Exploring the function of discussion forums in MOOCs: comparing data mining and graph-based approaches

Lorenzo Vigentini
Learning & Teaching Unit
UNSW Australia,
Lev 4 Mathews, Kensington 2065
+61 (2) 9385 6226
l.vigentini@unsw.edu.au

Andrew Clayphan
Learning & Teaching Unit
UNSW Australia,
Lev 4 Mathews, Kensington 2065
+61 (2) 9385 6226
a.clayphan@unsw.edu.au

ABSTRACT

In this paper we present an analysis (in progress) of a dataset containing forum exchanges from three different MOOCs. The forum data is enhanced because together with the exchanges and the full text, we have a description of the design and pedagogical function of forums in these courses and a certain level of detail about the users, which includes achievement, completion, and in some instances more details such as: education; employment; age; and prior MOOC exposure.

Although a direct comparison between the datasets is not possible because the nature of the participants and the courses are different, what we hope to identify using graph-based techniques is a characterization of the patterns in the nature and development of communication between students and the impact of the ‘teacher presence’ in the forums. With the awareness of the differences, we hope to demonstrate that student engagement can be directed ‘by-design’ in MOOCs: teacher presence should therefore be planned carefully in the design of large-scale courses.

Keywords

MOOCs, Discussion forums, graph-based EDM, pedagogy.

1. INTRODUCTION

In the past couple of years MOOCs (Massive Open Online Courses) have become the center of much media hype as disruptive and transformational [1, 2]. Although the focus has been on a few characteristics of the MOOCS – i.e. free courses, massive numbers, massive dropouts and implicit quality warranted by the status of the institutions delivering these courses – a rapidly growing research interest has started to question the effectiveness of MOOCS for learning and their pedagogies. If one ignores entirely the philosophies of teaching driving the design and delivery of MOOCs going from the the socio-constructivist (cMOOC, [4, 5]) to instructivist (xMOOC, [3]), at the practical level, instructors have to make specific choices about how to use the tools available to them. One of these tools is the discussion forum. Forums are one of the most popular asynchronous tools to support students’ communication and collaboration in web-based learning environments [6]. These can be deployed in a variety of ways, ranging from a tangential support resource which students can refer to when they need help, to a space for learning with others, driven by the activities students have to carry out (usually sharing work and eliciting feedback). The latter, in a sense, emulates class-time in traditional courses providing a space for structured discussions about the topics of the course. One could argue that like in face-to-face classes, the value of the interaction depends on the importance attributed to the forums by the instructors. This is an interesting point to explore teachers’ presence and the value of their input in directing such conversations. Mazzolini & Maddison characterize the role of the teacher and teacher presence in online discussion forums as varying from being the ‘sage on the stage’, to the ‘guide on the side’ or even ‘the ghost in the wings’ [7]. Furthermore they argue that the ‘ideal’ degree of visibility of the instructor in discussion forums depends on the purpose of forums and their relationship to assessment. There are also a number of accounts indicating that students’ learning in forums is not very effective [8, 9]. However if one looks at the data there are numerous examples indicating that behaviours in forums are good predictors of performance in the courses using them, particularly if forum activities are assessed [10,11,12,13]. Yet, forums in MOOCs tend to attract only a small portion of the student activity [14]. This is setting forums in MOOCs apart from ‘tutorial-type’ forums used to support students’ learning in online or blended courses in higher education. Furthermore, some argue that active engagement is not the only way of benefiting from discussion forums [15] and students’ characteristics and preferences could be more important than the course design in determining the way in which they take full advantage of online resources [16].

2. THE THREE MOOCS IN DETAIL

In order to investigate the way in which students use the discussion forums, we have extracted data from three MOOCS delivered by a large, research intensive Australian university. The three courses are: P2P (From Particles to Planets - Physics); LTTO (Learning to Teach Online); and INTSE (Introduction to Systems Engineering), which are broadly characterised in the top of Table 1. The courses were specifically designed in quite different ways to test hypotheses about their design, delivery and effectiveness.

