Visualisation of Learning Management System Usage for Detecting Student Behaviour Patterns

Thomas Haig **Katrina Falkner** **Nickolas Falkner**

School of Computer Science
The University of Adelaide
North Terrace, South Australia
Email: [thomas.haig,katrina.falkner,nickolas.falkner]@adelaide.edu.au

Abstract

Identifying "at-risk" students - those that are in danger of failing or not completing a course - is a crucial element in enabling students to achieve their full potential. However, with large class sizes and growing academic workloads, it is becoming increasingly difficult to identify students who require urgent and timely assistance. Efficient and easy to use tools are needed to assist academics in locating these students at early stages within their courses. A significant body of work exists in the use of student activity data, e.g. attendance, performance, participation in face-to-face and online sessions, to predict overall student performance and at-risk status. This is often built upon the considerable amount of student data within learning management systems. Manual data collection, including surveys and observation, which introduces additional workload is often required to extract relevant data meaning that it in large classes it is prohibitively difficult to apply such techniques.

In this paper, we introduce a framework for at-risk identification combining simple metrics, gathered from social network and statistical analysis domains, that have been shown to correlate with student performance and require slow amounts of manual data collection or additional expert analysis. We describe each of the metrics within our framework and demonstrate their usage. We use visualisation to enable easy interpretation of results. The application of our framework is demonstrated within the context of an advanced undergraduate computer science course.

Keywords: Student data, Learning Management Systems, Prediction, Visualisation

1 Introduction

In order to enable all of our students to succeed to their potential, academics must be able to identify students who are "at-risk" - those students who are likely to fail, or withdraw, from a course - within the early stages of their at-risk behaviour. Interventions can only be made if academics have efficient and clear facilities that enable them to identify at-risk students. At-risk behaviours are becoming increasingly difficult to detect within our overburdened higher education systems, with large classes, de-personalised administrative systems and separation from peer groups. Our classrooms are increasing in diversity (Biggs & Tang 2007), further complicating this issue by presenting us with an "unprecedentedly broad spectrum of student ability and background" (Ramsden 2003).

Early identification of at-risk students is of particular concern within the ICT discipline - within Australia, and globally, we have seen a recent dramatic drop in applications for ICT degree programs, poor progression and retention rates (Sheard et al. 2008).

In order to identify at-risk students we need to provide facilities to assist academics in finding these students. Any facilities or tools provided to assist academics must introduce minimal additional workload. Although true in every discipline, Computer Science and ICT academics face an increasing pressure to include more technical concepts in their curriculum (McGregor et al. 2000). In 1978 the ACM recommendations for undergraduate programs consisted of a 20-page document. In 2010, the current recommendations total 240 pages with a vast increase in the body of knowledge expected of an undergraduate curriculum (Becker 2008). These pressures, along with pressure from industry and accrediting bodies to focus more attention on the development of generic skills (Falkner & Falkner 2012), mean that ICT academics must find efficient and effective mechanisms to assist them in these tasks. Further, we must work within the available data sources that can be readily accessed by academics within their institutions.

There has been considerable work within the area of automated at-risk identification, using a variety of data sources, such as learning management systems, grade rosters, attendance records, and participation in online discussion forums. However, the majority of this work relies upon a blend of automated analysis and manual coding or recording of data. This includes the use of surveys and large-scale data collection to complement automatically available data. Even within a small cohort, these methods present an additional workload for academics, which becomes prohibitive within large classes, where these techniques are often most needed.

In this paper, we propose a framework for the identification of at-risk students using a combination of simple metrics, gathered from the domains of social network analysis and statistical analysis. These metrics, based upon data readily available within learning management systems, present an automated analysis framework requiring minimal manual interaction with the underlying data, and no additional data collection. We present a range of data visualisations that enable academics to easily and efficiently identify students who are exhibiting potential at-risk behaviours. We are able to gather the required data early on in a course without the requirement for additional assessments or surveys.

Using our obtained data and visualisations, we aim to find patterns of behaviour for successful and unsuccessful students for specific activities of a particular course. By identifying patterns of behaviour we may be able to identify those students who require additional assistance and intervention. We present visualisations for each metric that support the ease of identification of such patterns, and hence the subsequent identification of at-risk students.

We demonstrate the utility of our framework within a case study of an advanced undergraduate ICT course, using data available from the learning management system used within the course, the Moodle Learning Management System, which is used by multiple institutions around the world (Moodle 2007). We demonstrate each of the metrics within the framework applied to this case study, exploring visualisation potential and validating correlation for each metric and student performance.

The paper will be presented as follows. Section 1 presents an Introduction, Section 2 a review of related work, Section 3 our Methodology, Sections 4, 5 and 6 our Results, Section 7 a Discussion and Section 8 will present potential future work and conclusions.

2 Related Work

Obtaining measures of student engagement provides us with one method for determining at risk behaviour. However, traditional methods of ascertaining student engagement, such as attendance and participation in lectures and tutorials, become increasing difficult to use in large classes. With the increasing conflict between external pressures, such as work and family commitments, even engaged students may not be always able to make frequent face-to-face contact. Further, there is a time burden involved with taking these attendances and measuring participation, which academics may struggle to afford. We want to assess engagement automatically in order to predict results for students, so as that we may find those who may be at risk of failing a course, or dropping out, in order to attempt to prevent this occurring.

