On generalizability of MOOC models

Łukasz Kidziński, Kshitij Sharma, Mina Shirvani Boroujeni, Pierre Dillenbourg
Computer Human Interaction in Learning and Instruction
École polytechnique fédérale de Lausanne
{lukasz.kidzinski,kshitij.sharma,mina.shirvaniboroujeni,pierre.dillenbourg}@epfl.ch

ABSTRACT

The big data imposes the key problem of generalizability of the results. In the present contribution, we discuss statistical tools which can help to select variables adequate for target level of abstraction. We show that a model considered as over-fitted in one context can be accurate in another. We illustrate this notion with an example analysis experiment on the data from 13 university Massive Online Open Courses (MOOCs). We discuss statistical tools which can be helpful in the analysis of generalizability of MOOC models.

Keywords

Massive open online courses, MOOCs, bias-variance trade-off, generalizability

1. INTRODUCTION

The rapid growth of Massive Online Open Courses (MOOCs) has shown significant impact not only on the education but also on educational research. Over 100 world class universities partner with MOOC platforms to provide free education. Many of these universities, use data analytics to provide indicators to the policy makers, and valuable insights to the teachers and producers.

Researchers from emerging educational fields, such as learning analytics and educational data mining, attempt to make sense from the huge datasets from the MOOC providers (for example Coursera, Edx). These large datasets provide an opportunity to detect the slightest differences in the behaviour which are correlated to the students' performance.

However, the big data involves the risk of misinterpreting the results. The misinterpretations could surface mainly because of two reasons. First, the effect sizes are few orders of magnitude smaller then we used to expect in classical educational psychology studies; and the results are still significant due to the large sample. Second, "black-box" approaches like Support Vector Machines or Neural Networks give us great predictive power of models but do not explain the underlying processes.

Both of these reasons can lead to "overfitting" a model for a given context. Still, the same model can be be accurate in another context as illustrated in Figure 2. Choosing too specific descriptors could lead to the models which precisely describe one student but fail to generalize to new concepts. Too vague descriptors tend to generalize better but inform less about the specifics of the underlying processes. In statistical terminology this is often referred to as the "bias-variance trade-off".

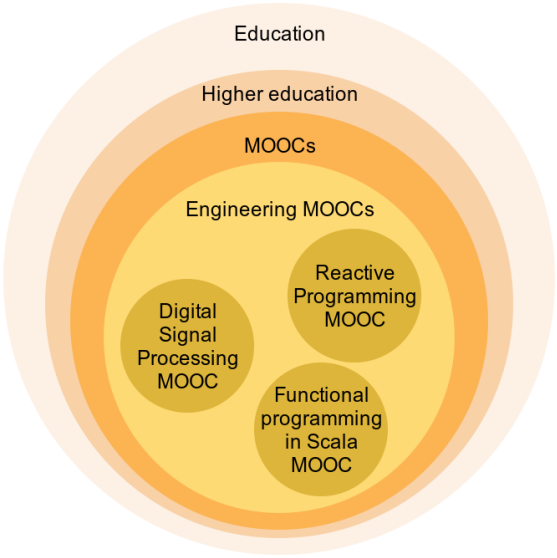

Figure 1: Example of layers to which we can draw conclusions from instances of MOOCs if the generalizability issues are addressed correctly.

The bias-variance trade-off is the central problem in statistical learning. It corresponds to the fact that one cannot minimize both quantities, "bias" and "variance", at the same time. A model with large bias is a smooth model not meant to fit sample points very closely but still captures the general trend in the data. Conversely, a model with large variance (not smooth) varies a lot for similar input parameters in order to fit well to each point in the dataset, often causing the so-called "overfitting".

The objective of this paper is to highlight the potential problem of closed-world context of MOOC research. We discuss techniques for leveraging existing models to more general context. We argue that designing context independent features is crucial for building generalizable models and we illustrate how variable selection process can be enhanced with statistical techniques. We illustrate a statistical technique which can be helpful in the choice of the important variables.

We address the following three research questions:

1. How to measure the extent to which the MOOC research as generalizable?
2. How to leverage predictive models in a MOOC to a broader context?
3. How to improve model's accuracy by restraining the scope of the variables used for prediction purposes?

2. RELATED WORK

2.1 Student Categorization

The common approach for finding generalizable patterns is to classify students into groups. To the best of our knowledge, there exist only a few categorisation schemes, mostly based on what emerges as a pattern of behaviour from MOOC students. These categories are based on the students' motivation [20], engagement patterns [10, 14, 16, 7] or demographics [5, 4].