In particular, P2P was designed emulating a traditional university course in a sequential manner. All content was released on a week-by-week basis dictating the pace of instruction. LTTO and INTSE, instead were designed to provide a certain level of flexibility for the students to elect their learning paths. All content was readily available at the start, however for LTTO, the delivery followed a week-on-week delivery focusing on the interaction with students and a selective attention to particular weekly topics (i.e. weekly feedback videos driven by the discussion forums as well as weekly announcements). Although announcements were used also in INTSE, the lack of weekly activities in the forums did not impose a strong pacing. In INTSE, the forums had only a tangential support value and were used mainly to respond to students' queries and to clarify specific topics emerging from the quizzes. Table 1 provides an overview of the different courses. This also shows that the forum activity in the various courses is a very small portion of all actions emerging from the logs of activity which has been reported in the literature [9].

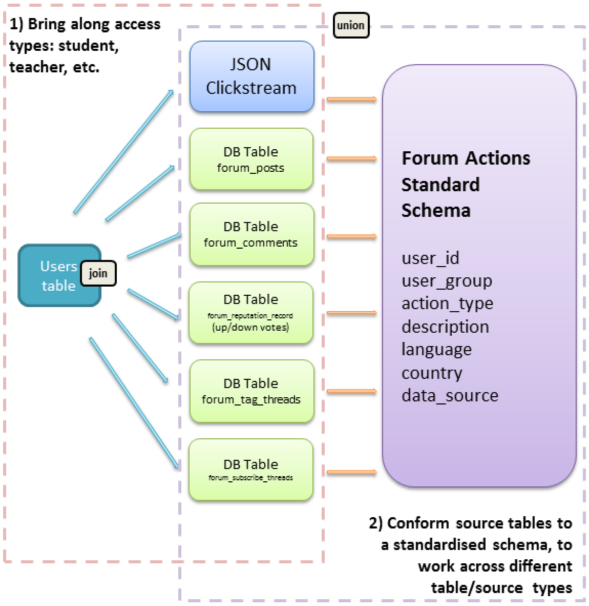
3. DETAILS OF THE DATASET

3.1 The dataset

The data under consideration is an export form the Coursera platform. Raw forum database tables (posts, comments, tags, votes) as well as a JSON based web clickstream were used. The clickstream events consist of a key which specifies action – either a 'pageview' or 'video' item. Forum clickstream events were identified by a common '/forum' prefix.

The clickstream was further classified into: browsing; profile lookups; social interaction (looking at contributions); search; tagging; and threads. From the classification it became evident the clickstream did not record all events, such as when a post or comment was made, or when votes were applied. For these, specific database tables were used. In order to manage different data sets and sources, a standardized schema was built, allowing disparate sources to feed into, but exposing a common interface to conduct analysis over forum activities. This is shown in Figure 1.

Figure 1. Forum data transformation process


1) Bring along access types: student, teacher, etc.
union
JSON Clickstream
DB Table forum_posts
DB Table forum_comments
DB Table forum_reputation_record (up/down votes)
DB Table forum_tag_threads
DB Table forum_subscribe_threads
Users table
join
Forum Actions Standard Schema
user_id
user_group
action_type
description
language
country
data_source
2) Conform source tables to a standardised schema, to work across different table/source types

3.2 An overview of forums activity

There are very interesting trends which require more detailed examination (bottom of table 1). As expected, in LTTO the forum activity is larger than in the other courses and this is probably due to the fact that students were asked to submit post in forums following the learning activities. The proportion of active students in forum is 4x in magnitude compared to the other courses. Yet, if we look at the average amount of posts or comments, the patterns are not straightforward to interpret, as the level of engagement is similar across the courses with 3 to 5 posts per student and 1 to 3 comments (i.e. replies to existing posts), but with P2P showing a higher level of engagement than the other courses. One possible explanation is the different target group of the different courses with INTSE including a majority of professional engineers with postgraduate qualifications, P2P focusing on high school student and teachers, and LTTO targeting a broad base of teachers across different educational levels.

