Computers & Education 54 (2010) 588–599

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Mining LMS data to develop an "early warning system" for educators: A proof of concept

Leah P. Macfadyen [a,*], Shane Dawson [b]

[a] Science Centre for Learning and Teaching, The University of British Columbia, 6221 University Boulevard, Vancouver, British Columbia, Canada V6T 1Z1
[b] Graduate School of Medicine, University of Wollongong, Wollongong, NSW 2522, Australia

ARTICLE INFO

Article history:
Received 21 May 2009
Received in revised form 31 August 2009
Accepted 3 September 2009

Keywords:
Collaborative learning
Evaluation methodologies
Learning communities
Teaching/learning strategies
Post-secondary education

ABSTRACT

Earlier studies have suggested that higher education institutions could harness the predictive power of Learning Management System (LMS) data to develop reporting tools that identify at-risk students and allow for more timely pedagogical interventions. This paper confirms and extends this proposition by providing data from an international research project investigating which student online activities accurately predict academic achievement. Analysis of LMS tracking data from a Blackboard Vista-supported course identified 15 variables demonstrating a significant simple correlation with student final grade. Regression modelling generated a best-fit predictive model for this course which incorporates key variables such as *total number of discussion messages posted*, *total number of mail messages sent,* and *total number of assessments completed* and which explains more than 30% of the variation in student final grade. Logistic modelling demonstrated the predictive power of this model, which correctly identified 81% of students who achieved a failing grade. Moreover, network analysis of course discussion forums afforded insight into the development of the student learning community by identifying disconnected students, patterns of student-to-student communication, and instructor positioning within the network. This study affirms that pedagogically meaningful information can be extracted from LMS-generated student tracking data, and discusses how these findings are informing the development of a customizable dashboard-like reporting tool for educators that will extract and visualize real-time data on student engagement and likelihood of success.

1. Introduction

Higher education institutions (HEIs) around the world are undergoing rapid changes as they adapt to the new realities of the knowledge society. While the development, maintenance and dissemination of knowledge have long been the primary goals of higher education institutions (Bloland, 1995), recent social and economic changes are forcing universities to adopt new approaches in the way these goals are achieved. Educators are being asked to demonstrate quality teaching practices with decreasing fiscal and human resources, whilst catering to the learning needs of a burgeoning student population that is increasingly diverse in many dimensions (Twigg, 1994, 1994a, 1994b).

Student demographics have radically shifted, and student enrollment numbers have dramatically increased (Patrick & Gaële, 2007). The traditional and idealized experience of higher education as "academically oriented living and learning communities" where "full-time students receive a good deal of faculty contact and many academic support services in the residential setting" (Volkwein, 1999, p. 14) relies upon intensive fiscal and labour resources. Unfortunately, the reality is that federal, state and provincial government funding commitments for higher education have failed to match the escalating sector resourcing requirements (Bates, 2000; Rossner & Stockley, 1997). To supplement reduced government input, institutions have sought to increase enrollments and tuition in order to maximize revenues. In addition, an increased emphasis on the value of higher education for future career prospects has prompted many governments to mandate increased enrollment in public higher education institutions as a matter of policy. Education and (re-)training of the adult workforce has also emerged as a priority. Colleges and universities are now facing increased demand from the workforce sector where ongoing up-skilling and re-training is required to maintain employment in a rapidly changing economic, social and technical world (Dolence & Norris, 1995). Extensive and detailed analysis of the changing context of higher education in the industrialized world can be found in past and current *Education at a Glance* reports published by the OECD (for example, OECD, 2008).

* Corresponding author. Tel.: +1 604 827 3001; fax: +1 604 822 4282.
E-mail addresses: leah.macfadyen@ubc.ca (L.P. Macfadyen), shane_dawson@uow.edu.au (S. Dawson).

0360-1315/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2009.09.008

Educators in higher education are therefore facing several major challenges. Class sizes are increasing, and they are now being asked to teach learners effectively across a much broader spectrum of ability, learning and study skills, prior learning and educational goals. In addition, a growing body of research on teaching and learning in higher education emphasizes the adoption of learner-centred pedagogical approaches and the critical importance of encouraging student–student interaction/collaboration, promoting active learning, providing prompt and detailed feedback, and stressing the need for time on task, whilst also respecting the diverse modes of learning (for example, Chickering & Gamson, 1987). Although the more traditional transmissive model of teaching may have its place in certain pedagogical contexts, it is increasingly perceived as a poor approach for facilitating student learning. At the same time, increased competition in the higher education sector and the rising cost of education means that students and parents are demanding higher quality teaching practices and greater transparency and accountability of expenditure (Mavondo, Zaman, & Abubakar, 2000; Varnham, 2001).

In response, educational institutions are implementing systems of evaluation and assessment to demonstrate their institutional commitment to efficiency, productivity, effectiveness and accountability (Volkwein, 1999). Educators are therefore under pressure to implement and demonstrate effective pedagogical practice, in a context where it is increasingly difficult to monitor student progress through personal contact (Smith, MacGregor, Matthews, & Gabelnick, 2004). At the same time, their own performance as educators is under increasing government, public, management, and student scrutiny.

1.1. The challenge of monitoring student progress

Given the contemporary education context, how can teachers effectively track the progress of their many students? How can they meaningfully assess the impact of newly implemented teaching strategies? While the traditional summative approaches for assessing student learning, such as written examinations and assignments, provide both student and teacher with the necessary feedback on learning progress, access to this information is frequently at a stage of course progression where minimal interventions and support can be implemented to overcome any identified problems. Moreover, summative assessment typically affords minimal insight into student learning strategies or study practices, the development (or not) of learning communities, or the actual degree of student engagement with peers and course materials (Coates, 2005; Richardson, 2005). Most worryingly, as class sizes continue to rise, students who are absent, disconnected or failing to engage sufficiently in coursework are increasingly overlooked, as educators struggle simply to attend to those students who do engage.

Instructors in the new world of higher education are therefore critically in need of new tools and strategies that will allow them to quickly identify at-risk students and devise ways of supporting their learning. Based on some preliminary findings, Wang & colleagues (Wang & Newlin, 2000, 2002; Wang, Newlin, & Tucker, 2001) proposed in 2002 that data on student online activity in a web-based Learning Management System (LMS) such as BlackBoard or Desire 2 Learn may provide an early indicator of student academic performance. More recently, John Campbell and colleagues (Campbell, DeBlois, & Oblinger, 2007; Campbell & Oblinger, 2007), have argued that the application of *academic analytics* (see Section 1.2.3) to institutional LMS data, can offer new insights into student success and identify students at-risk of attrition or course failure. For example, Campbell, Finnegan, and Collins (2006) conducted a regression analysis of student academic performance and selected online activity data. These authors demonstrated that while student SAT scores are mildly predictive of future student success, the inclusion of a second variable – LMS logins – tripled the predictive power of this model. They also presented data indicating that students entering a course with low to moderate SAT scores could achieve success (measured by final grade) through above-average levels of effort (as indicated by number of LMS logins). Drawing on this prior work, Campbell and others (Campbell & Oblinger, 2007; Goldstein & Katz, 2005) concluded that there is a strong relationship between LMS usage patterns and student achievement. Extending Wang and Newlin's proposition (2002), Campbell and colleagues proposed that institutions and teachers could draw direct benefit from the analysis of LMS data in order to develop "early warning" reporting tools that can flag at-risk students and allow instructors to develop early intervention strategies.

