"Turn on, Tune in, Drop out": Anticipating student dropouts in Massive Open Online Courses

Diyi Yang[1] **Tanmay Sinha**[2] **David Adamson**[1] **Carolyn Penstein Rose**[1]

[1] **Language Technologies Institute**
Carnegie Mellon University
Pittsburgh, PA 15213
{diyiy,dadamson,cprose}@cs.cmu.edu

[2] **School of Computer Science**
Vellore Institute of Technology
Chennai 600127, India
tanmay.sinha2010@vit.ac.in

Abstract

In this paper, we explore student dropout behavior in Massive Open Online Courses(MOOC). We use as a case study a recent Coursera class from which we develop a survival model that allows us to measure the influence of factors extracted from that data on student dropout rate. Specifically we explore factors related to student behavior and social positioning within discussion forums using standard social network analytic techniques. The analysis reveals several significant predictors of dropout.

1 Introduction

With the recent boom in development of educational resources both in industry and academia, Massive Open Online Courses(MOOC) have rapidly moved into a place of prominence in the media, in scholarly publications, and in the mind of the public. The hope is that this new surge of development will bring the vision of equitable access to lifelong learning opportunities within practical reach. MOOCs offer many valuable learning experiences to students, from video lectures, readings, assignments and exams, to opportunities to connect and collaborate with others through threaded discussion forums and other Web 2.0 technologies. Nevertheless, despite all this potential, MOOCs have so far failed to produce evidence that this potential is being realized in the current instantiation of MOOCs. Of particular concern is the extremely high rates of attrition that have been reported for this first generation of MOOCs. For example, Duke Universitys Fall 2012 offering of Bioelectricity had 12,175 students registered. However, only 7,761 students ever watched a video, 3,658 students took at least one quiz, 345 students attempted the final exam, and finally only 313 passed with a certificate [1, 4]. It is reported that general dropout rate is from 91% to 93% [2]. In this paper, we leverage a state-of-the-art statistical analysis approach to investigate this problem and form a vision for development of new paradigms for analysis of MOOC data. We work towards not only understanding the problem better but also proposing solutions that may make the next generation of MOOCs more successful.

One important hurdle that prevents MOOCs from reaching their transformative potential is that they fail to provide the kind of social environment that is conducive to sustained engagement and learning, especially as students arrive in waves to these online learning communities. The unique developmental history of MOOCs creates challenges that require insight into the inner-workings of massive scale social interaction in order to meet. In particular, rather than evolving gradually as better understood forms of online communities, MOOCs spring up overnight and then expand in waves as new cohorts of students arrive from week to week to begin the course. As massive communities of strangers that lack shared practices that would enable them to form supportive bonds of interaction, these communities grow in an unruly manner. While some students may successfully find birds of a feather with whom to bond and find support, when others come they may find an overwhelm-

ing amount of communication having already been posted that they feel lost in. Others may find themselves somewhere in between these two extremes. They may begin to form weak bonds with some other students when they join, however, massive attrition may create challenges as members who have begun to form bonds with fellow students soon find their virtual cohort dwindling. Early attempts to organize the community into smaller study groups may be thwarted by such periodic growth spurts paired with attrition, as groups that initially had an appropriate critical mass soon fall below that level and then are unable to support the needs of remaining students.

The existing research investigating the reasons for attrition do not address specifically the question of how to create a social environment that would be more conducive to promoting continued engagement. Much of this research focuses specifically on summative measures of attrition. They seek to identify factors that predict completion of the course, for example, by conducting a correlational analysis between course completion and click stream evidence of engagement with course activities. However, what we see is that attrition happens over time. While a large proportion of students who drop out either fail to engage meaningfully in the course materials at all or drop out after the first week of participation, a significant proportion of students remain in the course longer than that but then drop out along the way. This suggests that there are students who are struggling to stay involved. Supporting the participation of these struggling students may be the first low hanging fruit for increasing the success rate of courses. However, before we can do so, we need to understand better their experience of participation along the way as they struggle and then ultimately drop out. Current published analyses of participation in MOOCs have not provided the needed visibility into the social interactions between students within the MOOC context. In this paper, we address these limitations in two ways. First, we employ a statistical model referred to as a survival model in order to evaluate the factors that increase and decrease dropout along the way, rather than focusing specifically on course completion. Second, we include in our analysis indicators of social relatedness derived from social network analysis.

