.a shows, for each lecture, the total number of views. We can observe that *Lecture 02a* and *Lecture 05a* had the highest number of views. The lecturer determined that this behavior was expected, since *Lecture 02a* contained most of the definitions and knowledge that students needed to "remember" from

(a) Student distribution over compliance level by range. Compliance scores are relatively low.

(b) Dotted chart for all the students enrolled in the course. Video lecture usage peaks are highlighted with yellow vertical lines.
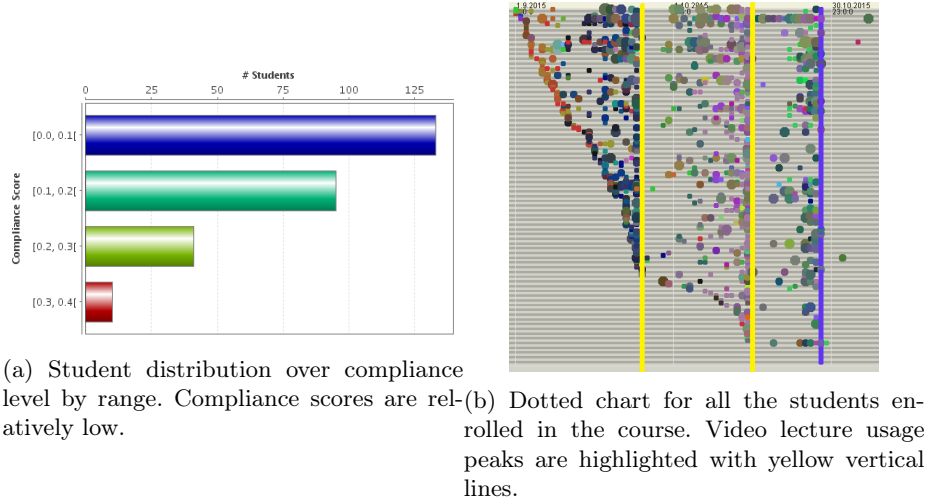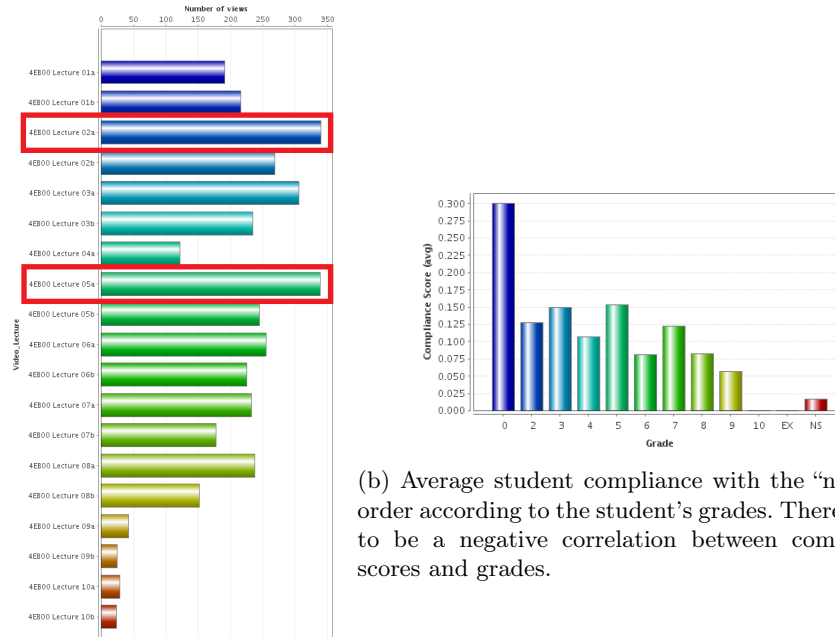
Fig. 9: Analysis results included in the report of the course 1CV00.

previous courses. On the other hand, *Lecture 05a* was related to *Entropy*, which was the most difficult topic of the course for students. Figure 10.b shows the average student compliance with the "natural" order according to the student's grades. We can observe that there is a negative correlation between the compliance scores and the grades. According to the lecturer: "A possible explanation of this could be that students with bad grades could have skipped face-to-face lectures and then needed to watch all the video lectures, while good students attended face-to-face lectures and only watched some video lectures if they needed to clarify something".