The increasing use of online learning management systems and online learning tools means that we now have alternative means for gathering information on student engagement, which supports the more flexible practices of the modern higher education sector.

Large amounts of data are readily available for analysis of student engagement. Bayer et al. (2012) have explored the use of a wide variety of data including "capacity to study" test scores, attained credits, average grades, gender and year of birth, to develop a model of social behaviour in order to predict potential drop-outs. Merceron & Yacef (2005) utilised course specific information, such as the types of mistakes made in individual assessment exercises, while El-Halees (2009) used preliminary course assessment results as a prediction of final grade, hence identifying at-risk students.

Norris et al. (2008) looked at the work of novice programmers in the BlueJay environment. This allowed them to log specific actions such as number of compilations made, amount of time spent working on a project and the amount of errors encountered. This level of data provided a log of students patterns of work while performing a programming task in a short closed session. Their aim was to gather patterns of work for successful students and compare these with unsuccessful students visually and intervene where necessary to get unsuccessful students into better patterns of work, for example compiling their code more often. While we are not able to log results at this level, it is useful to note that patterns of behaviour run to deep levels within a course, down to how students portion their time and work while programming.

Logging at a similar level is performed by Murphy et al. (2009) who also use a BlueJay or Eclipse plug-in to explore student programming habits. They logged the students time spent on assignment as well as their compilation errors. They then used this data to send recommendations to students about how they could improve, e.g. they are spending too long on an assignment, or making the same error too many times. Students are then able to reflect upon their patterns and this aims to move students towards more successful patterns of study, through self-intervention. Instructors were also able to use the data to make more meaningful interventions, specifically where students were making numerous errors it was shown that early intervention was able to help.

The work of Norris et al. (2008) was expanded upon by Fenwick Jr et al. (2009). Using the same ClockIt software, Norris et al. (2008) observed patterns of student behaviour in a programming task and how this potentially relates to cheating, as well as how much incremental work they put into their programming. They found that students that started the task later in general received a lower grade. As noted in their work "although this is what we have already been "preaching" to students, it is now based on objective analysis of quantitative data".

Edwards et al. (2009) analysed submission data and came to the same conclusions as Norris et al. (2008), in that students who start their work early, in their case as measured by their first submission, are more likely to perform better in a task. Of note from their study is that students who perform consistently well or consistently poorly may demonstrate similar behaviours, e.g. a good student may start consistently late and be able to perform under pressure.

Nandi et al. (2011) take the online participation of students in forums as a measurement of engagement and a possible predictor of grade. Their results show that students who input more into the course achieve a better grade. The course analysed was fully online, and hence students point of contact to course providers was through the forums, hence overall forum usage was high. This is in contrast to a blended learning course where students have more access to course providers without using the online environment, and hence overall participation is lower.

Tracking of student movement through an online learning website was performed by Ceddia et al. (2007) using web logs. They use these logs to track student behaviour and catagorise it as either purposeful or browsing behaviour. They found that as the course progressed students use the online system more purposefully, browsing less to get to their required goals and materials. They also used the logs to analyse the learning behaviours of students on the website. They used completion rates, duration, frequency to measure effectiveness, efficiency and explorational activities. They also used unusual results to find possible problems with the website interface.

Students self-managing their own timesheets was shown by Herbert & Wang (2007) to be an effective measure of students usage of an online learning system. Students self-evaluated their use of the system, and this was then contrasted with the actual usage data from the website. They sought to find behavioural patterns that showed students may work to deadlines, relate their time spent to marks available and to test if students could be induced to

start tasks early. They found student timesheets accurate enough to be able to critically analyse these behaviours, backing up anecdotal evidence that students do indeed work to deadlines and will only spend as much time as proportional to marks.

Although successful in identifying at risk potential, these studies utilise data that is not readily available across the sector, or not readily available in a timely fashion to perform early intervention.

Studies such as Sheard et al. (2003) and Georg (2009) supplement their automatically collected data with manually collected survey data. Sheard et al. (2003) utilised survey data to gain student ratings of how useful online course material is to them, and how useful the online site is to their studies. Georg (2009) used the Konstanz Student Survey, collected in Germany every two to three years, which collects data about students attitudes towards study and profession to analyse factors that lead to students dropping out of courses. However, the use of surveys is problematic. Black et al. (2008) suggest the burden of time on administration to create surveys is obvious and, further, students already suffer from"survey fatigue", where over-surveying causes data to become skewed due to students aiming to simply complete surveys, not give objective answers.

Data collection may be followed by a phase of addition, where the data is inspected manually and supplemented with expert evaluations. This is used in studies such as Lopez et al. (2012), where forum posts are viewed and analysed by experts within the field in order to evaluate their worth. This expert rating is then added to the data as a measure of how useful a resource will be to a student, and hence how much potential benefit they may receive from using it. This re-analysis of data, after its collection, is a further demand on academics time as the manual inspection of data is extremely time consuming. We would also argue that students are able to be their own "expert evaluators" of the data that is the most relevant to them. They are able to identify resources that are useful and hence successful students will access and use these resources more frequently.

