MOODLE FORUMS: EXPLORATION INTO LOGGED DATA AND APPLYING MACHINE LEARNING TO GET FURTHER INSIGHT.

Lappeenranta-Lahti University of Technology LUT

Master's Program in Business analytics, Master's Thesis

2023

Ilpo Kainulainen

Examiners: Lassi Roininen
Uolevi Nikula

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Sciences
Business analytics

Ilpo Kainulainen

Moodle forums: Exploration into logged data and applying machine learning to get further insight.

Master's thesis
2023
51 pages, 32 figures, 1 table, 2 appendices
Examiners: Lassi Roininen and Uolevi Nikula

Keywords: Data analysis, Moodle, Clustering, Forum usage, Course module analysis, Forum analysis, Social network analysis, Classifying

This work investigates the data logged into Moodle logs in relation to the usage of forums, which while not containing the text itself contains information about who did what and where. The data was from a programming basics course forums which focused on tasks given on the course. The data was gathered through the courses logs and prepared and analysed through Python with libraries such as pandas and seaborn and finally machine learning was applied with MATLAB. The objectives were to do an exploratory analysis, building on top of the log data and trying of different types of machine learning to gather further insight into the usage of forums. As for results of grouping: there were 2 groups on both courses or in 2021 fall course five groups and in 2022 fall course six groups using forums while in classifying the accuracy was low and some of the attempted methods only guessed single classification for every sample.

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Sciences
Business analytics

Ilpo Kainulainen

Moodle Foorumit: Tutkimus tallennettuun dataan ja koneoppimisen soveltaminen ymmärryksen lisäämiseksi

Diplomityö
2023
51 sivua, 32 kuvaa, 1 taulukko, 2 liitettä
Tarkastajat: Lassi Roininen ja Uolevi Nikula

Hakusanat: Data-analyysi, Moodle, klusterointi, foorumien käyttö, kurssimoduulin analyysi, foorumi analyysi, sosiaalisten verkkojen analyysi, luokittelu
Keywords: Data analysis, Moodle, Clustering, Forum usage, Course module analysis, Forum analysis, Social network analysis, Classifying

Tässä työssä tutkitaan mitä dataa Moodlen loki pitää foorumeiden käyttöön liittyen. Vaikka loki ei sisällä foorumeille laitettua tekstiä se sisältää vastauksen kysymyksiin: kuka, mitä ja missä. Tutkittu data tuli ohjelmoinnin perusteet kurssin foorumilta, joka keskittyy kurssilla annettuihin tehtäviin. Data kerättiin kurssin lokeista, valmisteltiin ja analysoitiin Pythonin avulla käyttäen kirjastoja kuten pandas ja seaborn ja lopulta MATLAB:in avulla koneoppimista sovellettiin. Tarkoituksena oli tehdä eksploratiivinen tutkimus, lisätä dataa lokien lisäksi ja yrittää erilaisia koneoppimismalleja saadaksemme lisää ymmärrystä foorumien käyttöön. Ryhmittelyn tuloksena oli että kummankin syksyn 2021 ja 2022 kurssilla oli kaksi ryhmää tai 2021 syksyllä oli viisi ryhmää ja 2022 syksyllä kuusi ryhmää käyttämässä foorumeita. Luokittelussa tarkkuus oli matala ja jotkut yritetyistä malleista arvasivat vain yhtä luokitusta kaikille näytteille.

ACKNOWLEDGEMENTS

I would like to thank my supervisors and co-workers for helping me with this thesis and friends and family for moral support.

Lappeenranta, May 2, 2023

Ilpo Kainulainen

LIST OF ABBREVIATIONS

EDM	educational data mining
ML	machine learning
MOOC	Massive Open Online Courses
PCA	Principal Component Analysis
PLSR	Partial Least Squares regression
REGEX	Regular expression
SVM	Support Vector Machine

CONTENTS

1 INTRODUCTION

1.1 Background

LUT university offers programming courses in Finnish, these courses have two forums: Anonymous and staff forums. On anonymous forums the students can make their own posts, see other student's posts and can reply to them anonymously, while on the staff forums students cannot see other students' posts. The anonymous forum is for students to get help from their peers and the only limitation given to them is that they try to limit the amount of code posted to prevent plagiarism. The forums are anonymous to try to limit the amount of anxiety of the students as Griffin & Roy [1] found course forums as anxiety inducing.

The staff forum is so that the students can ask help from the staff, and they can post the full code as there is no fear of plagiarism as the post are only visible to the student who made the post and the staff. The forums have recently changed platforms and during this change of platform the teacher responsible felt that there seemed to be fall in participation on forums in relation of students helping each other. This led to a request of data analysis on the course forums to see if this can be seen in the numbers and what else can be found in the data. Also, the connection of forum usage and the grade student received by the end of the course was looked into.

The data for this analysis is to be gathered through Moodle's course log, which are extensive. It is hoped that this allows us to investigate even how long the user stays on the forums or in particular discussions. But during the time of writing this is not confirmed to be true on part of the forums, as even the Moodle's own wiki does not go into detail about the logs.

Also, it was considered if more information could be gathered through machine learning (ML), for example what kind of forum usage groups could be identified or could it be used to find connection between the forum usage and the students grade?

1.2 Objectives

The goals of this thesis are to answer the following questions:

1. What information is gathered in logs and what can be seen from it, in relation to the Moodle forums?
2. How can the data gained from Moodle logs be used to analyse the forum usage?
3. Can the users be grouped by their usage of the forums and if they can be what kinds of groups does this raise?
4. Can ML be used to find connections between forum usage data and the grade received by the student?

1.3 Structure of the thesis

The structure of this study approximately follows the process of the actual research. Starting with literacy review looking into other studies and what they have found. Then researching what the Moodle logs contain, what can be extracted from it, what are its limitations and can it be augmented through some processes. Then a look into the data. After this the study goes into the processing of the data: clean-up and extracting useful features. Then the cleaned data will be looked at and it will be searched for more useful data points and visualised to bring the useful data into view. After that, the study looks into different kinds of ML that could be used and what they produced on this data. And finally, discussion about this study itself, its limits, what was found and further need for analysis or research, and the conclusion of this study.

2 RELATED WORK

Online forum analysis is a long-studied subject and there is no shortage of related work. With a simple Google scholar search for "online forum analysis" producing approximately 2.7 million results as of September of 2022. But this does not give a good picture for the purposes of this thesis as for example the first search page has studies relating to mental health effects, sentiments on forums among other really specified topics. Another related area of research is educational data mining (EDM) which focuses on gaining insight from data relating to education. Romero *et al.* give a good definition of EDM's primary goal as "to use large-scale educational data sets to better understand learning and to provide information about the learning process" [2] and Scheuer *et al.* [3] defines EDM's background as relatively new, while an older survey by Romero *et al.* [4] shows that the area has been slowly developing for a long time. This work by Romero *et al.* also lists few points that this work mainly focuses on, mainly: statistics, visualisation, clustering and classifying. It also lists text mining which, as talked in the next part, seems to be popular.

2.1 Studies relating to usage of forums in education

Onah *et al.* [5] in their study focus on Massive Open Online Courses (MOOC)'s forum but this course also uses Moodle's forums. They found that they had about 13% of the students make a post on the forums but had over 7000 views with 552 students. In their study they also found difference in the levels of participation when they compared it to another MOOC's report and due to open nature of the studied course when compared to the course looked at in this thesis there might be a difference in the levels of participation. And in Coetzee *et al.* [6] found that the students who use forums usually achieve higher grades.

Hirschel's [7] work looks into students' opinions on forums, quizzes and glossaries in studying English. In it, they classified survey responses from students, and they found that forums were especially beneficial for reserved students as it allowed comprehensible output and interaction. While Suviste *et al.* [8] go into classifying the posts into seven categories, though in this approach ML is not used and then they do questionnaires to look more into the users and the use cases of the forum. This kind of classification could be approached with ML and to see if this classifying could be done with clustering or classifying approaches.

2.2 Studies relating to using forums as a part of educational datamining

This thesis is not some innovative approach to the analysis and similar work has been done earlier at other places of study, as a good example Baruque *et al*. [9] could be considered for guidance but due to only the abstract being in English, using it as guideline would be difficult. Meanwhile Pong-inwong & Rungworawut [10] go through Moodle forums and classify the posts to two different classes positive and negative on the perception of education. This study does give promising ideas about how you can approach this kind of work, but is missing some important points, for example: how they extracted the data from the forums is missing. The content of these discussions can affect the grade of the participants as found out by Wise *et al*. [11], even though the effect was small in their case, they also found that "addition of social centrality measures did not significantly improve the variance explained by the model" [11].

M. I. Lopez *et al*. [12] look into can they use clustering as a way to try predicting the final mark of the student based on the students' participation on the forums. They created their own module to Moodle that collected the information they were going to use and extracted the information through that. Even though this approach will not be used, they give an idea of what kind attributes one can use when classifying or clustering the users based on forum usage. They also investigate how well their clustering and classification approaches work, depending on the approach and algorithm they got accuracy of 53% to 89%. This shows that there might be a need to use few different algorithms to try to gain better understanding of how easily the data can be classified or clustered. But as López-Zambrano *et al*. [13] found, it is hard to transfer ML models from a course to another course without the loss of accuracy.

3 HANDLING OF DATA

3.1 Raw data

Moodle platforms through its website may offer teachers way to check its logs, depending on the settings set on the specific site, these logs are very extensive about what happens on the Moodle courses page. For example, these logs contain every view of the courses page or every link clicked that takes them off the Moodle platform. The logging event in our case has the following information: time, users full name, affected user, event context, component, event name, description, origin and IP address. The time column has the timestamp of the log event with accuracy of a minute. Users full name is self-explanatory but might be missing in case the logging event was caused by someone not logged in. Affected user holds the users name if the action was directed at some user for example teacher editing someone's post or grading someone's work. Event context contains the non-user target of the action and its type for example in our case as the course uses Open forum and the forums name is "Kysy Kaikilta" the Event context is "Open forum: Kysy kaikilta". Component has the type of the of the nonuser target of the action so for example File or Open forum. Event name is the actual action for example viewed, created or deleted. Description usually gives the same information as earlier ones but in text format and instead of names it uses identification numbers for almost everything. The last two, origin and IP address, holds information about through what the log event was caused for example through web and what was the IP address of the source of logging event.

This introduces problems for our thesis, the logs do not have the information about what was posted but it contains link to the forum discussion and when the logs are downloaded the links are removed from the downloaded file. The description field is useful as it holds identification numbers. For example: "The user with id '132245678' has viewed the discussion with id '1234' in the forum with the course module id '123456'." or "The user with id '12345678' has created the post with id '1234' in the discussion with id '1234' in the forum with the course module id '123456'." This can be used to build the connections between posts or see how many checked certain discussions. It seems that for actual contents for posts some other approach is used. One possible approach here is to download the web pages of forums and as their source contains the identification number of the post it could be used to connect the post to the log event, but in this approach removed post cannot be analysed. Further searching of the Moodle pages showed that the add-on used for the forums contained option to download the posts, this could be used to divide the posts to classes as in the work of Suviste *et al*. [8] or used for sentiment analysis

as in the work of Pong-inwong & Rungworawut [10], and these could be used to further analyse the connection to the grade received by the student. Looking into the contents of these posts was considered during the process but it could have led to the work expanding beyond the original planned breadth of the work.

3.2 Building on top of raw data

As an example, for this subsection, the data used comes from the courses studied and graphs provided use the data from the fall of 2021. The data was first processed through Python with the following packages and their purpose: pandas for handling of the data, seaborn for graph generation for general data and NetworkX for analysis and graphing of social networks. Figure 1 contains simplified version of the data process up to the applying machine learning part. As seen in Figure 1, first thing done was calculating the time between two log events as a way of approximating the time used on the page or action. The values were limited to maximum of 30 minutes as a way to prevent to cases of user leaving and then some longer time later returning and thus having huge amount of time used on a action. This limit was originally chosen by intuition and if the time between the log events is drawn as in Figure 2, we can see that most of the events happen within few minutes from each other but sometimes the event are over 30 minutes and thus are limited to the maximum of 30 minutes.

After the time calculation, the data was filtered to the actions happening on the course forums, to lower the amount of data and to prevent actions outside of the forums affecting the analysis later. There is also a way to limit the data given in the extraction tool but for time calculation purposes, everything was extracted from the Moodle page. The courses studied had two forums: "Kysy kaikilta" and "Kysy henkilökunnalta", the first would translate to something like "Ask anyone" and the latter to "Ask staff". The filtering was done by simple search for "Open forum: Kysy kaikilta" in event context. Only the ask anyone forum which is anonymous was used in the analysis as the staff forums would raise the staff to central focus points as on the staff forums only staff can see the students' messages. When only forum events are left in the log Regular expression (REGEX) was used to extract the discussions' and the posts' identification numbers as they are part of the description field only and do not have their own fields.

One of the ways the original data was expanded, was through building the connections through the forums to be used in social network analysis. This could provide insight into the networks on the forums or be used to find persons that are central to the forum, which

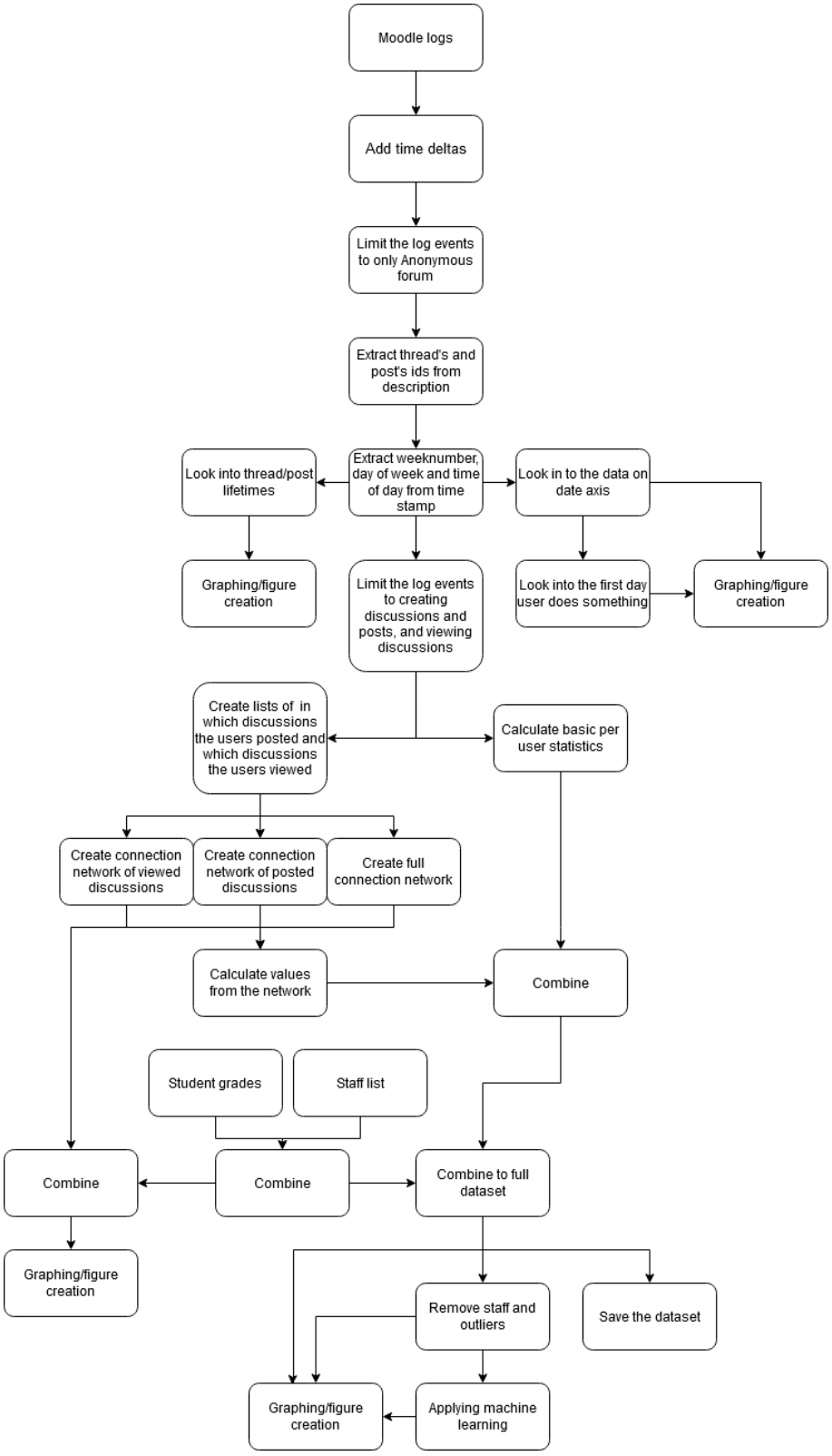

Figure 1. Simplified dataflow graph from files downloaded from Moodle to applying ML.