There are many categorisation schemes depending on the engagement patterns. [10] categorised the students in Completing, Auditing, Disengaging and Sampling students based on their activities which range from watching majority of lectures and submitting all the assignments (Completing) to watching only one or two lectures and no assignment submissions (Sampling). In a connectivist MOOC setting, [14] categorised students into Active (students who adapt well to the connectivest pedagogy), Passive (frustrated ones) and Lurkers (who actively follow the course but do not interact with anyone). Phil Hill first categorised MOOC students into Lurkers (ones who only enrol or sample the course), Active (fully engaged with the course material, quizzes and forums), Passive (only consume the content, did not participate in forums) and Drop-ins (consumed only a part of the course as an Active student) [8]. Later he revised his categories and divided the Lurkers into No-shows and Observers [7].

These schemes are either defined by hard-coded thresholds or by unsupervised learning techniques. For that reason, they remain robust in terms of generalizability within the MOOC's context, but they are hard to generalize outside of it. In this study, we will rather discuss regression than classification/clustering, keeping in mind that similar observations can be done in both contexts.

2.2 Performance and engagement prediction

Student's performance is one of the key metrics analyzed in MOOCs. Many studies chose performance as an indicator for showing the value of the categorization methods. Massive datasets allow us to discover relation between performance and even the smallest factors like the number of pauses during watching a MOOC video or ratio of a video replayed [12]. Performance is also a crucial indicator for policy makers and MOOC practitioners. Reports focus on performance of MOOCs as a function of performance of students [13].

Previous studies on performance often concern a small set of MOOCs [1, 17, 9]. These studies provide insights about a large cohort of students and generalize to another cohorts, however the studies encounter lack of generalizability due to a small sample in the sense of course variability. In other studies, authors used time spent on lecture video, lecture quiz, homework, forum, quiz, assignments to predict students' learning gain [3, 11, 21, 3]. Lauria et al. [11] used the amount of content viewed, forum read, number of posts, assignments and quizzes submitted, to predict the performance and the engagement of the students. Wolff et al. [21] used the temporal clickstream data to predict students' performance.

These studies risk having high bias towards the courses in context and thus might lack the generalizability to be extended to courses with different content and/or courses from different domain. However in the aforementioned works, it is difficult to confirm our claim due to small number of MOOCs being analyzed. An example with generalizable set is shown by [2], where authors used the weekly time series data with 2-, 3-, 4-, and 5- grams to predict the final grades of the students. They experienced issues with the predictive models being generalisable - the model accuracy decreases as the authors used the same course session, to a different session from the same course, to a different course.

3. PROBLEM STATEMENT

In the MOOC context, models with large variance might correspond to the cases where one includes specific information about users, which are characterising only the sample at hand. For example, a model which includes exact timing of actions into account, could fit precisely to the data, since it identifies the user by the time of his actions, but it provides no generizability to new samples. Conversely, models with high bias correspond to situations when one considers general indicators like only the number of forum activities in a MOOC - thus, the model will fit worse to specific users but is more likely to generalize.

In practice, it is impossible to make both variables small, i.e. to retain both good fitness and smoothness. We need to choose the complexity of the model such that the sum of these two quantities is minimised. One could show that for any statistical learning method, the error can be decomposed to variance and bias terms. For a given target value y, predictors x and the estimator $\hat{f}$, the error of the model can be depicted as:

$$\mathrm{E}\Big[\big(y - \hat{f}(x)\big)^2\Big] = \mathrm{Bias}\big[\hat{f}(x)\big]^2 + \mathrm{Var}\big[\hat{f}(x)\big] + \sigma^2, \qquad (1)$$

where σ is the standard deviation of the residuals,

$$\mathrm{Bias}\big[\hat{f}(x)\big] = \mathrm{E}\big[\hat{f}(x)\big] - f(x)$$

and

$$\mathrm{Var}\big[\hat{f}(x)\big] = \mathrm{E}\Big[\big(\hat{f}(x) - \mathrm{E}[\hat{f}(x)]\big)^2\Big].$$

In other terms, bias is the squared distance between the real output $f(x)$ and the average prediction for given x, i.e. the $\mathrm{E}\big[\hat{f}(x)\big]$. The bias gets large whenever the average of predictions x differs highly from $\mathrm{E}\big[\hat{f}(x)\big]$. Conversely, the variance, expressing how do prediction vary from average around x, gets large whenever the variability is high.