Visualisation of a network in a learning management system has been carried out in studies such as Dawson et al. (2010). Dawson et al. (2010) showed the structure of the social network created between students when they interact on the forums in a learning management system. The use of student patterns as predictors of final result has been studied in Zhang et al. (2007) and Casey et al. (2010) however they do not present their results visually and give their results in a more "raw" format, which requires a degree of statistical knowledge and insight to understand and work with.

3 Methodology

We aim to create an "at-risk identification" framework by combining simple, automated metrics and simple visualisations for assessing student behaviours. We gather methods from a range of areas, including social network analysis, statistical analysis and data visualisation.

Studies such as Zhang et al. (2007) and Casey et al. (2010) show that successful students are more frequently and regularly participating and engaged in online activities. Much of the work within the area of at-risk identification addresses the early identification of students that exhibit conflicting behaviours, i.e. they are not engaged in the course and are not actively participating. One of the most accurate measures of student engagement is student performance in assessment activities, but this may not promote timely interventions, as assignment work may come too late in a course. We utilise data available in learning management systems, such as access to online course resources and participation in forums, as measures of engagement.

Measures such as frequency of access are somewhat coarse and require more detailed analysis to determine engagement. A student may be accessing many resources but they may not be relevant to the activities currently at hand, or alternatively a student may be accessing only a small number of resources, but those that are the most directly relevant to their work. We would argue that the latter student is the more successfully engaged in their studies, and hence measures of the frequency of accesses only presents a partial picture, and may incorrectly categorise students.

Accordingly, we propose a framework that tracks students activities over time, combined with their frequency of accesses.

We utilise three distinct methods to explore student engagement:

- Social Network Analysis (SNA) - SNA techniques enable us to explore relationships within our "network", which consists of data contained within the LMS, such as forum postings and course resources, the student cohort, and connections from students to the data, i.e. a student may read a forum message posted by another student.
- Frequency of Access Analysis - analysing patterns of access for individual resources and student access patterns over time.
- Measure of Distance Analysis - analysing patterns of access behaviour and similarity between student access patterns both visually and quantitatively.

Using the combination of these metrics, we are able to identify patterns of behaviour based upon participation in online course activities. We have developed visualisations for each approach that can be used by academics to identify students who are demonstrating patterns of at-risk behaviour within the context of their course.

We discuss a students success based upon their grades received in the course. There are five grades awarded:

- High Distinction - A grade of or over 85%. Due to our small sample size we only had one High Distinction student in the course, as such their result has been put with the next grade band down, to form a larger "Distinction" grade band of students.
- Distinction - A grade of or over 75%.
- Credit - A grade of or over 65%.
- Pass - A grade of or over 50%. This is the lowest acceptable passing grade.
- Fail - A grade below 50%. This is the only course failure grade.

3.1 Social Network Analysis

In order to show the data visually we use Social Network Analysis (SNA) techniques such as in Dawson et al. (2010).

We use SNA to give a simple display of the network of students participating in a course. The forum network on an LMS can be considered to be a social network, where students "socialise" by accessing the same materials or interacting asynchronously on the forums. Analysing this network allows us to see the extent with which students are engaging with each other and the course materials.

SNA creates a network made up of a set of "actors" who have a relationship with each other. The actors in the course are students who interact with each other via resources. They are related when two of the same actors have accessed the same resource. It is not necessary for all actors in the network to be related, as both present and absent connections are taken into account. SNA is used to explain the network of actors and the effect of the relationships in the network. In our case we can use absent connections to find those who are disengaged from the network and hence likely to be disengaged from the course.

3.2 Statistical Measures

We will analyse student engagement by measuring student access to the LMS in two ways, access to individual resources and frequency of access to resources. We believe that students who are more engaged with a course will seek to access a wide array of materials and that they will access them frequently as required. We shall count distinct accesses to resources as the number of times a resource has been accessed and a date access as days on which students actively participated with the LMS, by accessing materials or engaging on the forum.

We begin by looking at binary-yes-or-no access counts, with a resource having been accessed or an access occurring on a date. We then use a box-and-whisker plot to show the differences between these accesses frequencies in comparison to other students by grade band. When presented with data from a new student an academic would be able to check their accesses frequencies and give a high level assessment of their engagement. For a more fine-grained analysis we will then look at the distinct number of accesses made to each resource and the number of accesses made on a given day. We will visualise these results using a "heat map", which shows gradients of access frequencies. The heat map shows, using colour gradients, the intensity of activity on this resource or on a particular date. The darker the colour on the heat map the more frequent the activity occurring, i.e. darker points indicate more intense activity. From this we expect to see patterns of student behaviour that we can compare with other students or new data.

After a course has been run we are able to create an averaged pattern of accesses, i.e. what action the majority of students have taken. This may be useful after multiple iterations of the same course as an academic could compare new data to the "average" pattern.

