Predicting Success in Massive Open Online Courses (MOOCs) Using Cohesion Network Analysis

Predicting Success in Massive Open Online Courses (MOOCs) Using Cohesion Network Analysis

Scott A. Crossley, Georgia State University, scrossley@gsu.edu
Mihai Dascalu, University Politehnica of Bucharest, mihai.dascalu@cs.pub.ro
Danielle S. McNamara, Arizona State University, dsmcnamara1@gmail.com
Ryan Baker, University of Pennsylvania, ryanshaunbaker@gmail.com
Stefan Trausan-Matu, University Politehnica of Bucharest, trausan@gmail.com

Abstract: This study uses Cohesion Network Analysis (CNA) indices to identify student patterns related to course completion in a massive open online course (MOOC). This analysis examines a subsample of 320 students who completed at least one graded assignment and produced at least 50 words in discussion forums in a MOOC on educational data mining. The findings indicate that CNA indices predict with substantial accuracy (76%) whether students complete the MOOC, helping us to better understand student retention in this MOOC and to develop more actionable automated signals of student success.

Introduction

Massive Open Online Courses (MOOCs) open a number of educational opportunities for traditional and non-traditional learning. However, the size of classes, which easily reaches into the thousands of students, requires educators and administrators to reconsider traditional approaches to instructor intervention and the manner in which student engagement, motivation, and success is assessed, especially since attrition rates in MOOCs is notoriously high (Ramesh, Godwasser, Huang, Daume, & Getoor, 2014). The uniqueness of MOOCs and the difficulties associated with them has opened new research areas, especially in predicting or explaining completion rates and general student success. Research has mainly focused on predicting success using click-stream data (i.e., student interactions within the MOOC software). Other recent approaches include the use of Natural Language Processing (NLP) tools to gauge students' affective states (Wen, Yang, & Rose, 2014b, 2014a), measure the sophistication and organization of students' discourse within a MOOC (Crossley et al., 2015; Crossley, Paquette, Dascalu, McNamara, & Baker, 2016, and a combination of click-stream and NLP data (Crossley et al., 2016). In this study, we examine new NLP approaches grounded in text cohesion and Social Network Analysis (SNA) to predict success in a MOOC related to educational data mining. Social interaction has long been recognized as an important component of learning (Vygotsky, 1978). However, while the relationship between language and social participation has been studied in MOOCs (Dowell et al., 2015), social interaction reflected through the language produced by MOOC students has not been investigated within large-scale, on-line learning environments.

The variables used in this study are based on Cohesion Network Analysis (CNA), which can be used to analyze discourse structures within collaborative conversations (Dascalu, Trausan-Matu, McNamara, & Dessus, 2015). CNA indices estimate cohesion between text segments based on similarity measures of semantic proximity. We hypothesize that students who produce forum posts that are on topic, are more related to other student posts, are more central to the conversation, and are more collaborative will be more likely to complete the MOOC than those that are not. We focus specifically on student completion rates because they are an important component of student success within the course, as well as after its completion (Wang, 2014). We assess links between completion and CNA indices because CNA indices afford a wide array of opportunities for better understanding student success in terms of collaboration. Using CNA indices to better understand student completion rates has the potential to inform pedagogical interventions that provide individualized feedback to MOOC participants and teachers regarding social interactions such as collaboration. Ultimately, our objective is to enhance participation and active involvement, to increase completion rates, as well as to increase our understanding of the factors associated with MOOC completion.

MOOC analysis

MOOCs have become an important component of education research for both instructors and researchers because they have the potential to increase educational accessibility to distance and lifelong learners (Koller, Ng, Do, & Chen, 2013). Researchers examine links between click-stream data in MOOCs and academic performance because MOOCs provide a tremendous amount of data via click-stream logs containing detailed records of the students' interactions with the course content. The measures typically computed from click-stream data that have been used in MOOC analyses include variables related to counts of the different possible types of actions, the timing of

actions, forum interactions and assignments attempts among others (Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014).