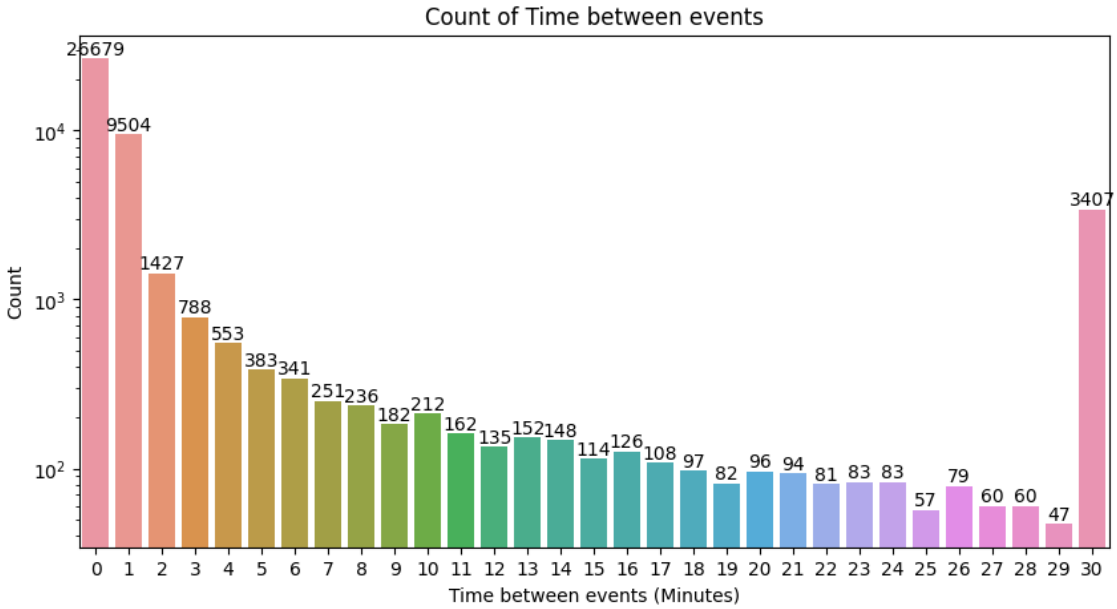

Figure 2. Count of time between events on logarithmic scale. (fall of 2021)

could probably bring some more information about the usage of forums. As stated earlier the data only points to that someone created a discussion or posted to a discussion, but we do not know the flow of messages inside those threads. To get around this, two-mode social network analysis methods as listed by Borgatti [14] can be used.

First two kinds of networks were created from the data: those who watched the same discussions and those who posted to same discussions. This was done by doing a check if the users watched or posted to the same discussions and if they did then they were considered connected. Later through searching the package used for network analysis, it was also found to include bipartite networks, also known as two-mode networks, and through that full network was created by listing who posted to which discussion and which discussion was watched by who. The first networks even though not fully showing how the networks formed were later found out be useful and thus were kept in the data flow.

Figure 3 represents the posting network of the course during the fall of 2021, each point represents a person, with colour representing either the grade received or them being a staff member and connections mean that they participated in same discussion. It can be seen that the staff have pretty central role but also some students can be in the middle. Also, it seems that there might be a connection between grade received and the position on the graph, as it seems that there are more green colours in middle.

Figure 3 shows that staff members are pretty central to the discussions, but this is simply

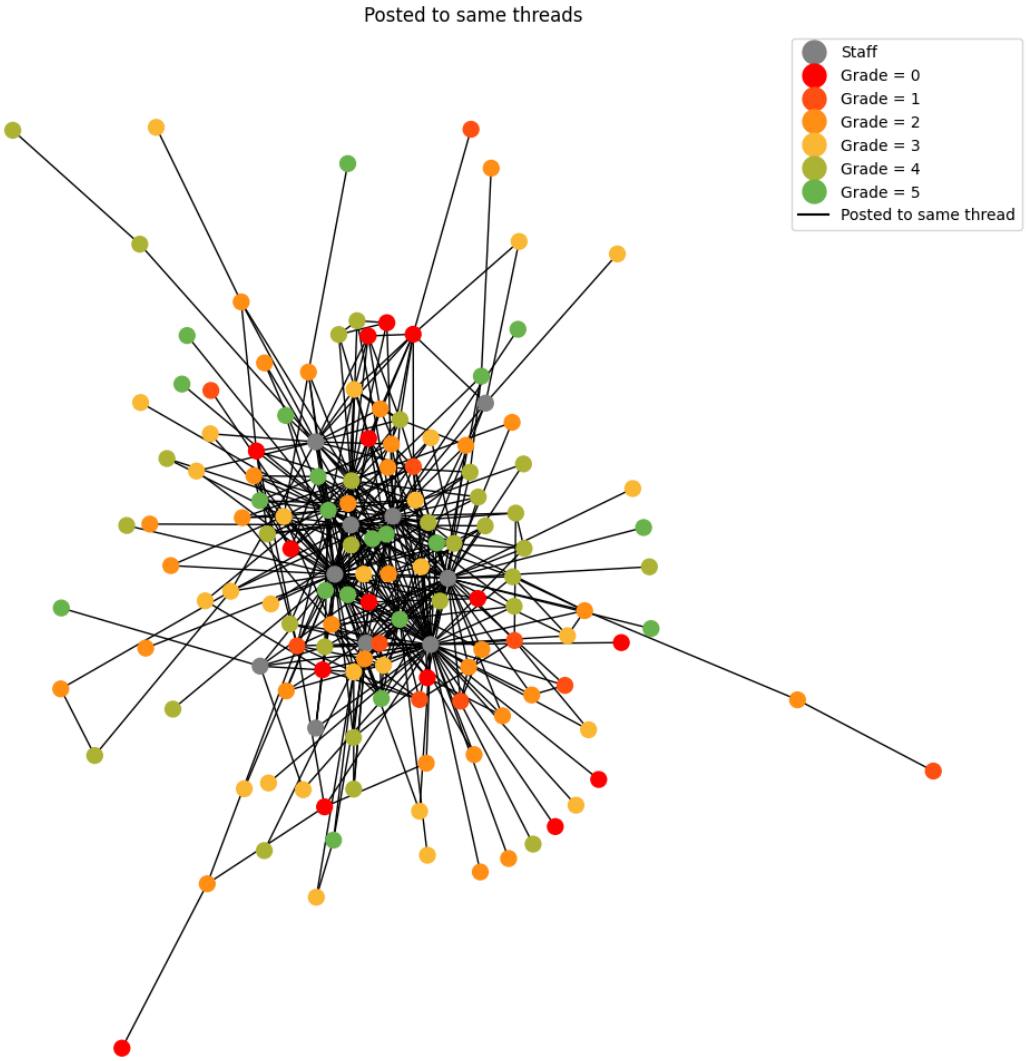

Figure 3. Persons participating in same discussions. (fall of 2021)

due to them responding if no one else answered before them noticing the post or to give the staffs' perspective on the question. Also, a group of central figures seems to rise from the network with many connections, while most of the students have only a connection or two. Same kinds of figures could be built from the viewed or full networks but due to the amount of students and their connections, the graphs get too cluttered for beneficial visual analysis, but if the amount of discussions are limited to 10 most popular discussions by the amount we can see the difference in amount of users simply viewing discussions instead of taking part in them as seen in Figure 4. This also shows the problem with the missing data that will be explained in sub-section "Course during the fall of 2021".

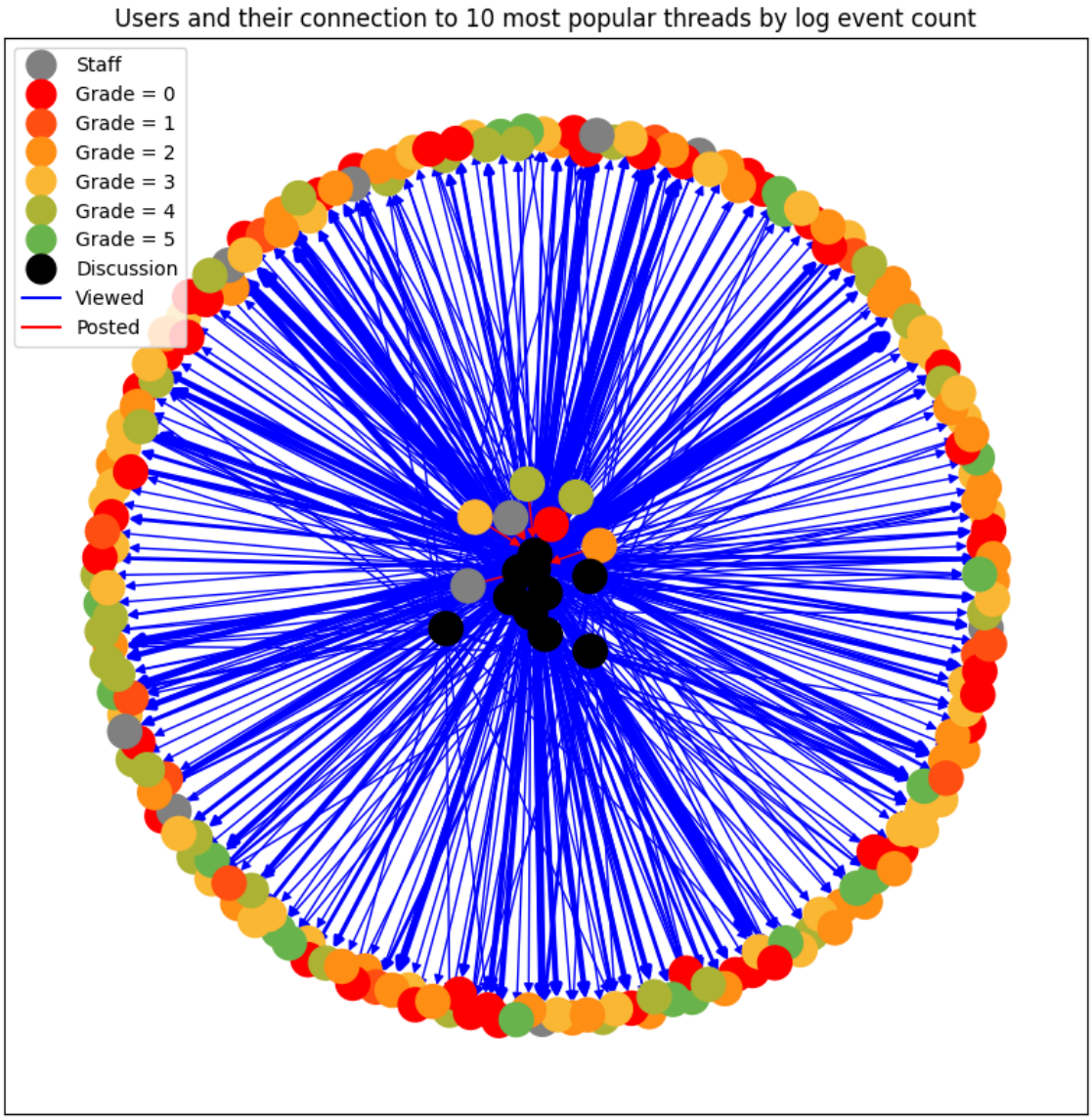

Figure 4. Actions surrounding the 10 most popular threads. (fall of 2021)

Also worth noting is that there were some threads with only single user posting. Upon closer inspection it was found that there were three reasons for these: missing data, discussions deleted or more informational posts rather than questions or discussions. This led to question being raised about deleted discussions and decision to look into the lifetime of removed threads and posts. When looking into it was found for every "Post removed" action that there was not a responding "Post created" log event, this was found out to be caused by discussion creation also creating a post but only the "Discussion created" event being logged into the log.

From these networks different kinds of values were calculated, simplest being number

of connections as a way of calculating with how many students did the student watch or post same threads. Other kinds of values relating to social network analysis were managed through NetworkX functions to calculate values for degrees, degree centrality, betweenness centrality and eigenvector centrality in the viewed, posted and full network. Outside of social network analysis values, the simpler values were calculated from the logs. here is the full list of the values calculated and how they were calculated:

1. **Discussions viewed**: Count of "Discussion viewed" log events for the student.
2. **Discussions created**: Count of "Discussion created" log events for the student.
3. **Posts created**: Count of "Post created" log events for the student.
4. **Most common weekday**: Most common weekday in the log events for the student.
5. **Most common day of time**: Most common hour in the log events for the student.
6. **Average times posted to same thread**: Average of how many times the student posted to the same discussion.
7. **How many discussions viewed**: How many different threads the student viewed.
8. **Time spent on viewed**: Total time used for "Discussion viewed" log events for the student.
9. **Time spent on Posted**: Total time used for "Discussion created" or "Post created" log events for the student.
10. **Average time spent on viewed**: Average time used for "Discussion viewed" log events for the student.
11. **Average time spent on posting**: Average time used for "Discussion created" or "Post created" log events for the student.
12. **Average times viewed the same thread**: Average of how many times the student visited the same discussion.
13. **How many discussions posted to**: How many different discussions the student posted to or started.
14. **Viewed degrees**: To how many students/staff is the student connected in the view network.
15. **View degree centrality**: Fraction of the overall students/staff in the view network the student is connected to.

16. **View betweenness centrality**: Sum of the fraction of shortest paths that pass through this node in view network.
17. **View Eigenvector centrality**: The importance of the neighbours [15].
18. **Post degrees**: Same as fourteen but from post network.
19. **Post degree centrality**: Same as fifteen but from post network.
20. **Post betweenness centrality**: Same as sixteen but from post network.
21. **Post Eigenvector centrality**: Same as seventeen but from post network.
22. **Full network Bipartite degree centrality**: Fraction of the overall discussions the student is connected to.
23. **Full network Bipartite closeness centrality**: Distance to all other students and discussions.
24. **Full network Bipartite betweenness centrality**: Same as sixteen but from bipartite network.
25. **Full network Bipartite clustering coefficient centrality**: Measure of local density [16].

4 EXPLORING THE DATA

As there are two courses studied in this work, first focus will be on the course hold during the fall of 2021 and will be used to set the baseline and then used to reflect on the following years course's data. There is some difference between the data and this will be explored more in discussion. In this work, the focus will be on three different kinds of events: "Discussion viewed", "Discussion created" and "Post created". The first is pretty self-explanatory, the second means that the user starts a new discussion and chooses the title of the discussion while the last means that user responds or writes to a discussion that already exists.

4.1 Course during the fall of 2021

The course started on 6th of September but the first date on the logs was on 28th of September, so about 3 weeks of data were lost. This loss can have significant impact and is visible in Figure 4 as some of these discussions were started before the time from where we have the logs from and thus nobody in the graph posted to these threads.

If the names are counted in the log events it can be seen that there were log events for 592 persons but when the filters are set to forums only 484 persons are left, meaning that over 100 persons who had at least visited the course site did not visit the forums during the time the logs are from. Other filtered values when per student figures are looked at are students who had less than 6 "Discussion viewed" events, were removed to limit non active users, and two outliers which had over 2000 minutes of viewing time also for classifying purposes staff was removed from the data and only added back to some Figures. If the count of the event types is looked at as in Figure 5, the event "Discussion viewed" is the most common event type while subscription related events are really rare.

There is also a huge amount of "Post deleted" events, around 20.1% of all posts are removed and 16.3% of discussions. When the time between removal and posting time is looked at, it can be seen that the huge majority of these are within two minutes of the posting event, we can also see same kind of spike in the first minutes of starting a new discussion and removal of discussion, but there is also smaller peak around five to ten minutes as seen in Figure 6, few removals seem to be nearing the 30 minute limit at which point the student cannot anymore edit or remove their own post or discussion. This will be more explored in the discussion as the reasons for these removals cannot be inferred

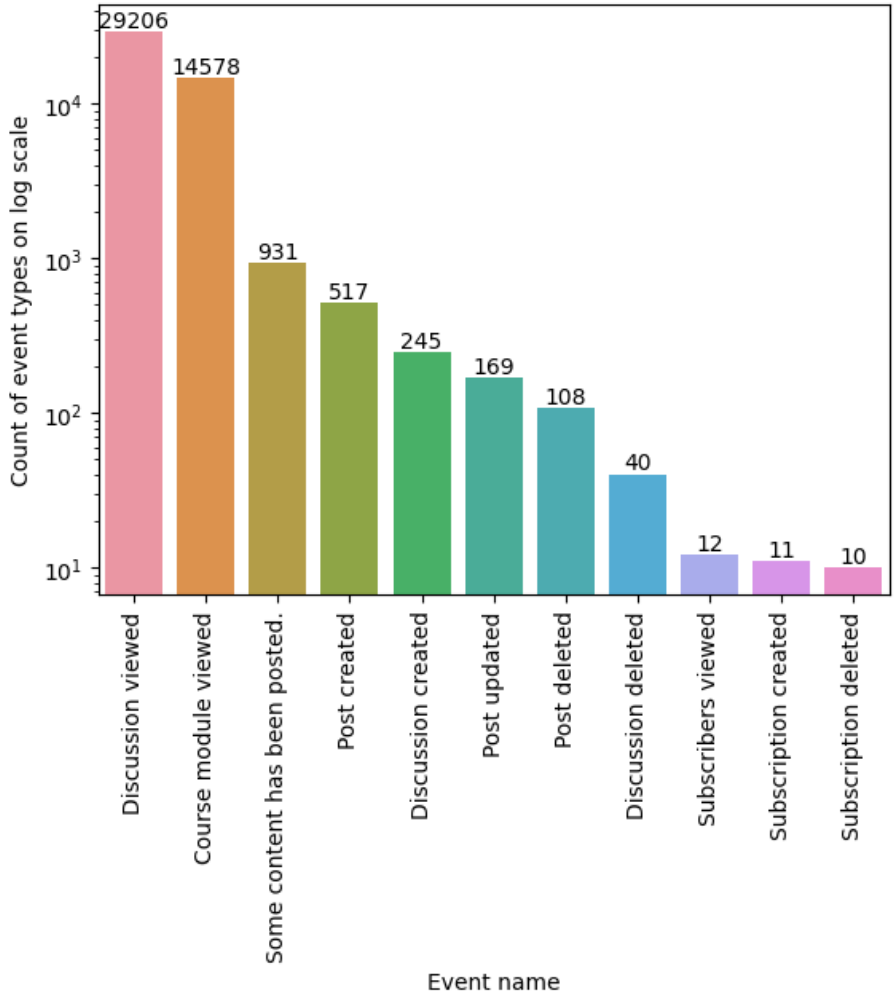

Figure 5. Count of event types on log scale. (fall of 2021)

only from logs.

Another important thing to look at is the distribution of the grades as seen in Figure 7. Few persons were dropped from this graph and from the later parts as their accounts were deactivated during their course leading to a missing grade and thus could be problematic to classify. The graph almost seems like skewed distribution with mean of two but with zero and one switched around and low number of staff.