3.3 Implementation

We have implemented our framework using data from the Moodle LMS. Moodle logs a large amount of student participation data in a CSV file, which is readily available in all Moodle installations. The data contains four entries for each "action" performed on the system, these are:

- **Full Name** of the student or lecturer accessing the system.
- **Date** on which the action occurred.
- **Access Type**, a shorthand for the type of action performed.
- **Information**, the name of the resource on which the action was performed.

The set of actions that can be taken on the system is extensive and we shall not cover all possible actions here. The set of actions are broken into five categories of access; User, Course, Resource, Administrative and Forum. Of relevance to us for this research are Resource and Forum accesses, of which all possible actions are:

- **Resource Access**

  Resource View: view a course resource. The information then contains the name of this resource.

- **Forum Access**

  Forum View Forums: View all available forums, i.e. a list of sub-forums if such forums exist.

  Forum View Forum: View all discussion headings in a specific forum.

  Forum View Discussion: View a particular discussion on the forum, also included is the title of the forum.

  Forum Add Post: Add a new post to a discussion forum.

  Forum Update Post: Update a post that was previously created.

Relevant Moodle data from the system is imported into an external MySQL database to support queries. MySQL is a ubiquitous and free resource that is easily installed and is the most popular open source database system in the world (MySQL 2012). The database also contains grade information for the students.

The SQL database is then integrated with a Python program which allows us to extract and process the data. Python is a free, ubiquitous open source product (Python 2012). The output from the Python program is stored which allows the visualisations to be re-run as required without a large amount of space overhead.

The file produced from the Python program is then read into R. R is a statistical program that features many built in libraries capable of helping to visualise the data, as well as run statistical analysis and once again is free and open source (R Core Team 2012). We use R to produce the final visualisations which we use to identify patterns in the data.

The outlined process of creating these visualisations is able to be automated using scripts, which run each of the required programs in order, negating the need to perform all of these steps manually.

The data which we have used is from a typical third year course run at the University of Adelaide. The data represents a course that contained 47 enrolled students, with 44 completing the course and receiving a final grade, and has 22,320 unique data entries logged on the Moodle system for this course.

The spread of student grades were from a high score of 90, down to a low of 38 out of 100. We experimented on data from another course to check the validity of our results and found our results to be typical of the courses under study. We present data from a single course for clarity.

4 Network Visualization

We aim to show the network as a whole and show key points of the network, such as its density and make-up. A view of how grades are spread in the network relative to the density or sparsity of links between students will show if there are any distinct communities of grades within our network.

Students are categorized as having a "link" in the network if they have read a discussion that has been started by another student. We take all relations to be didactic, as directionality is not important, students who are engaging in the course will be represented as having a link with other students accessing the same materials.

In Figure 1 we see the visualization of the network as a whole. This gives us a view of the engagement with the course by showing students linking with the materials.

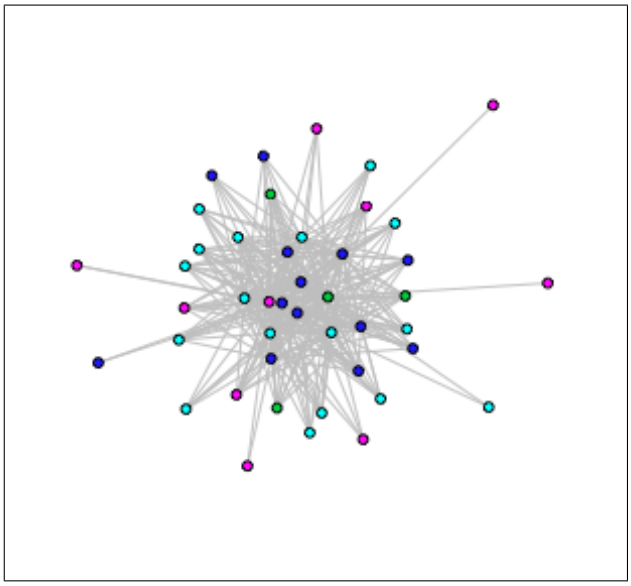

Figure 1: Visualization of the overall Network of students.

We see an incredibly dense network, even for this very small cohort. We see that students on the outer edges of the network are those that are less engaged with the course, and are more likely to fail. This behaviour is supported by similar analysis in Dawson et al. (2010). This gives a good initial guide into the overall structure of the network, and is a quick method to identify students who are potentially at risk as labels are able to be retrieved.

By using this network analysis, educators are able to get a "snapshot" of their class as a whole, potentially allowing them to target students on the fringes of the network for assistance and to get them re-engaged with the course.

From this we can pick out the students who are not well connected to the network. These are the students for whom intervention is needed, as they are not accessing materials. This can be monitored before any assignments have been marked and requires no additional assessment or materials. Further, we can view changes over time to see if students who have been encouraged to interact begin to do so. We cannot, however, see when students have become disengaged who were engaged previously. These students may begin to drift more towards the fringe of the network, but the drift will be slow and likely not noticeable, hence we need different metrics to identify this type of case.