More recently, researchers have applied NLP tools to MOOC data (Chaturvedi, Goldwasser, & Daume, 2014; Wen, Yang, & Rose, 2014a, 2014b; Crossley et al., 2015; Crossley et al., 2016). Traditional usage of NLP tools in this context focus on a text's syntactic and lexical properties. The simplest approaches count the length of words or sentences, or use pre-existing databases to compare the word properties in a single text to that of a larger, more representative corpus of texts. More advanced NLP tools measure linguistic features related to the use of rhetorical structures, syntactic similarity, text cohesion, topic development, and sophisticated indices of word usage. Such tools have been used to examine text complexity (e.g., cohesion, lexical, and syntactic complexity) in forum posts and the degree to which these indicators are predictive of MOOC completion. For instance, Crossley et al. (2015) found that language related to forum post length, lexical sophistication, situational cohesion, cardinal numbers, trigram production, and writing quality were significantly predictive of whether a MOOC student completed the course (reporting an accuracy of 67%). In a follow up study, Crossley et al. (2016) combined click-stream data and NLP approaches to examine if students' on-line activity and the language they produced in the on-line discussion forum was predictive of MOOC completion. They found that click-stream variables (e.g., weekly lecture coverage and how early students submitted their assignments) were the strongest predictors of MOOC completion but that NLP variables (e.g., the number of entities in a forum post, the post length, the overall quality of the written post, the linguistic sophistication of the post, cohesion between posts, and word certainty) significantly increased the accuracy of the model. In total, click-stream and NLP indices predicted which students would complete the course with 76% accuracy. Combined, these findings indicate that students who are more involved in the course and demonstrate more advanced linguistic skills, are more likely to complete a MOOC.

Current study

The goal of the study is to test new indices that measure social integration and collaboration using Cohesion Network Analysis in order to examine student success in a MOOC. Thus, we perform a longitudinal analysis on the weekly timeline evolution of CNA indices to predict MOOC success and examine if students who engage in greater social interaction, that is on topic and central to the MOOC, are more successful (i.e., complete the course).

Method

The MOOC: Big data in education

In this paper, we evaluate course completion in the context of the Big Data in Education MOOC (BDEMOOC), using the data from the first iteration on this course, offered through the Coursera platform in 2013. This is the same MOOC investigated by Crossley et al. (Crossley et al., 2015; Crossley et al., 2016). The course was designed to support students in learning how to apply a range of educational data mining (EDM) methods to conduct education research questions and to develop models that could be used for automated intervention in online learning, or to inform teachers, curriculum designers, and other stakeholders. This course was targeted to the postgraduate level, and covered material comparable to a graduate course taught by the instructor. The MOOC ran from October 24, 2013 to December 26, 2013, and included several lecture videos in each of the 8 weeks, and one assignment per week.

In each of the weekly assignments, students conducted a set of analyses on a given data set and answered questions about the analyses. All assignments were automatically graded, and students had up to three attempts to complete each assignment successfully. Students received a certificate by obtaining an overall average grade of 70% or better on at least 6 of the 8 assignments. The course had an official enrollment of over 48,000 at the time of the course's official end. 13,314 students watched at least one video, 1,242 students watched all videos, 1,380 students completed at least one assignment, and 710 made a post in the discussion forums. Of those with posts, 426 students completed at least one class assignment while 638 students completed the online course and received a certificate. As such, some students earned a certificate for BDEMOOC without ever posting to the discussion forums.

Student completion rates

We selected completion rate as our variable of success because it is one of the most common metrics used in MOOC research (He, Bailey, Rubinstein, & Zhang, 2015), and correlates to future career participation (Wang, 2014). For this study, completion was based on a smaller sample of forum posters as described below. "Completion" was pre-defined as earning an overall grade average of 70% or above. The overall grade was calculated by averaging the 6 highest grades extracted out of the total of 8 assignments.