Even though the event type count Figure 5 earlier had the y scale on logarithmic scale, graphing the events against time axis shows that they really are not even on the same scale. As the documentation of the logs was really limited, what causes the discussion viewed is not well defined, does reloading a page cause a discussion viewed event to be logged or does clicking link on discussion to show who the person responded to count as one? By limiting the events so that a single user is limited to one log event of event type per

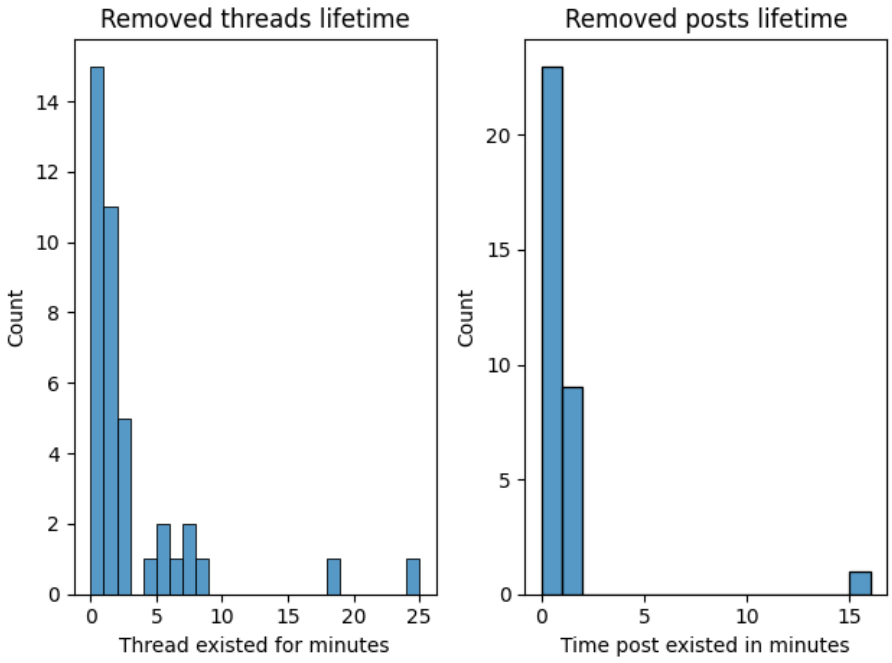

Figure 6. Removals of discussions and posts. (fall of 2021)

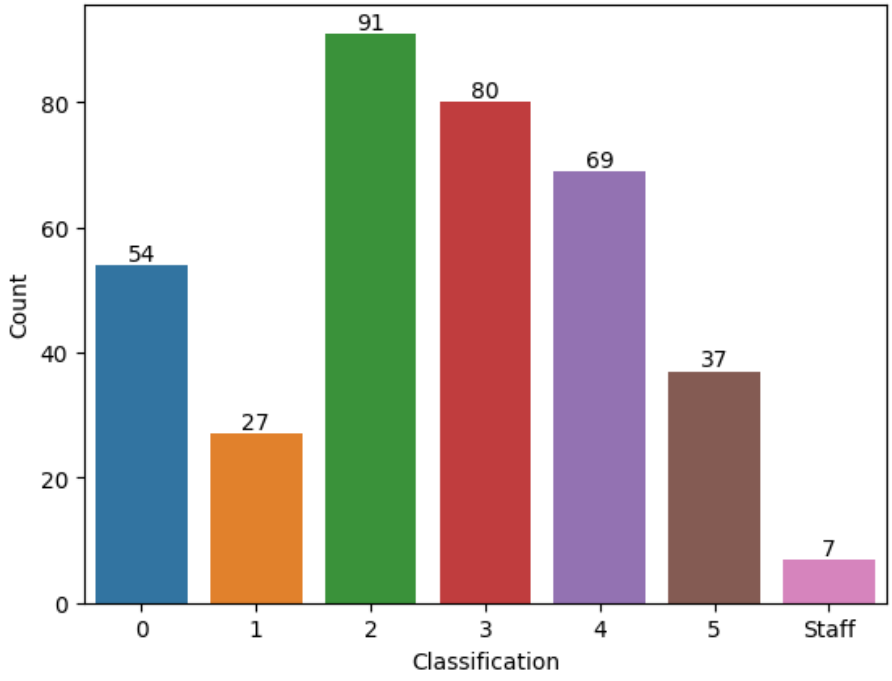

Figure 7. Bar-graph of the grade distribution. (fall of 2021)

day per discussion leads to the huge amount of the Viewed discussion events to disappear while other events suffer smaller decrease as seen in Figure 8.

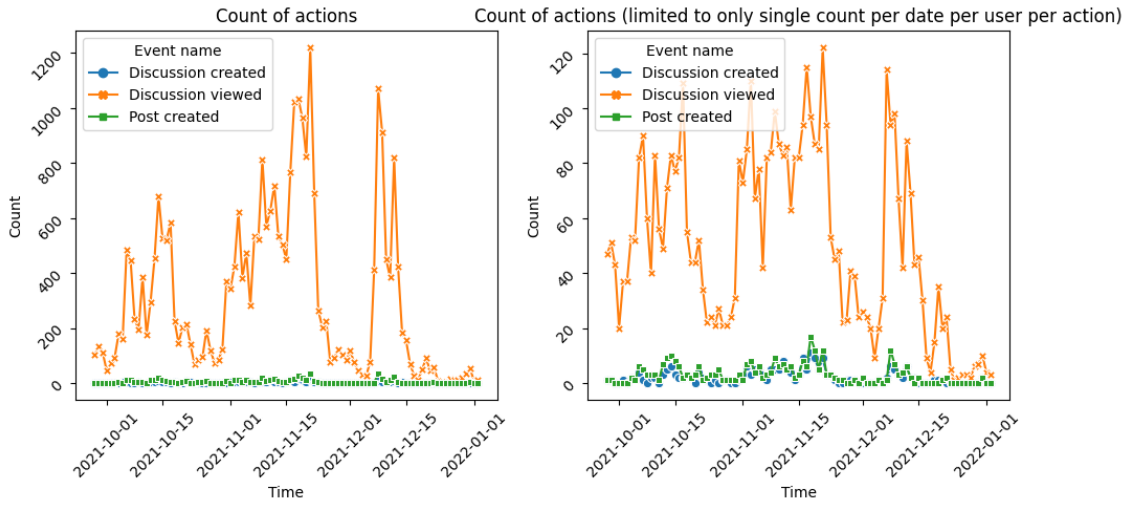

Figure 8. Timeline of with daily event counts. (fall of 2021)

This shows that purely following the thread viewed action as a measure of the usage of the forums can be misleading, and some other measure such as unique users per day might be a better measure or if the first time user does something on the forum is counted as in Figure 9 the important dates become even more visible. On the course studied the assignment can have three phases: first phase is noted in the Figure as "Assignment return is possible" and during it the students can return the assignment for grading, after this phase the assignments are graded and if they have serious mistakes or problems, they are returned to the student in the beginning of second phase, noted in the Figure as "Fixed assignment return is possible". After this phase they are again graded and if they still have mistakes or problems are given back to the students in the beginning of "Final assignment return is possible". In Figure 9 these phases relating to the final assignment and the increases in the first-time users doing some action be seen though this might be tainted little bit by the missing data in the beginning of the course.

As seen in the Figure 4 most of the students do not write on the forums and only look what others say there, and if simple histograms of how many times the students view threads and posts to threads are compared, this becomes even more visible as seen in Figure 10, though the graphs have similar shape the scales are way different. if the discussions created vs posts created Figure 11 is looked at, the most of the staff can be very easily identified by high amount of posts versus discussions created and in Figure 12 it can be seen that on staffs median count of discussions viewed is high when compared to students.

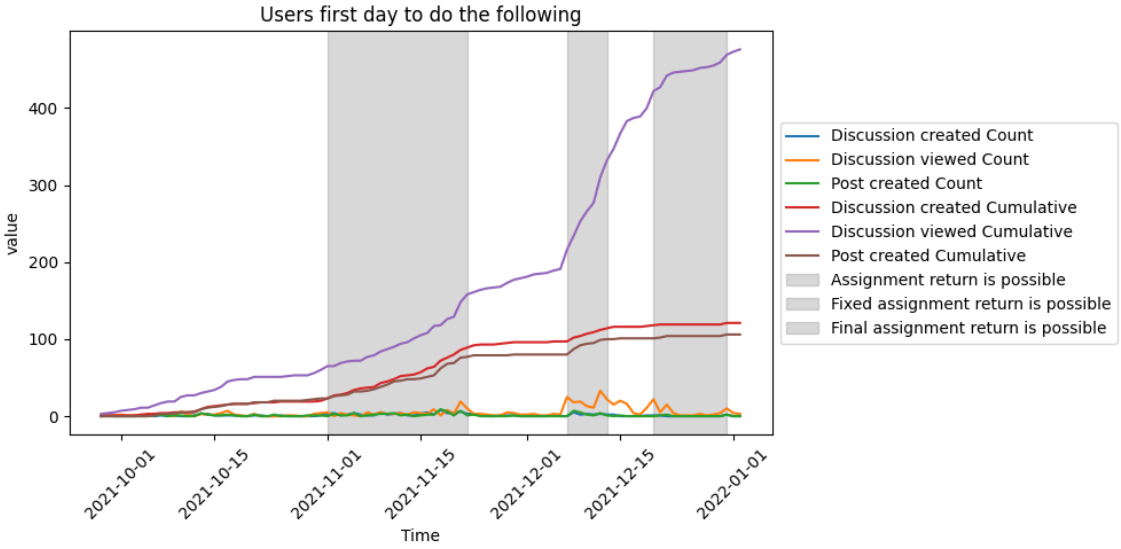

Figure 9. Users first time doing a action. (fall of 2021)

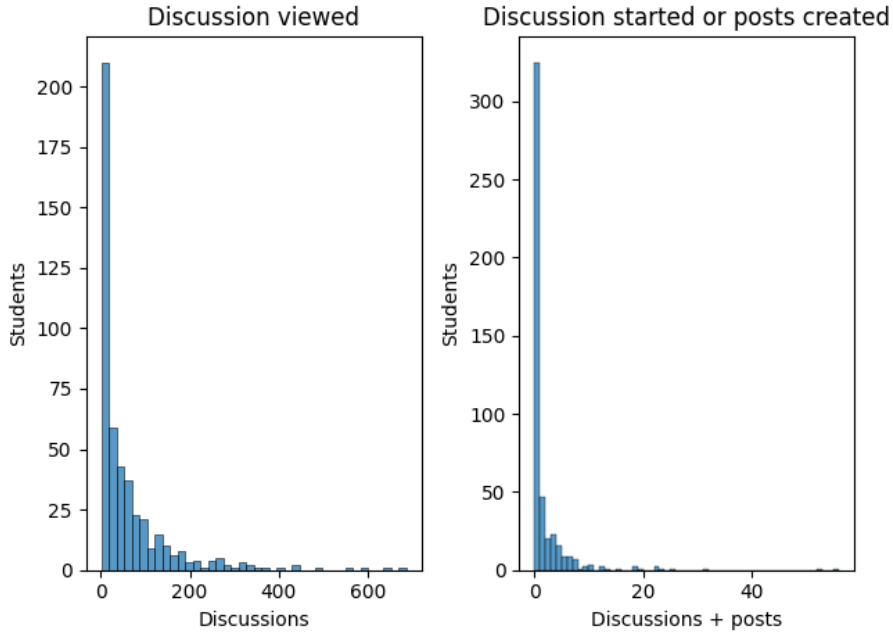

Figure 10. Histograms of viewed discussions and discussions started and posted to. (fall of 2021)

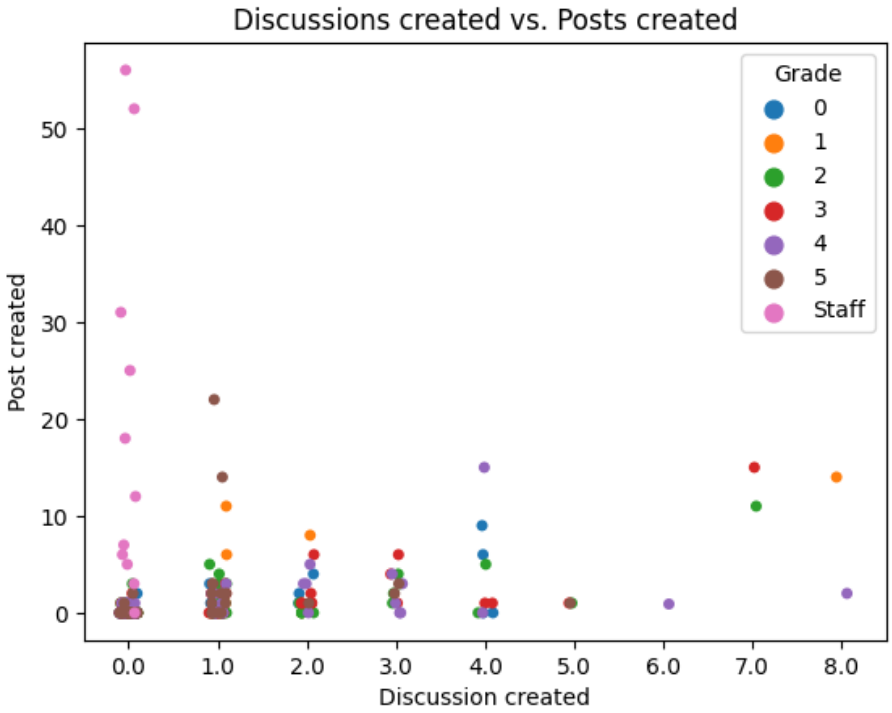

Figure 11. Amount of created discussions vs posts. (fall of 2021)

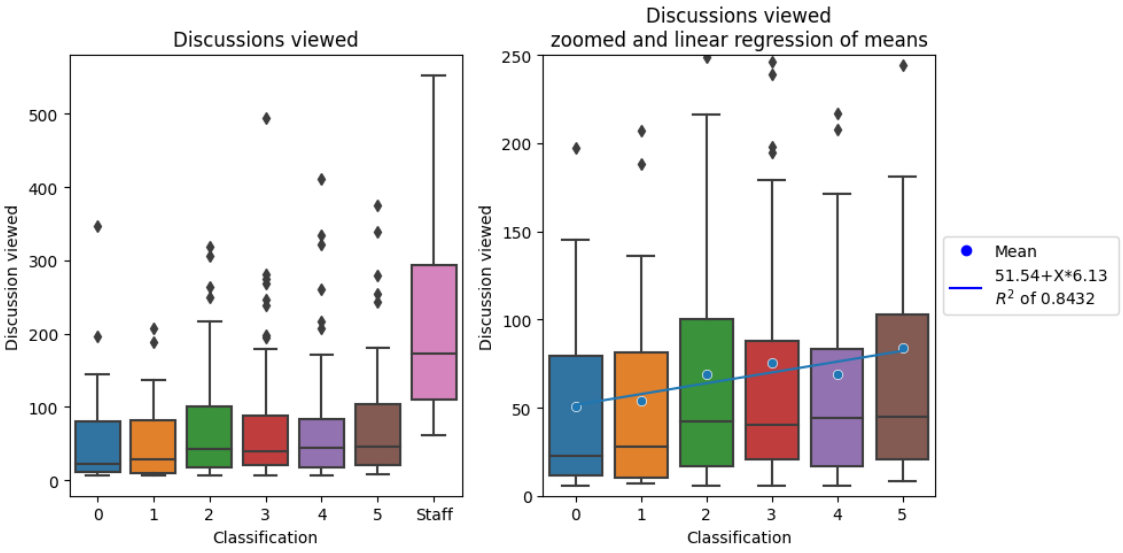

Figure 12. boxplot of discussions viewed and linear regression. (fall of 2021)

In Figure 13 it can be seen that there are two peaks of high usage, Sunday-Monday night and Wednesday around 5 p.m. These two peaks also differ little bit from each other as the posting peak is little bit higher during the Wednesday peak, this might be caused by the deadline as it is on Monday morning, meaning that the users on Sunday are rushing to get the coding tasks done and are just searching the forums for answers and do not post as much. This could also be partly caused by staff mainly answering during the weekdays and working hours guiding the discussion time and partly being part of the discussion during weekdays.

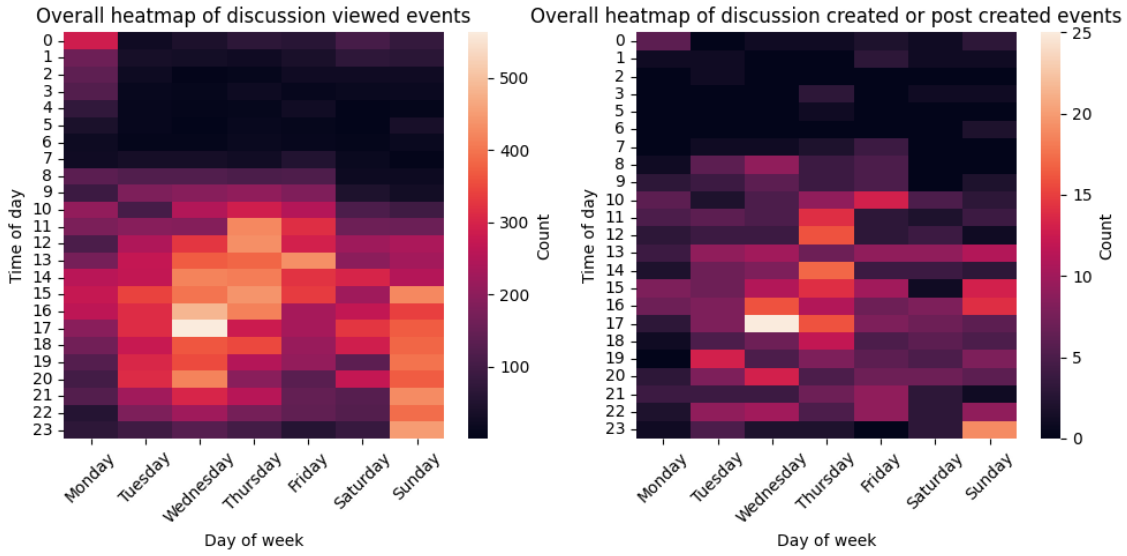

Figure 13. Heat maps of forum usage. (fall of 2021)

When the data is divided into following 7 "classes" : grades from zero to five and staff, for later classification and looked through simple box-plots show that some of the variables on average have a relation with grade received but most of the time these are not linear or the classes easily distinguishable from each other. For example, the when the discussions viewed is drawn as box-plots and then averages are added, one can do a simple linear regression of the means, in this case it shows a simple linear regression between grade and discussions viewed with the following values as seen in Figure 12. The high R^2 value shows that the regression could be considered strong but with low multiplier its overall effect is very small and this only shows that on average and due to the high variance inside the classes this cannot be used to draw conclusions about specific people only about the classes as general.