In this paper we report on an international research project whose end-goal is to develop a dashboard-like (Campbell & Oblinger, 2007) visualization tool for educators. This evaluative resource will draw upon statistically relevant LMS tracking data, to facilitate more meaningful real-time pedagogical analysis that will allow educators to more easily monitor student engagement and learning progression, as well as evaluate the impact of implemented learning and teaching activities. This paper presents sample data from the exploratory research phase of the project, and describes how these findings are informing the development of a dashboard interface.

1.2. Foundations of an "*early warning system*"

1.2.1. Internet and communication technology (ICT) integration into teaching and learning

Integration of ICTs into teaching and learning has accelerated in the past decade; this has been driven by both pedagogical goals and the need for enhanced flexibility of content delivery and engagement with course materials. The broad number of ICT resources now readily available within HEIs presents numerous pedagogical advantages to both educators and learners. For instance, educators have access to a variety of tools that can assist in the design and delivery of learner-centred courses, and students have greater access to more flexible options for engaging with peers and instructors.

The high level of adoption of ICTs is well documented, with current LMSs ranking in the top 10 technologies for higher education (Yanosky, Harris, & Zastrocky, 2004). Currently, most LMSs are web-based platforms that bring together tools and materials to support learning, including: content files and multi-media resources relevant to the course of study; assessment tools that may permit students to complete online quizzes or submit assignments; communication tools such as mail, chat and asynchronous discussion forums; course administration tools that allow instructors to record and store grades, make announcements and display course deadlines; and 'learning management' tools that allow students to review grades and track their progress. A recent survey of United States (US) higher education institutions indicated a greater than 70% adoption rate of campus-wide LMS (Campus Computing Project, 2008). 2007 figures for the US suggest that greater than 3.9 million students (>20% of all US higher education students) enrolled in at least one online course during the Fall 2007 term (Allen & Seaman, 2008). Even more significant is the uptake of LMS to support 'traditional' classroom-based courses, in formats better described as 'hybrid' (or mixed-mode) and 'web-supported'. For example, while more than 4500 students enrolled in 109 fully online courses

offered by The University of British Columbia (UBC) during 2008, this figure represents a fraction of the total LMS usage across the institution. During 2008, more than 225,000 separate student enrollments were logged in 3169 different LMS-supported course sections taught by more than 3800 instructors (Personal communication, UBC Office of Learning Technologies). Clearly, and as in many HEIs around the world, Blackboard (BB) Vista™ – UBC's enterprise LMS – is a heavily utilized tool and a critical resource for supporting teaching and learning across the institution.

1.2.2. Increased availability of LMS tracking data

LMSs such as BB Vista™ capture and store large amounts of sophisticated user activity and interaction data. User tracking variables include measures such as the number and duration of online sessions (student visits to the online course site), LMS tools accessed, messages read or posted, and content pages visited. These data sets are captured in real time, and can in principle be mined at any stage of course progression. The collection of such data by the LMS is non-intrusive and requires no faculty or staff intervention. Importantly, this data may represent aspects of learner behaviour that are difficult or impossible to apprehend by other means: study patterns, engagement, and degree and mode of participation in learning networks (Dawson, McWilliam, & Tan, 2008).

To date, however, investigators have only been able to access, analyze, visualize and interpret this data via slow and cumbersome manual processes, and current LMSs offer very limited data reporting options (Dawson et al., 2008; Mazza & Dimitrova, 2007). Moreover, very little research exists, and no guidance is available for educators, to indicate which (if any) of the captured tracking variables may be pedagogically meaningful – that is to say, which of the many available data points are indicative of student participation in educationally purposeful activity that may contribute to their learning and achievement in a course.

1.2.3. Emergence of academic analytics

The new millennium has, however, seen the emergence of a new approach to system-wide data harvesting and analysis that has the potential to unlock the value of the vast data sets captured by institutional LMS. Initially developed in the corporate sector as 'business intelligence', Goldstein and Katz (2005) have called the application of these analytical tools and processes to educational systems *academic analytics*. Analytics brings together procedures for capturing, selecting and organizing, storing and reporting on enterprise-wide data. Most significantly, an analytics approach merges data collation with statistical techniques and predictive modelling that can be subsequently used to inform both pedagogical practice and policy.

Until recently, there has been limited interest in analytics within the academy (Goldstein & Katz, 2005). The few projects undertaken within the higher education sector have primarily focused on the analysis of institutionally-collected data on student demographics and overall academic performance, with the goal of understanding and improving institutional recruitment and retention. There are few examples that demonstrate the successful and systematic application of academic analytics across an institution in order to inform and enhance teaching and learning practices.

1.2.4. Increased attention to the social nature of learning

In recent years, educators have increasingly recognized the pedagogical benefits associated with learning design that embraces socio-constructivist principles. Spurred by the seminal works of theorists such as John Dewey (1966), Jean Piaget (1952) and Lev Vygotsky (1962), socio-constructivist pedagogies emphasize learner-centred instructional design, and a recognition of the social nature of learning that calls for dynamic learner interaction with peers, learning materials and teachers (Gabelnick, MacGregor, Matthews, & Smith, 1990; Levine Laufgraben & Shapiro, 2004). The development and support of *learning communities* has therefore become a common goal among educators in their attempt to facilitate student learning (Cho, Lee, Stefanone, & Gay, 2005; Shapiro & Levine, 1999). Further supporting the call for community centric practices, numerous authors have now described the capacity of ICTs to facilitate learner-to-learner communications and engagement, thereby promoting the development of social networks and sense of community (Brook & Oliver, 2003; Hew & Cheung, 2003; Palloff & Pratt, 1999).

In parallel, new approaches are emerging that allow educators to evaluate the learning impact of designed activities in terms of measures of learner community, interaction and engagement. For example, Dawson and colleagues (Dawson, 2006; Dawson, Burnett, & O'Donohue, 2006; Dawson et al., 2008) have demonstrated that LMS tracking variables indicating learner communication behaviours are significant indicators of their 'sense of community'. To date, however, little research has been published that might demonstrate how or if student online communicative practices and sense of community map to their eventual academic success.

These four key developments – ICT integration into teaching and learning, the increased availability of LMS tracking data, the emergence of academic analytics and the increased focus on involving students in effective learning communities – underpin our project.