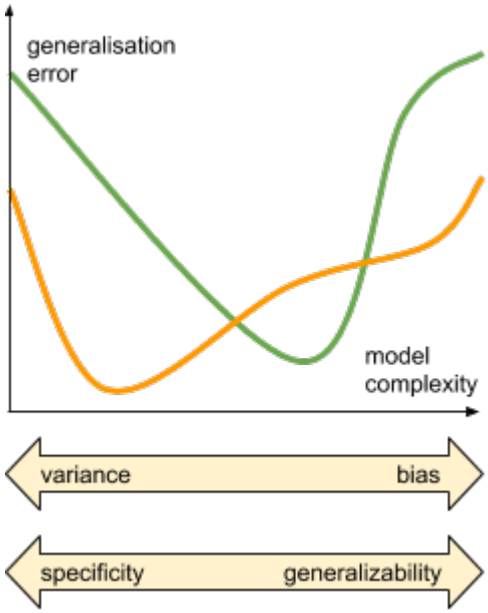

Figure 2: Influence of bias-variance trade-off on the generalzation error - illustrative conceptual drawing.

The ideal model would have both quantities $\mathrm{Bias}\big[\hat{f}(x)\big]^2$ and $\mathrm{Var}\big[\hat{f}(x)\big]$ equal to zero, but, as we mentioned before, it is not practically possible. However, we can control this error, as both quantities depend on the complexity of the model. For example, a linear model with large number of parameters has high variance and thus the error term increases. On the contrary, if one chooses low complexity (small number of variables), the model might have high error due to the high bias. The "best" model is somewhere in the middle, as illustrated by the green curve in Figure 2.

What is often missed in the analysis of the bias-variance plot, is that the error depends also on the context in which we generalize. Particularly in the MOOC context, in Figure 2 the green curve corresponds to generalization to another instance of the same MOOC, whereas the error follows a different pattern (orange curve) if we change the context to another MOOC.

4. MATERIALS AND METHODS

As we focus on the concept of generalizability of models and robustness of variables, we investigate our approach on several different MOOCs. We used data from 13 MOOCs, from EPFL, from both coursera and edX platforms. The dataset contains 1 MOOC which had 3 sessions in 3 consecutive semesters and 2 MOOCs which had 2 sessions in 2 consecutive semesters, as indicated in Table 4.

This setup allows us to investigate several aspects of generalizability. We investigated the fit of a model in correspondance to: 1) the course itself; 2) another instance of the same course; 3) another engineering course.

4.1 Setup

In order to attain a generalizable model, the setup must be consistent between the training data and the test data. Thus, we use the variables which could be defined for all the courses. Additionally, all the scores are normalized to the same range (0 - 100). Since courses have different lengths, we focus only on student activities in the first week. Finally, since 95% of the students did not submit any assignments and significantly bias linear models, we analysed only those students who got at least 1 point as their final grades. Note, that the context we are defining serves mainly as an illustration, thus we choose a relatively simple setup for transparency.

As the measure of performance of a model we take the Normalized Mean Squared Error (`NMSE`), defined as

$$NMSE = \mathrm{Var}(y - \hat{f}(x))/\mathrm{Var}(y),$$

where y is the dependent variable to predict, $\hat{f}$ is the estimator of the relation between y and independent variable x and Var corresponds to the sample variance.

4.2 Example method

In the linear regression, the main source of complexity is due to the number of variables in the model. Classical statistics provide us with robust tools for variable selection, such as ANOVA, Akaike Information Criteria. These techniques are useful for their inferential value, however, they do not guarantee the best generalizability in terms of prediction.

One of the techniques, where the complexity is controlled using a parameter that also affects the performance of the model, is regularized linear regression. In classical statistics, called ridge regression, the standard linear model is extended with an additional, regularizing term. This regularising term controls the parameters of the model with respect to the performance measure based on the prediction, by decreasing the importance of variables which do not account for the prediction.

In particular, given the independent variables $X_1, X_2, ..., X_d$ and the dependent variable Y we build a model minimizing

$$\mathrm{E}\big\|Y - \beta_1 X_1 - \beta_2 X_2 - ... - \beta_n X_d\big\|^2 + \lambda \sum_{i=1}^{k} \beta_i^p, \quad (2)$$

where d is the number of variables, $\beta_1, \beta_2, ..., \beta_d$ are the parameters of the model and $p = 2$.