The third lecturer was responsible for the course 5ECC0 - Electronic Circuits 2, provided by the Electrical Engineering department. In this course, all the topics were related, every topic built-up on the previous one. Figure 7.a showed the average student compliance with the "natural" order according to the student's grades. We can observe a positive correlation between compliance scores and grades. The lecturer was positively surprised by this finding, but he considered that the correlation was not strong. Figure 11 shows a fragment of the sequence model with frequency deviations for two different groups of students of the course (i.e., those with a grade lower than 5, and those with a grade equal to 6 or 7). We can observe in Figure 11.a that *Lecture 01c* is being skipped by 13% of the students that watched the *Lecture 01b*. This behavior does not occur for students with higher grades (shown in Figure 11.b). The lecturer then considered this finding as "unexpected, but positive", since *Lecture 01c* consists of the basic topics from the previous course (i.e., Electronic Circuits 1) and it was meant to refresh student's knowledge. According to the lecturer, the fact that students did not need to watch it is positive.

(a) Number of views of each video lecture of the course. Lectures 02a and 05a were the most watched by students (highlighted in red).

(b) Average student compliance with the "natural" order according to the student's grades. There seems to be a negative correlation between compliance scores and grades.

Fig. 10: Analysis results included in the report of the course 4EB00.

The fourth lecturer was responsible for the course 5XCA0 - Fundamentals of Electronics, provided by the Electrical Engineering department. This course considers topics are relatively independent from each other. Figure 12.a shows, for each lecture, the total number of views. It is interesting to notice that the *Lecture 05a*, the video lecture most watched by students (highlighted in red) is an *instruction* lecture (i.e., a lecture that consists of exercises instead of topics). Given this finding, the lecturer has expressed the intention of splitting that video lecture, for the next executions of the course, into a series of 10-minute web lectures comprehending all the different types of exercises covered in the video lecture. Figure 12.b shows the student distribution over ranges of compliance score. It is clear from the chart that most students have a very low compliance score w.r.t. the "natural" viewing order. This was justified by the lecturer through the following statement: "The topics are relatively disconnected, and it seems that most students would watch only specific lectures".

The general comments that we received from lecturers are summarized as follows. "It would be interesting to see the correlation with face-to-face lectures to see if students use video lectures as a replacement or as a complement for them". "Video lectures are very good for the *middle* students. good students do
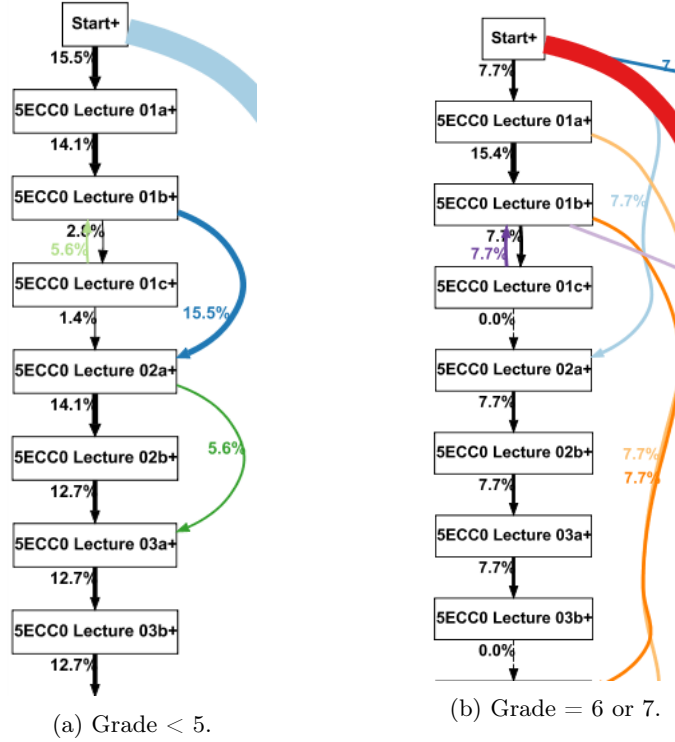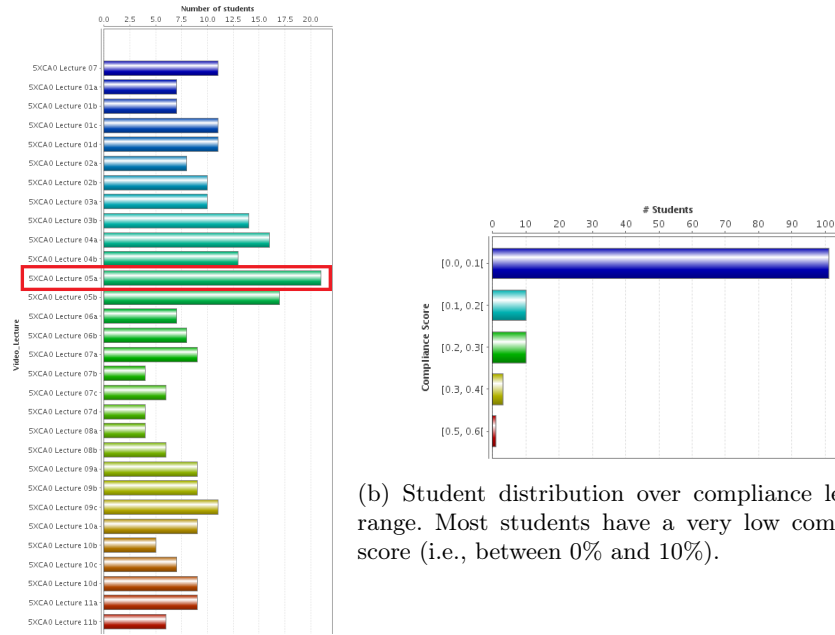
(a) Grade < 5.