2. Approach

2.1. Research questions

Initial exploratory research was undertaken to identify the data variables that would inform the development of a data visualization tool for instructors. This involved the extraction of all LMS tracking variables for selected BB Vista™-based course sections at The University of British Columbia, Canada. In so doing, the study aimed to address the following research questions:

- Which LMS tracking data variables correlate significantly with student achievement?
- How accurately can measures of student online activity in a BB Vista™-based course site predict student achievement in the course under study?
- Can tracking data recording online student communication patterns offer pedagogically meaningful insights into development of a student learning community?

2.2. Study population and context

This paper reports on the analysis of BB Vista™ LMS tracking data from three terms (five classes) of a fully online undergraduate biology course offered at The University of British Columbia during 2008. A 'fully online' course is defined as one in which all content delivery, communication and assessment is carried out via the institutional LMS. The course under study is the online version of a core second year course in the undergraduate life sciences program at UBC. This course provides the prerequisite context for subsequent courses in biochemistry, microbiology, physiology, genetics and molecular biology, and thus is a core offering for all students preparing for a career in the health sciences. This comprehensive online course makes extensive use of available BB Vista™ tools to provide access to course content (content pages, learning modules, web links, self-assessment quizzes), promote student communication and engagement (discussion forums, chat), assess student learning (quizzes, assignments), and allow instructors and students to self-manage their learning (mail, calendar, MyGrades, MyProgress, announcements). In order to provide a valid data set for analysis of LMS tracking data, only students completing all coursework were included in the study. Students who failed to complete some or all of the required assignments, graded quizzes, examinations or other graded tasks were subsequently removed from the biology and chemistry data sets. This resulted in a sample size of $N_{students} = 118$ 'completers'.

2.3. Data collection and procedures

2.3.1. Analysis of LMS tracking data

The data analyzed in this exploratory research was extracted from the course-based instructor tracking logs and the BB Vista™ production server. For example, tracking data variables that related to the use of tools implemented in all sections for each course were retrieved from the BB Vista™ production server using the Blackboard PowerSight kit. This kit provides access to the server logs in order to extract a greater set of user tracking information than is currently available within the BB Vista™ instructor 'Tracking' tool. Data collected on each student included 'whole term' counts for frequency of usage of course materials and tools supporting content delivery, engagement and discussion, assessment and administration/management. In addition, tracking data indicating total time spent on certain tool-based activities (assessments, assignments, total time online) offered a total measure of individual student time on task. In some cases only one indicator of several available for a given activity was selected. For example, while this study opted for 'chat room entered' as the key variable for chat room use, the PowerSight kit offers seven other chat-related variables that records user participation in the chat resource. In other cases, variables were combined to give a more accurate and complete measure of tool usage. For instance, counts per student for 'announcements viewed as pop-ups' and 'announcement list viewed' were aggregated to provide a complete score for student use of the announcements tool. Table 1 shows the initial set of course tracking variables examined in this study for relationship with student success in the course. Data was exported into an Excel spreadsheet and merged with final course grade data received from course instructors. This complete student data set was later imported into SPSS for further statistical analysis. As the extracted variables are measured using a variety of scales, the 'transform data' function of SPSS was used to standardize variable data as Z scores and allow assessment of covariance. Sample descriptive data for this online course are presented in Table 2.

2.3.2. Network analysis of course discussion forums

Total discussion forum data was extracted from one section of this online course ($N_{students} = 36$) after course completion, in order to analyze and visualize student learning networks. Forum data was extracted using SNAPP (Bakharia & Dawson, 2008) a JavaScript using Greasemonkey (Greasespot weblog, 2009), a Mozilla Firefox browser extension (Mozilla website, 2009). Greasemonkey allows users to install user-developed scripts to better customize published html pages. When enabled, SNAPP extracts and tabulates all interactions in a BB Vista™ discussion forum and presents this data in vna or xml file format for export (Bakharia & Dawson, 2008). The tabulated data is then used to construct network relationships based on the established forum interactions. (For example, Student A posts a message to the forum and

Table 1
Course tracking variables selected for further analysis.

Total number of online sessions	# Uses of the 'Compile' tool	# Assessments started
Total time online	# Uses of the 'Search' function	# Assessments finished
# Mail messages read	# Visits to MyGrades tool	Time spent on assessments
# Mail messages sent	# Visits to MyProgress tool	# Assignments read
# Discussion messages read	# Uses of the 'Who is online' viewer	# Assignments submitted
Total # discussion messages posted	# Visits to course chat area	Time spent on assignments
# New discussion messages posted	# Files viewed	
# Reply discussion messages posted	# Web links viewed	

Table 2
Sample descriptive indicators for the course under study.

Descriptor variable	Count	SD
N (completers)	118	n/a
Average final grade	60%	13%
Average online sessions/student/term	153	77
Average hours online/student/term	102	56
Average discussion messages read/student/term	4589	8620
Average messages posted/term	72	58
Average # files viewed/term	826	374

Student B replies to that initial post – a relationship is established between Students A and B). The forum relationship data was then imported into NetDraw (Borgatti, 2002), a third party social network visualization tool. NetDraw renders the relationship data to visualize a sociogram of the learning network (see, for example, Fig. 3). It also provides the capacity to undertake various calculations of network properties such as ego-networks (the sub-network of connections established by an individual within a larger network), and measures of network properties such as, betweenness, closeness and degree centralities. For further information, Wasserman and Faust (1994) provide an excellent overview of social network analysis.

3. Results

3.1. Simple (bivariate) correlations of LMS tracking variables with final grade

The development of scatter plots is a useful initial approach for identifying potential correlational trends between variables under investigation (Field, 2005). Fig. 1 shows representative scatter plots of selected LMS variables versus student final grade for this course. Prior studies by Morris, Finnegan, and Wu (2005) and Campbell (2007) indicated that a significant relationship exists between LMS variables and academic performance. Similarly, analysis of scatter plots for the online course in this study demonstrates a positive correlation between a number of LMS data variables and student final grade.

To further interrogate the significance of selected variables as indicators of student achievement in this course, a simple correlation analysis of each variable with student final grade was undertaken. Of the 22 BB Vista™ variables examined, 13 demonstrate a positive and statistically significant correlation with student final grade ($p < .05$) (Table 3). Within the significant subset of the LMS variables, seven demonstrate a medium-large effect size ($r = .30–.50$), with each variable explaining between 9% and 27% of the variance in student final grade. The remaining six variables have a small-medium effect size ($r = .10–.30$), with each explaining from 5% to 9% of variance in student final grade.

As previously highlighted by Morris et al. (2005), this research does to some extent appear to be "documenting the obvious" (p. 229). The results indicate that engaged and discursive students are more likely to complete the course successfully than their less interactive peers.