Discussion posts

Discussion posts are of interest within research on student participation in MOOCs because they are one of the core methods that students use to participate in social learning (Ramesh, Goldwasser, Huang, Daume, & Getoor, 2014). Discussion forums provide students with a platform to exchange ideas, discuss lectures, ask questions about the course, and seek technical help, all of which lead to the production of language in a natural setting. Such natural language can provide researchers with a window into individual student motivation, linguistics skills, writing strategies, and affective states. This information can in turn be used to develop models to improve student learning experiences (Ramesh, Goldwasser, Huang, Daume, & Getoor, 2014). In BDEMOOC, students and teaching staff participated actively in weekly forum discussions. Each week, new discussion threads were created for each week's specific content, including both videos and assignments under sub-forums, each with corresponding discussion threads. Forum participation did not count toward student's final grades. For this study, we focused on the forum participation in the weekly course discussions. For this study, we extracted all forum posts and corresponding comments from the MOOC environment for all 426 students who both made at least a forum post and completed an assignment. We removed all data from instructors and teaching assistants. We analyzed data from those students who produced at least 50 words in their aggregated posts (n = 319). Fifty words was used as a cut-off to ensure sufficient linguistic information. Of these 319 students, 132 did not successfully complete the course while the remaining 187 completed the course.

Cohesion network analysis

In Computer Supported Collaborative Learning (CSCL) environments, Cohesion Network Analysis analyzes discourse structure by combining NLP approaches with SNA (Dascalu, Trausan-Matu, McNamara, & Dessus, 2015). In CNA, cohesion is computationally represented as an average value of similarity measures (or an aggregated score) between semantic distances (Budanitsky & Hirst, 2006) using *WordNet* (Miller, 1995) *Latent Semantic Analysis* (Landauer & Dumais, 1997) and *Latent Dirichlet Allocation* (Blei, Ng, & Jordan, 2003). We used the Touchstone Applied Science Associates (TASA) corpus (approximately 13 million words; http://lsa.colorado.edu/spaces.html) together with a collection of articles extracted from the Learning Analytics & Knowledge dataset (652 Learning Analytics and Knowledge and Educational Data Mining conference papers and 45 journal papers; https://www.w3.org/TR/REC-rdf-syntax/) to train dedicated LSA and LDA semantic models. The resulting corpora covered both the curricula of the MOOC course and provided also a general knowledge background. Before training, the texts were preprocessed such that stop-words were removed and all words were lemmatized.

A *cohesion graph* (Dascalu, Trausan-Matu, McNamara, & Dessus, 2015) was generated using cohesion values in order to determine connections between discourse elements. This graph represents a generalization of the utterance graph (Trausan-Matu, Stahl, & Sarmiento, 2007) and can be used as a proxy for the semantic content of discourse. The cohesion graph is a multi-layered structure containing different nodes (Dascalu, 2014) and the links between them. A central node, representing the conversation's thread, is divided into contributions, which are further divided into sentences and words. Links are then built between nodes in order to determine a cohesion score that denotes the relevance of a contribution within the conversation, or the impact of a word within a sentence or contribution. Other links are generated between adjacent contributions, which are used to determine changes in the topics or of the conversation's thread. These changes are reflected by cohesion gaps between units of texts. Explicit links, created using an interface functionality such as the "reply-to" option, are contained within the cohesion graph as well. In addition, cohesive links determined using semantic similarity techniques are added between related contributions within a timeframe of maximum 20 successive contributions, which can be considered the maximum span for these type of cohesive links (Rebedea, 2012).

Cohesion scoring mechanism

The cohesion graph determines the active engagement in terms of participation in the MOOC. This is computed quantitatively based on relations established between nodes from the cohesion graph. The contributions are analyzed to determine their importance in relation to the discussion's thread, coverage of topics, and their relatedness to other contributions. The relevance score of a node in the cohesion graph is based on the relevance of underlying words and on its relation to other components. For example, a contribution's relevance score is computed as the sum of its constituent words based on *statistical presence* and the *semantic relatedness* (Dascalu, Trausan-Matu, Dessus, & McNamara, 2015). Statistical presence represents the word frequency within the text, while semantic relatedness refers to semantic similarity between the word and the entire conversation thread that contains it. Keywords for the whole conversation are determined by considering the aggregated score of the two factors.