5 Frequency of Access

The materials with which a student engages, and the frequency with which they access them, are more informative than simple measures of overall engagement. Due to the large amount of resources in any one course it may be easy for students to access irrelevant materials, however students who access relevant materials, at the correct time, are those who are most likely to succeed. As such we look at measures of access frequencies and patterns, exploring the behaviours of both successful students and those who fail the course. We shall look at frequency of access in two ways, both of which are engagement through examination:

- By resource, where it is shown how many course resources a student has viewed.
- By date, where it is shown how many access to resources on a particular date are made.

We firstly look at the raw counts of student accesses and compare these in grade bands using box-and-whisker plots. Within these plots our grade bands are numbered, specifically:

- 1 - High Distinction and Distinction grade students.
- 2 - Credit grade students.
- 3 - Pass grade students.
- 4 - Failing grade students.

Hence, the first three boxes all refer to passing grades, while the fourth refers to a failing grade.

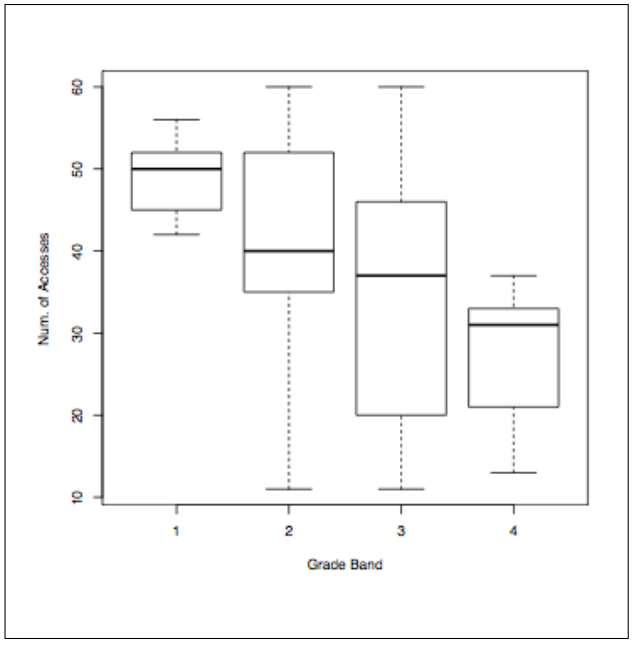

Figure 2: Box and Whisker plot of results by Resource

Figure 2 provides us with a box-and-whisker plot of resource access counts and shows that students in the highest grade bands have a greater mean number of resource accesses, accessing around 50 distinct resources. Credit students have a mean of around 40, similar to that of passing students, however have a much smaller Inter-Quartile Range, or dispersion between students, showing that Credit students on average have a higher access rate than passing students. Finally, failing students have the lowest mean number of resource accesses, with around 30. The bound on the 75th percentile for failing students falls below the mean for passing students (in fact the longest whisker of the failing students is still below the mean for passing students). This shows a very evident difference in behaviour and that there is correlation between frequency of access and a students final grade.

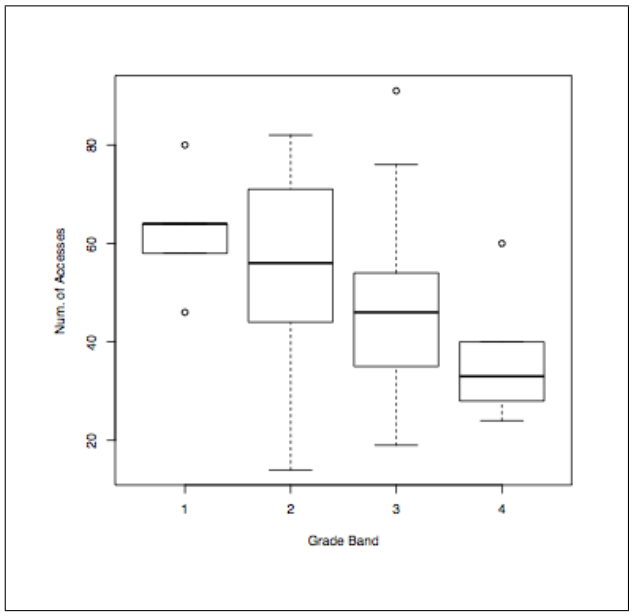

Figure 3: Box and Whisker plot of results by Date

In Figure 3 we look at accesses by date. When viewing by date similar patterns to that of by resource are evident. Again the means increase with grade. Credit students have a large Inter-Quartile Range meaning that their accesses varied, however the mean was still greater than that for passing students. What is clearly evident, is that lower achieving students have a lower rate of access than higher achieving students. Failing students access the forums on around 30 days of a course that ran for over 100 days, in comparison to a mean of 45+ dates for passing students and 60 for Distinction students.

From this we can check an individual students access levels against the rest of the cohort. If their accesses are in the lower ranges then there is a possibility that they may be at-risk. The box and whisker plots give a raw numerical output and do not show more timely accesses or accesses to more important resources.