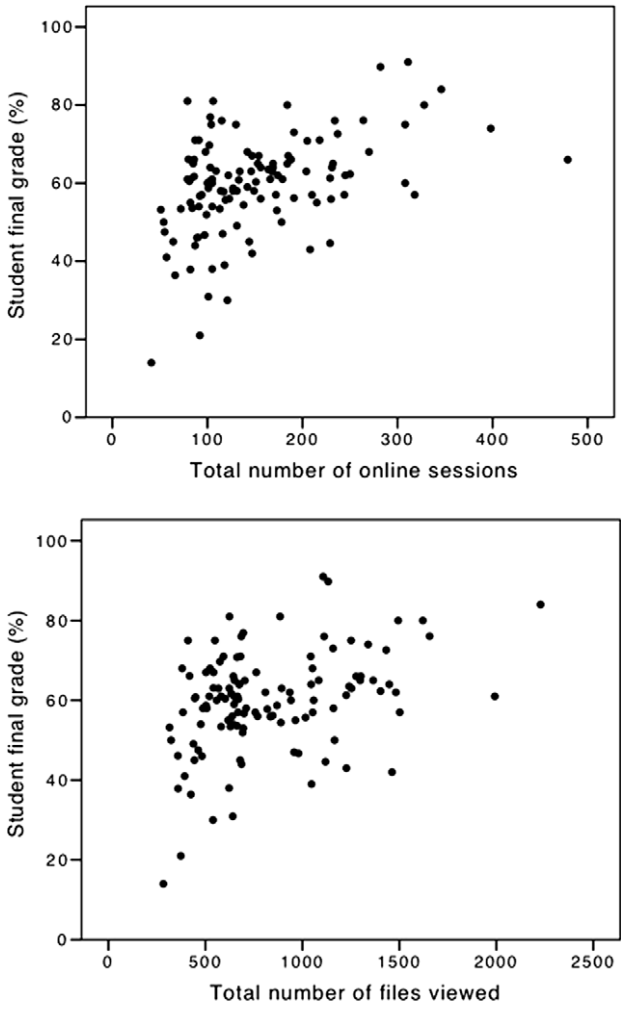

Fig. 1. Scatterplots of selected course LMS tracking variables with student final grade.

Table 3
Simple correlation of relevant course LMS tracking variables with student final grade.

Variable	r_s	r^2	p
Total # discussion messages posted	.52	.27	.00
Total number of online sessions	.40	.16	.00
Total time online	.34	.11	.00
# Files viewed	.33	.11	.00
# Assessments finished	.31	.10	.00
# Assessments started	.31	.09	.00
# Reply discussion messages posted	.30	.09	.00
# Mail messages sent	.28	.08	.00
# Assignments submitted	.26	.07	.00
# Discussion messages read	.25	.06	.00
# Web links viewed	.25	.06	.00
# New discussion messages posted	.24	.06	.01
# Mail messages read	.22	.05	.01
Time spent on assignments	.14	.02	.06
# Visits to MyGrades tool	.14	.02	.06
# Assignments read	.10	.01	.15
# Visits to course chat area	.05	.00	.30
# Uses of the 'Search' function	.01	.00	.45
# Uses of the 'Compile' tool	.01	.00	.48
Time spent on assessments	.00	.00	.48
# Uses of the 'Who is online' viewer	−.02	.00	.43
# Visits to MyProgress tool	−.09	.01	.18
# Views of announcements	−.10	.01	.17

3.2. Multiple regression

Although 13 LMS variables for this course appear to show significant correlation with student final grade, it would be erroneous to rely too heavily on the predictive power of simple correlations. Students do not show simple univariate patterns of online behaviour within course websites, but instead undertake complex composite behaviours in which they make decisions to give more or less time to different tools and activities. *Some* combinations of online activities are likely to translate into effective learning strategies but 'more time spent on online activities' does not simply predict higher achievement. For example, while students in the lowest quartile of achievement (by course grade) in the course investigated here tended to spend, on average, slightly more time online per week, and students in the highest quartile of achievement tended to spend on average slightly less time online per week, students in the middle two quartiles showed no simple or consistent difference in average time online per week (Fig. 2). Indeed, no consistent pattern of average time online in relation to course final

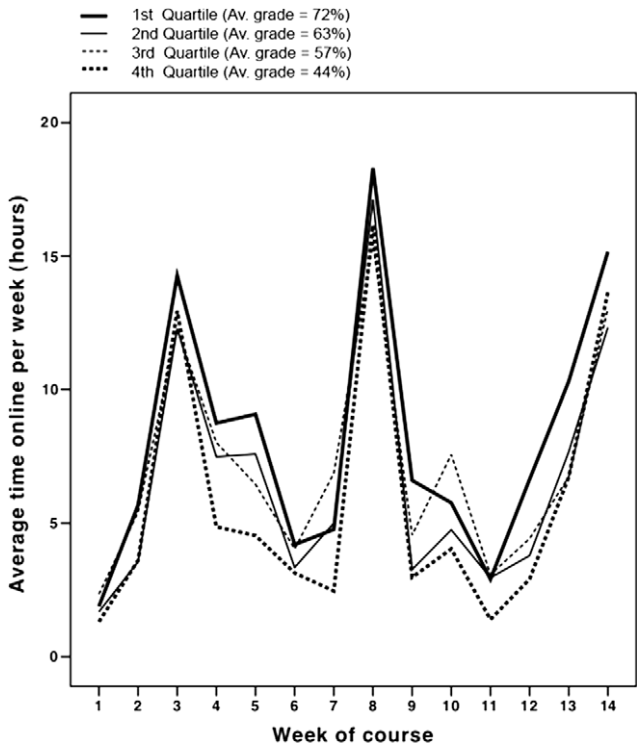

Fig. 2. Weekly average time online by 'achievement quartiles' of students (as determined by final course grade).

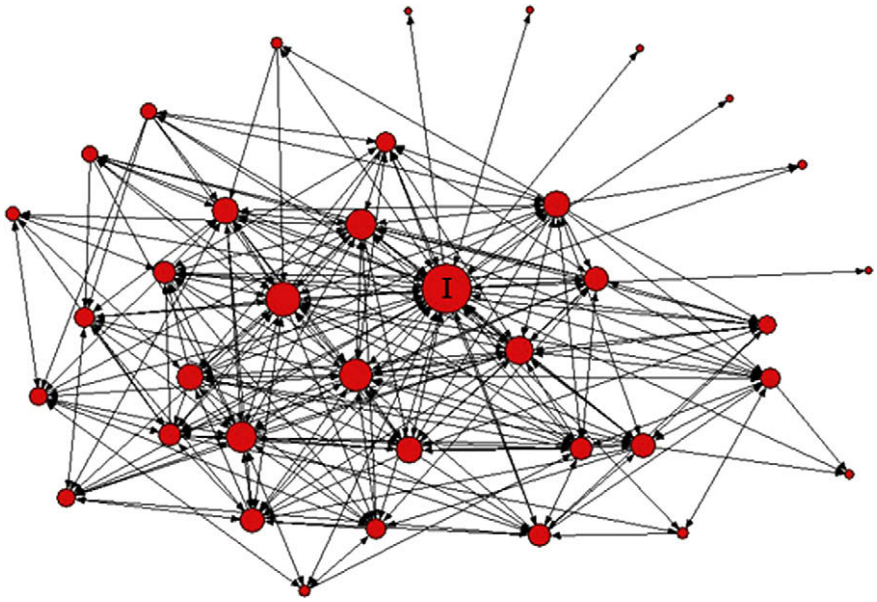

Fig. 3. Sociogram of discussion forum communications in one section of the course. Cohort = 36 students; course discussion forums contained 3145 messages. Size of node represents relative number of connections established by individual.

grade is evident at all. In other words, some students are making more effective strategic decisions about time use within the virtual classroom that is not adequately represented by simple correlations with time online.