	INTSE	LTTO	P2P
Target group	Engineers	Teachers at all levels	High school and teachers
Course length	9 weeks	8 weeks	8 weeks
Forums	54 (14 top level)	105 (17 top-level)	63 (15 top-level)
Design mode	All-at-once	All-at-once	Sequential
Delivery mode	All-at-once	Staggered	Staggered
Use of forums	**Tangential**	**Core activity**	**Support**
N in forum	422 (2.1%)	1685 (9.3%)	293 (2.8%)
Tot posts	1361 (avg=3.3)	6361 (avg=3.8)	1399 (avg=4.8)
Tot comments	285 (avg=0.7)	2728 (avg=1.7)	901 (avg=3.1)
Registrants	32705	28558	22466
Active students[1]	60%	63%	47%
Completing[2]	4.2% (0.3% D)	4.4% (2.4 D)	0.7% (0.2%)

Table 1. Summary of the courses under investigation. NOTE: 1. Active students are those appearing in the clickstream; 2. Completing students achieve the pass grade or Distinction (D)

The type of activity is summarised in Figure 2. In the chart, the five categories refer to the following: View corresponds to listing forums, threads and viewing posts; Post is the writing of a post or start of a new thread; Comment is a reply to an existing post; Social refers to all actions engaging directly with other's status (up-vote, down-vote and looking at profiles/reputation); Engage refers to the additional interaction with forums content (searching, tagging, 'watching' or subscribing to posts or threads).

The viewing behaviour is the most prominent for both the student and instructor groups and the figures are pretty much similar across the board. A two-way ANOVA (2x5, role by activity) on the percentage of distributions, shows that there is no significant difference between students and instructors, but there is an obvious difference between views and the other types of behaviour ($F(4,29) = 1656.3$, $p < .01$).

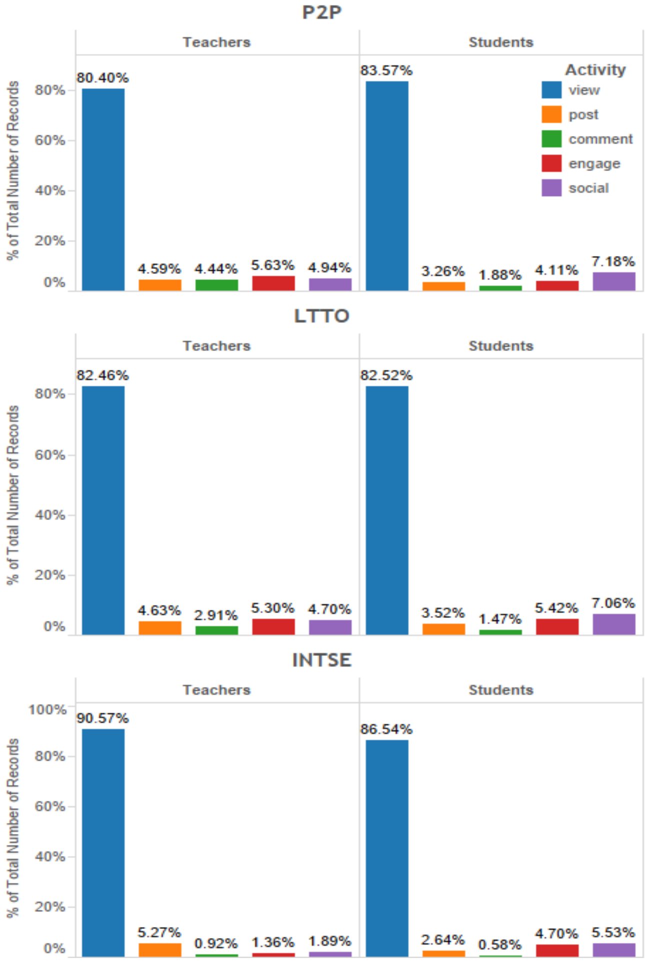
If we consider the engagement over the timeline and compare the type of activities carried out by students and instructors, Figure 3 (end of the paper) shows the patterns for the three courses. The most striking pattern is that there doesn't seem to be an obvious one. For what concerns posts and views in all the three courses there is a sense of synchronicity between the two groups, however from this chart it is not possible to understand in more detail what are the connections between what students and teachers do.

Instructors' comments are slightly offset, possibly as a reaction to students' posts. An interesting aspect is the amount of 'social' engagement in the P2P course that merits further analysis.

4. DIRECTIONS AND OPEN QUESTIONS

From this coarse analysis it is apparent that there seem to be minimal behavioural differences in the way students and instructors interact in the different courses, however more analysis is required to tackle questions about the individual differences in students' and instructors' patterns of interaction and their interrelations. Furthermore little can be said about how the nature of interactions drives the development of communication and engagement. However a number of questions like the following remain open and unanswered: how do discussions develop over time? How teacher presence affects the development of discussions? Is the number of forums affecting how students engage with them (i.e. causing disorientation)?