Another two variables that could have an impact on grade and the connection between the grade and them is visible on histograms is most common time of day and the most

common weekday as in Figure 14. When looked at and a kernel density estimation is drawn on top it shows that there seems to be clear movement of the peak of the distribution to earlier in the week and earlier in the day. Especially the grade 0 shows huge peak on Sunday. But with lower number of samples, certain grades become more likely to be affected with outliers. The full description of the data can be seen in Appendix 1

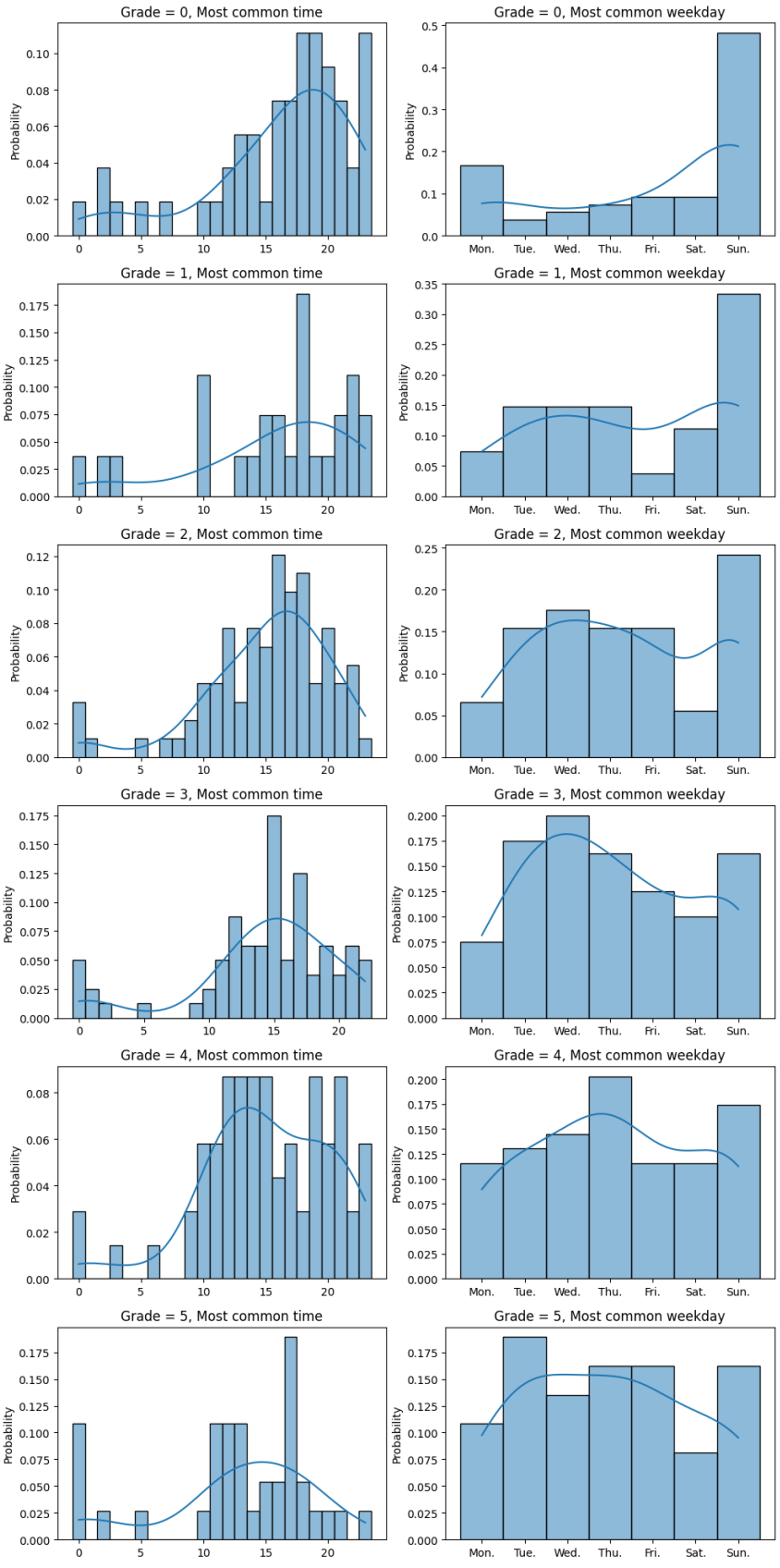

Figure 14. Histograms of most common time of day and weekday by grade. (fall of 2021)

4.2 Course during the fall of 2022

The amount of participants the course had during the year fall of 2022 had grown, with amount of unique names in log events going from 592 in previous year to 777 in the fall of 2022, and when filtered to only contain log events relating to forums this time 167 names were dropped, which had also grown from the 108 names in the previous year. One unexpected statistic was almost similar counts of log events as seen in Figure 15, only the subscription and course module updates have huge differences, this was unexpected due to the mentioned growth in student count and due to the loss of data from the first few weeks of the fall of 2021. Figure 15 also suggests that at least during the first course the forum module was opened more but as the discussions viewed was similar, the students checked less discussion per opening of the module.

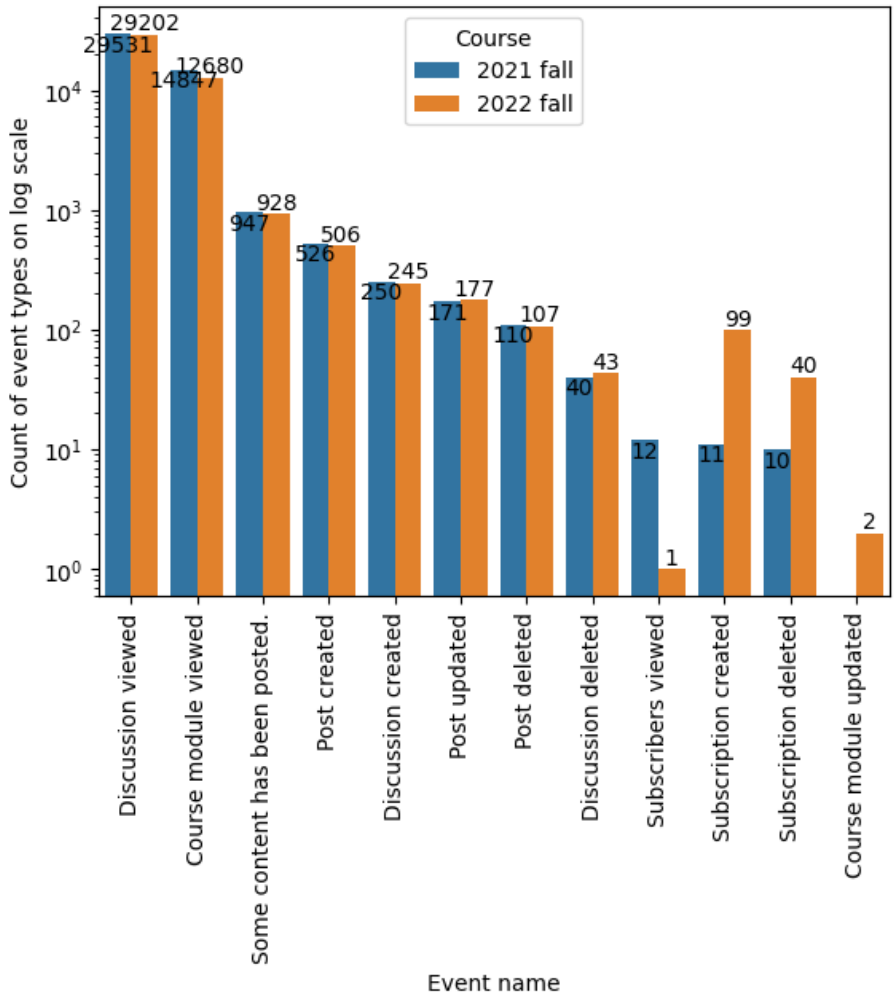

Figure 15. Log event counts during both courses on log scale.

Another graph that shows how similar the courses were considering the forums is the first-time users do something on the forums, as seen in Figure 16 and when compared to Figure 9 it can be seen that the figures seem very similar. The biggest change seems that even though the gradients seem similar, the big growth seen on the earlier course during the assignment returns was possibly partly exaggerated by the missing data from the beginning of the course.

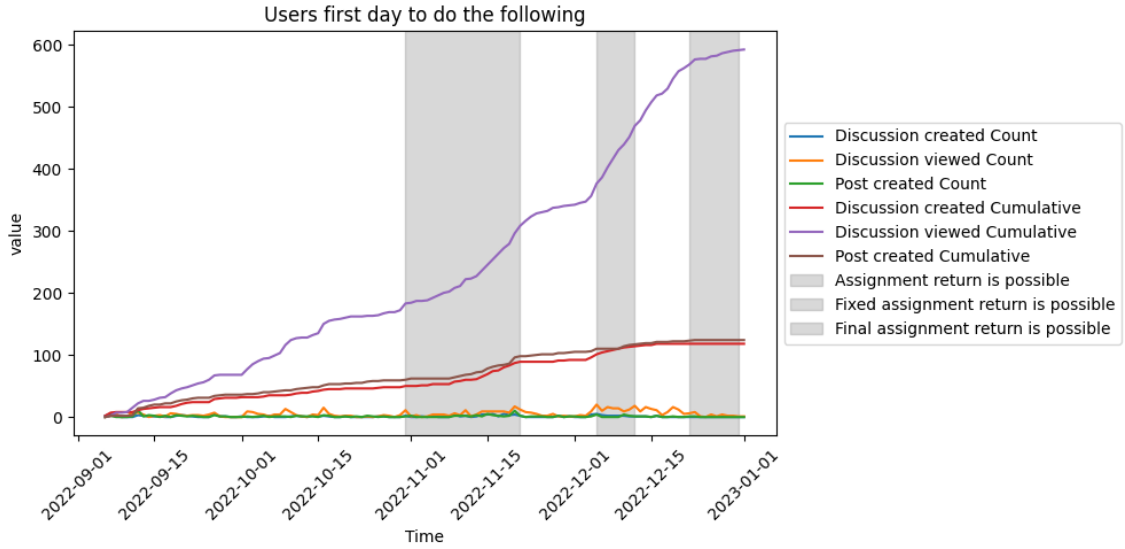

Figure 16. Users first time doing an action during the fall of 2022.

Figure 17 shows the grades received on both of courses and a shift in the received grades is visible. This shift in grades is discussed more in the discussion part. Figure 18 shows the people who participate in the same discussion and when compared to the previous year's Figure 3 the grades seem more spread out and there is a tight group of students with lower grades in the fall 2022 course.

Most of the statistics outside of the grade distribution seem similar but due to the lower number of samples in some grades they are far more susceptible to outliers affecting them, especially the large growth of grade 0 can make the other classes harder to distinguish from the noise caused by the grade 0 students. One change was easily visible when comparing the graphs, the most common days of the week users were active had some change. As seen in Figure 19 the most common days during the fall of 2022 seeming to shift towards earlier in the week and the Sunday rush to finish the tasks is again clearly visible in the view events when compared to the discussion or post created events. This time the amount people posted also seemed to be lower on Sunday when compared to the earlier year. Figure 21 shows similar shift from grade 0 visiting forums on Monday and

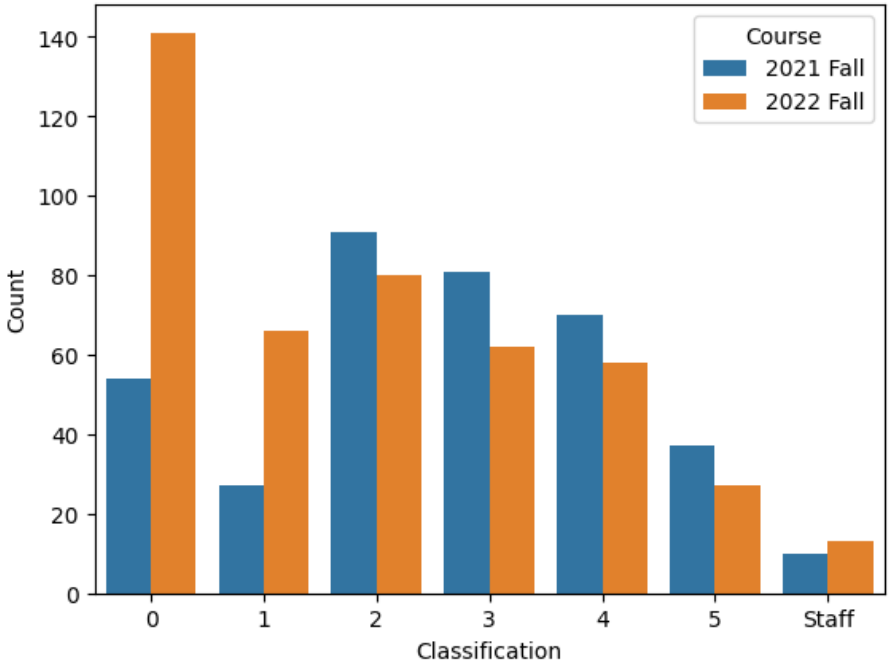

Figure 17. Grade distributions on both years

Sunday to it being to more common on rest of the weekdays with the exception of grade 4 where Monday still is the most common by far, in other dates the shift is not prominent as in the previous year's course as seen in Figure 14.

Figure 20 shows discussions viewed and their means' linear regression for both years, the fall 2022 course still has same kind of correlation but this time the it is quite bit lower and the strength of the correlation is really low, with R^2 only being 0.3317 when compared to the previous years 0.8927. This lower correlation strength might be explained by lower number of samples in some grades causing them to be more susceptible for outliers to change the mean. But with both years showing some correlation between the average of discussion viewed and grade, it is highly likely to exist, though the strength of this correlation is debatable especially with the low R^2 value of the second course.

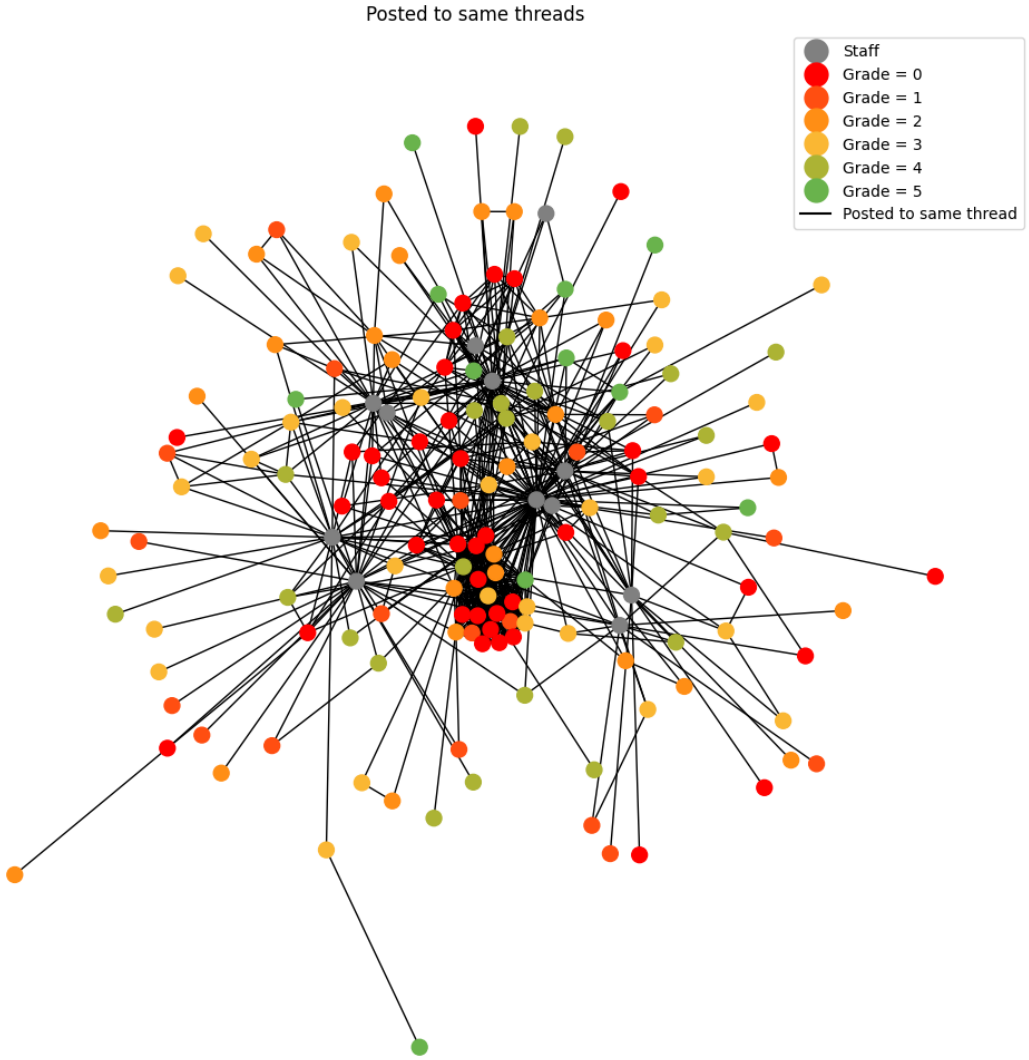

Figure 18. Persons participating in same discussions during the fall 2022 course.