Given the deficiencies of simple univariate correlations for this type of study/analysis, we sought to develop a predictive model that incorporates all pertinent and available data sources regarding student online activity. From the set of significantly correlated online biology course LMS variables (Table 3), eight potentially significant indicator variables were identified for inclusion in a multiple regression analysis. All 'assignment' variables were excluded, because although essays submitted via the BB Vista™ 'assignments' tool contribute a significant portion of student final grade in this particular course, these writing tasks involve significant periods of offline reading, research and writing. As such, the BB Vista™ tracking of 'time spent on assignments' is unlikely to accurately reflect student effort in this regard. Moreover, several assignments in this course are completed by student groups and submitted by only one group member; therefore, counts for 'assignments submitted' cannot accurately reflect individual student completion of assignment tasks. In addition the following redundant variables were excluded to avoid problems of multi-collinearity: number of new discussion messages posted, number of reply discussion messages posted, and number of assessments started.

In ideal situations where known predictors have been identified in previous or published work, regression models are generally developed using hierarchical or blockwise approaches. The previously identified predictors are entered into the model in order of their importance in determining the outcome variable (Field, 2005). However, in the absence of such information, a stepwise approach for entering potentially significant variables into a model is a robust and valid approach. In these instances a forward or backward stepwise regression analysis is undertaken. However, a forward regression does present a higher risk in terms of excluding potential predictors that are involved in suppressor effects and thus may generate Type II errors (that is, accidentally excluding significant predictors) (Field, 2005). Consequently, this study elected to adopt a backwards stepwise method, in which variables that are not statistically significant in relation to the predictive power of the model are removed.

A linear multiple regression analysis was therefore conducted, in order to develop a predictive model in which 'Student final grade' was the continuous dependent variable. As shown in Table 4, this process generated a 'best predictive model' of student final grade (F (18.73), $p = .00$) as a linear combination of the LMS tracking data variables measuring only three online activities: *total number of discussion messages posted*, *total number of mail messages sent,* and *total number of assessments completed*. All three variables are statistically significant contributors ($p < .05$). The multiple squared correlation coefficient for this model is .33, indicating that some 33% of the variability in student achievement in this course can be explained by this combination of student online activities within the course site.

To further validate the observed predictor variables a forced entry approach incorporating the significantly correlated variables (Table 3) was undertaken. A forced approach generated a comparable multiple squared correlation coefficient of ~.33 and also determined the same three predictor variables as identified through the backwards stepwise approach.

Table 4
Multiple regression analysis summary for course LMS tracking variables ($N_{students} = 118$).

	Unstandardized coefficients		Standardized coefficient
Variable	*B*	*SE B*	*β*
(Constant)	−3.46 E−16	.08	
Total # discussion messages posted	.44	.08	.44*
# Assessments finished	.18	.08	.18*
# Mail messages sent	.17	.08	.17*

$r = .58$, $r^2 = .33$.
* $p < .05$.

3.3. Logistic regression

A binary logistic regression analysis was conducted to test the reliability of the model in predicting whether or not an individual student is considered 'at risk of failure'. Individual students with course final grade <60% were coded as 'at risk' (0), while students with final grade ⩾60% were coded as 'performing adequately or better' (1). In the UBC grading scheme, <60% represents a grade of C- or poorer; <50% is considered a failing grade (University of British Columbia, 2009). We selected this division point to include students whose final grade indicates that they barely passed the course and may have benefitted from earlier support and intervention.

Details of the regression model are shown in Table 5. For the purposes of this study, the 'hit rate' or predictive power of the model is of greatest significance. Overall, the logistic regression model accurately placed individual students in either the 'at risk' or 'performing adequately' category 73.7% of the time (Table 6). The model resulted in 'Type II' errors (classifying an 'at risk' student as 'performing adequately') at a rate of only 12.7%: 15 students out of 118 were predicted to be performing adequately, while their final course grade placed them in the 'at risk' category. However, of these 15 students, only four actually failed the course (achieving a final grade of <50%), representing a 'predictive failure' rate of only 3.4% (four students out of 118). The logistic model also resulted in Type I errors 13.6% of the time by placing 16 students in the 'at risk' category even though these students eventually passed the course (achieving grades of >60%). However, given the importance of identifying students at risk of failure early in their course progression, the occurrence of Type I errors is of less concern. More simply put, it is better to mistakenly identify a student as at risk of failure then to neglect a student requiring additional learning support. In sum, logistic modelling effectively identified the majority of the students who failed or almost failed this course and who would have been considered 'at risk of failure' by instructors if this data had been accessible earlier in the term.

Table 5
Logistic regression analysis summary for course LMS tracking variables ($N_{students}$ = 118).

Included	95% CI for Exp *b*			
	β (SE)	Lower	Exp *b*	Upper
Constant	.46 (.24)		1.58	
# Mail messages sent	.74* (.35)	1.05	2.10	4.19
Total # discussion messages posted	1.02* (.39)	1.30	2.78	5.98
# Assessments finished	.66* (.26)	1.16	1.93	3.22

Note: r^2 = .31 (Cox & Snell, 1989), .32 (Nagelkerke, 1991). Model χ^2 = 9.59.
* $p < .05$.

Table 6
'Risk of failure' classification results for students in course ($N_{students}$ = 118).

Observed	Predicted		
	At risk	Not at risk	Percentage correct
At risk	38	15	71.7
Not at risk	16	49	75.4
Overall percentage			73.7

'At risk' = final grade <60%; 'Not at risk' = final grade >60%.