Afterwards, the cohesion scoring mechanism assigns contribution scores by multiplying each word's previously determined score with its normalized term frequency (Dascalu, 2014), estimating an on-topic relevance of the utterance. Links with other contributions, stored within the CNA are further used to improve contribution scores. Each contribution's local relevance is then calculated with regards to related contributions. Thus, each textual element's score can be viewed as its importance within the discourse, covering both the topic and the semantic relatedness with other elements.

Collaboration assessment

Social knowledge-building (KB) processes (Bereiter, 2002) are derived through collaboration (i.e., scores calculated on the inter-animation of interactions between different participants). Social KB refers to the external dialog between at least two participants supporting collaboration, while inner dialogue is reflected by the continuation of ideas or explicit, referred contributions belonging to the same speaker.

Each contribution has a previously defined importance score and an effect score in term of both personal and social KB. The personal score is initially assigned as each utterance's importance score, while the social score is initially assigned a zero. By analyzing the links from the cohesion graph, these scores are augmented. If a link is established between contributions belonging to the same speaker, the knowledge (personal and social) from the referred contribution is transferred to the personal dimension of the current contribution through the cohesion score. If the link is established between different users, only the social dimension of the currently analyzed contribution is increased by the cohesion measure. This enables a measurement of collaboration perceived as a sum of social KB effects that consist of each contribution's score, multiplied by the cohesion value to related contribution (Dascalu, Trausan-Matu, McNamara, & Dessus, 2015).

Interaction modeling and integration of multiple CNA graphs

The *sociogram* reflects information exchanges between users and represents the central structure for modeling interaction and information transfer between participants (Dascalu, 2014). The nodes represent users, while the edges represent interchanged contributions. This graph considers not only the number of exchanged contributions, but weights each utterance as a sum of social KB effects to other MOOC participants. Specific SNA metrics are further computed starting from the sociogram in order to measure centrality or involvement (Dascalu, 2014). Some examples include the number of links to (out-degree) and from (in-degree) other participants for a specific user. Betweenness centrality (Bastian, Heymann, & Jacomy, 2009) is computed to determine central nodes and highlights the information exchange between participants who, if eliminated, would highly reduce communication. The participant's connection to other nodes, called closeness centrality (Sabidussi, 1966), is computed as the inverse distance to all other nodes. A higher value represents a participant's stronger connection to all other discussion thread participants. The maximal distance between a node and all other nodes, called eccentricity (Freeman, 1977), shows the closeness of a user to other participants. These models were extended to facilitate the evaluation of not only a single discussion, but of an entire MOOC by considering the aggregation of multiple discussion threads. Such a global analysis was used to build a social network consisting of all involved participants and their contributions, thus enabling the evaluation of participation at a macroscopic level, not only for specific discussions, but for the entire MOOC. The sociogram between all participants was generated considering the sum of contribution scores per discussion thread within the forum. The overview of different user goals, distributions, and interactions provides a broader perspective of a participants' evolution within the MOOC.

Longitudinal analysis

We performed a longitudinal analysis by measuring the distribution of each participant's involvement throughout the duration of the MOOC which enabled us to quantify the evolution of learners' participation, collaboration and interaction patterns across time. In order to generate each participant's time distribution, specific sociograms were built for incremental weekly timeframes and CNA-derived quantitative indices were evaluated, covering the following elements, as discussed above: a) cumulative utterance scores per participant (i.e., the sum of individual contribution importance scores that were uttered by a certain participant), b) social KB effect as the cumulative effect of a participant's contribution in relation to other speakers, and c) specific SNA metrics (i.e., in-degree, out-degree, betweenness, closeness and eccentricity centrality measures) computed on the CNA interaction graph.