In the remainder of the paper we first offer a more detailed review of current research related to attrition in MOOCs. Next, we describe our exploratory work that motivated our statistical analysis. We then outline our method for extracting indicators of course engagement along the way, including both social network measures and measures of engagement in discussion forums. We then present results from a survival analysis model to investigate the factors that contribute to course dropout along the way. We conclude with discussion and our vision for future research, particularly focused on opportunities for modeling using machine learning and graphical modeling techniques.

2 Related Work

In this section we outline the perspectives on students attrition that have been explored so far in the emerging literature on MOOCs. Much of this work successfully leverages machine learning and other advanced statistical methods.Some of this work has grown out of research on distance education that preceeded the emergence of MOOCs as an online learning paradigm.

Some prior work has focused on questions related to the preparation of students for engaging in MOOC learning. For example, Tinto et al. [25] explain student dropout in as a socio-psychological processes that occur as students transition from their life prior to university participation and their new engagement in university life. Thus, this work has focused on factors that affect readiness to engage in this new stage of life, including gender roles and expectations, financial resouces, etc. Other work in this area focuses on environmental factors within the university learning community itself. Questions focus on how factors related to student rediness interact with these environmental factors. Related to this, Gerben et al. [23] conducted a case study in which they used machine learning techniques to predict student success using features extracted from student pre-university academic records. Similarly, Marquez-vera et al. [24] published results from a model predicting school failure from grade point average, scores in each subject, age, etc. This work suggests that student attrition is predictable to some extent from information that can be quantified, which paves the way for significant machine learning work in this area.

It is important to note, however, that despite the substantial history of research in online education from the past, the rise of the MOOCs raises new questions. In the context of MOOCs, students have substantially more freedom to determine what, when, where, and how they will learn. The materials are freely available, which creates the opportunity for students to brows, pick and choose, and follow their own agenda in ways that were not feasible in earlier models of online education. The barrier to

entry is low, and there is no penalty for dropout. In the Wang et al. [7] comparison between dropout in MOOCs relative to more traditional forms of instruction, dropout in MOOCs was far steeper than expected.

The environmental factors just discussed both dramatically affect the external forces that affect student motivation as well as the meaning and significance of dropout itelf as a construct. In an attempt to better understand the individual differences and environmental factors that influence student disengagement in MOOCs, Kizilcec et al. [6] applied an unsupervised machine learning approach to characterize patterns of engagement and disengagement in three computer science courses of varying difficulty levels. They identified emerging clusters of learners characterized as completing, auditing, disengaging and sampling learners. In their work, they evaluated factors such as demographic information from surveys, geo-location of learners, stated intentions upon enrolling in courses, and preliminary forum activity. Recommendations based on observed learning trajectories within the clusters was provided. Clow et al. [26] have characterized the pattern of attrition in MOOCs as a funnel of participation. This concept of a funnel is strongly related to our own effort to understand how attrition happens along the way as students participate in a course. Ispot, Cloudworks and OpenED MOOC platforms corroborated the large drop off at each stage of the evolution of the MOOC. These evolutionare stages include potential learners coming to know that the MOOC exists, a small proportion of those who were aware signing up for the MOOC, a fraction of those registered going on to engage in some activity or other, and some of those making meaningful learning progress. In the spirit of these types of historical analyses, and more similar to our own approach, Balakrishnan et al. [9] have focused on student behavior in an edX course and examined student histories of actions during a course in order to predict dropout from one week to the next. Their goal, like ours, was to understand the factors that affect dropout along the way in order to motivate development of interventions that might address these shortcomings. Huang et al. [8] analyzed students' submission behaviors in MOOCs, mainly focusing on the syntax and functional similarity of the submissions in order to explore the extent to which these characteristics make predictions about committment to the course over time. While this work provides a conceptual foundation for our own efforts, it does not address our specific questions related to the connection between social relations within the discussion threads and how that affects dropout.

3 Data and Exploratory Analysis

In preparation for engaging in a partnership with an instructor team for a Coursera MOOC that was launched in Fall of 2013, we were given permission by Coursera to scrape and study a small number of other courses. Our goal was to gain insights that would enable us to develop tools for instructor support. We began by informally examining the interactions on discussion threads in a literature course offered in June 2013.