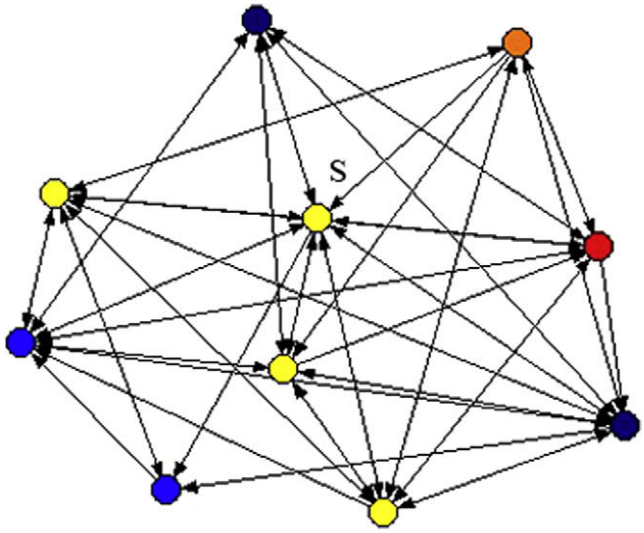

Fig. 4. Ego network of a low-performing student Nodes are colour-coded to indicate individual student final grade. Red = A; Orange = B; Yellow = C; Blue/black = D or Fail. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

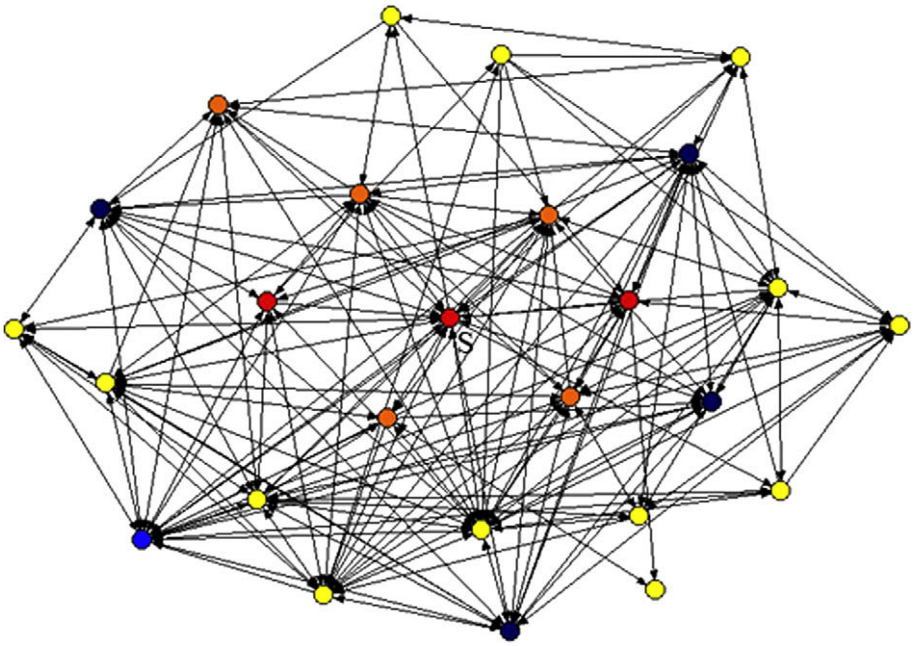

Fig. 5. Ego network of a high-performing student Nodes are colour-coded to indicate individual student final grade. Red = A; Orange = B; Yellow = C; Blue/black = D or Fail. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. Network analysis of asynchronous discussion forums

Exploratory extraction of discussion forum data for analysis highlights the value to educators of representing the social network in an easily interpretable graphic. Visualization of networks affords easy identification of students who are peripheral to or absent from the learning network (Fig. 3). Each node in this sociogram represents a student or instructor participating in the course forum discussions, and relative node size represents 'degree centrality' – the relative number of direct connections made by each user. Although student names have been removed from the network, the sociogram clearly indicates the overall degree of engagement among the cohort, and for each individual student.

This data can be further refined by examining individual or 'ego' networks. Analysis of ego-networks interrogates the specific relationships an individual student has formed via the forum-mediated discussions. Who, exactly, is this student forming communicative relationships with? Figs. 4 and 5 show ego networks for two selected students from one section of the course under study here; nodes have been colour coded to indicate the final grade achieved by individual students. Fig. 4 shows the ego network for a student who achieved a C- grade in this course – in accordance with UBC's grading scheme (University of British Columbia, 2009). This student established communicative connections with only nine other students in a class of 36, over the entire term; four students in his/her ego network achieved a D or F grade, and three others also achieved only a C grade. In other words, this student's network is primarily comprised of low-performing peers. Conversely, Fig. 5 shows the ego network of a student in this section who achieved a final course grade of 84% (A). This high-performing student's much denser ego network clearly illustrates that the student established connections with 25 of their peers, including seven students who achieved an A or B grade in the course. In other words, this student established a complex network that included a core group of high-performing peers. With such information, educators can implement strategic learning interventions to manipulate social structures and promote increased network density and diversity of relationships – hallmarks of an active and functional learning community.

Moreover, this mode of analysis and network visualization affords contemporary teaching staff early opportunities to adapt their teaching practices in order to meet the changing learning dynamics of a given student cohort. For example, Fig. 3 also illustrates the centralized position of the instructor in this learning network. If the desired pedagogical outcome was to generate a student-led learning community, this information would suggest that the instructor may need to modify their activity or intervene strategically to generate additional student to student relationships (although of course it is also feasible that the desired peer to peer interactions have been accomplished in other ways). Generation of network visualizations can therefore also provide instructors with information regarding their own position in the emerging learning network, allowing them to assess the impact of their activity on network behaviour, and whether the learning networks are developing in line with their intended pedagogical outcomes.

4. Discussion and conclusions

4.1. LMS data variables are significant indicators of student success in a course

In this exploratory study we extracted and analyzed relevant student tracking variables from UBC's BB Vista™ production server for a selected online course. Our findings indicate that a regression model of student success, developed using tracking variables relevant to the instructors' intentions and to online course website design (tools implemented to allow content delivery, and/or student engagement, and/or assessment & grading, and/or administration) combined with measures of time on task (variables indicating number of logins and time spent online) explains more than 30% of the variation in student final grade. For the representative course presented here, regression

modeling suggested that three measures of student online activity (number of forum postings, mail messages sent, and assessments completed) function as significantly predictive variables of student final grade. Logistic modelling demonstrated that a predictive model developed using these tracking variables correctly identified students at risk of failure with 70.3% accuracy (correctly classifying 38 of 54 students who achieved a final grade of <60%), and correctly flagged as 'at risk' 80.9% of students who *actually failed* the course (17 of 21 students who achieved a failing grade).

While past studies have focused on 'time online' as a measure of student engagement and effort, our current investigation indicates that measures of total time spent online correlate only weakly with student final grade, and that total time online is not a significant predictor variable of student grade in regression models. The single most significant predictive variable in the biology course reported here, with a regression coefficient (β) of .44 is the tracking variable measuring total student contribution of messages to course discussion forums. These results support the proposition that learning is a social process, and further confirm that the degree of student engagement with peers is an important indicator of success in a course. In the course examined in this study, the assessment tool (often used to administer graded quizzes) was used exclusively as a means of offering students optional (ungraded) self-test quizzes. Thus, in this case, the second significantly predictive LMS variable 'number of assessments finished' represents student engagement with the learning content, rather than their effort in completing graded course assessments. The final significantly predictive variable observed in this study related to the BB Vista™ mail tool. In this course, the mail tool is primarily for administrative communications between instructor and individual students. Hence, data related to mail messages sent can be seen as a measure of student attention to administrative matters related to their learning. In other words, this model suggests that students in this course who take the opportunity to engage with peers via discussion, actively engage with course materials and stay on top of administrative details relating to their participation, achieve higher overall final grades. While this revelation is not surprising, the value of the study lies in the capacity to rapidly determine which students are not engaging with course materials, instructors and fellow peers in a manner that is indicative of an effective online learning strategy.