Figure 2. Distribution of forum activities by role



4.1 The DM and graph-based approaches

A possible way to answer the questions about the types/patterns of behaviours, the structure and development of networks and the growth of groups/communities over time might be using data mining and graph-based approaches. For example, [6] used a combination of quantitative, qualitative and social network information about forum usage to predict students' success or failure in a course by applying classification algorithms and classification via clustering algorithms. In their approach the activity of students in the forums is organized according to a set of commonly used quantitative metrics and a couple of measures borrowed from Social Network Analysis (table 2). Although this seems to be a promising approach, there are two issues with this methodology in the MOOCs: 1) only a tiny proportion of students can be considered active and 2) it is hard to scale the instructor's evaluation. The first problem is not easily resolved and it is an issue in the literature reviewed [17, 18]; non-posting behavior is considered as an index of disengagement, partly because this is easy to measure. In principle the latter could be substituted by peer evaluation (up-vote, down-vote), but there is no easy way to ensure consistency.

Indicator	Type	Description
Messages	Quantitative	Number of messages written by the student.
Threads	Quantitative	Number of new threads created by the student.
Words	Quantitative	Number of words written by the student.
Sentences	Quantitative	Number of sentences written by the student.
Reads	Quantitative	Number of messages read on the forum by the student.
Time	Quantitative	Total time, in minutes, spent on forum by the student.
AvgScoreMsg	Qualitative	Average score on the instructor's evaluation of the student's messages.
Centrality	Social	Degree centrality of the student.
Prestige	Social	Degree prestige of the student.

Table 2. Possible indicators characterising forum engagement

An alternative method that can be explored is graph-based approaches. For example, Bhattacharya et al. [19] used graph-based techniques to explore the evolution of software and source branching providing an insight in the process. Kruck et al. [20] developed GSLAP, an interactive, graph-based tool for analyzing web site traffic based on user-defined criteria.

Kobayashi et al. [18] used a method to quickly identify and track the evolution of topics in large datasets using a mix of assignment of documents to time slices and clustering to identify discussion topics. Yang et al [21] integrated graph-based clustering to characterize the emergence of communities and text-based analysis to portray the nature of exchanges. In fact, students move in the various sub-forums taking different roles or stances as they engage with different subsets of students. As the reasons to engage in these discussions are partly determined by different interests, goals, and issues, it is possible to construct a social network graph based on the post-reply-comment structure within threads. The network generated provides a possible view of a student's social participation within a MOOC, which may indicate some detail about their values, beliefs and intentions.

Furthermore, Brown et al [22] have already shown the value of exploring the communities in discussion forums in MOOCs particularly for what concerns the homogeneity of performance but dissimilarity of motivations characterizing student hubs.

4.2 Discussion points

The examples above provide evidence of the potential for using graph-based methods to obtain better insights into the process and content analysis for our dataset and to extend its applicability to MOOCs, however there are a number of contentious points to raise which will provide opportunities for discussion.

Firstly the number of students who are actively involved in discussion is a very small proportion of the *active* participants. This means that the subset may not be representative at all. One could argue that these students are already engaged or desperately need help. Previous literature [21, 22, 23] focused on the ability to predict performance and on the peer effect which can emerge from the analysis of the graphs/social networks.

Secondly, one could question the value of the communities in xMOOCs: especially when courses are designed with an instructivits approach leading to mastery, by definition this is an individualistic perspective focused on the testing of one's own skills/learning. Of course in cMOOCs -connectivists by design- the importance of the development of social support is essential. This seems to be supported by Brown et al [22]: they were not able to uncover a direct relation between stated goals and motivations with the participation in forums, and attributed this to pragmatic needs. However, as the authors suggested earlier, the instructors might play a fundamental role in shaping the communities based on the value attributed to forums in their plans/design and the level of engagement/interaction. Considering the split between cMOOCs and xMOOCs again, interesting work might come out of the experiment conducted by Rose' and colleagues in the DALMOOC in which automated agents were deployed to support students' conversations. In Coursera the deployment of 'community mentors' will be an interesting space to explore, given that the importance of design seems to be removed from instructors in the 'on-demand' model.