To achieve a more fine-grained view of student accesses we observe, on an individual level, the engagement with the forums and find if this has a bearing upon the final grade of a student. We visualise our results using a heat map which allows us to observe broad student access patterns, including how frequently students return to view a resource. Key resources may be accessed more frequently, for example a forum post with a long, relevant discussion. In order to show the results clearly we limit the number of "return" events to nine. A student did access some resources 50+ times, however this is considered an outlier and would make it difficult to see lower number of accesses. It would be expected that students that have a higher rate of access are more engaged with the course.

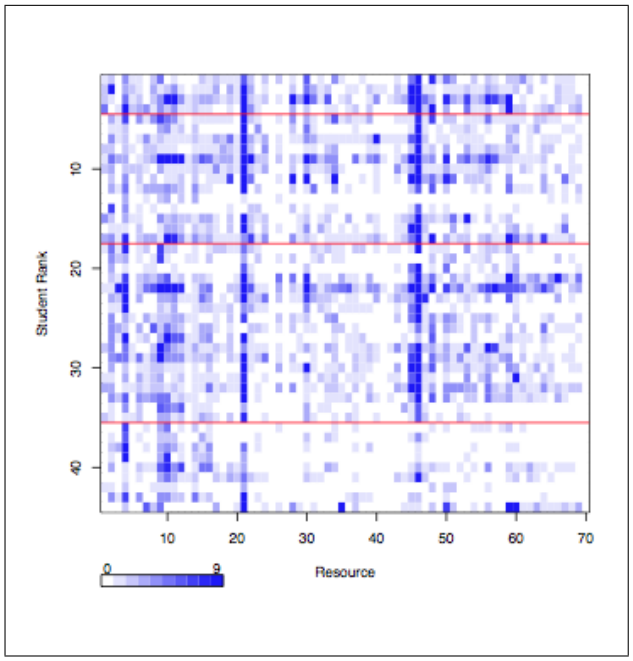

Figure 4: Results of Frequency of Access by Resource. Results are ordered by final grade from top to bottom. Resources are ordered in date of posting.

In our figures we show the edges of the grade bandings with heavy horizontal lines, which are ranked from Distinction to Credit, Pass and Fail. Figure 4 gives us evidence that lower grade students have more sparse access levels that high grade students. The lower access rate for failing students is indeed very clear, hence if we were presented with a student with a similarly low level of engagement we would expect them to obtain a similar result. Similarly, we can see if students are not accessing a key resource as frequently as their peers. For example resource "46" has been accessed multiple times by students in the highest grade bands, however the same result is not seen for students who go on to fail the course. This type of information is critical when evaluating student performance against their peers, and gives likely indicators of why their results may be unsatisfactory.

We now look at the frequency of access by date in Figure 5. Again, we would expect that students who are more successful would return to the forums more frequently over a larger range of days. More over it is expected that they access the materials at relevant times, much like accessing relevant resources.

The matrix clearly shows periods of more intense access and that students with a higher grade are accessing the forums more frequently during these periods. It shows a "ramping up" of accesses at certain periods that correspond to when assignments are due as well as the date of the final examination. As an example, between days 70 and 80, high-achieving students access the forums almost daily while failing students have a considerably lower amount of accesses. Overall we can see that failing students have a lower rate of access.

When presented with a new student an academic can check their raw access counts first, which will give

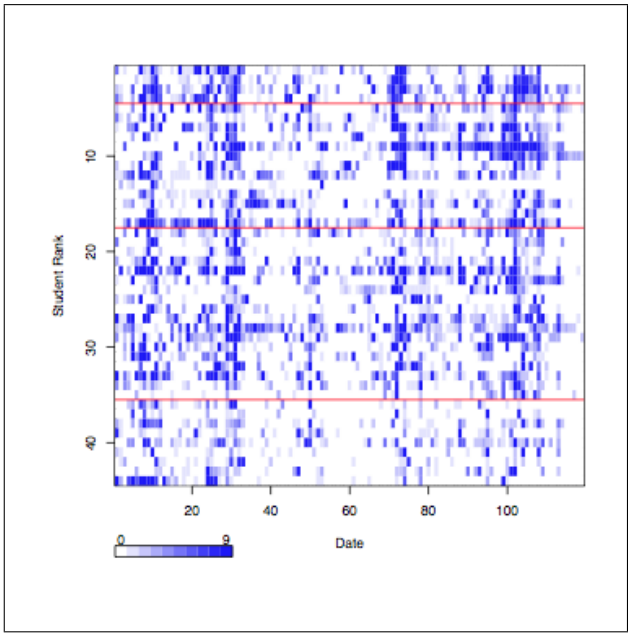

Figure 5: Results of Frequency of Access by Date. Results are ordered by final grade from top to bottom. Dates are from the first day of the course to the end of the examination period.

a good guide as to a students possible result, and then obtain more accurate detail using the heat maps.

6 Distance Measures

Finally, we begin grouping students into grade bands to find if there are particular behavioural patterns over each grade band. We have shown previously that there is correlation between behaviour and the student grade result, we now aim to quantify these differences and examine the behaviour of each grade band. This would be most useful after multiple cohorts have run through the same materials, allowing quick comparisons to multiple years worth of student data. It also shows the relative "distances" between grade bands of students. When presented with a new piece of data an academic could compare it to the averaged results of each grade band, and find which band the new student is closest to, which would be the likely predictor of their grade.