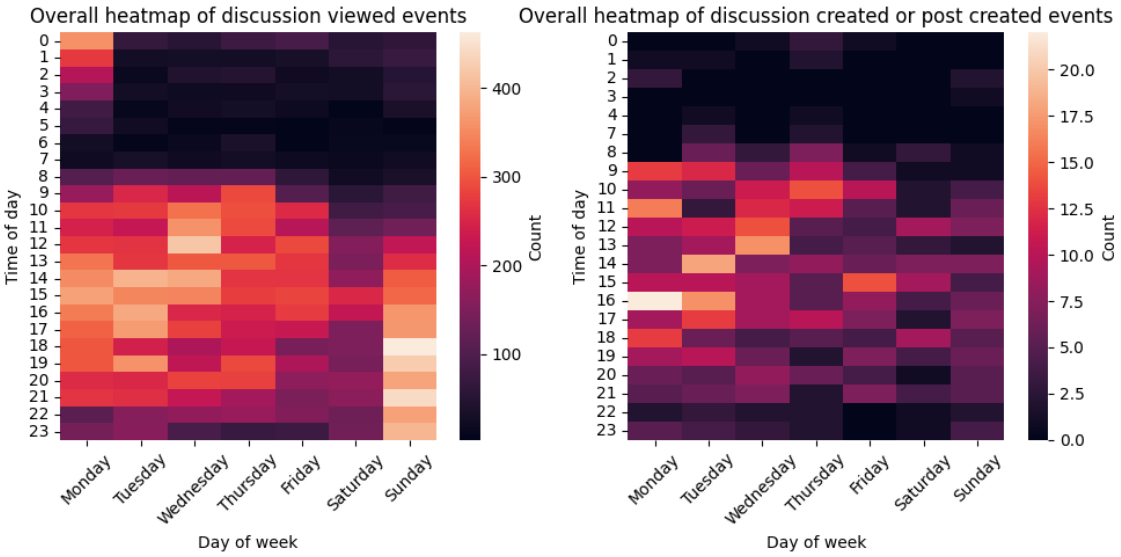

Figure 19. Heatmaps of usage during the fall of 2022 course.

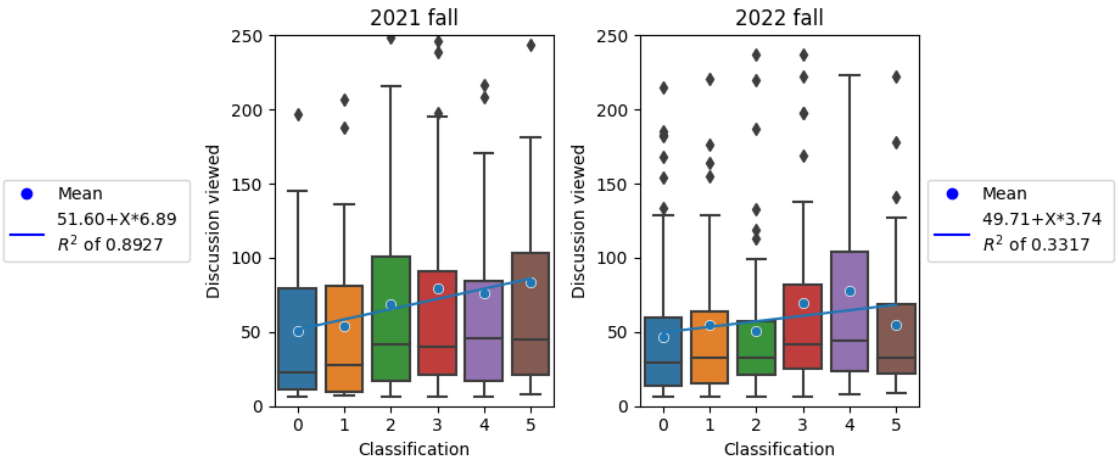

Figure 20. boxplot of the discussions viewed and linear regression of the means. (both courses)

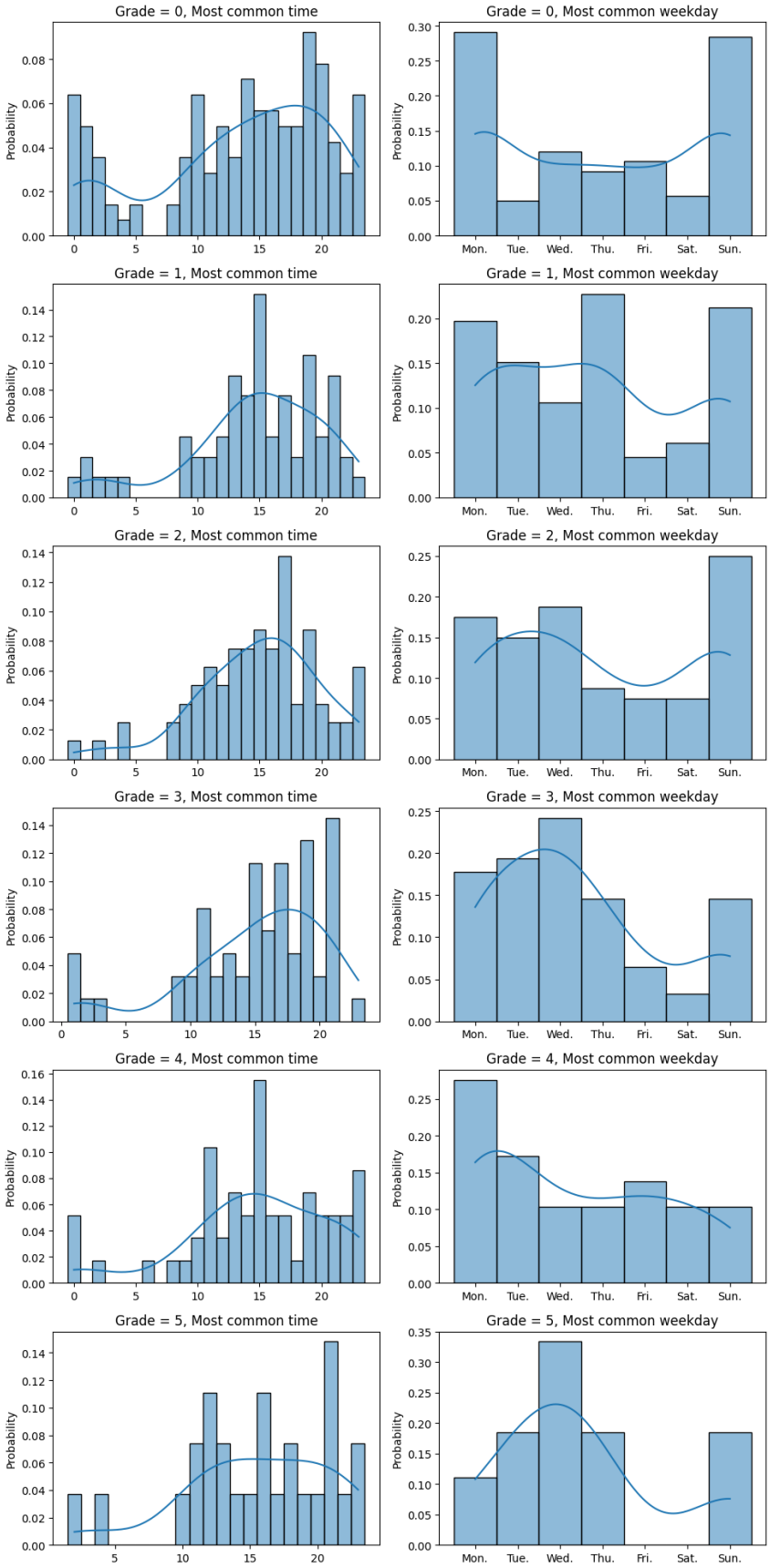

Figure 21. Histogram of most common times of day and weekday by grade. (fall of 2022)

5 APPLYING MACHINE LEARNING

This section focuses on ML approaches and as stated in the goals of this thesis there are two main areas of approach, clustering and classifying. All of these approaches were done using MATLAB. First area to approach was clustering and this was done through k-means algorithm. After clustering some Principal Component Analysis (PCA) was done to see if the data set's dimensionality could be dropped from the current twenty-six dimensions and to see how the variation is distributed among the principal components. After PCA many different classifying approaches were attempted either through applets or functions offered in MATLAB's toolboxes. For example, these approaches include but are not limited to: ensemble of learners, Support Vector Machine (SVM) and Partial Least Squares regression (PLSR).

5.1 Clustering and analysis of course during the fall of 2021

For k-means algorithm 6 dimensions of the data were chosen: posts created, discussions created, discussions viewed, time spent viewing discussions, how many different discussions the user viewed and posted to. These dimensions were then standardized, such that the mean equals to zero and standard deviation would be one. For calculation MATLAB's evalclusters function was used with evaluation done for 2 to 15 clusters and random seed being set to 655321 between each run. With the fall of 2021 data this approach measured on silhouette, Gap and Calinski-Harabasz index did not converge on any certain solution with each one providing alternative answers as seen Figure 22, with Calinski-Harabasz recommending 5 clusters with 2 clusters scoring little bit lower, while Silhouette recommending in order: four and two, and gap recommending 14 and having smaller local maxima of 5.

In order to see what kind of clusters rise from these, the following cluster amounts were selected to see what their cluster centres are: two and five. The centres were then multiplied by the standard deviation of the original data set and the mean added to transform them back from the standardization. The centres of the clusters can be looked at Table 2 which also contains the clusters found from fall of 2022 course. As one might expect in the trivial case of two clusters, one of the clusters seems to represents the students who mainly only watch the threads and the other those who participate, but in 5 cluster case there are more complex behaviour groups:

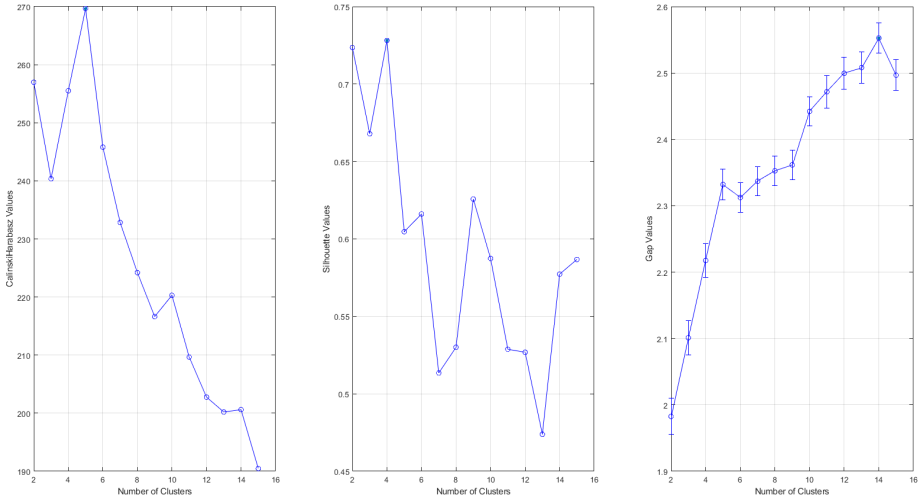

Figure 22. Result of k-means evaluation. (fall of 2021)

1. **1st group (N=218):** on average does not participate in threads but looks at few threads.
2. **2nd group (N=6):** on average very active group, takes part in many threads and spends high amount of time on forums.
3. **3rd group (N=79):** on average little more likely to participate than 1st group but around four times more time spent and discussions viewed.
4. **4th group (N=32):** on average starts many discussions and participates in them but on other statistics similar to 3rd group.
5. **5th group (N=23):** on average starts a discussion, highest number of discussions and different discussions viewed, but not so much time spent.

For PCA, MATLAB was used. When the fall of 2021 data set is run through it immediately gives a warning that some of the columns are linearly dependent and thus it uses twenty-one first components in calculations. This means that the data set is actually 21 dimensional through the 25 columns. When the cumulative variance per principal component is graphed as in Figure 23, it can be seen that the first principal component achieves already over 95% of the variance while the second principal component together with the first capture almost 99% of the variance and with three first principal components over 99.75% of variance is explained. The Bi-plot shows that most of the variables move the values among the first principal component leading to their variable names fusing together

and being unintelligible but the variables relating to viewing add movement among the second principal component. This shows that there is quite high correlation between most of the variables, as could be expected as if the user posts their connection count raises, their centrality raises, etc. And View degrees, discussion viewed count and viewed threads also seem to correlate as viewing more discussions means that they are more likely to be connected with other viewers and viewing unique threads.

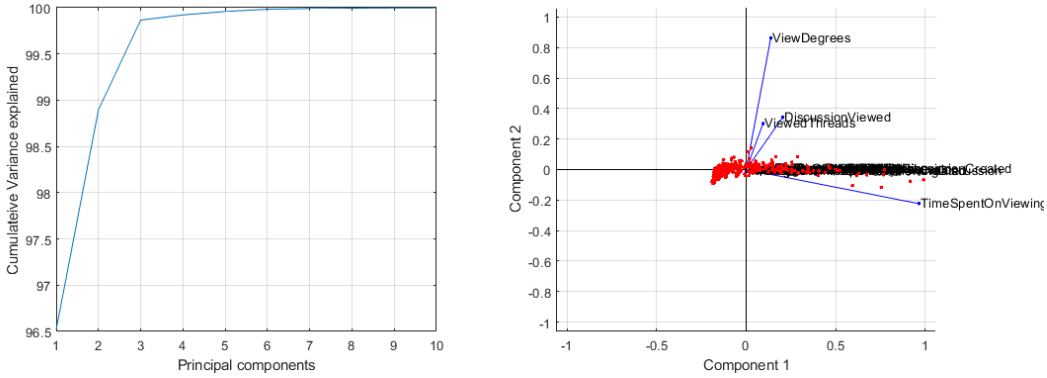

Figure 23. Cumulative variance explained per principal component and bi-plot of the two first principal components. (fall of 2021)

5.2 Clustering and analysis of the course during the fall of 2022

Again the same dimensions were run through evalclusters in MATLAB, and again the methods for evaluation of cluster counts seemed to disagree with how many clusters there were as seen in Figure 24. Especially standing out is the Calinski-Harabasz evaluation at thirteen clusters, Calinski-Harabasz seems to suggest two, five or thirteen clusters, while Silhouette value seems to suggest 3 being optimal, with two, four and six clusters following with lowering values and Gap shows slow growth with increasing cluster count with six and 14 clusters standing out.

Then the clusters were formed with the trivial case of two and non-trivial case of six clusters. When Table 2 with both courses clusters is looked at, it shows similar clusters even with the different number of clusters. In the case of two clusters again is simply explained by the division of users to those who participate and those who do not participate actively but in the case of six clusters the groups central seemed to be following:

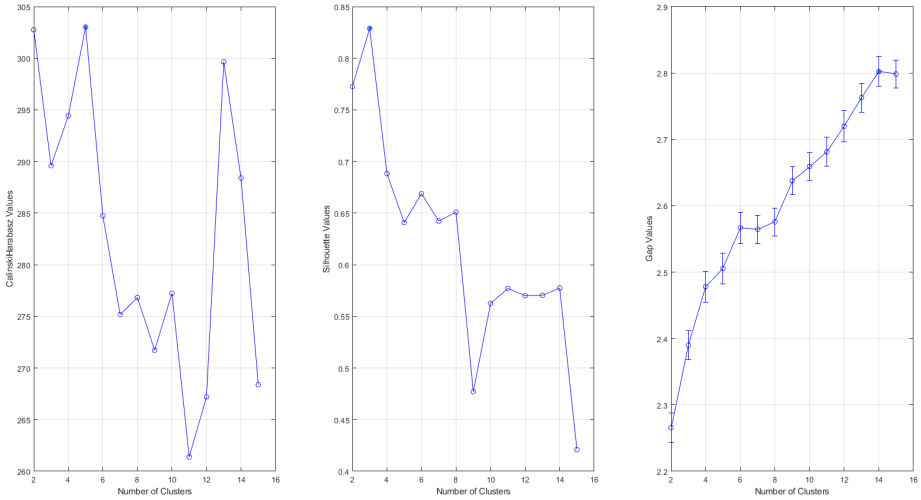

Figure 24. Results of k-means evaluation. (fall of 2022)

1. **1st group (N=274):** on average does not participate in threads but looks at few threads.
2. **2nd group (N=3):** on average very active group, takes part in many threads and spends high amount of time on forums.
3. **3rd group (N=77):** on average little less likely to participate than 1st group but around four times more time spent and discussions viewed.
4. **4th group (N=39):** on average starts two discussions and participates in three but on other statistics similar to 3rd group.
5. **5th group (N=26):** on average not likely to start a discussion, highest number of discussions and different discussions viewed.
6. **6th group (N=15):** on average starts many discussions and makes many posts to different threads, medium number of discussions and different discussions viewed.

If the PCA is looked at, the biplot Figure 25 looks very similar to the previous year's Figure 23, with only values slightly changing. While there was only small change to the biplot, the cumulative variance explained shows that the variance is slightly more focused on the two first principal components as they explain over 99.5% of the variance, with three first principal components explaining again over 99.75%.

Table 2. Table of the k-means centres for both courses.