As expected due to attrition, a large discrepancy was observed in terms of the density of the interaction graphs found between the first and last week of the course, denoting a significant decrease in density. The values of each CNA index per timeframe were used to create individual time series reflecting each participant's evolution throughout the course. Afterwards, the longitudinal analysis indices presented in Table 1 were used to model the trends of the time series generated per participant and per CNA quantitative index. This approach creates an in-depth NLP-centered perspective of our longitudinal analysis built on top of CNA.

Statistical analysis

CNA indices that yielded non-normal distributions were removed. A multivariate analysis of variance (MANOVA) was conducted to examine which indices reported differences between students who completed or did not complete the MOOC. The MANOVA was followed by a stepwise discriminant function analysis (DFA) using CNA indices that were normally distributed and demonstrated significant differences between students who completed the course and those who did not. CNA indices were also checked for multicollinearity ($r > .90$). In the case of multicollinearity between indices, the index demonstrating the largest effect size in the MANOVA was retained in the analysis. The DFA was used to develop an algorithm to predict group membership through a discriminant function coefficient. A DFA model was first developed for the entire corpus of student forum posts. This model was then used to predict group membership (completers v. non-completers) for the student forum posts using leave-one-out-cross-validation (LOOCV) in order to ensure that the model was stable across the dataset.

Table 1: Longitudinal analysis indices applied on students' social media contributions across time

Name	Description
Avg. & St. Dev.	Average and standard deviation of the considered CNA quantitative index within all timeframes
Slope	The degree of the slope corresponding to the linear regression applied on the time series. The slope indicates whether students became more actively involved (slope > 0), had a uniform involvement (slope = 0), or lost their interest throughout the semester (slope < 0).
Entropy	Considering the probability of posting within each timeframe, Shannon's entropy formula (Shannon, 1948) grasps the discrepancies or inconsistencies in participation patterns. For example, if students are active in only one timeframe, their entropy is 0, whereas if they have a constant activity throughout the course, their entropy converges towards the maximum value of log(n), where n is the number of timeframes
Uniformity	Degree of uniformity is measured using Jensen Shannon dissimilarity (JSD) (Manning & Schütze, 1999) to a uniform distribution of 1/n. The JSD is a symmetric function based on the Kullback–Leibler divergence and is used to measure the similarity between two distributions, in our case the student's time series and an ideal, uniform participation in each week
Local extreme points	The number of local extreme points determined as the number of timeframes for which the inflection or the direction of the evolution of the CNA index changes. This reflects the monotony degree of the evolution or inconsistency in participation or collaboration - if multiple spikes are encountered, these will be identified as local minimum or maximum points; therefore, more local extreme points will be identified within the time series evolution
Average & standard deviation of recurrence	Recurrence is expressed as the distance between timeframes in which the learner had at least one contribution in the time series. This is useful for identifying and quantifying pauses as adjacent weeks without any activity. If each timeframe has at least one event, recurrence is 0, whereas if students take long pauses that inherently generate timeframes with 0 events, recurrence increases (e.g., if they post every 2 weeks, recurrence becomes 1, and so forth).

Results

A MANOVA was conducted using the CNA indices as the dependent variables, and whether the student completed or did not complete the MOOC as the independent variable. Of the 56 indices, 15 indices were not normally distributed and were removed. Of the remaining 41 indices, 27 indices did not demonstrate multicollinearity and were retained. Of these 27 indices, 26 of them demonstrated significant differences between students who completed the MOOC and students who did not complete the MOOC (see Table 2 for details). These indices demonstrated that MOOC completers produced posts that were on topic, were more related to other posts, demonstrated greater collaboration, and were more central to the conversation. These indices were used in the subsequent DFA.

A stepwise DFA using the 26 indices selected through the MANOVA retained three variables: Standard deviation of recurrence (Overall Score), Slope degree (Closeness), and Average (Closeness). The results demonstrate that the DFA using these three indices correctly allocated 243 of the 319 forum posts in the total set, $\chi^2(df=1) = 86.325$, $p < .001$, for an accuracy of 76.2%. For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 242 of the 319 students for an accuracy of 75.9%. See Table 3 for recall, precision,

and F1 scores for this analysis. The Cohen's Kappa measure of agreement between the predicted and actual class label was .518, demonstrating moderate agreement.

