Pauline Chow

Draft Report for EduTech Data Analysis

Working Title: Social Network Analysis (and Machine Learning/Data Science/X, Y, Z Model) (to Pinpoint Student Success): Student Participation and Interaction Metrics in Online Learning Environments

[[add table of contents]]

A. Introduction

Online learning through Massive Open Online Courses (MOOCs) and larger scale accredited programs are changing opportunities for higher education. MOOCs are technology assisted classrooms that allow a large number of access courses and courses of study over the internet free of charge. Between 2012 and 2015 over 25 million global users have engaged in MOOCs on Coursera, EdX and other similar platforms. The initial excitement for MOOCs impact on the delivery and dissemination of education was diminished when course administrators realized that only a small percentage of students were actually finishing these courses. On one-hand the lower than expected “success” of students can be seen as a failure by a platform for learning. While on the other hand, it is not surprising that non-traditional learning environments like MOOCs and large scale learning environments do not conform with metrics used for traditional education programs. Students of MOOCs perceive an impact to their education (72%) and provided career benefits (61%) whether or not they finished the registered course.

Benefits, outcomes, and perceptions in larger scaled online learning environments can differ from traditional experiences. Georgia Institute of Technology’s Online Masters in Computer Science (OMSCS) program is not categorized as a MOOC. Yet OMSCS is positioned in an interesting place relative to similar online computer science degrees at the time of writing this report. For context of the OMSCS program, it provides an exclusively online experience for 2,841 students as of Fall 2015. The estimated cost of attendance for OMSCS is $7,000 USD, while the range for online computer science degrees ranges from $27,660 to $7,000.

[[develop section more with citations]] In an online learning context, success, satisfaction, and progress of students may look different from traditional educational programs. Analytics practices for scaled online learning, including MOOCS and accredited programs, must be reconceptualized and adapted to different curriculums, objectives, learning principles, and technologies. Optimized data collection approaches in online learning brings a massive influx of information in the form of structured, unstructured, and semi-structured data sets. Studies continue to explore and investigate potential online learning tracking and evaluation metrics, especially from unstructured discussion board text.

Literature on online learning analytics acknowledge the lack of metrics to fully serve online learning environments and students engaging in these free and paid for services. Online learning programs continue to base key metrics on summary statistics of the student body, technology, and course material. [site sources] The target feature or objective remains focused on traditional means, such as retention, enrollment, and grades. [site sources]

The focus of this paper is to recommend metrics for evaluating and observing student participation and interaction in online learning environments. This paper can also be characterized as a data report that compare and contrast students from two online learning classes from Georgia Institute of Technology’s Online Masters in Computer Science (OMSCS) program. The case study approach will explore structured and unstructured data from Piazza posts, surveys, student grade data, and peer feedback forms, and merge relevant features for implementing data science models, including unsupervised learning clustering (i.e. kmeans and hierarchical clustering) and supervised learning (i.e. decision trees). Data for two classes is provided by Professor Joyner in the Fall 2016 Educational Technology course: (1) project based course Educational Technology (EduTech), (2) programming or theory based course knowledged based artificial intelligence (KBAI).

The comparison of metrics in different contexts will help advance understanding of instructional teams and students. Leveraging access to OMSCS piazza posts as starting points for constructing metrics will contribute to the literature about online learning spaces and their implementation. Potentially these insights from a smaller scale can be applied to larger scale learning environments. Additionally, I hope to contribute my code, visualizations and insights to the open source community as a github repository.

B. About GaTech’s OMSCS Program and Classes

Within the OMSCS program, data from two courses were compared:

EduTech curriculum and objectives

KBAI curriculum and objectives

Two classes compared are from the same professor at GaTech

C. Data Source

[[Describe data sources, columns fields, etc]] No control over collection and download of dataset. Datasets are collected throughout the semester. Student information is anonymized so that there is no individually identifiable information related to students.

Create data set from:

Student Surveys

Gradebook

Piazza Forum discussion board

Users

Class Content

Peer feedback data (maybe??)

About the created dataset:

An important step is to clean and integrate provided data sources into datasets that be used as inputs for machine learning/data science models. Dataset 1 is constructed to evaluate the final features of users and the class. As in Dataset 1 is analysis on the status of the course and students at the end of the course, after all items are collected and completed.

Metric

Description

Type

String

Categorical: Student, Teacher, Misc, Guest

Student - part of the student body, participates in the course, hands in assignments, and evaluated for the course

Teacher - part of the instructor team, such as professor, teaching assistant, graders, and administration

Guest - [[inquire]]

Misc - [[inquire]]

[[build out table]]

D. Data Exploration

This section of data exploration and analysis provides the detailed step-by-step approach taken to understand and integrate the collected data from two OMSCS courses in Fall 2015, Educational Technology (EduTech) and Knowledge Based Artificial Intelligence (KBAI). The objective of this section is to guide readers through the thought process and analysis in creating this report and recommendations. Data exploration here takes a deductive approach, where analysis starts from a high level and then drills down into the details.