In the 7th week of the course when we scraped the data, there were 771 users who had ever posted at least once, and our constructed social network graph from interactions within the discussion forums contained a total of 3848 edges. In our preliminary work, we began by applying a variety traditional social network analysis measures to the graph at different time points with an eye towards examining how cohorts of students who started the course at different times interacted with the community as it existed at the time of their entry. At this stage, our work was exploratory rather than hypothesis driven. A central theme in our exploratory work was to leverage the concept of a cohort, which we defined as the set of students who began work on the course within a particular week after the official launch date of the course. Thus, all students who began work within the first week were labeled as cohort 1, while all students who began in the second week were labeled as cohort 2, and so on. Our goal was to understand what was different in the experiences of these different waves of students.

Our preliminary work revealed distinctly different trajectories through the course for cohorts who began at different times. The earliest cohorts completed more of the course and were less likely to drop out. Students from later cohorts, perhaps in an effort to catch up, appeared to frequently adopt the strategy of beginning their active engagement with the materials of the course during the week of material prior to the week in which they joined rather than beginning with the week 1 materials. What appeared most significant to us in our exploratory work was that the later cohort students appeared to have trouble getting integrated into the community discussion. The few high centrality participants were mainly from the earliest cohort. Members from the early cohort were also more likely than others to continue participating in discussion forums dedicated to earlier portions of the course long after they advanced past that material. Nevertheless, the students who joined during

later cohorts and posted in the same subforums where they were active were not highly likely to engage in discussions with them. Students in later cohorts were more likely to remain at the periphery while students from the earlier cohorts continued interacting with one another. Later cohort students appeared to adopt a less intense style of participation. They posted at a lower rate than the earlier cohorts. And they rarely returned to discussions in earlier weeks once they had advanced to subsequent weeks of material. This pattern points either to motivational differences between students who join early and students who join late or to challenges for students who join the course late in getting integrated properly into the course. In order to formalize these findings and understand this pattern better, we employed a well known statistical technique that has been used previously to understand participation patterns in other types of online communities, namely a survival model[10, 11, 12]. Survival models are a form of proportional log odds logistic models for time series analysis that aim to answer quantify the impact certain factors have on the rate of drop out over time. Such a model might offer some answers to questions such as what kinds of students might be more likely to persist in a course past a particular hurdle, or how important is it that students receive immediate responses to their queries, or does it matter whether answers to queries come from the instructor, a fellow student, or a bot? Where this approach has been applied to online medical support communities [13], for example, a survival model was used to quantify the impact of receipt of emotional and informational support on longevity within the community. The analysis was able to reveal that whereas both information and emotional support are important, emotional support appears to be far more critical to formation of committment to remain active in the community.

4 Method: Exploring Factors Affecting Dropout through Operationalizations of Social Positioning

Our exploratory analysis suggested that student behavior in the discussion forum might predict attrition. This makes sense intuitively. For example, a student who has decided to finish an online course might be more likely to dig in to the details of each assignment and lecture, which might make him/her more inclined to actively post questions, reply or comment in the discussion forum. In an attempt to operationalize these factors, we define metrics related to posting behavior and social positioning within the resulting reply network. We also discuss here control variables used in our subsequent analysis as well as more details about the survival analysis method we use.

4.1 Posting Behavior

The features we considered for each person each week include:

- Thread starter- Refers to whether a student has started a thread within the particular week or not (binary value of 0/1); Sub thread starter- Refers to whether a student has started a sub thread within the particular week or not. Such people are actually discussion initiators within threads whose posts generate comments greater than a particular threshold. We arbitrarily choose the threshold value as 3 comments (binary value of 0/1)
- Post length- Refers to the number of posts for a particular user; Post density- Refers to the Post length divided by Post duration for the weeks a student survived; Post Duration refers to the time difference between the first post and last post in current week.
- Content length- Refers to the number of characters spoken on the discussion forum; Content density- Refers to the Content length divided by Post length for the weeks a student survived

The motivation behind examining the above forum features is to gain insights into : 1)whether thread starters and sub thread starters differ in their probabilities of surviving, as opposed to people who only reply to threads/sub threads; 2)whether the pattern of posting makes any difference to students' survival; 3)whether survival of students is affected by starting/not starting threads/sub threads and then engaging/not engaging in active discussion afterwards. For example, when there are a stream of bursty posts from a student, is it a potential indicator of their interestedness in the course, and are they are more or less likely to dropout afterwards?

4.2 Social Network Behavior

To fully understand the structured discussion forum we explore a wide range of standard social network analytic measures that capture aspects of social positioning within the resulting reply networks. In our network formalization, thread starters or thread initiators have an outward link to all

people who have posted or commented within that particular thread. If people post more than once within a thread, we count them as having a stronger tie strength to the thread starter. We use directed links in our network construction. Because we are interested in how behavior within a week affects survival to the next week of the course, we construct a separate network specifically for each week of participation and then extract the measures from that network. We do this for each student in each week of their active participation. This representation of each student week of behavior is then input to the survival model.