If λ is large, we put more weight to the sum of βs. Therefore, the number of parameters will be reduced and the model will have a low bias. On the other hand, if λ is small, the model corresponds to linear regression and the variance is high since we use all the variables.

We chose this model for our analysis since it allowed us to control both the bias and the variance with a single parameter λ. Moreover, changing the value of p from 2 to 1 is (2), gives better results in many setups. Hence, we choose to use $p = 1$. The model is known in the machine learning literature as LASSO [19]. The complete algorithm, for those interested, can be found in [19]. Here we are refraining ourselves to the basic description as this is not the main focus of the paper.

4.3 Variables

For illustrating the problem, we chose the students' final grade in the course as the dependent variable. Following are the features that we extracted from the data for modeling this value.

1. **Counts**: We counted different online activities exhibited by the students. 1) *Lectures*: lecture view, lecture re-view, lecture download and lecture re-download; 2) *Quiz*: quiz submission, quiz re-submission, here we differentiated between the quizzes as an exercise, in-video quizzes and the surveys; 3) *Assignments*: assignment submission and assignment re-submission; 4) *Forums*: thread launches, upvotes, downvotes, subscriptions, views, comments and posts.

2. **Delays**: We computed the time difference between the different events in the MOOC structure and students' activities. 1) *First View Delay*: the time difference between the first view or first download of the lecture and the time when the lecture was online; 2) *Overall View Delay*: the average first view delay for all the lecture views and downloads; 3) *Between Lecture Delay*: the time difference between the views or downloads of two different lectures; 4) *Within Lecture Delay*: the average time difference between two views and/or downloads of the same lecture; 5) *First Quiz Attempt Delay*: the time difference between the first submission for a quiz and the time when the quiz was online; 6) *Within Quiz Time*: the time difference between two attempts for the same quiz; 7) *Overall Quiz Attempt Delay*: the average first quiz attempt delays for all the quizzes.

3. **Progress**: We computed the score difference between the two consecutive attempts to the same quiz or the same assignment.

4. **2-way Transitions**: We labeled the different activities as L, A, Q and F for lectures, assignments, quizzes, and forums respectively. Further, we constructed a time-series of the actions and counted how many times the action pairs (for example, AA, AL, AF, LQ, FL, 16 pairs) occur in the time series for each student.

5. **3-way Transitions**: using the same time series, as to compute the 2-way transitions, we counted how many times the action triples (for example, AAA, FAL, QAF, LLQ, FLL, 64 triples) occur in the time series for each student.

5. RESULTS

Using the variables, defined in the Section 4.3, we illustrate the setup for modelling the data. As we mentioned in the Section 4.1, we considered only the activities from the first week of the courses and from those students who scored at least 1. We would also like to emphasize here that the main aim of this contribution is not to present a model that has the least error, but to show how we can build generalizable models taking into account the bias-variance trade off.

In the proposed setup, we demonstrate how generalizable a model is to: 1) the students from the same course (separate test set of 20% of observations), 2) the students from another instance of the same course, 3) to a different course.

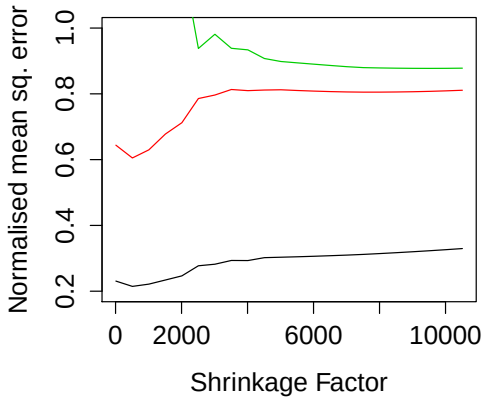

Figure 3: Prediction error (NMSE) for the test samples, for the different values of the shrinkage factor λ in (2), using all the variables.

First, we analyze the model fit to the first session of the *Numerical Analysis* course and test it on: itself, another session of *Numerical Analysis* and *House Water Treatment Systems* a course from a different domain. We illustrate the results in Figure 3. We observed that the model which had highest predictive power on the test set in the session 1 (black curve) has the worse predictive power for another instance of the same course (red curve), but still performs well. The optimal shrinkage factor (λ in equation (2)) turns out to be close to 0 in both cases. This shows that almost all the variables we introduced are included in the model. We could conclude that the model generalizes to another instances of the same course.