Cluster	Discussions started	Discussions Viewed	Posts Created	Time spent viewing	Different threads viewed	Different threads posted to	Number of students (%)
2021 1/2	0,27	38,38	0,21	161,82	26,82	0,37	281 (78.5%)
2022 1/2	0,23	36,22	0,25	186,78	25,19	0,37	367 (84.6%)
2021 2/2	2,18	175,69	3,04	748,00	92,36	3,57	77 (21.5%)
2022 2/2	2,31	167,54	2,78	963,66	88,69	3,45	67 (15.4%)
2021 1/5	0,18	24,15	0,16	98,23	18,16	0,27	218 (60.9%)
2022 1/6	0, 15	23,27	0,13	114,79	17,19	0,23	274 (63.1%)
2021 2/5	4,67	247,17	15,17	1143,50	111,17	13,33	6 (1.7%)
2022 2/6	7,67	165,67	15,33	1199,33	71	15	3 (0.7%)
2021 3/5	0,54	102,84	0,47	461,73	63,47	0,75	79 (22.1%)
2022 3/6	0,17	81,62	0,27	437,44	53,71	0,32	77 (17.7%)
2021 4/5	3,38	98,13	3,16	475,94	52,44	4,41	32 (8.9%)
2022 4/6	1,92	83,10	1,94	437,72	48,08	2,85	39 (9.0%)
2021 5/5	1,04	274,00	1,26	1003,70	144,78	1,74	23 (6.4%)
2022 5/6	0,46	248,19	1,19	1416,5	129,38	1,23	26 (6.0%)
2022 6/6	4,93	111,13	4,53	698,67	59,13	6,13	15 (3.5%)

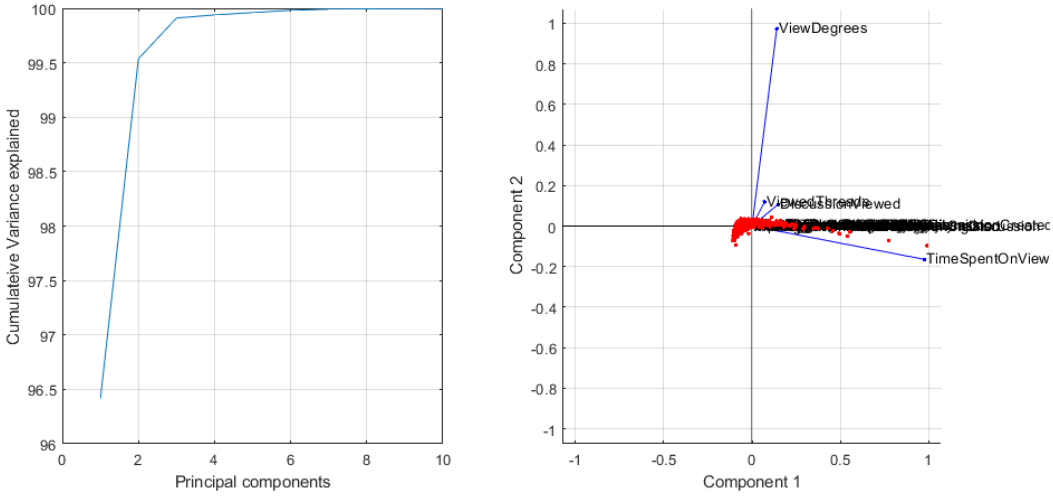

Figure 25. Cumulative variance explained per principal component and biplot of the two first principal components for the data of 2022 fall.

5.3 Classifying fall of 2021

For classifying the grade received at the end of the course is used as the target variable. For training and testing split due to low number of students, 258 for during the fall of 2021, a 90% for training and 10% for testing split was decided upon. Stratified sampling was used to keep the distribution of grades in training and testing split as close matches as possible. Approaches used through code were binary classification tree [17], regression tree [17], knn classifier with k equalling from one to four, ensemble of learners, SVM [18] and PLSR [19]. And from the classification learner applet all the classifiers were attempted.

Some of the classifiers just give up and guess only one of the grades and thus achieve accuracy of 20-25% while some of them actually try with wide range of success, some of them are as bad as guessing same number while some of them perform little bit better. The accuracy with these approaches changed every time the samples were re sampled pointing to the fact that there is small number of samples and most of them are hard to distinguish from each other. Best results have been achieved through the app with ensemble of learners and cubic knn or through code with PLSR, with them at best achieving a 37.1% of accuracy. With the PLSR as with all classifiers we can look at the confusion matrix as in Figure 26, which shows a different run with lower accuracy but also the variables importance in the projection can be drawn as in Figure 27, as a way of looking into which variables are more important for the PLSR.

Figure 27 shows how much each of the variables affect the projection in the model. Especially important seems time spent on viewing discussions, with the number of viewed threads following and most common time of day coming as third. While the social network analysis values seem very low. Overall, the values over one can be considered important.

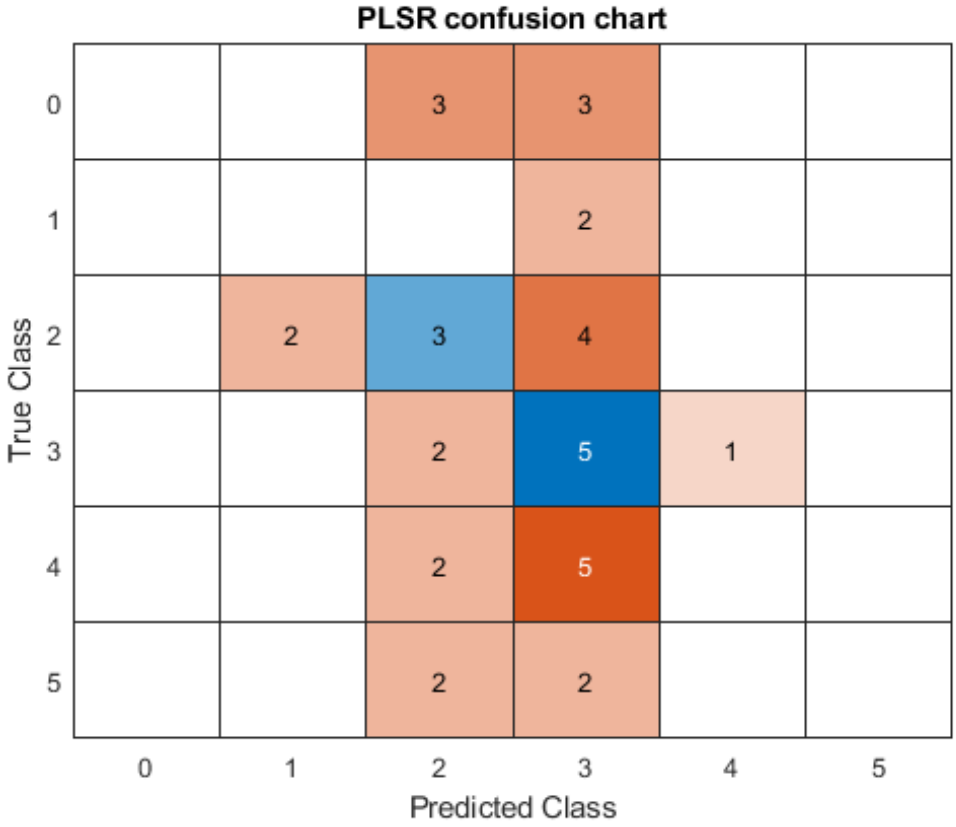

Figure 26. confusion matrix of PLSR. (fall of 2021)

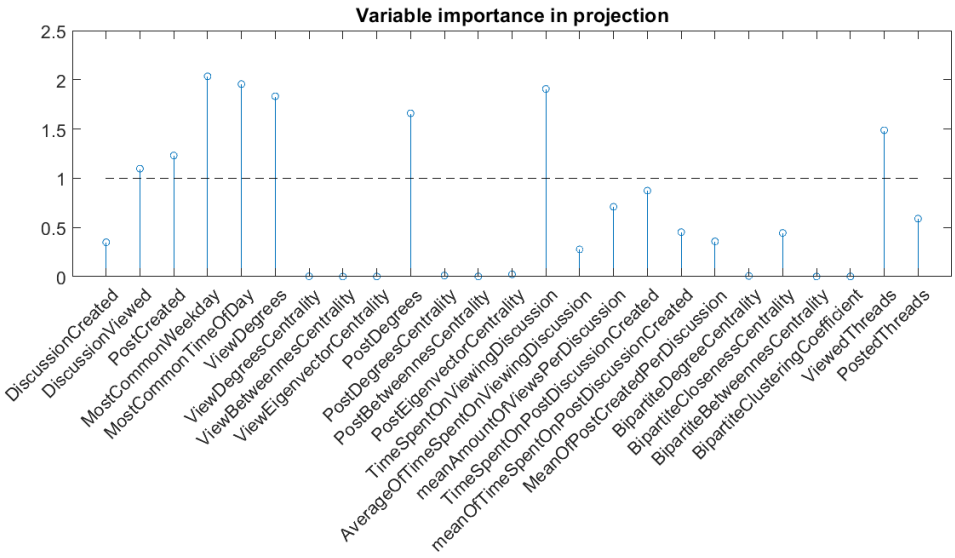

Figure 27. Variable importance in projection of PLSR. (fall of 2021)

5.4 Classifying fall of 2022

From Figure 17 can be seen that there was huge amount of 0 grades, this exaggerated the problem with the classifying 2021-2022 course with most classifiers simply giving up and grading everything as 0 and thus achieving accuracy of around 35% depending on the split of training-test data. When using PLSR an accuracy of 16.3% was achieved with the following confusion chart Figure 28, showing that the PLSR actually tries to predict the class and does not predict zeros.

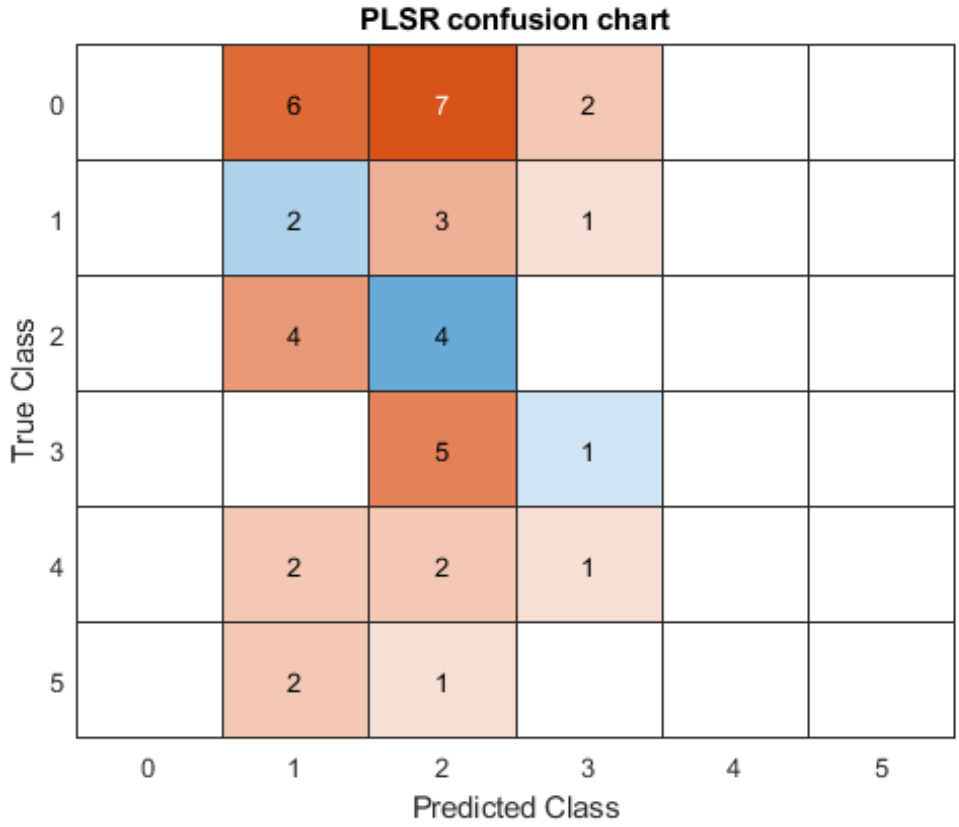

Figure 28. confusion matrix of PLSR. (fall of 2022)

Figure 29 shows variables importance in projection in PLSR and from it can be seen that there is some similar values when compared to the earlier courses Figure 27. When cross-referenced the following variables seem more important than the rest: Most common day of week, Time spent on viewing discussions and Viewed discussions. It also seems that centrality measures have very little effect on the grade received at least when these courses are considered.

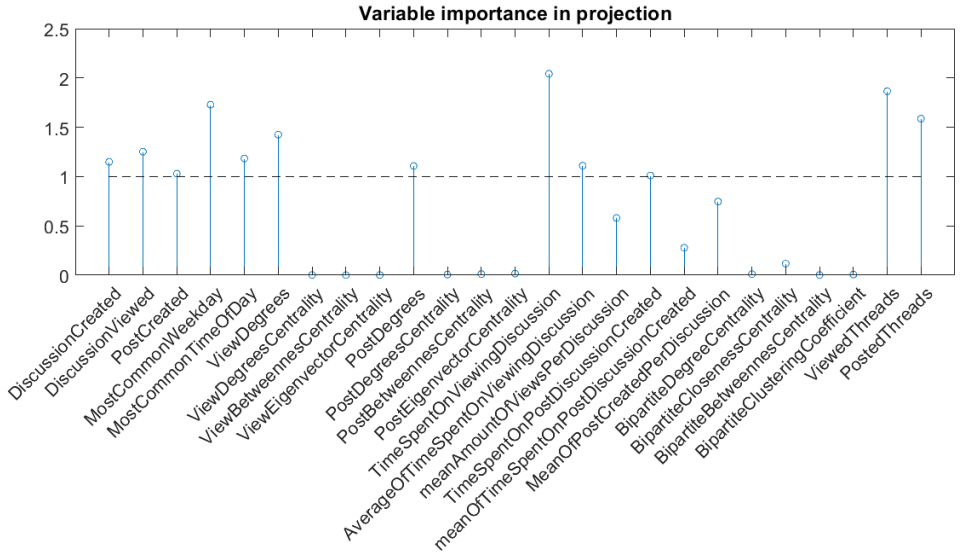

Figure 29. Variable importance in projection of PLSR. (fall of 2022)

5.5 Trying to improve the accuracy by time shifting the fall of 2022 data

With the most common weekday having high importance in projection as seen in Figure 27 and Figure 29, and with activity peak seeming to continue on Sunday evening to Monday night as seen in Figure 19 and Figure 13 it was decided to see the impact of shifting of timestamps by six hours. This was achieved by simply creating a copy of the timestamps column and subtracting the six hours and then counting the most common day again. Figure 30 shows how the most common days especially on lower grades shifted. And when the dataset is driven through the same processes in MATLAB, most of the results stayed the same or had a minor improvement. There was an improvement in the accuracy of the PLSR with the accuracy rising from 16.3% to 25.6% as seen in Figure 31 showing PLSR's confusion matrix when compared to the earlier Figure 28. Figure 32 shows how the importance of the time shifted variable is higher than the normal most common weekday.

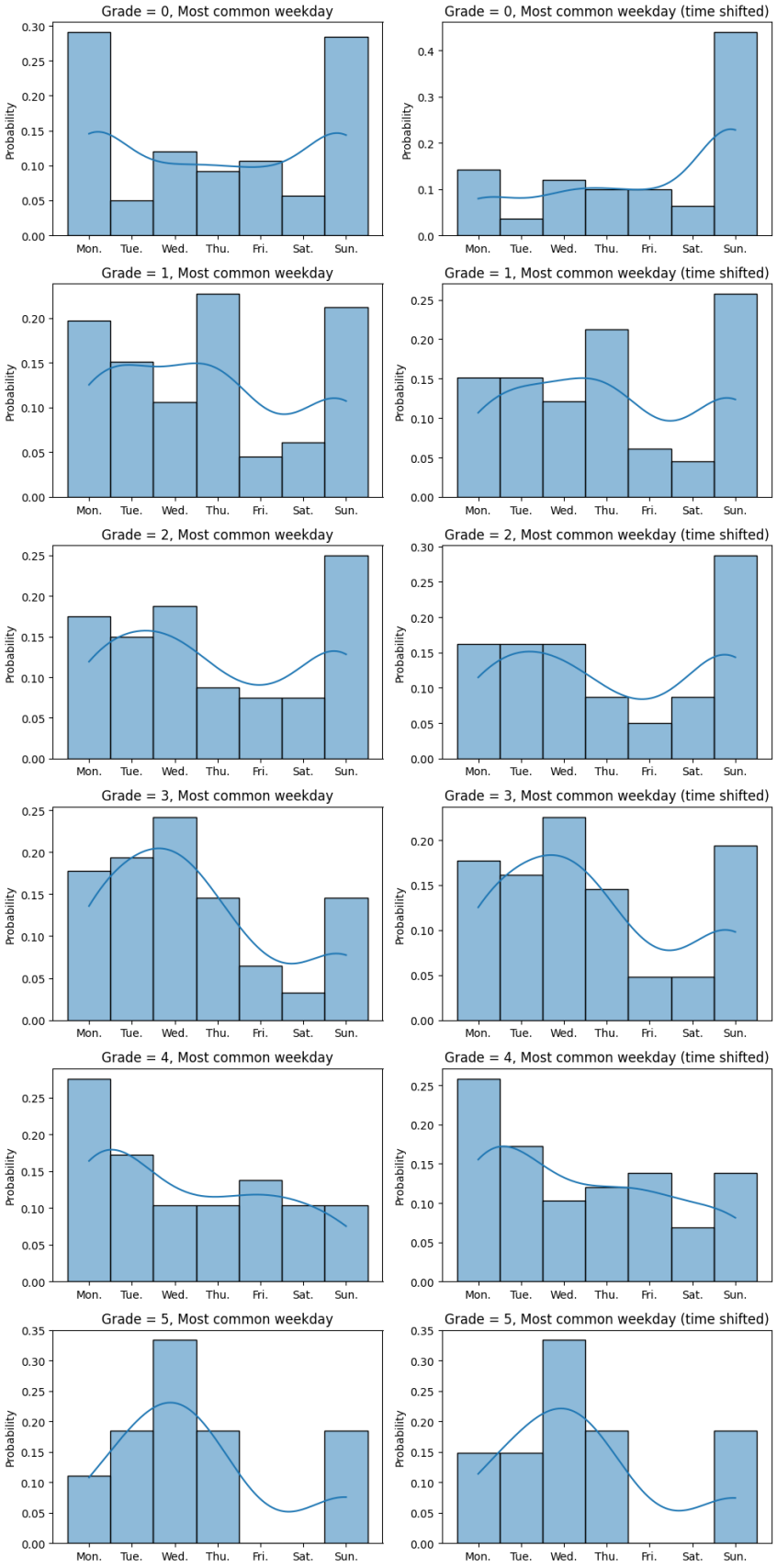

Figure 30. Most common day of week by grade and time shifted most common day of week. (fall of 2022 with time shift)