- Centrality measures such as Degree (which is the average value of number of inlinks and outlinks), Eigenvector centrality (which measures node importance in a network based on a node's connections), Betweenness (which measures how often a node appears on the shortest path between nodes in the network) and Closeness centrality (which measures average distance from a given starting node to all other nodes in the network). Using a similar analogy as in Borgatti et al. [14], we can say that Eigenvector centrality will capture the long term direct and indirect effect of a student's interaction patterns and the implication of their connections in the MOOC network, while Degree centrality will capture more immediate effects.
- Average Clustering coefficient [15]- Indicates how nodes are embedded in their neighborhood, which can be thought of as an overall indication of the "small world" effect or clustering in the network. The motivation for examining this factor as a potential indicator of dropout is because, if there is absence of cliquishness in discussion forums, students would not find enough active partners to engage in discussions. So, due to lack of support and fruitful means of engagement, they would be more inclined to leave the course. Having more tightly knit neighbors influences structural location of students in the network, which inturn facilitates discussion and motivates students to remain in the course.
- Eccentricity [16]- Indicates the distance from a given starting node to the farthest node in the network. As an extreme measure of centrality, importance of examining this variable in the MOOC network is to intuitively monitor how response time affects student's participation. In the MOOC, students with low eccentricity values will primarily be at center of the graph and therefore more receptive to information, as compared to students having high eccentricity who will belong to the periphery of the network. And if students are partially cut off from the influence of the core or central group of students in the network, their chances of dropout might be increased.
- Authority and Hub scores [17]- These indicate how valuable information stored in a node is and the quality of the node's links. In a MOOC, students with a good authority scores are those who engage other students in discussions. Students with good hub scores are those are who get engaged in discussions initiated by many active learners such as thread starters or sub thread starters. We find a strong correlation between these two measures in our data.

4.3 Control Variables

While our goal is to investigate the effect of week by week participation patterns, we must acknowledge that there are stable characteristics of students that remain with them throughout their participation. In our work, we have included only one control variable, namely which cohort a student belongs to. In a MOOC, learners can join the course at any point of time till the course is actively running. Cohort number indicates which week a student began participating in the course.

4.4 Survival Model

Survival model could be regarded as type of regression model, which captures the changes of probability of survival over time. Survival analysis is known to provide less biased estimates than simpler techniques (e.g., standard least squares linear regression) that do not take into account the potentially truncated nature of time-to-event data (e.g., users who had not yet left the community at the time of the analysis but might at some point subsequently) From a more technical perspective, a survival model is a form of proportional odds logistic regression, where a prediction about the likelihood of a failure occurring is made at each time point based on the presence of some set of predictors. The estimated weights on the predictors are referred to as hazard ratios. The hazard ratio of a predictor indicates how the relative likelihood of the failure occurring increases or decreases with an increase or decrease in the associated predictor.

In our model, the dependent measure is Dropout, which is 1 on a student's last week of active participation, and 0 on other weeks. Formally, we represent cohort in our model as a set of binary indicators that identify a student as being part of a cohort or not. We began with seven such indicators, one for each week of the course. The other factors were numeric scale indicators. For running our Survival models, we use the statistical software package Stata [18]. The survival analysis is conducted via the incorporation of a subset of discussion forum and social network metrics. The results of the survival analysis (assuming a Weibull distribution of survival times) are discussed in the next section.

5 Results: Survival Prediction and Results Analysis

A qualitative sense of the social network and its relation to attrition can be viewed in the following graphs. First, as an indication of the growth of social connectedness from week to week, see Figure 1 below. We conducted our survival analysis in a stage-based way in order to identify just those factors that had the strongest effect on student attrition. The results of the final model that contains only effects greater than 5% is displayed in the Table 1. Note that a hazard ratio of less than 1 indicates that the factor leads to less attrition where a hazard ratio of more than 1 indicates that a factor leads to more attrition.