Lastly, more research is needed in the time-based dimension of development of forums in MOOCs. Questions like how students bond and create stable relations, how they become authoritative and what motivates them to contribute over time are all open questions which the analysis of graphs over time might be able to address.

5. REFERENCES

[1] Dirk Jan van den Berg and Edward Crawley. Why MOOCS Are Transforming the Face of Higher Education. Retrieved April 12, 2015 from http://www.huffingtonpost.co.uk/dirk-jan-van-den-berg/why-moocs-are-transforming_b_4116819.html

[2] Chris Parr. The evolution of Moocs. Retrieved April 12, 2015 from http://www.timeshighereducation.co.uk/comment/opinion/the-evolution-of-moocs/2015614.article

[3] C. Osvaldo Rodriguez. 2012. MOOCs and the AI-Stanford Like Courses: Two Successful and Distinct Course Formats for Massive Open Online Courses. *European Journal of Open, Distance and E-Learning* (January 2012).

[4] George Siemens. 2005. Connectivism: A learning theory for the digital age. *International journal of instructional technology and distance learning* 2, 1 (2005), 3–10.

[5] Stephen Downes. 2008. Places to go: Connectivism & connective knowledge, Innovate.

[6] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, and Sebastián Ventura. 2013. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education* 68 (October 2013), 458–472. DOI: http://dx.doi.org/10.1016/j.compedu.2013.06.009

[7] Margaret Mazzolini and Sarah Maddison. 2007. When to jump in: The role of the instructor in online discussion forums. *Computers & Education* 49, 2 (September 2007), 193–213. DOI:http://dx.doi.org/10.1016/j.compedu.2005.06.011

[8] M.j.w. Thomas. 2002. Learning within incoherent structures: the space of online discussion forums. *Journal of Computer Assisted Learning* 18, 3 (September 2002), 351–366. DOI:http://dx.doi.org/10.1046/j.0266-4909.2002.03800.x

[9] Daniel F.O. Onah, Jane Sinclair, and Russell Boyatt. 2014. Exploring the use of MOOC discussion forums. In *Proceedings of London International Conference on Education*. London: LICE, 1–4.

[10] Alstete, J.W. and Beutell, N.J. Performance indicators in online distance learning courses: a study of management education. *Quality Assurance in Education 12*, 1 (2004), 6–14.

[11] Cheng, C.K., Paré, D.E., Collimore, L.-M., and Joordens, S. Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers & Education 56*, 1 (2011), 253–261.

[12] Palmer, S., Holt, D., and Bray, S. Does the discussion help? The impact of a formally assessed online discussion on final student results. *British Journal of Educational Technology 39*, 5 (2008), 847–858.

[13] Patel, J. and Aghayere, A. Students' Perspective on the Impact of a Web-based Discussion Forum on Student Learning. *Frontiers in Education Conference, 36th Annual*, (2006), 26–31.

[14] Jacqueline Aundree Baxter and Jo Haycock. 2014. Roles and student identities in online large course forums: Implications for practice. *The International Review of Research in Open and Distributed Learning* 15, 1 (January 2014).

[15] Vanessa Paz Dennen. 2008. Pedagogical lurking: Student engagement in non-posting discussion behavior. *Computers in Human Behavior* 24, 4 (July 2008), 1624–1633. DOI:http://dx.doi.org/10.1016/j.chb.2007.06.003

[16] René F. Kizilcec, Chris Piech, and Emily Schneider. 2013. Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. LAK '13. New York, NY, USA: ACM, 170–179. DOI:http://dx.doi.org/10.1145/2460296.2460330

[17] Dennen, V.P. Pedagogical lurking: Student engagement in non-posting discussion behavior. *Computers in Human Behavior 24*, 4 (2008), 1624–1633.

[18] Kobayashi, M. and Yung, R. Tracking Topic Evolution in On-Line Postings: 2006 IBM Innovation Jam Data. In T. Washio, E. Suzuki, K.M. Ting and A. Inokuchi, eds., *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2008, 616–625.

[19] Bhattacharya, P., Iliofotou, M., Neamtiu, I., and Faloutsos, M. Graph-based Analysis and Prediction for Software Evolution. *Proceedings of the 34th International Conference on Software Engineering*, IEEE Press (2012), 419–429.