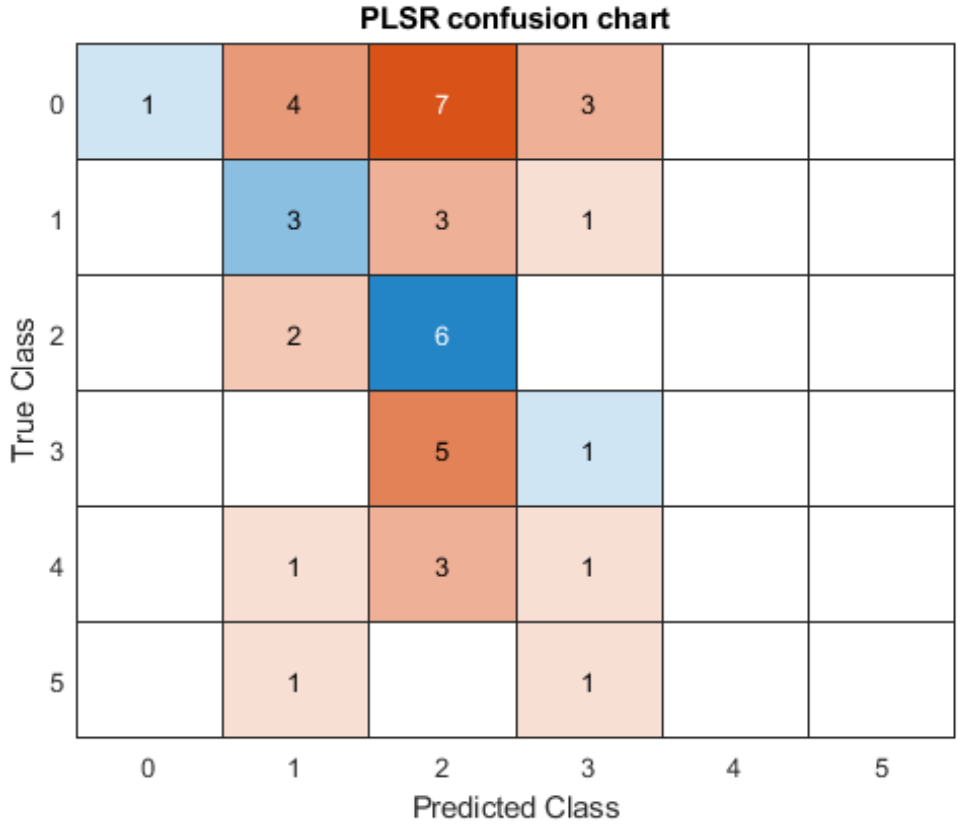

Figure 31. confusion matrix of PLSR. (fall of 2022 with time shift)

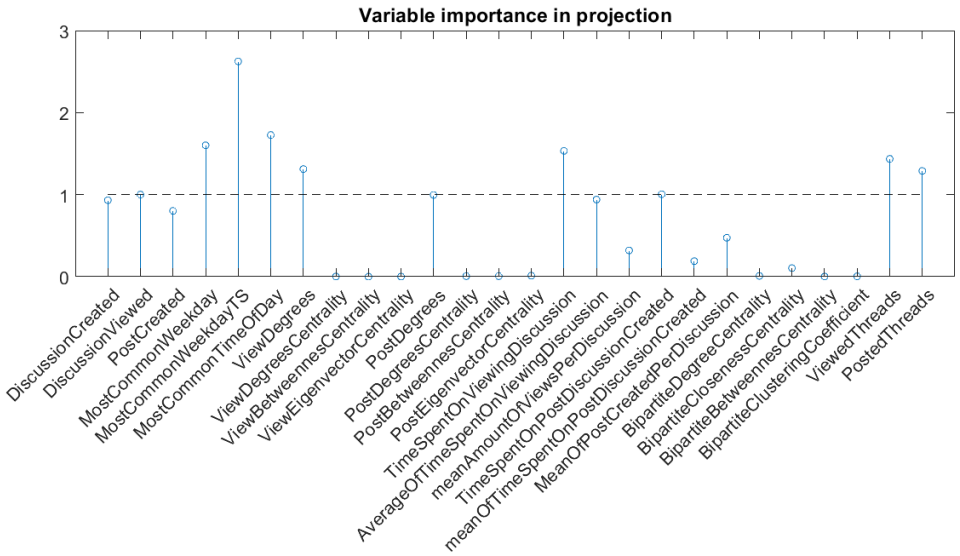

Figure 32. Variable importance in projection of PLSR. (fall of 2022 with time shift)

6 DISCUSSION

6.1 Current study

This study has pretty limited findings. Due to the breadth of the work and dataset, a well-structured deep look into all of the variables and their connections would take a series of studies, as even this thesis starts to go over 40 pages with most of the variables being only mentioned and then being ignored. As the study focused on only on forums and their connection to the final grade and tried to keep the measures of users generalised so that the same measures can be measured from other courses and compared, the final accuracy was pretty low. Also, the overall thought of keeping this generalised and studying forums that are not central part of the course limited the accuracy. There might also be some other measures that might have more effect on the final grade.

From the look into related work, it seemed that big part of these studies or related work seem to focus on MOOCs, this thesis does not. The problem with MOOC versus non-MOOC is the amount of data, with MOOC you can have thousands of participants but with our course it is more limited, though the course studied is for first year students and a part of multiple study programs. One else approach most of the studies that try to predict the grades takes, that is not used in this thesis, is the classification or analysis of the text in the discussions. In some cases, studies create their own add-ons to Moodle to gather the information analysed.

The missing data during the fall 2021 course together with Figure 4 of most popular threads might point to a problem with the platform, the 10 most popular threads are missing who posted to them which would suggest that the threads are from the time where we are missing data, so from the beginning of the course, while the users first time doing something growth happens during the later part of the course. One could assume that when the user goes to the forum during the assignment, especially if they are going for the first time, they would be looking for information about the assignment, so threads about these should grow, but they do not reach the action count of the first threads even when these first threads have lost some of the data.

Due to lack of documentation the meaning of some of these events is not fully known, for example what is the difference between "Some content has been posted" versus "Post created"? The event "Some content has been posted." seems to be always associated with either "Discussion created", "Post created" or "Post edited" event and the sum of

the previously mentioned events seem to match the count of "Some content has been posted events" as seen in Figure15. Through a look of the logs, it can be hypothesized that the event "Some content has been posted" is some part of the process where the text or content is created while the "Post created" is the event where the post itself is made visible. This was tested by creating a forum with same settings and a user doing selected actions and the following connections were drawn from this: "discussion viewed" event seems to be caused by user loading the page, be it by reload, opening new tab or opening it normally, and "some content has been posted" is always before "discussion created" or "post created".

When the high number of removals was looked at it was theorized that one of the reasons someone would remove posts or discussion could be person finding answer to their question after some searching, but this does not explain the huge peak in the first minutes of post or discussion going live. As there is no reason asked for the removal of the posts or discussions, this was later explored as a part of a survey after the 2022 fall course, 14 people answered to the question: "Why did you remove or edit your post on forums", with most of the answers relating to rephrasing/clarifying the post, three people answered that they solved the problem themselves, 2 answered that they found the answer on the forums and one thought that they accidentally posted with their name being visible.

The course studied is always trying to improve and with this comes some changes for example on the newest version of the course, with the change of code grading platform there were some delays with tasks and some other teething problems, that might have affected the student behaviour. This might partly explain the huge change in the grades.

There was some correlation between grade and average of discussions viewed by the grade. Some other variables also seemed to have also connection with the grade, but with huge variability inside the classes, the classes were inseparable because of this and this led to low accuracy in the prediction phase.

Some could argue that the accuracy of around 30-40% achieved during the first course is great considering that the messages contents were not considered, as some might theorize that the people asking questions could be the lower performing and the ones answering would be the better performing ones. Due to the generalised nature of the code created for this thesis it can be used with small modifications for monitoring of forum usages and viewing the same statistics of forums as explored in this thesis.

A way to increase the accuracy of the ML could be to consider creating a variable or set of variables that would measure the time before a deadline, this was not looked at in this

thesis due to the generalised nature, but some hints of this might be seen in Figure 14 and with the effect of time shifting. Other ways of specialising the code to the course could be considered for example expanding the log analysis from forum logs to general logs but that would also require more specialising the code to the course.

6.2 Future work

To increase the usefulness of this study one could consider some kind of variables that consider for example the order messages arrive in the discussions. Also, as in this thesis only forum logs were considered. Someone could use the whole log but at that point it would become more specialised instead of general analysis as the contents of courses Moodle page can change according to the wishes and skills of the courses Moodle pages' creator. Also, there could be further deep dive in to the data as in this thesis only scratches the surface to see if there is some other ways to increase the accuracy or insight gained. Also, this thesis findings could be used as a comparative point for other studies or looks into use of the forums on courses, especially on the course studied in this thesis in future.

Most studies that try to do similar prediction of seem to include some kind of analysis of the contents of the messages, this is not possible straight from the logs as mentioned, but also as mentioned it might be possible to do this through the module itself. This could be one way to increase the accuracy of the predictions but would require some kind of classifying or creating a variable that tries to measure the content or the context of the text itself. As a simple way someone might just to do this is to look for specific words or for example question marks, and this way trying to classify the text.

7 CONCLUSION

To conclude this thesis, there is a huge amount of data in Moodle logs that can be beneficial for analysis of student behaviour. Through datamining this information can be accessed, even in this case as we only looked at the logs focusing on the forums there is too much data to go into deep dive of each one of variables we chose. This thesis can be used as a kind of a starting point and as a comparison point at least for the forum parts.

There seems to be some correlation at average between discussions viewed and the grade the student receives at the end of these courses at least when the 2021-2022 course is considered. The later course also had same kind of but lower correlation with lot lower strength of correlation, but these might be partly caused by low number of samples in some of the grades. A possible reason for the huge change in the distribution of grades is the change of platform used on the coding tasks. Another data point that seemed to stand out was the most common day of week and the most common time of day, with huge changes in the distributions of these but they were still not enough for classifying.

The clustering approach with k-means seems to suggest that there is the simple case of two clusters: those who participate and those who just watch, or more complex case of five to six clusters where the clusters reflect more complex behaviour, for example those who just watch seems to be divided into those who watch a lot of different discussions or those who just watch few different discussions. Higher number of clusters are also possible but at that point it starts to become harder to explain the differences between them and with five to six clusters some clusters already had less than 10 students.

For classifying part, these variables did not really fit with the data being too noisy and with the distributions being too overlapped. At best, a PLSR result of 37.1% was achieved on the 2021-2022 course but accuracy of only 16.3% on the 2022-2023 course. With PLSR we also gained some insight which variables at least PLSR considered important for the prediction of grades: Most common time of day, Time spent on viewing discussions and Viewed discussions.

REFERENCES

[1] L. Griffin and J. Roy. A great resource that should be utilised more, but also a place of anxiety: student perspectives on using an online discussion forum. *Open Learning: The Journal of Open, Distance and e-Learning*, 37(3):235–250, July 2022. Publisher: Routledge _eprint: https://doi.org/10.1080/02680513.2019.1644159.

[2] C. Romero, S. Ventura, M. Pechenizkiy, and R. SJd Baker. *Handbook of educational data mining*. CRC press, 2010.

[3] O. Scheuer and B. M. McLaren. Educational data mining. *Encyclopedia of the Sciences of Learning*, 13, 2012.

[4] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, July 2007.

[5] D. F. O. Onah, J. E. Sinclair, R. Boyatt, and J. Foss. Massive open online courses: learners participation. *ICERI2014 Proceedings*, pages 2348–2356, 2014. Conference Name: 7th International Conference of Education, Research and Innovation ISBN: 9788461724840 Meeting Name: 7th International Conference of Education, Research and Innovation Place: Seville, Spain Publisher: IATED.

[6] D. Coetzee, A. Fox, M. A. Hearst, and B. Hartmann. Should your MOOC forum use a reputation system? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, CSCW '14, pages 1176–1187, New York, NY, USA, February 2014. Association for Computing Machinery.

[7] R. Hirschel. Moodle: Students' perspectives on forums, glossaries and quizzes. *The JALT CALL Journal*, 8:95–112, August 2012.

[8] R. Suviste, H. L. Reponen, E. Tõnisson, M. Lepp, and P. Luik. Forum usage in massive open online courses on the example of a programming course "About programming". *EDULEARN20 Proceedings*, pages 3855–3864, 2020. Conference Name: 12th International Conference on Education and New Learning Technologies ISBN: 9788409179794 Meeting Name: 12th International Conference on Education and New Learning Technologies Place: Online Conference Publisher: IATED.

[9] C. B. Baruque, M. A. Amaral, A. Barcellos, J. C. da Silva Freitas, and C. J. Longo. Analysing users' access logs in Moodle to improve e learning. In *Proceedings of the 2007 Euro American conference on Telematics and information systems*, EATIS '07, pages 1–4, New York, NY, USA, May 2007. Association for Computing Machinery.

[10] C. Pong-inwong and W. Rungworawut. Teaching evaluation using data mining on moodle LMS forum. In *2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM2012)*, pages 550–555, October 2012.

[11] A. F. Wise and Y. Cui. Unpacking the relationship between discussion forum participation and learning in MOOCs: content is key. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 330–339, Sydney New South Wales Australia, March 2018. ACM.

[12] M. I. Lopez, J. M. Luna, C. Romero, and S. Ventura. *Classification via Clustering for Predicting Final Marks Based on Student Participation in Forums*. International Educational Data Mining Society, June 2012. Publication Title: International Educational Data Mining Society.

[13] J. López-Zambrano, J. A. Lara, and C. Romero. Towards Portability of Models for Predicting Students' Final Performance in University Courses Starting from Moodle Logs. *Applied Sciences*, 10(1):354, January 2020.

[14] S. P. Borgatti. Social Network Analysis, Two-Mode Concepts in. In R. A. Meyers, editor, *Encyclopedia of Complexity and Systems Science*, pages 8279–8291. Springer New York, New York, NY, 2009.

[15] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, Oxford ; New York, 2010. OCLC: ocn456837194.

[16] M. Latapy, C. Magnien, and N. D. Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, January 2008.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees*. Routledge, 1 edition, October 2017.

[18] E. Allwein and R. Schapire. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1, 01 2001.

[19] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, March 1993.

Appendix 1. Descriptive statistics for the data set of 21 fall rounded to 2 decimals

Stat	0	1	2	3	4	5	Staff
Min Discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Discussion viewed	6.0	7.0	6.0	6.0	6.0	8.0	61.0
Min Post created	0.0	0.0	0.0	0.0	0.0	0.0	3.0
Min Most common weekday	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Min Most common time of day	0.0	0.0	0.0	0.0	0.0	0.0	8.0
Min View Degrees	165.0	214.0	164.0	149.0	169.0	206.0	388.0
Min View Degrees centrality	0.35	0.45	0.35	0.31	0.36	0.43	0.82
Min View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min View Eigenvector centrality	0.02	0.03	0.02	0.02	0.03	0.03	0.05
Min Post Degrees	0.0	0.0	0.0	0.0	0.0	0.0	4.0
Min Post Degrees centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.03
Min Post Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Post Eigenvector centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.03
Min Time spent on viewing discussion	2.0	3.0	3.0	3.0	2.0	10.0	132.0
Min Average of time spent on viewing discussion	0.33	0.43	0.27	0.43	0.33	1.11	0.77
Min mean amount of views per discussion	1.0	1.0	1.0	1.0	1.0	1.0	1.27
Min Time spent on post discussion created	0.0	0.0	0.0	0.0	0.0	0.0	3.0
Min mean of time spent on post discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.06
Min Mean of post created per discussion	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Min Bipartite Degree Centrality	0.01	0.02	0.01	0.02	0.02	0.03	0.17
Min Bipartite Closeness Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.49
Min Bipartite Betweennes Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Bipartite Clustering coefficient	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Viewed threads	3.0	5.0	3.0	4.0	4.0	8.0	41.0
Min Posted threads	0.0	0.0	0.0	0.0	0.0	0.0	2.0
Mean Discussion created	0.46	0.63	0.7	0.66	0.83	0.73	0.0
Mean Discussion viewed	50.87	53.96	68.81	75.12	68.72	83.65	227.71
Mean Post created	0.57	1.59	0.6	0.62	0.77	1.62	15.71
Mean Most common weekday	4.09	3.59	3.31	3.04	3.12	2.97	2.0
Mean Most common time of day	16.3	15.7	15.12	14.34	15.12	12.84	14.57

Continued on next page

Appendix 1. (continued)

(Continued)

Mean View Degrees	337.0	350.22	367.26	377.75	368.29	375.11	445.0
Mean View Degrees centrality	0.71	0.74	0.77	0.8	0.78	0.79	0.94
Mean View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean View Eigenvector centrality	0.05	0.05	0.05	0.05	0.05	0.05	0.06
Mean Post Degrees	0.98	1.85	1.46	1.38	2.0	3.86	21.57
Mean Post Degrees centrality	0.01	0.01	0.01	0.01	0.01	0.03	0.15
Mean Post Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.08
Mean Post Eigenvector centrality	0.01	0.02	0.02	0.01	0.02	0.04	0.14
Mean Time spent on viewing discussion	244.43	210.33	264.22	321.23	298.48	374.41	598.0
Mean Average of time spent on viewing discussion	4.67	3.73	4.05	4.05	4.54	4.93	2.64
Mean mean amount of views per discussion	1.53	1.49	1.53	1.47	1.46	1.47	1.7
Mean Time spent on post discussion created	0.8	3.33	1.43	1.88	1.74	1.95	9.71
Mean mean of time spent on post discussion created	0.2	0.31	0.36	0.25	0.41	0.42	1.6
Mean Mean of post created per discussion	0.4	0.64	0.51	0.48	0.62	0.7	1.15
Mean Bipartite Degree Centrality	0.12	0.13	0.17	0.18	0.17	0.2	0.54
Mean Bipartite Closeness Centrality	0.15	0.23	0.23	0.16	0.2	0.27	0.56
Mean Bipartite Betweennes Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.04
Mean Bipartite Clustering coefficient	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean Viewed threads	29.3	33.07	41.4	45.69	42.17	49.76	123.43
Mean Posted threads	0.67	1.07	1.01	0.89	1.17	1.89	14.86
Max Discussion created	4.0	8.0	7.0	7.0	8.0	5.0	0.0
Max Discussion viewed	347.0	207.0	319.0	494.0	411.0	375.0	553.0
Max Post created	9.0	14.0	11.0	15.0	15.0	22.0	52.0
Max Most common weekday	6.0	6.0	6.0	6.0	6.0	6.0	3.0
Max Most common time of day	23.0	23.0	23.0	22.0	23.0	23.0	22.0
Max View Degrees	445.0	456.0	474.0	472.0	474.0	466.0	474.0
Max View Degrees centrality	0.94	0.96	1.0	0.99	1.0	0.98	1.0
Max View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Max View Eigenvector centrality	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Max Post Degrees	8.0	11.0	16.0	18.0	29.0	34.0	56.0