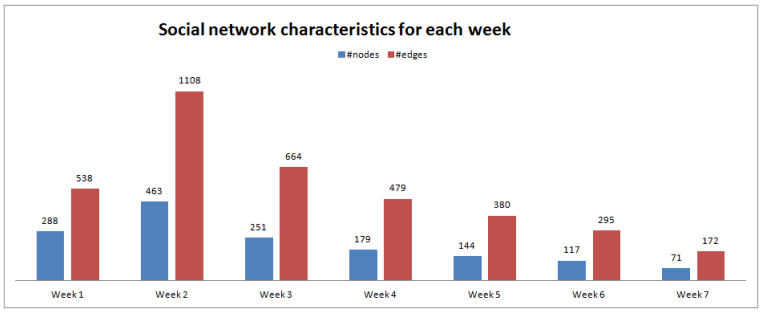

Figure 1: Social Network characteristics

We began with a basic model investigating just the cohort effect. The finding was that the later cohorts each had progressively higher attrition than the earlier ones, but the only statistically significant effect was that members of the first cohort had significantly lower dropout than others. In the final model, therefore, we only include a binary indicator of membership in Cohort 1. That factor has a hazard ratio of .67, which indicates that members of the first cohort are 33% less likely than others to drop out on a given week.

In the next stage, we tested a model including the control variable Cohort 1 along with the behavioral variables related to forum behavior, specifically, Thread Starter, Subthread Starter, Number of Posts, Post Density, Post Duration, Content Length, and Content Density. Of those, only Post Duration met our criterion of leading to a significant impact of at least 5%. When included in our final model, the hazard ratio was measured at .4, which indicates that students whose posts spanned a duration of 1 S.D higher than average were 60% less likely to drop out on that week than students with an average Post Duration.

Next, we evaluated a model containing the Cohort 1 variable, the Post Density variable, and subsets of the network measures of degree, In degree, Out degree, Eccentricity, Closeness, Betweenness, Authority, Hub, Eigenvector Centrality, and Clustering Coefficient. Many of these measures were highly correlated, and thus we were not able to include them all in a model simultaneously. Due to skewness in distribution, network measures were log transformed and standardized prior to inclusion in the model. Our finding from this analysis demonstrated that only Hub and Authority met our criteria for inclusion. Since they were highly correlated, we were only able to include one in the final model. Since Authority made a stronger prediction, we included that. In the final model, the hazard ratio assigned to Authority is .67, which indicates that Authority scores that are 1 standard deviation higher than average indicate that the student is 33% less likely to drop out on that week than students with an average value.

Variable	Hazard Ratio	Standard Error	Z	p
Cohort 1	.67	.06	-4.1	.001
Post Duration	.4	.08	-4.7	.001
Authority	.67	.04	-6.1	.001

Table 1: Survival Analysis Results

6 Conclusion

In this paper we have begun to lay a foundation for research investigating the social factors that affect dropout along the way during participation in MOOCs. Our results suggest that these social factors do indeed affect dropout along the way and provide a potentially valuable source of insight for design of MOOCs that may be more conducive to social engagement that promotes committment and therefore lower attrition.

In our work going forward, we seek to understand more about how bonds begin to form in the interactions with the discussion threads as we search for answers for how to better structure discussion forums so that students who join the course late are as able as the early cohorts to form lasting bonds and get integrated into the course. In order to gain better visibility into bond formation and evolving community structure, we are developing probabilistic graphical models for modeling subcommunity structure formation over time. Specific questions that guide our work going forward include: What experiences in interaction between students predict continued social engagement between them? To what extent does emergent subcommunity structure explain attrition patterns in MOOCs? For example, what variables distinguish subcommunities with lower attrition from those with higher attrition? To what extent do patterns of subcommunity participation (i.e., consistently interacting within a specific subcommunity versus participation in multiple subcommunities) predict patterns in attrition? In our modeling work, we make use of mixed membership social network partitioning models [27] that compute soft partitions of social networks[29], where each partition represents a subcommunity, and individual members can belong to and thus be influenced by the norms associated with different ones at different times. We similarly draw from work integrating text mining techniques with social network analysis in order to form representations of text that reflect the community structure [28]. Our goal will be to understand the emergence of shared practices within the patterns of social interactions in MOOCs so that we can use this understanding to develop support for healthy and prolonged engagement in learning in that context.

Acknowledgement

This work was supported in part by NSF grant OMA-0836012.

References

[1] Catropa, Dayna (24 February 2013). Big (MOOC) Data. Inside Higher Ed. Retrieved 27 March 2013.

[2] MOOCs on the Move: How Coursera Is Disrupting the Traditional Classroom (text and video). Knowledge @ Wharton. University of Pennsylvania. 7 November 2012. Retrieved 23 April 2013.