(b) Grade = 6 or 7.

Fig. 11: Fragment of the sequence model with frequency deviations for all students. In (a), Lecture 1c is being skipped. These charts were included in the report of the course 5ECC0 - Electronic Circuits 2.

not seem to need them as much". "I should split the most visited video lectures into a series of web lectures (i.e., 10 minute recordings of specific topics) so I could really know which topics are the most difficult for the students". "Students tend to use exercise lectures much more intensively than the actual theory. They seem to be exam-oriented, as they prepare mostly watching exercises".

Regarding the report itself, again we received suggestions to incorporate face-to-face lecture attendance. As mentioned in Section 6, it is very difficult to record face-to-face attendance of students for technical reasons. Other lecturers suggested that we incorporate student feedback into the report. We certainly recognize the potential that incorporating the students' feedback on the report could have in the insights that the lecturer can obtain from it. We plan to incorporate this feedback and its potentially positive effects in the reports for the next quartile.

(a) Number of views of each video lecture of the course. Lecture 05a was the most watched by students (highlighted in red).

(b) Student distribution over compliance level by range. Most students have a very low compliance score (i.e., between 0% and 10%).

Fig. 12: Analysis results included in the report of the course 5XCA0.

8   Conclusion

This paper has illustrated the benefits of combining the complementary approaches of process cubes and analytic workflows in the field of process mining. In particular, the combination is beneficial when process mining techniques need to be applied on large, heterogenous event data of multidimensional nature.

To demonstrate such benefits, we applied the combined approach in a large scale case study where we provide reports for lecturers. These reports correlate the grades of students with their behavior while watching the available video lectures. We evaluated the usefulness of the reports in two evaluation rounds. The second evaluation round presented an improved report, which was modified based on the feedback obtained in the first evaluation round. Unlike existing *Learning Analytics* approaches, we focus on dynamic student behavior. Also, descriptive analytics would not achieve similar analysis results because they do not consider the process perspective, such as the ordering of watching video lectures.

Educational data has been analyzed by some disciplines in order to understand and improve the learning processes [7,9,8,5,6], even employing process

cubes [20]. However, these analyses were mostly focused on individual courses. No research work has previously been conducted to allow large-scale process mining analysis where reports are automatically generated for any number of courses. Our approach has made it possible by integrating process mining with analytic workflows, which have been devised for large-scale analysis, and process cubes, which provide the capabilities needed to perform comparative analyses.

As future work, the report generation will be extended to *Massive Open Online Courses* (MOOCs) given by Eindhoven University of Technology. This type of courses are particularly interesting due to the fact that face-to-face lectures are not used: video lectures are the main channel used by students for accessing the course topics. For example, over **100.000** people from all over the world registered for the two executions of the MOOC *Process Mining: Data science in Action*.[11] We also plan to apply this analysis to the courses provided by the *European Data Science Academy* (EDSA).[12]

References

1. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. 1st edn. Springer-Verlag Berlin Heidelberg (2011)
2. van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H.M.W., Weijters, A.J.M.M., van der Aalst, W.M.P.: The ProM framework: A new era in process mining tool support. In: Applications and Theory of Petri Nets. Volume 3536 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2005) 444–454
3. Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M., Shan, M.C.: Business process intelligence. Computers in Industry **53**(3) (2004) 321 – 343 Process / Workflow Mining.
4. Castellanos, M., Alves de Medeiros, A.K., Mendling, J., Weber, B., Weijers, A.J.M.M.: Business process intelligence. In Cardoso, J., van der Aalst, W.M.P., eds.: Handbook of Research on Business Process Modeling. IGI Global, Hershey, PA, USA (2009) 456–480
5. Romero, C., Ventura, S.: Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics (Part C: Applications and Reviews) **40**(6) (Nov 2010) 601–618
6. Gorissen, P.J.B.: Facilitating the use of recorded lectures: analysing students' interactions to understand their navigational needs. PhD thesis, Eindhoven University of Technology (2013)
7. Siemens, G.: Learning analytics: The emergence of a discipline. American Behavioral Scientist **57**(10) (2013) 1380–1400
8. Ferguson, R.: Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning **4**(5-6) (2012) 304–317
9. Siemens, G., Baker, R.S.J.d.: Learning analytics and educational data mining: Towards communication and collaboration. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. LAK '12, New York, NY, USA, ACM (2012) 252–254