Table 2: Longitudinal analysis indices applied on students' social media contributions across time

Index	Did not complete: Mean (SD)	Completed: Mean (SD)	F	η^2
Standard deviation of recurrence (Overall score)	2.433 (0.839)	1.395 (0.994)	95.666**	.232
Local extremes (Overall score)	2.106 (1.134)	3.401 (1.550)	66.842**	.174
Slope degree (Closeness)	0.006 (0.011)	0.024 (0.022)	71.637**	.184
Slope degree (Eccentricity)	0.084 (0.115)	0.281 (0.252)	69.91**	.181
Local extremes (Out-degree)	1.864 (1.247)	3.198 (1.678)	60.045**	.159
Degree of uniformity (Overall score)	0.639 (0.099)	0.518 (0.169)	54.739**	.147
Entropy (Overall score)	0.277 (0.349)	0.634 (0.542)	44.412**	.123
Standard deviation of recurrence (In Degree)	2.113 (0.949)	1.338 (0.996)	48.713**	.133
Standard deviation of recurrence (Out Degree)	2.207 (1.117)	1.434 (1.062)	39.325**	.110
Average (Closeness)	0.063 (0.056)	0.118 (0.093)	36.965**	.104
Local extremes (In-degree)	2.265 (1.313)	3.166 (1.492)	31.108**	.089
Entropy (Closeness)	0.309 (0.432)	0.702 (0.654)	36.333**	.103
Average recurrence (Overall score)	2.628 (0.949)	1.856 (1.456)	28.548**	.083
Local extremes (Betweenness)	1.409 (1.266)	2.369 (1.709)	29.997**	.086
Degree of uniformity (Closeness)	0.606 (0.123)	0.494 (0.198)	33.153**	.095
Degree of uniformity (In-degree)	0.613 (0.110)	0.522 (0.165)	30.736**	.088
Entropy (Out-degree)	0.162 (0.290)	0.416 (0.469)	30.461**	.088
Entropy (In-degree)	0.319 (0.372)	0.598 (0.534)	26.909**	.078
Average recurrence (Out-degree)	3.181 (1.593)	2.232 (1.724)	24.941**	.073
Degree of uniformity (Out-degree)	0.646 (0.098)	0.574 (0.143)	24.844**	.073
Standard deviation of recurrence (Betweenness)	1.905 (1.394)	1.318 (1.129)	17.236**	.052
Entropy (Betweenness)	0.123 (0.287)	0.284 (0.416)	14.893**	.045
Standard deviation (Closeness)	0.121 (0.083)	0.155 (0.087)	12.586**	.038
Average recurrence (In-degree)	2.449 (1.392)	1.889 (1.640)	10.219*	.031
Degree of uniformity (Betweenness)	0.626 (0.110)	0.583 (0.128)	9.909*	.030
Average recurrence (Betweenness)	3.999 (1.931)	3.320 (2.239)	7.956*	.024

* $p < .010$, ** $p < .001$

Table 3: Recall, precision, and F1 scores for LOOCV DFA

Count	Did not complete	Completed
Recall	.687	.820
Precision	.765	.754
F1-score	.724	.786

Discussion and conclusion

Previous MOOC studies have investigated completion rates though click-stream data or NLP techniques or a combination of both. Our interest in this study was to focus on language indices related to social interaction and collaboration, which are important components of learning, both inside and outside the classroom (Vygotsky, 1978). This study examined MOOC completion rates using novel Cohesion Network Analysis indices to estimate connections between discourse elements in order to develop models of the underlying semantic content of the MOOC forum posts. The findings from this study indicate that CNA indices are important predictors of student completion rates with students who produce more on-topic posts, posts that are more strongly related to other posts, or posts that are more central to conversation. Thus, the results support the notion that students who collaborate more are more likely to complete the MOOC. These findings have important implications for how students' interactions within the MOOC in reference to collaboration and social integration can be used to predict success.