[3] Kolowich, Steve (8 April 2013). Coursera Takes a Nuanced View of MOOC Dropout Rates. The Chronicle of Higher Education. Retrieved 19 April 2013.

[4] Jordan, Katy. MOOC Completion Rates: The Data. Retrieved 23 April 2013.

[5] MOOC Interrupted: Top 10 Reasons Our Readers Didnt Finish a Massive Open Online Course. Open Culture. Retrieved 21 April 2013.

[6] Ren F. Kizilcec; Chris Piech, Emily Schneider. Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. LAK conference presentation. Retrieved 22 April 2013.

[7] Yuan Wang. Exploring Possible Reasons behind Low Student Retention Rates of Massive Online Open Courses: A Comparative Case Study from s Social Cognitive Perspective". In Proceedings of the 1st Workshop on Massive Open Online Courses at the 16th Annual Conference on Artificial Intelligence in Education, 2013.

[8] Jonathan Huang, Chris Piech, Andy Nguyen, Leonidas Guibas. Syntactic and Functional Variability of a Million Code Submissions in a Machine Learning MOOC. In Proceedings of the 1st Workshop on Massive Open Online Courses at the 16th Annual Conference on Artificial Intelligence in Education, 2013.

[9] Balakrishnan Girish. Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models. EECS Department, University of California, Berkeley, May, 2013.

[10] Singh R, Mukhopadhyay K. Survival analysis in clinical trials: Basics and must know areas. Perspect Clin Res 2011;2:145-8.

[11] Bruin, J. 2006. newtest: command to compute new test. UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/stata/ado/analysis/.

[12] Richards, S. J. A handbook of parametric survival models for actuarial use. Scandinavian Actuarial Journal. 2012;4:233-257.

[13] Wang, Yi-Chia, Robert Kraut, and John M. Levine. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. ACM, 2012.

[14] Borgatti, Stephen P. Centrality and network flow. Social networks 27.1 (2005): 55-71.

[15] Watts, Duncan J., and Steven H. Strogatz. Collective dynamics of small-worldnetworks. nature 393.6684 (1998): 440-442.

[16] Wilson, Christo, et al. User interactions in social networks and their implications. Proceedings of the 4th ACM European conference on Computer systems. Acm, 2009.

[17] Kleinberg, Jon M. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) 46.5 (1999): 604-632.

[18] Stata Corporation. Stata Statistical Software Release 7.0: Programming. Stata Corporation, 2001.

[19] Mayfield, E., and C. P. Ros. LightSIDE: open source machine learning for text. Handbook of automated essay evaluation. Routledge, New York (2013).

[20] Demar, Janez, et al. Orange: From experimental machine learning to interactive data mining. Springer Berlin Heidelberg, 2004.

[21] Jakulin, aleks. Machine learning based on attribute interactions. Diss. Univerza v Ljubljani, 2005.

[22] Kotsiantis, Sotiris B., C. J. Pierrakeas, and Panayiotis E. Pintelas. Preventing student dropout in distance learning using machine learning techniques. Knowledge-Based Intelligent Information and Engineering Systems. Springer Berlin Heidelberg, 2003.

[23] Dekker, Gerben, Mykola Pechenizkiy, and Jan Vleeshouwers. Predicting Students Drop Out: A Case Study. EDM 9 (2009): 41-50.

[24] Marquez-Vera, Carlos, Cristobal Romero, and Sebastin Ventura. Predicting School Failure Using Data Mining. EDM. 2011.

[25] Tinto, V. Limits of theory and practice in student attrition, Journal of Higher Education 53, p. 687-700, 1982.

[26] Clow, Doug. MOOCs and the funnel of participation. Proceedings of the Third International Conference on Learning Analytics and Knowledge. ACM, 2013.

[27] Airoldi, E., Blei, D., Fienberg, S. Xing, E. P. (2008). Mixed Membership Stochastic Blockmodel, Journal of Machine Learning Research, 9(Sep):1981–2014, 2008.

[28] McCallum, A., Corrada-Emmanuel, A., Wang, X. (2004). The Author-Recipient-Topic Model for Topic and Role Discovy in Social Networks: Experiments with Enron and Academic Email, Technical Report UM-CS-2004-096, University of Massachusetts, Amherst.

[29] Sim, Y. C., Smith, N., Smith, D. (2012) Discovering factions in the computational linguistics community. In Proceedings of the ACL Workshop on Rediscovering Fifty Years of Discoveries, Jeju, Korea, July 2012.