In summary, this exploratory work has shown that some (but not all) LMS variables are useful predictors of student achievement in an LMS-supported course, but also that the predictive utility of many variables is dependent upon course site design and pedagogical goals. An important finding is that knowledge of actual course design and instructor intentions is critical in determining which variables can meaningfully represent student effort or activity, and which should be excluded. Best-fit predictive models may identify and select different combinations of statistically significant contributor variables depending on LMS course website design (which tools are implemented, and to what end). The initial findings presented in this paper strongly indicate, however, that in order to offer instructors meaningful indicators of student performance, dashboard-style visualizations of student tracking data for a selected course must be highly customizable to reflect pedagogical intent.

4.2. Network analysis of course discussions offers insight into student engagement

Complementing the insights available from analysis of LMS tracking variables, network analysis of communications in course discussion forums permitted further exploration of the learning communities that developed in this course.

It is now widely accepted that student learning is enhanced through ongoing and diverse social interactions with fellow community members. The introduction of Web 2.0 technologies into the educational sphere has further emphasized the importance of promoting ongoing and diversified interactions between students, to better facilitate the learning process. This principle has been further elaborated by investigators such as Astin (1993) and Light (2001), who have demonstrated that robust and diverse peer networks are an important influencing factor on student study persistence and overall academic success. However, not all attempts at developing highly engaging social interactions are effective. To date, evaluation of teaching practices designed with the goal of promoting community development has largely been confined to retrospective reflections or to simple measures of the quantity of discussion postings generated. While, such data can provide some valuable information, they do not allow educators or readers to accurately ascertain real levels of student engagement or assess the complexity of developing student learning networks. A key element in *this* project, however, has been the generation of a tool that allows visualization and exploration of network relationships fostered through student forum discussions (Bakharia & Dawson, 2008), and which can therefore offer a more accurate representation of student engagement in a course. Monitoring the evolution of student networks, can offer educators a comprehensive representation of how the student learning community is progressing, even in very large classes. Closer analysis of ego-network data can reveal important underlying details of peer interactions between students. For example, examination of student network development in the course under study here confirms the phenomenon reported by Dawson (2009): as student networks develop, and in the absence of instructor intervention, 'like flocks to like'. High-performing students predominantly liaise with students of similar academic ability; the lowest performing group of students also tend to cluster with each other in online discussions. The development of tools that allow educators to monitor the development of peer relationships provides an opportunity for staff to introduce learning interventions that better promote student interaction, and that actively manipulate and re-engineer the network to better promote relationship diversity.

4.3. Limitations of this study

There are a number of limitations that impact the overall generalizability and interpretation of the findings of this preliminary study. For example, the implications of the study are limited by its focus on data derived from a fully online course within one institution. In fully online courses, it is reasonable to expect that the only venue for student interaction with peers, instructors and course content is via the LMS. This suggests that indicators of student online activity will represent a large proportion of their overall course-related learning activity. Future studies should be directed towards the investigation and analysis of potential significant LMS tracking indicators and network measures in relation to student success for alternate pedagogical designs and course delivery modalities.

4.4. Proof of concept

In this paper we have presented sample findings from the exploratory phase of our study investigating the potential for LMS tracking data variables to be used as predictors and lead indicators of student academic success in a course. We are mindful that correlations do not

necessarily indicate causality: it cannot, for example, be concluded that 'posting more messages' is causally connected to 'achieving a higher grade'. However, the significance of this study lies not in causation but in *indication*.

Our findings confirm and extend earlier propositions (Campbell et al., 2007; Campbell & Oblinger, 2007; Morris et al., 2005; Wang & Newlin, 2002) that pedagogically meaningful information can be extracted from LMSs and made available to educators via a dashboard-like interface that incorporates predictive models and network visualization tools. Student tracking data are readily accessible, scalable and non-intrusive and provide sound lead indicators of eventual student achievement or failure. Patterns of asynchronous communication can indicate to instructors whether socio-constructivist instructional goals are being achieved. In apparent contrast to our findings, Campbell (2007) has reported that the inclusion of LMS data added only minimally to the power of his model for predicting student academic success. We suggest, however, that this weak predictive power may be due to the very broad scope of his model. Campbell sought to predict student overall success across an entire academic program, and consequently was unable to consider the differential intentions and differential tool implementation of individual instructors within individual courses. Our findings suggest that for the purposes of monitoring student activity and achievement, predictive models must be developed at the course level. Furthermore, future developments of any evaluative and data visualization resources must be highly customizable to cater to instructor differences for adopting LMS tools and their overarching pedagogical intent. A 'one size fits all' dashboard tool will not accurately reflect both pedagogical intention and the subsequent observed online behaviour in a manner that affords instructors the capacity to predict student success. However, in the rapidly evolving context of higher education, the availability of a *customizable* reporting tool that extracts and visualizes real-time data on student engagement and likelihood of success – indicating which students are on track and which may need additional help – will be an invaluable resource for all contemporary educators.

Acknowledgements

Support for this publication has been provided by the Australian Learning and Teaching Council Ltd., an initiative of the Australian Government Department of Education, Employment and Workplace Relations. The views expressed in this publication do not necessarily reflect the views of the Australian Learning and Teaching Council. Additional support for this project has been provided by the Teaching and Learning Enhancement Fund at The University of British Columbia.

References

Allen, I. E., & Seaman, J. (2008). *Staying the course. Online education in the United States, 2008*. Newburyport, MA: The Sloan Consortium. <http://www.sloan-c.org/publications/survey/pdf/staying_the_course.pdf>.

Astin, A. (1993). *What matters in college: Four critical years revisited*. San Francisco: Jossey-Bass.

Bakharia, A., & Dawson, S. (2008). *Social network adapting pedagogical practice* (SNAPP). <http://research.uow.edu.au/learningnetworks/seeing/action/index.html>. Version 1.1).

Bates, A. W. (2000). *Managing technological change*. San Francisco: Jossey-Bass Publishers.

Bloland, H. G. (1995). Postmodernism and higher education. *Academic Matters: The Journal of Higher Education, 66*(6), 521–559.

Borgatti, S. P. (2002). *NetDraw: Graph visualization software*. Harvard, MA: Analytic Technologies (Software).

Brook, C., & Oliver, R. (2003). Online learning communities: Investigating a design framework. *Australian Journal of Educational Technology, 19*(2), 139–160.