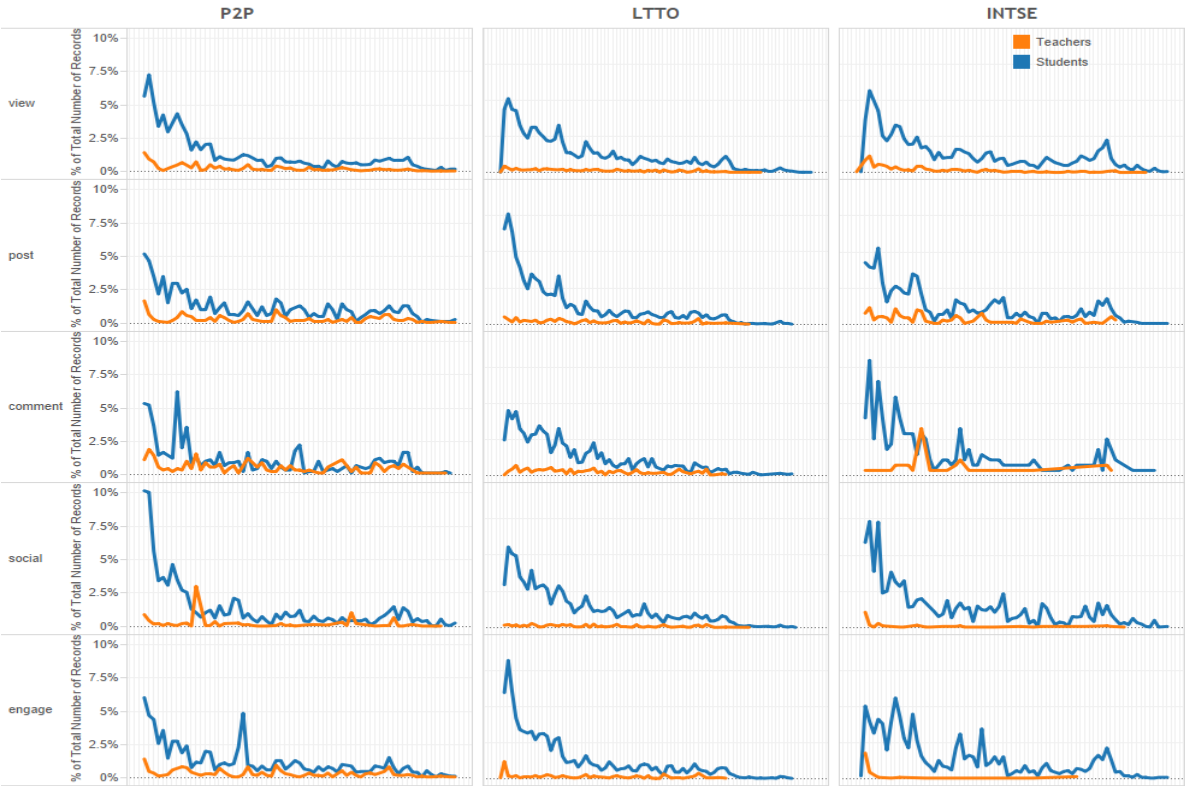
[20] Kruck, S.E., Teer, F., and Jr, W.A.C. GSLAP: a graph-based web analysis tool. *Industrial Management & Data Systems 108*, 2 (2008), 162–172.

[21] D. Yang, M. Wen, A. Kumar, E. P. Xing, and C. P. Rose, “Towards an integration of text and graph clustering methods as a lens for studying social interaction in MOOCs,” *The International Review of Research in Open and Distributed Learning*, vol. 15, no. 5, Oct. 2014.

[22] R. Brown, C Lynch, Y. Wang, M. Eagle, J. Albert, T. Barnes, R. Baker, Y. Bergner, D. McNamara. Communities of performance and communities of preference. GEDM 2015, in press.

[23] M. Fire, G. Katz, Y. Elovici, B. Shapira, and L. Rokach, “Predicting Student Exam’s Scores by Analyzing Social Network Data,” in *Active Media Technology*, R. Huang, A. A. Ghorbani, G. Pasi, T. Yamaguchi, N. Y. Yen, and B. Jin, Eds. Springer Berlin Heidelberg, 2012, pp. 584–595.

Figure 3. Time sequence of activity in the forums in the three courses by students and instructors grouped by activity type


P2P
LTTO
INTSE
Teachers
Students
view
post
comment
social
engage
% of Total Number of Records