**1. There is a linear relationship between effort and achievement for both KBAI and EduTech, even when the underlying structures are significantly different given social network visualizations of user interactions.**

First, classroom and user data is analyzed at the highest level, with the complete history of interactions at the course and when all values are finalized. Analyzing the total user piazza activity and final grades at the end of class demonstrate a linear relationship between effort and result for both EduTech and KBAI. See Figure 1a and 1b. Students with the highest total posts, asks, shares, and/or views attained with higher grades. Although there are a few outliers (state users) visualizations show that viewing more posts, asking more questions, spending more time on the discussion boards, and/or viewing more information can make better students, in the traditional grade sense. The impact of these efforts will be analyzed in a later section using data science modes. This is an important observation but only the initial view of the OMSCS online classrooms. See Figure 2a and 2b.

In addition to individual learning, the central space for facilitating sharing, collaboration, and interaction comes from engaging with peers and instructional team in Piazza discussion boards. These boards are integral in supporting the dynamic experiences of students. Counts from Piazza interactions, such as posts, answers, asks, and time on the site can only tell us superficial information about the community on Piazza. Thus during data exploration Piazza data will be converted to social network analysis (SNA) visualizations and metrics. SNA builds a different perspective and understanding of user behavior and relationships.

SNA nodes (i.e. users, posts or topics) and edges (connection between nodes) describe the social structure within the EduTech and KBAI communities. It is also a quantitative way to gather metrics about entities in Piazza, such as users, posts, and topics. SNA graphs are compared and contrasted below to pinpoint the best measure to extract for further analysis and modeling. See Figure 3a and 3b.

Visualizations:

**Pairwise relationships** of User Metrics Piazza Behavior and Final C­ourses Grades: 2015EduTech (Figure 1a) and 2015 KBAI (Figure 1b). There is a clear linear relationship between higher activities to higher achievement by students. Correlation is important to glean from these pairs, where obviously correlated pairs can **confound/confuse** data science models and should be considered for removal from the model.

**Continuous Integration graphs** to demonstrate relationship between views, posts and final course grades. 2015EduTech (Figure 2a) and 2015 KBAI (Figure 2b). Continual confirmation from this set of visualizations that increased effort (posts) contributes to increased achievement based on student user’s final grade. From this chart, the “influence” of users with higher achievement and/or effort can be identified and tracked throughout our analysis.

Evolution of Social Network Graph for EduTech

**Social Network Graph** **of Interactions between Users with the instructional team**. Data for these graphs are extracted at the conclusion of respective courses, which are course with significantly different curriculums, objectives, and implementation styles. The underlying structure \_\_\_\_\_. 2015EduTech (Figure 3a) and 2015 KBAI (Figure 3b).

**Social Network Graph** **of Interactions between Users without the instructional team**.

TODO:

SNA

* Add achievement to the input of nodes
* users only with instructors
* users only WIHTOUT instructors
* post SNA (with and without? Instructors)

**2. Tracking Users with a high level of achievement and effort in the Final SNA visualization and metrics…..**

Achievement and effort should also be viewed in the perspective of contribution to the online classroom experience. For instance, what are the effects of students with the highest achievement and greatest level of effort? Are students achieving as the result of positive and constructive interactions online? Are higher achieving students giving back to the community online and if so, is this done at a higher rate?

TODO:

Pinpoint highest achieving and active students, track the progress of this student over time in SNA, as well as metrics on these students in initial evaluation of the network (at the final stage)

Also view graphs…

3. Leveraging LDA Topic Modeling to construct SNA visualization for users interaction with topics

4. Achievement and effort related to topics (SNA) visualization

5. Evaluation of SNA metrics – relevant for the next step of modeling. What do they mean and why are they interesting.

- SNA with time as a factor to show the underlying structure within the classes, comparison

- importance of collaboration and participation in this model of learning

Network Graphs for Education Technology

Student interactions overall, aggregation/count of the number of times student interact with each other in piazza.

Compare with structure of interactions by post

Add success or characteristics

Output of these metrics and to use in graphs/analytics to determine success

Example of gephi zooming in and showing key interactions and relationships with Teacher2 node

Source: https://webdocs.cs.ualberta.ca/~zaiane/postscript/BC13Collaborative.pdf

E. Unsupervised Learning Models

1. Clustering with these values – as target values?
2. PCA with these values – if we did not want to do feature selection

F. Supervised Learning Models

1. Discuss the value of target values. What is your questions and the choice of a target value for models.
   1. Simple linear regression to discover which effort pays off the most (posts, views, asks, days, connection with peers/connectivity)
   2. Decision Tree / Random Forests -- feature selection; What is the most informative for achievement // other type of target value
   3. Logistic Regression / More than 2 classes -- finding contribution of features to success of students

G. Metrics recommendations, integration into tools for instructors

Finally - Link to Github repo with cleaned data (ok to share because anonymized) and final repo; Create Presentation with visualizations