The results indicate that overall contribution scores showed the strongest differences between those that completed the MOOC and those that did not (see MANOVA results in Table 2). In addition, overall contribution

scores, which reflect an estimate of on-topic relevance for each utterance made by each participant, were a significant predictor in the DFA model. The mean scores (see Table 2) show that participants who produced a greater number of on-topic posts (i.e., were more engaged with the topic of the MOOC) were more likely to complete the course. The next strongest predictors of whether students completed or did not complete the course were related to *closeness* and *eccentricity* applied on weekly CNA interaction graphs. These indices reflect how strongly a student's posts are related to other posts made by other students (i.e., strength of connection to other posts). The results indicate that students are more likely to complete the MOOC if their posts share semantic commonalities with posts made by other students. Two indices related to closeness were included in the final DFA model. After closeness and eccentricity indices, the next strongest indices were related to *in-degree* and *out-degree*. These indices are also computed based on interaction graphs and measure the number and the semantic strength of links to and from other students. The findings show that students who complete the MOOC have a greater number of semantically related links to and from other students in the MOOC. Lastly, a number of *betweenness* indices demonstrated significant differences between students who completed the MOOC and those that did not. Betweenness is a measure of how central a node is to communication in term of the information exchanged between participants. Importantly, betweenness indices indicate how much information would be reduced if participants were eliminated from the conversation. The findings from this study indicate that participants who were more critical to forum discussion threads were more likely to complete the MOOC.

In terms of comparison to previous findings, our CNA indices alone are as powerful as the ones employed in previous studies that combined both NLP and click-stream data (Crossley et al., 2016) with accuracies of 76% in both cases, and more powerful than using NLP indices alone (67% with NLP indices compared to 76% with CNA indices used in the longitudinal analysis; (Crossley et al., 2015). More importantly, the indices indicate that patterns of collaboration and social interaction are important for understanding success, going beyond individual linguistic differences and click-stream patterns. Thus, the findings help to provide support to the basic notion that cognitive engagement during learning is a key component of learning and success (Corno & Mandinach, 1983) and that cooperative work may lead to greater learning gains (Johnson & Johnson, 1990). More importantly, these theories of collaboration within learning environments can be extended to large scale on-line classrooms, such as MOOCs. Even in MOOCs, it appears that those students who deviate less from the expected content (Standard deviation of recurrence [Overall Score]), and have higher and stronger connections to other participants (Slope degree and Average [Closeness]) are more likely to be successful. Other CNA indices that were not included in the DFA, but demonstrated significant differences between students who completed the course and those that did not, indicated that more successful students had more links to and from other students (in- and out-degree), were central within the community (low eccentricity) and facilitated conversation among students (betweenness).

The models presented in this paper could be employed to monitor and support students less likely to complete the course by providing timely and personalized feedback in order to increase MOOC engagement and long-term completion. However, much of this depends on the availability of textual traces, which are not always available in many MOOCs. While we focused on forum posts in this study, the employed mechanisms should generalize and, as such, could be applied on other text traces such as participation in collaborative chats, written assignments that are scored in terms of effectively summarizing course lectures, responses to open answer questions which are automatically assessed. In all cases, the results reported here need to be substantiated in follow up studies that evaluate the applicability of the introduced CNA indices in the analysis of MOOCs from other domains and on MOOCs built on other platforms. The LSA and LDA spaces developed for this study may need to change based on new domains, although this needs to be tested. In addition, the CNA indices introduced here could be combined with more traditional NLP indices, click-stream variables, and individual difference measures to further enhance our understanding of student success in on-line classes.

References

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Int. AAAI Conf. on Weblogs and Social Media* (pp. 361–362). San Jose, CA: AAAI Press.

Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics, 32*(1), 13–47.

Chaturvedi, S., Goldwasser, D., & Daume, H. (2014). Predicting instructor's intervention in MOOC forums. In *52nd Annual Meeting of the ACL* (pp. 1501–1511). Baltimore, MA: ACL.