We obtain the distances between students or bands of students as a Hamming Distance, where we calculate the difference in magnitude of two binary "behavioural vectors". These behavioural vectors are obtained by filling a vector with a binary 1 or 0 related to the behaviour, with a 1 either being an access to a particular resource or an access on a particular date, a 0 is the absence of this behaviour. Hamming Distance finds the difference in characteristics between two vectors. The Hamming Distance is equivalent to the count of 1s in s_i XOR s_j. A smaller Hamming Distance will be found for bands with similar access patterns and a large distance for bands with different behaviour patterns. We use a heat map to visualize these differences. For the grade bands we take the average behaviour for that grade band, e.g. if two of three students access a resource it is counted as accessed, if one of three accesses it, it is not. The Hamming Distance is calculated on these averages to find the differences between grade bands. A darker colour indicates a greater similarity, and grade bands are labelled as earlier.

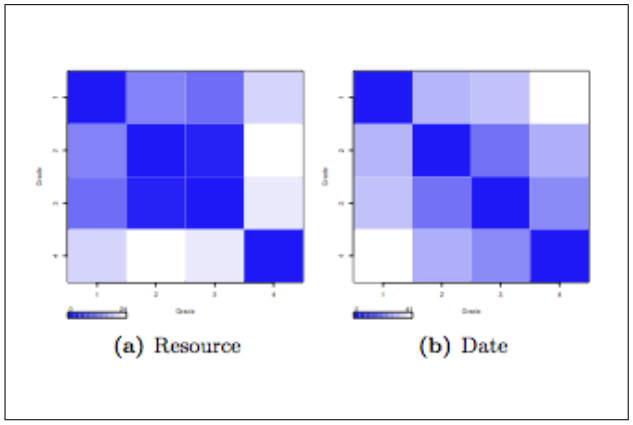

Figure 6: Results grouping by grade banding. Grades run from Distinction to Fail, top to bottom, left to right.

When taking the results averaged over grade band in Figure 6(a) we see differences between high grade and failing students. When looking at resource accesses we see that:

- Credit and pass students are most similar
- Credit and pass students are more similar to distinction students than to failing students.
- Failing students behave differently to all other grades.

When looking at accesses by date in Figure 6(b) we see similar results:

- Failing students are very different to Distinction students.
- Credit and Pass students are still most similar
- Credit students are almost as similar to Distinction students as to Passing students.

From this, when presented with a new student we can combine these two metrics to get the best predictor of their likely result. There may be some difficulty separating Credit and Pass students, however we are, in a large number of cases, able separate students who are failing from those who are passing the course.

7 Discussion

We have created a framework to assist academics in assessing whether a student is at risk, without imposing large additional burdens in time and administration. We have shown that surveys or the manual inspection of data may not be necessary as we were able to achieve useful results using data easily obtained from a commonly used LMS.

We find that a student's pattern of study, as measured by LMS usage, correlates with their final grade in a course. Our results are supported by studies such as Lopez et al. (2012) and Morris et al. (2005), who obtained similar results.

Students who are more engaged with the course perform better in terms of their final grade. This result is not unexpected, however, finding a way to measure engagement and participation is difficult with larger class sizes and current administrative burden. Our framework will provide a method with which academics can measure and visualise participation more

easily. Further, we also find that students who perform better in the course access the most relevant materials at the appropriate times. This is shown by Credit students accessing the same materials as Passing students, but doing so on more appropriate dates.

We present the data obtained visually, allowing it to be more widely understood, including by those without a strong statistical background. This allows academics to identify students who are at risk of failure, if they are found to be demonstrating patterns that correlate to this outcome. We have created this resource without the need for a large amount of manual overhead on the part of the academic and provided them with a simple medium with which to make predictions.

We use this data to guide the creation of an automatic system for detection of failing students, providing an indicator of the types of flags that may be put up as warning signs that students are at-risk.

8 Conclusion and Future Work

We have shown that LMS usage can be correlated with grade, and that we can use simple, highly automated metrics and visualisations to capture students who are potentially at-risk. By doing this we hope that we can increase retention rates as well as course performance and overall facilitate better learning outcomes for students. The system provides an automated tool in order to create these results.

There is the potential for this type of tool to be built directly into Moodle, as shown with SNAPP (Dawson et al. 2010) decreasing the amount of burden to the educator. Further, we can increase automation by checking correlations automatically and creating alerts for course co-ordinators to contact students when necessary, rather than the educator performing this step manually.

Overall, we have shown we are able to find patterns for successful, as well as at risk students, and use these to make predictions about likely outcomes. Doing so may be a step toward decreasing failure rates and increasing retention rates.

References

Bayer, J., Bydzovská, H., Géryk, J., Obšıvac, T. & Popelınskỳ, L. (2012), Predicting drop-out from social behaviour of students, *in* 'Proceedings of the 5th International Conference on Educational Data Mining-EDM 2012', pp. 103–109.