Campbell De Blois, P. B., & Oblinger, D. G. (2007). Academic analytics. *A New Tool for a New Era. EDUCAUSE Review, 42*(4), 42–57.

Campbell, & Oblinger, D. (2007). *Academic analytics*. Washington, DC: EDUCAUSE Center for Applied Research. <http://connect.educause.edu/library/abstract/AcademicAnalytics/45275>.

Campbell, J. (2007). *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study*. Doctoral thesis, Purdue University, Indiana, USA.

Campbell, J. P., Finnegan, C., & Collins, B. (2006). *Academic analytics: Using the CMS as an early warning system*. Paper presented at the WebCT impact conference 2006. <http://www.alt.usg.edu/publications/impact2006/campbellfinnegancollinsgage_impact06.ppt>.

Campus Computing (2008). *The 2008 Campus computing survey*. Encino, CA: The campus computing project. <http://www.campuscomputing.net/survey-summary/2008>-campus-computing-survey.

Chickering, A. W., & Gamson, Z. (1987). Seven principles for good practice in undergraduate education. *AAHE Bulletin, 39*(7), 3–7.

Cho, H., Lee, J.-. S., Stefanone, M., & Gay, G. (2005). Development of computer-supported collaborative social networks in a distributed learning community. *Behaviour and Information Technology, 24*(6), 435–447.

Coates, H. (2005). The value of student engagement for higher education quality. *Quality in Higher Education, 11*(1), 25–36.

Cox, D. R., & Snell, D. J. (1989). *The analysis of binary data* (2nd ed.). London: Chapman & Hall.

Dawson, S. (2006). A study of the relationship between student communication interaction and sense of community. *The Internet and Higher Education, 9*, 153–162.

Dawson, S. (2009). 'Seeing' the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology*. Available online 2 June 2009.

Dawson, S., Burnett, B., & O'Donohue, M. (2006). Learning communities: An untapped sustainable competitive advantage for higher education. *The International Journal of Educational Management, 20*(2), 127–139.

Dawson, S., McWilliam, E., & Tan, J. P.-L. (2008). *Teaching Smarter: How mining ICT data can inform and improve learning and teaching practice.* Paper presented at the ASCILITE 2008, Melbourne. Australia. <http://www.ascilite.org.au/conferences/melbourne08/procs/dawson.pdf>.

Dewey, J. (1966). *Democracy and education*. New York: Free Press.

Field, A. (2005). *Discovering statistics using SPSS*. London: Sage.

Dolence, M. G., & Norris, D. M. (1995). *Transforming higher education: A vision for learning in the 21st century*. Ann Arbor, MI: Society for College and University Planning.

Gabelnick, F., MacGregor, J., Matthews, R., & Smith, B. (1990). *Learning communities: Creating connections among students faculty and disciplines*. San Francisco: Jossey-Bass.

Goldstein, P. J., & Katz, R. N. (2005). *Academic analytics: The uses of management information and technology in higher education*. Washington, DC: EDUCAUSE Center for Applied Research.

Greasespot weblog (2009). The weblog about Greasemonkey. <http://www.greasespot.net/>.

Hew, K. F., & Cheung, W. S. (2003). Models to evaluate online learning communities of asynchronous discussion forums. *Australian Journal of Educational Technology, 19*(2), 241–259.

Levine Laufgraben, J., & Shapiro, N. (2004). *Sustaining and improving learning communities*. San Francisco: Jossey-Bass.

Light, R. J. (2001). *Making the most of college: Students speak their minds*. Cambridge, MA: Harvard University Press.

Mavondo, F., Zaman, M., & Abubakar, B. (2000). *Student satisfaction with tertiary institution and recommending it to prospective students*. Paper presented at the Australia New Zealand Marketing Academy, Griffith University, Gold Coast, Australia.

Mazza, R., & Dimitrova, V. (2007). CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses. *International Journal of Human–Computer Studies, 65*, 125–139.

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

Mozilla (2009). Firefox information and download site. <http://www.mozilla.com/en-US/firefox/firefox.html>.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

OECD. (2008). *Education at a Glance 2008: OECD Indicators*. Paris: Directorate for Education, Organisation for Economic Co-operation and Development. <http://www.oecd.org/edu/eag2008>.
Palloff, R., & Pratt, K. (1999). *Building learning communities in cyberspace*. Jossey-Bass: San Francisco.
Patrick, C., & Gaële, G. (2007). Exploring Access and Equity in Higher Education: Policy and Performance in a Comparative Perspective. *Higher Education Quarterly, 61*(2), 136–154.
Piaget, J. (1952). *The origins of intelligence in children*. New York: International University Press.
Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education, 30*(4), 387–415.
Rossner, V., & Stockley, D. (1997). Institutional perspectives in organizing and delivering web-based instruction. In B. Kahn (Ed.), *Web-based instruction* (pp. 333–336). Englewood Cliffs, NJ: Educational Technology Publications.
Shapiro, N., & Levine, J. (1999). Introducing learning communities. *About Campus, 4*(5), 2–10.
Smith, B., MacGregor, J., Matthews, R., & Gabelnick, F. (2004). *Learning communities: Reforming undergraduate education*. San Francisco, CA: Jossey-Bass.
Twigg, C. A. (1994a). The changing definition of learning. *Educom Review, 29*(4), 23–25.
Twigg, C. A. (1994b). The need for a national learning infrastructure. *Educom Review, 29*(5), 17–20.
Twigg, C. A. (1994c). Navigating the transition. *Educom Review, 29*(6), 21–24.
University of British Columbia (2009). *Calendar 2009/2010. Policies and regulations.* <http://www.students.ubc.ca/calendar/index.cfm?tree = 3,42,96,0>.
Varnham, S. (2001). Straight talking, straight teaching: Are New Zealand tertiary institutes potentially liable to their students under consumer protection legislation? *Education and the Law, 13*(4), 303–317.
Volkwein, J. F. (1999). *The four faces of institutional research. What is institutional research all about? New directions for institutional research, 104*. San Francisco: Jossey-Bass.
Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.
Wang, A. Y., & Newlin, M. H. (2000). Characteristics of students who enroll and succeed in web-based psychology classes. *Journal of Educational Psychology, 92*(1), 137–143.
Wang, A. Y., & Newlin, M. H. (2002). Predictors of performance in the virtual classroom: Identifying and helping at-risk cyber-students. *The Journal of Higher Education. Academic Matters, 29*(10), 21–25.
Wang, A. Y., Newlin, M. H., & Tucker, T. L. (2001). A discourse analysis of online classroom chats: Predictors of cyber-student performance. *Teaching of Psychology, 28*, 221–225.
Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.
Yanosky, R., Harris, M., & Zastrocky, M. (2004). *Top technology priorities for higher-education e-learning.* Stamford, CT: Gartner Research Group. <http://www.mtholyoke.edu/lits/ris/CourseMgmt/Gartner/top_tech_priority-g.pdf>.