Corno, L., & Mandinach, E. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist, 18*, 88–100.

Crossley, S. A., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). Language to completion: Success in an educational data mining massive open online class. In *8th Int. Conf. on Educational Data Mining* (pp. 388–392). Madrid, Spain.

Crossley, S. A., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining Click-Stream Data with NLP Tools to Better Understand MOOC Completion. In *6th Int. Conf. on Learning Analytics & Knowledge (LAK '16)* (pp. 6–14). Edingurgh, UK: ACM.

Dascalu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence* (Vol. 534). Cham, Switzerland: Springer.

Dascalu, M., Trausan-Matu, S., McNamara, D. S., & Dessus, P. (2015). ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism. *International Journal of Computer-Supported Collaborative Learning, 10*(4), 395–423. doi: 10.1007/s11412-015-9226-y

Dowell, N., Skrypnyk, O., Joksimovic, S., Graesser, A., Dawson, S., Gasevic, D., de Vries, P., Hennis, T., & Kovanovic, V. (2015). Modeling Learners' Social Centrality and Performance through Language and Discourse. In *EDM 2015* (pp. 130–137). Madrid, Spain: International Educational Data Mining Society.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*(1), 35–41.

He, J., Bailey, J., Rubinstein, B. I.P., & Zhang, R. (2015). Identifying at-risk students in massive open online courses. In *AAAI 2015* (pp. 1749–1755). Austin, Texas: AAAI Press.

Johnson, D. W., & Johnson, R. T. (1990). Cooperative learning and achievement. In S. Sharan (Ed.), *Cooperative learning: Theory and research* (pp. 23–37). New York, NY: Praeger.

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and Intention in Massive Open Online Courses. *EDUCAUSE Review, 48*(3), 62–63.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Ramesh, A., Goldwasser, D., Huang, B., Daume, H., & Getoor, L. (2014). Understanding MOOC Discussion Forums using Seeded LDA. In *9th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 28–33). Baltimore, MA: ACL.

Rebedea, T. (2012). *Computer-Based Support and Feedback for Collaborative Chat Conversations and Discussion Forums.* (Doctoral dissertation), University Politehnica of Bucharest, Bucharest, Romania.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika, 31*, 581–603.

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM, 57*(4), 58–65.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27*, 379–423 & 623–656.

Trausan-Matu, S., Dascalu, M., & Dessus, P. (2012). Textual complexity and discourse structure in Computer-Supported Collaborative Learning. In S. A. Cerri, W. J. Clancey, G. Papadourakis & K. Panourgia (Eds.), *11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012)* (pp. 352–357). Chania, Grece: Springer.

Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2007). Supporting polyphonic collaborative learning. *E-service Journal, Indiana University Press, 6*(1), 58–74.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wang, Y. (2014). MOOC Leaner Motivation and Learning Pattern Discovery. In J. Stamper, Z. Pardos, M. Mavrikis & B. McLaren (Eds.), *7th Int. Conf. on Educational Data Mining* (pp. 452–454). London, UK.

Wen, M., Yang, D., & Rose, C. P. (2014a). Linguistic Reflections of Student Engagement in Massive Open Online Courses. In *Int. Conf. on Weblogs and Social Media*.

Wen, M., Yang, D., & Rose, C. P. (2014b). Sentiment Analysis in MOOC Discussion Forums: What does it tell us. In J. Stamper, Z. Pardos, M. Mavrikis & B. M. McLaren (Eds.), *7th Int. Conf. on Educational Data Mining* (pp. 130–137). London, UK.

Acknowledgments

This research was partially supported by the FP7 2008-212578 LTfLL project, by the 644187 EC H2020 *Realising an Applied Gaming Eco-system* (RAGE) project, by University Politehnica of Bucharest through the "Excellence Research Grants" Program UPB–GEX 12/26.09.2016, as well as by the NSF grant 1417997 to Arizona State University.