Becker, K. (2008), 'The use of unfamiliar words: writing and cs education', *Journal of Computing Science in Colleges* **24**(2), 13–19.

Biggs, J. & Tang, C. (2007), *Teaching for Quality Learning at University, 3rd edition*, The Society for Research into Higher Education.

Black, E., Dawson, K. & Priem, J. (2008), 'Data for free: Using lms activity logs to measure community in online courses', *The Internet and Higher Education* **11**(2), 65–70.

Casey, K., Gibson, P. & Paris, I. (2010), 'Mining moodle to understand student behaviour', *International Conference on Engaging Pedagogy 2010 (ICEP10), National University of Ireland Maynooth* .

Ceddia, J., Sheard, J. & Tibbey, G. (2007), Wat: a tool for classifying learning activities from a log file, *in* 'Proceedings of the ninth Australasian conference on Computing education-Volume 66', Australian Computer Society, Inc., pp. 11–17.

Dawson, S., Bakharia, A. & Heathcote, E. (2010), Snapp: Realising the affordances of real-time sna within networked learning environments, *in* 'Proceedings of the 7th international conference on networked learning, Aalborg 3-4th May'.

Edwards, S., Snyder, J., Pérez-Quiñones, M., Allevato, A., Kim, D. & Tretola, B. (2009), Comparing effective and ineffective behaviors of student programmers, *in* 'Proceedings of the fifth international workshop on Computing education research workshop', ACM, pp. 3–14.

El-Halees, A. (2009), 'Mining students data to analyze learning behavior: A case study', *Department of Computer Science, Islamic University of Gaza PO Box* **108**.

Falkner, N. & Falkner, K. (2012), A fast measure for identifying at-risk students in computer science, *in* 'Proceedings of the ninth annual international conference on International computing education research', ACM, pp. 55–62.

Fenwick Jr, J., Norris, C., Barry, F., Rountree, J., Spicer, C. & Cheek, S. (2009), 'Another look at the behaviors of novice programmers', *ACM SIGCSE Bulletin* **41**(1), 296–300.

Georg, W. (2009), 'Individual and institutional factors in the tendency to drop out of higher education: a multilevel analysis using data from the konstanz student survey', *Studies in Higher Education* **34**(6), 647–661.

Herbert, N. & Wang, Z. (2007), Student timesheets can aid in curriculum coordination, *in* 'ACM International Conference Proceeding Series', Vol. 239, pp. 73–80.

Lopez, M., Luna, J., Romero, C., Ventura, S., Molina, M., Luna, J., Romero, C., Ventura, S., Cano, A., Luna, J. et al. (2012), Classification via clustering for predicting final marks based on student participation in forums, *in* 'Proceedings of the 5th International Conference on Educational Data Mining, EDM 2012', Vol. 42, pp. 649–656.

McGregor, H., Saunders, S., Fry, K. & Tayler, E. (2000), 'Designing a system for the development of communication abilities within an engineering context', *Australian Journal of Communication* **27**, 83–94.

Merceron, A. & Yacef, K. (2005), Educational data mining: a case study, *in* 'Proceeding of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology', IOS Press, pp. 467–474.

Moodle (2007), 'version 1.9'.
URL: *http://moodle.org*

Morris, L., Finnegan, C. & Wu, S. (2005), 'Tracking student behavior, persistence, and achievement in online courses', *The Internet and Higher Education* **8**(3), 221–231.

Murphy, C., Kaiser, G., Loveland, K. & Hasan, S. (2009), Retina: helping students and instructors based on observed programming activities, *in* 'ACM SIGCSE Bulletin', Vol. 41, ACM, pp. 178–182.

MySQL (2012), 'Mysql: the world's most popular open source database'.
URL: *http://www.mysql.com*

Nandi, D., Hamilton, M., Harland, J., Warburton, G., Hamer, J. & de Raadt, M. (2011), How active are students in online discussion forums?, *in* 'Proceedings of the Australasian Computing Education Conference (ACE 2011)', Australian Computer Society Sydney, pp. 125–134.

Norris, C., Barry, F., Fenwick Jr, J., Reid, K. & Rountree, J. (2008), Clockit: collecting quantitative data on how beginning software developers really work, *in* 'ACM SIGCSE Bulletin', Vol. 40, ACM, pp. 37–41.

Python (2012), 'Python'.
URL: *http://www.python.org/*

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Ramsden, P. (2003), *Learning to Teach in Higher Education*, RoutledgeFalmer, London.

Sheard, J., Carbone, A., Markham, S., Hurst, A., Casey, D. & Avram, C. (2008), Performance and progression of first year ict students, *in* 'Proceedings of the Tenth Australasian Computing Education Conference (ACE 2008)'.

Sheard, J., Ceddia, J., Hurst, J. & Tuovinen, J. (2003), 'Inferring student learning behaviour from website interactions: A usage analysis', *Education and Information Technologies* **8**(3), 245–266.

Zhang, H., Almeroth, K., Knight, A., Bulger, M. & Mayer, R. (2007), Moodog: Tracking students online learning activities, *in* 'Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications', pp. 4415–4422.