However, as we hypothesized, the full model did not fit at all to a course from a different domain. Only with a large value of the shrinkage factor, which removed 97 variables

from the model, we obtain a model with some informative value for a course from another domain. Furthermore, the errors become similar for all the courses, illustrating that the model has lower variance. It generalizes better to another course but it lost its fit to the Numerical Analysis course.

We conducted the identical analysis (see Table 1) on all the courses mentioned in the dataset. In all the cases, generalizability to another course required significant decrease in the complexity, using the shrinkage factor. Removing certain variables from the model turns out to be crucial for the performance. Since we started with 134 variables, to further analyze the ability to generalize, we restricted ourselves to a simpler case with the first three (counts, delays and progress) groups of variables introduced in Section 4.3.

The same patterns were observed in this simpler case. The optimal model for prediction in the same instance and in another instance of the course have the lowest error if the complexity (variance) is high. However, the model with such a high complexity exhibits poor performance in another course, from the same domain, i.e. the linear optimization.

As hypothesized, variables which were removed by LASSO, are course-structure dependent. The most generalizable models contain the variables related to the lecture, forum and quiz activities. These variables provide the required generalizability to the model and hence we observe that as we increase the shrinkage factor, the predictive power of the model increasingly became similar for the different courses.

6. CONCLUSION

We demonstrated through examples that in the terms of bias-variance trade-off, achieving both the specificity and generalizability is not possible while modelling student behaviour. Through the statistical methods available, one can only achieve one of the two goals, or find an optimum solution that is specific to one course and only reveals the surface learning behaviour of the students from a course from another domain, or vice verse.

Similar validation framework, analysing fitness in the same course, another session of the same course and another course was previously introduced [2] in literature. Results from this work are equivalent to ours with some predefined and fixed complexity parameter. In our work we show that practitioners can modulate the complexity and generalizabiity by selecting a subset of variables.

Previous works, have small sample size in terms of number of MOOCs. It is therefore difficult to assess their generalizability. For example, Social Network Analysis (as shown by [18, 15, 6]) is based on the motivation of the student - if the students are sharing the exact answers (or revealing them in some other ways) forum view can play a big role in achievement. Clickstreams (as shown by [21]) in a video are highly dependent on the content. Finally, from the methodological perspective generalizability is also a design choice - for example - if we choose a smaller number of clusters in unsupervised learning, we may obtain more robust results (smaller variance higher bias).

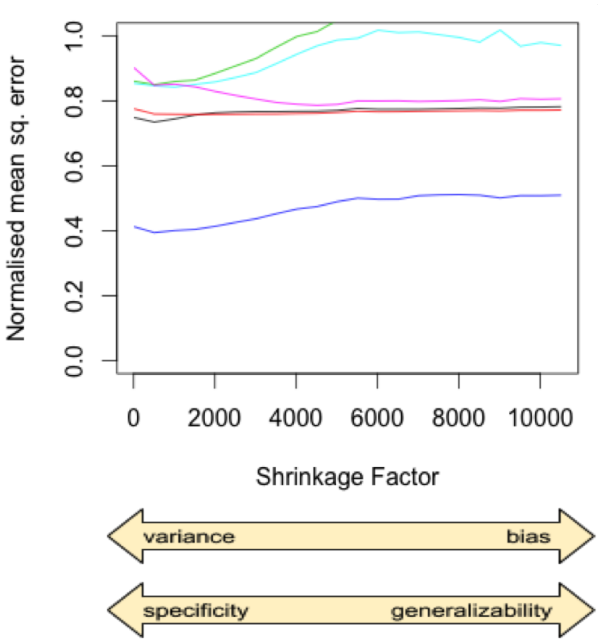

Figure 4: Illustration of bias-variance trade-off from engineering courses. Prediction error (NMSE) for the test samples, for the different values of the shrinkage factor λ in (2)

7. DISCUSSION

Our goal was to illustrate the generalizability issue which we encounter in any machine learning or learning analytics setups. We did not compare multiple algorithms, but we used a simple one to support our claims. It is worth mentioning that the same phenomenon is encountered in any other machine learning method.

Moreover, the same analysis can be performed with any regularized regression algorithms, i.e., consisting a parameter to control the complexity of the model, like SVM, logistic regression, neural networks, etc. In each of these methods regularization selects the optimal sets of parameters.