---

[11] `http://www.coursera.org/course/procmin`
[12] `http://edsa-project.eu`

10. Günther, C.W., van der Aalst, W.M.P.: Fuzzy mining – adaptive process simplification based on multi-perspective metrics. In: Proceedings of the 5th International Business Process Management Conference. Volume 4714 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2007) 328–343

11. Trcka, N., Pechenizkiy, M., van der Aalst, W.M.P.: Process mining from educational data. In: Handbook of educational data mining. CRC Press, London (2010) 123–142

12. Ly, L.T., Indiono, C., Mangler, J., Rinderle-Ma, S. In: Data Transformation and Semantic Log Purging for Process Mining. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 238–253

13. Jaeger, E., Altintas, I., Zhang, J., Ludäscher, B., Pennington, D., Michener, W.: A scientific workflow approach to distributed geospatial data processing using web services. In: Proceedings of the 17th International Conference on Scientific and Statistical Database Management (SSDBM'2005), Berkeley, CA, US, Lawrence Berkeley Laboratory (2005) 87–90

14. Turner, K., Lambert, P.: Workflows for quantitative data analysis in the social sciences. International Journal on Software Tools for Technology Transfer **17**(3) (2015) 321–338

15. Bolt, A., de Leoni, M., van der Aalst, W.M.P.: Scientific workflows for process mining: building blocks, scenarios, and implementation. International Journal on Software Tools for Technology Transfer (2015) . To appear. Doi: 10.1007/s10009–015–0399–5

16. Mans, R.S., van der Aalst, W.M.P., Verbeek, H.M.W.: Supporting process mining workflows with RapidProM. In: Proceedings of the BPM Demo Sessions 2014 Colocated with the 12th International Conference on Business Process Management (BPM). Volume 1295 of CEUR Workshop Proceedings., CEUR-WS.org (2014) 56–60

17. van der Aalst, W.M.P.: Process cubes: slicing, dicing, rolling up and drilling down event data for process mining. In: Proceedings of the First Asia Pacific Conference on Business Process Management. Volume 159 of Lecture Notes in Business Information Processing., Springer International Publishing (2013) 1–22

18. Ribeiro, J., Weijters, A.: Event cube: Another perspective on business processes. In: On the Move to Meaningful Internet Systems: OTM 2011 (Proceedings of the Confederated International Conferences: CoopIS, DOA-SVI, and ODBASE 2011). Volume 7044 of Lecture Notes in Computer Science., Springer Berlin Heidelberg (2011) 274–283

19. Vogelgesang, T., Appelrath, H.J.: Multidimensional process mining: A flexible analysis approach for health services research. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops. EDBT '13, New York, NY, USA, ACM (2013) 17–22

20. van der Aalst, W.M.P., Guo, S., Gorissen, P.J.B.: Comparative process mining in education: An approach based on process cubes. In: Proceedings of the 4th international Symposium on Data-driven Process Discovery and Analysis. Volume 203 of Lecture Notes in Business Information Processing., Springer Berlin Heidelberg (2015) 110–134

21. Mamaliga, T.: Realizing a process cube allowing for the comparison of event data. Master's thesis, Eindhoven University of Technology, Eindhoven, the Netherlands (2013)

22. Bolt, A., van der Aalst, W.M.P.: Multidimensional process mining using process cubes. In: Enterprise, Business-Process and Information Systems Modeling.

Volume 214 of Lecture Notes in Business Information Processing. Springer International Publishing (2015) 102–116

23. van der Werf, J.M.E.M., van Dongen, B.F., Hurkens, C.A.J., Serebrenik, A.: Process discovery using integer linear programming. In van Hee, K.M., Valk, R., eds.: Applications and Theory of Petri Nets. Volume 5062 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2008) 368–387

24. Leemans, S.J.J., Fahland, D., van der Aalst, W.M.P.: Exploring processes and deviations. In Fournier, F., Mendling, J., eds.: Business Process Management Workshops. Volume 202 of Lecture Notes in Business Information Processing. Springer International Publishing (2015) 304–316

25. van der Aalst, W.M.P., Adriansyah, A., van Dongen, B.F.: Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(2) (2012) 182–192

26. Song, M., van der Aalst, W.M.P.: Supporting process mining by showing events at a glance. In: Proceedings of the 17th Annual Workshop on Information Technologies and Systems (WITS). (2007) 139–145