Continued on next page

Appendix 1. (continued)

(Continued)

Max Post Degrees centrality	0.06	0.08	0.11	0.13	0.2	0.24	0.39
Max Post Betweennes centrality	0.01	0.03	0.03	0.03	0.1	0.1	0.25
Max Post Eigenvector centrality	0.08	0.12	0.14	0.16	0.2	0.25	0.32
Max Time spent on viewing discussion	1308.0	955.0	1121.0	1580.0	1921.0	1542.0	1631.0
Max Average of time spent on viewing discussion	13.89	9.0	15.4	10.75	12.08	10.56	3.69
Max mean amount of views per discussion	3.38	3.83	3.07	2.99	2.5	2.34	2.66
Max Time spent on post discussion created	21.0	48.0	43.0	67.0	35.0	20.0	23.0
Max mean of time spent on post discussion created	4.0	2.33	10.75	6.25	5.33	5.0	3.83
Max Mean of post created per discussion	4.0	4.0	2.5	4.0	4.0	3.0	1.5
Max Bipartite Degree Centrality	0.46	0.43	0.81	0.69	0.8	0.67	1.01
Max Bipartite Closeness Centrality	1.65	1.65	1.65	0.56	0.6	0.61	0.7
Max Bipartite Betweennes Centrality	0.02	0.0	0.02	0.02	0.03	0.05	0.17
Max Bipartite Clustering coefficient	0.01	0.01	0.01	0.01	0.02	0.01	0.01
Max Viewed threads	115.0	111.0	206.0	174.0	203.0	160.0	208.0
Max Posted threads	6.0	8.0	15.0	13.0	15.0	18.0	51.0
Median Discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median Discussion viewed	22.5	28.0	42.0	40.0	44.0	45.0	172.0
Median Post created	0.0	0.0	0.0	0.0	0.0	0.0	7.0
Median Most common weekday	5.0	3.0	3.0	3.0	3.0	3.0	2.0
Median Most common time of day	18.0	18.0	16.0	15.0	15.0	13.0	14.0
Median View Degrees	334.0	343.0	371.0	388.0	377.0	394.0	463.0
Median View Degrees centrality	0.7	0.72	0.78	0.82	0.79	0.83	0.97
Median View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median View Eigenvector centrality	0.05	0.05	0.05	0.05	0.05	0.05	0.06
Median Post Degrees	0.0	0.0	0.0	0.0	0.0	1.0	14.0
Median Post Degrees centrality	0.0	0.0	0.0	0.0	0.0	0.01	0.1
Median Post Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.02
Median Post Eigenvector centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Median Time spent on viewing discussion	120.0	80.0	183.0	173.5	175.0	224.0	502.0

Continued on next page

Appendix 1. (continued)

(Continued)

Median Average of time spent on viewing discussion	4.97	3.41	4.05	3.79	4.22	4.43	2.95
Median mean amount of views per discussion	1.41	1.2	1.44	1.39	1.42	1.39	1.57
Median Time spent on post discussion created	0.0	0.0	0.0	0.0	0.0	0.0	9.0
Median mean of time spent on post discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.58
Median Mean of post created per discussion	0.0	0.0	0.0	0.0	0.0	1.0	1.04
Median Bipartite Degree Centrality	0.07	0.1	0.11	0.13	0.12	0.14	0.47
Median Bipartite Closeness Centrality	0.0	0.0	0.0	0.0	0.0	0.45	0.53
Median Bipartite Betweennes Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.02
Median Bipartite Clustering coefficient	0.0	0.0	0.0	0.0	0.0	0.0	0.01
Median Viewed threads	17.0	25.0	27.0	32.5	29.0	34.0	116.0
Median Posted threads	0.0	0.0	0.0	0.0	0.0	1.0	6.0
Std Discussion created	1.13	1.57	1.26	1.3	1.52	1.07	0.0
Std Discussion viewed	60.74	56.14	70.01	85.02	81.87	95.84	172.48
Std Post created	1.61	3.69	1.58	1.99	2.05	4.2	17.62
Std Most common weekday	2.33	2.17	1.97	1.9	1.97	1.98	0.82
Std Most common time of day	5.79	6.34	4.98	5.47	5.14	6.07	4.39
Std View Degrees	79.59	70.01	65.33	65.15	69.45	62.67	36.26
Std View Degrees centrality	0.17	0.15	0.14	0.14	0.15	0.13	0.08
Std View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Std View Eigenvector centrality	0.01	0.01	0.01	0.01	0.01	0.01	0.0
Std Post Degrees	2.16	3.38	2.9	3.13	4.16	7.15	20.88
Std Post Degrees centrality	0.02	0.02	0.02	0.02	0.03	0.05	0.15
Std Post Betweennes centrality	0.0	0.01	0.0	0.01	0.01	0.02	0.09
Std Post Eigenvector centrality	0.02	0.04	0.03	0.03	0.04	0.06	0.12
Std Time spent on viewing discussion	307.37	253.72	261.54	364.37	360.84	388.7	534.89
Std Average of time spent on viewing discussion	2.56	2.44	2.25	2.03	2.11	2.24	1.0
Std mean amount of views per discussion	0.48	0.6	0.42	0.37	0.35	0.36	0.45

Continued on next page

Appendix 1. (continued)

(Continued)

Std Time spent on post discussion created	3.18	9.74	5.39	8.87	5.42	4.42	6.9
Std mean of time spent on post discussion created	0.71	0.69	1.32	0.97	1.06	0.99	1.6
Std Mean of post created per discussion	0.83	1.03	0.66	0.81	0.88	0.74	0.21
Std Bipartite Degree Centrality	0.11	0.12	0.16	0.17	0.17	0.19	0.31
Std Bipartite Closeness Centrality	0.29	0.37	0.35	0.23	0.24	0.25	0.07
Std Bipartite Betweennes Centrality	0.0	0.0	0.0	0.0	0.0	0.01	0.06
Std Bipartite Clustering coefficient	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Std Viewed threads	26.44	29.83	39.31	41.51	42.3	46.21	67.55
Std Posted threads	1.44	2.02	2.0	1.98	2.38	3.55	17.53

Appendix 2. Descriptive statistics for the data set of 22 fall rounded to 2 decimals

Stat	0	1	2	3	4	5	Staff
Min Discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Discussion viewed	6.0	6.0	6.0	6.0	8.0	9.0	18.0
Min Post created	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Most common weekday	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Most common weekday TS	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Most common time of day	0.0	0.0	0.0	1.0	0.0	2.0	0.0
Min View Degrees	125.0	214.0	241.0	273.0	288.0	226.0	427.0
Min View Degrees centrality	0.21	0.36	0.41	0.46	0.49	0.38	0.72
Min View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min View Eigenvector centrality	0.01	0.02	0.03	0.03	0.03	0.02	0.04
Min Post Degrees	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Post Degrees centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Post Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Post Eigenvector centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Time spent on viewing discussion	3.0	2.0	10.0	8.0	5.0	6.0	114.0
Min Average of time spent on viewing discussion	0.36	0.33	1.0	0.4	0.42	0.6	2.5
Min mean amount of views per discussion	1.0	1.0	1.0	1.0	1.0	1.0	1.2
Min Time spent on post discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min mean of time spent on post discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Mean of post created per discussion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Bipartite Degree Centrality	0.01	0.02	0.02	0.02	0.03	0.03	0.06
Min Bipartite Closeness Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Bipartite Betweennes Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Bipartite Clustering coefficient	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min Viewed threads	3.0	3.0	4.0	4.0	7.0	8.0	15.0
Min Posted threads	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean Discussion created	0.36	0.39	0.51	0.65	1.14	0.56	0.0
Mean Discussion viewed	46.34	55.17	50.29	69.65	77.9	55.0	267.15
Mean Post created	0.54	0.27	0.6	0.79	1.16	0.7	16.92

Continued on next page

Appendix 2. (continued)

(Continued)

Mean Most common weekday	2.98	2.8	2.96	2.4	2.38	2.52	1.77
Mean Most common weekday TS	3.93	3.05	3.11	2.65	2.47	2.48	1.69
Mean Most common time of day	13.49	14.68	14.92	15.13	14.84	15.59	11.08
Mean View Degrees	442.01	450.21	449.88	471.95	480.16	459.96	553.54
Mean View Degrees centrality	0.75	0.76	0.76	0.8	0.81	0.78	0.94
Mean View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean View Eigenvector centrality	0.04	0.04	0.04	0.05	0.05	0.04	0.05
Mean Post Degrees	2.79	1.41	2.31	2.52	2.4	2.41	22.46
Mean Post Degrees centrality	0.02	0.01	0.01	0.02	0.01	0.01	0.14
Mean Post Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.09
Mean Post Eigenvector centrality	0.02	0.01	0.01	0.01	0.01	0.01	0.05
Mean Time spent on viewing discussion	248.44	285.27	284.14	386.13	424.48	295.0	1386.38
Mean Average of time spent on viewing discussion	5.19	5.13	5.53	5.22	5.18	4.61	5.24
Mean mean amount of views per discussion	1.49	1.44	1.49	1.48	1.5	1.41	2.01
Mean Time spent on post discussion created	1.27	0.27	1.25	2.24	1.67	1.15	15.62
Mean mean of time spent on post discussion created	0.29	0.1	0.35	0.38	0.24	0.26	1.04
Mean Mean of post created per discussion	0.38	0.37	0.47	0.63	0.62	0.55	1.12
Mean Bipartite Degree Centrality	0.12	0.14	0.13	0.18	0.2	0.15	0.58
Mean Bipartite Closeness Centrality	0.16	0.19	0.2	0.23	0.23	0.2	0.54
Mean Bipartite Betweennes Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.04
Mean Bipartite Clustering coefficient	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean Viewed threads	28.65	34.33	31.25	43.02	46.98	36.59	128.62
Mean Posted threads	0.63	0.55	0.8	1.0	1.6	0.89	14.0
Max Discussion created	6.0	9.0	6.0	7.0	14.0	3.0	0.0
Max Discussion viewed	298.0	345.0	357.0	337.0	505.0	222.0	593.0
Max Post created	8.0	4.0	7.0	10.0	22.0	5.0	58.0
Max Most common weekday	6.0	6.0	6.0	6.0	6.0	6.0	4.0
Max Most common weekday TS	6.0	6.0	6.0	6.0	6.0	6.0	3.0
Max Most common time of day	23.0	23.0	23.0	23.0	23.0	23.0	17.0

Continued on next page

Appendix 2. (continued)

(Continued)

Max View Degrees	578.0	586.0	581.0	591.0	591.0	582.0	591.0
Max View Degrees centrality	0.98	0.99	0.98	1.0	1.0	0.98	1.0
Max View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Max View Eigenvector centrality	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Max Post Degrees	31.0	23.0	30.0	26.0	47.0	25.0	82.0
Max Post Degrees centrality	0.19	0.14	0.19	0.16	0.29	0.16	0.51
Max Post Betweennes centrality	0.02	0.0	0.02	0.01	0.12	0.02	0.37
Max Post Eigenvector centrality	0.2	0.19	0.2	0.2	0.22	0.19	0.25
Max Time spent on viewing discussion	1490.0	1582.0	2679.0	2671.0	3326.0	1964.0	3549.0
Max Average of time spent on viewing discussion	13.05	15.86	11.45	11.38	11.89	11.03	9.7
Max mean amount of views per discussion	3.08	3.11	2.75	2.37	2.97	1.94	2.94
Max Time spent on post discussion created	62.0	8.0	30.0	54.0	33.0	13.0	60.0
Max mean of time spent on post discussion created	10.33	2.67	8.0	6.5	2.8	3.33	2.67
Max Mean of post created per discussion	4.0	3.0	3.0	2.33	2.0	2.0	1.67
Max Bipartite Degree Centrality	0.53	0.66	0.59	0.78	0.83	0.64	1.03
Max Bipartite Closeness Centrality	1.71	1.71	1.71	0.58	1.14	0.57	0.73
Max Bipartite Betweennes Centrality	0.01	0.01	0.01	0.01	0.03	0.01	0.17
Max Bipartite Clustering coefficient	0.06	0.06	0.03	0.06	0.01	0.01	0.01
Max Viewed threads	126.0	162.0	138.0	191.0	201.0	155.0	202.0
Max Posted threads	7.0	9.0	9.0	8.0	18.0	5.0	49.0
Median Discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median Discussion viewed	29.0	33.0	32.5	41.5	44.5	33.0	212.0
Median Post created	0.0	0.0	0.0	0.0	0.0	0.0	13.0
Median Most common weekday	3.0	3.0	2.0	2.0	2.0	2.0	2.0
Median Most common weekday TS	5.0	3.0	3.0	2.0	2.0	2.0	2.0
Median Most common time of day	15.0	15.0	15.0	16.0	15.0	16.0	10.0
Median View Degrees	462.0	465.5	463.0	489.5	497.5	470.0	578.0
Median View Degrees centrality	0.78	0.79	0.78	0.83	0.84	0.8	0.98
Median View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median View Eigenvector centrality	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Continued on next page

Appendix 2. (continued)

(Continued)

Median Post Degrees	0.0	0.0	0.0	0.0	0.0	0.0	15.0
Median Post Degrees centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.09
Median Post Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.06
Median Post Eigenvector centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.03
Median Time spent on viewing discussion	158.0	156.0	174.0	209.0	181.5	132.0	996.0
Median Average of time spent on viewing discussion	5.03	4.99	5.3	5.17	5.1	4.0	4.72
Median mean amount of views per discussion	1.39	1.43	1.45	1.39	1.44	1.32	1.89
Median Time spent on post discussion created	0.0	0.0	0.0	0.0	0.0	0.0	8.0
Median mean of time spent on post discussion created	0.0	0.0	0.0	0.0	0.0	0.0	0.6
Median Mean of post created per discussion	0.0	0.0	0.0	0.0	0.0	0.0	1.2
Median Bipartite Degree Centrality	0.09	0.1	0.09	0.14	0.14	0.09	0.64
Median Bipartite Closeness Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.58
Median Bipartite Betweennes Centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.01
Median Bipartite Clustering coefficient	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median Viewed threads	21.0	24.5	22.0	31.0	32.5	23.0	138.0
Median Posted threads	0.0	0.0	0.0	0.0	0.0	0.0	10.0
Std Discussion created	0.95	1.21	1.21	1.23	2.39	0.97	0.0
Std Discussion viewed	48.06	64.46	55.89	74.38	88.86	54.04	168.21
Std Post created	1.32	0.76	1.35	1.79	3.6	1.35	16.5
Std Most common weekday	2.42	2.17	2.25	1.95	2.11	1.91	1.09
Std Most common weekday TS	2.24	2.17	2.31	2.1	2.14	1.95	0.95
Std Most common time of day	6.95	5.4	4.83	5.37	5.82	5.41	4.25
Std View Degrees	90.89	92.88	76.68	81.55	79.1	86.46	52.43
Std View Degrees centrality	0.15	0.16	0.13	0.14	0.13	0.15	0.09
Std View Betweennes centrality	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Std View Eigenvector centrality	0.01	0.01	0.01	0.01	0.01	0.01	0.0
Std Post Degrees	6.75	4.21	6.17	5.43	6.96	5.51	22.18

Continued on next page

Appendix 2. (continued)

(Continued)

Std Post Degrees centrality	0.04	0.03	0.04	0.03	0.04	0.03	0.14
Std Post Betweennes centrality	0.0	0.0	0.0	0.0	0.02	0.0	0.11
Std Post Eigenvector centrality	0.05	0.03	0.04	0.04	0.03	0.04	0.06
Std Time spent on viewing discussion	273.43	327.06	376.71	488.75	568.72	421.37	1086.27
Std Average of time spent on viewing discussion	2.48	2.55	2.26	2.57	2.05	2.19	2.13
Std mean amount of views per discussion	0.4	0.36	0.33	0.32	0.37	0.3	0.57
Std Time spent on post discussion created	6.47	1.13	4.28	7.99	5.25	3.18	20.67
Std mean of time spent on post discussion created	1.33	0.4	1.19	1.24	0.59	0.72	0.95
Std Mean of post created per discussion	0.68	0.62	0.74	0.75	0.75	0.73	0.39
Std Bipartite Degree Centrality	0.1	0.13	0.11	0.16	0.18	0.14	0.3
Std Bipartite Closeness Centrality	0.29	0.35	0.34	0.25	0.28	0.25	0.18
Std Bipartite Betweennes Centrality	0.0	0.0	0.0	0.0	0.01	0.0	0.05
Std Bipartite Clustering coefficient	0.01	0.01	0.01	0.01	0.0	0.0	0.0
Std Viewed threads	24.19	31.87	26.49	38.95	42.96	33.04	65.81
Std Posted threads	1.32	1.34	1.66	1.59	3.51	1.42	13.71