Finally, the choice of the feature set should be based on the desired outcome of modelling student behaviour in a MOOC. If the goal is to attain high predictability in a small variety of courses, one could choose to include course-structure related variables. On the other hand, if the modelling requirement is to have a decent genralizability over a wide variety of courses, one has to compromise the predictability over a set of courses and select only the course-structure-independent variables.

8. REFERENCES

[1] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton. Studying learning in the worldwide classroom: Research into edxâĂŹs first mooc. *Research & Practice in Assessment*, 8(1):13–25, 2013.

[2] C. Brooks, C. Thompson, and S. Teasley. A time series

Table 1: Results from the identical analysis done on all the other courses as shown in Figures 3. The courses with N/A in the second column had only one session. The errors reported are NMSE. The values in the perenthesis are the optimal shrinkage factors in given context.

Course Name	Testing on the same course	Testing on other session	Testing on different course
Digital signal processing	0.76 (10)	0.99 (10)	0.35 (2010)
Geomatics	0.67 (10)	N/A	0.35 (2010)
House water treatment systems	0.58 (10)	N/A	0.48 (510)
Linear optimisation	0.67 (10)	N/A	0.36 (3010)
Mechanics	0.68 (10)	N/A	0.59 (1010)
Sanitation	0.80 (510)	N/A	0.73 (1010)
Structures	0.95 (10)	0.93 (2010)	0.84 (4510)
Micro-controllers	0.35 (2010)	0.35 (2010)	0.35 (2010)

interaction analysis method for building predictive models of learners using log data. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 126–135. ACM, 2015.

[3] J. Champaign, K. F. Colvin, A. Liu, C. Fredericks, D. Seaton, and D. E. Pritchard. Correlating skill and improvement in 2 moocs with a student's time on tasks. In *Proceedings of the first ACM conference on Learning@Scale conference*, pages 11–20. ACM, 2014.

[4] G. Christensen, A. Steinmetz, B. Alcorn, A. Bennett, D. Woods, and E. J. Emanuel. The mooc phenomenon: who takes massive open online courses and why? *Available at SSRN 2350964*, 2013.

[5] J. DeBoer, G. S. Stump, D. Seaton, and L. Breslow. Diversity in mooc students? backgrounds and behaviors in relationship to performance in 6.002 x. In *Proceedings of the Sixth Learning International Networks Consortium Conference*, 2013.

[6] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth. Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 146–150. ACM, 2015.

[7] P. Hill. Emerging student patterns in moocs: A graphical view, 2013.

[8] P. Hill. The four student archetypes emerging in moocs. *E-Literate. March*, 10:2013, 2013.

[9] S. Jiang, A. Williams, K. Schenke, M. Warschauer, and D. O'dowd. Predicting mooc performance with week 1 behavior. In *Educational Data Mining 2014*, 2014.

[10] R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 170–179. ACM, 2013.

[11] E. J. Lauría, J. D. Baron, M. Devireddy, V. Sundararaju, and S. M. Jayaprakash. Mining academic data to improve college student retention: An open source perspective. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 139–142. ACM, 2012.

[12] N. Li, Ł. Kidziński, P. Jermann, and P. Dillenbourg. Mooc video interaction patterns: What do they tell us? In *Design for Teaching and Learning in a Networked World*, pages 197–210. Springer International Publishing, 2015.

[13] A. McAuley, B. Stewart, G. Siemens, and D. Cormier. The mooc model for digital practice. 2010.

[14] C. Milligan, A. Littlejohn, and A. Margaryan. Patterns of engagement in connectivist moocs. *MERLOT Journal of Online Learning and Teaching*, 9(2), 2013.

[15] W. C. Paredes and K. S. K. Chung. Modelling learning & performance: a social networks perspective. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 34–42. ACM, 2012.

[16] T. Petty and A. Farinde. Investigating student engagement in an online mathematics course through windows into teaching and learning. *Journal of Online Learning and Teaching*, 9(2):261–270, 2013.

[17] S. Rayyan, D. T. Seaton, J. Belcher, D. E. Pritchard, and I. Chuang. Participation and performance in 8.02 x electricity and magnetism: The first physics mooc from mitx. *arXiv preprint arXiv:1310.3173*, 2013.

[18] D. Rosen, V. Miagkikh, and D. Suthers. Social and semantic network analysis of chat logs. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, pages 134–139. ACM, 2011.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[20] J. Wilkowski, A. Deutsch, and D. M. Russell. Student skill and goal achievement in the mapping with google mooc. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 3–10. ACM, 2014.

[21] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek. Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 145–149. ACM, 2013.