Data Mining and Machine Learning in Education with Focus in Undergraduate CS Student Success

William Gregory Johnson
Department of Computer Science
Georgia State University
Atlanta, GA 30303 USA
wjohnson6@student.gsu.edu

ABSTRACT

Computer science (CS) enrollments are at an all-time high, [1] and successful undergraduate CS graduations are indisputably important. With a student population of approximately 51,000, Georgia State University is a USA based state university which is diverse and forms a rich big data footprint as students navigate pathways to graduation. Quoted in a July 2017 article from HigherEd.com, "Georgia State's extensive predictive analytics efforts are leading to better grades and student retention – and more minorities graduating from STEM programs." This doctoral project builds upon current data mining and modeling, machine learning applications, and learning analytics for predicting student success that is beyond retention. Gaining knowledge of CS student learning, developing better alerting models for success, and discovering behavioral indicators from learning analytics reporting is the goal of this research. Using this knowledge as evidence based data for improving the CS student experience will aid in performance improvements and increase pathways to graduation. My supporting research project is building CS student datasets to represent the student as directed graphical models, investigating their relationships using machine learning frameworks, and complex mathematical computations (tensors or gradient boosting) along with graph data mining techniques.

CCS CONCEPTS

• **Social and professional topics → Computer science education**;

KEYWORDS

STEM student retention; graduation pathways; educational data mining; graph data mining

ACM Reference Format:
William Gregory Johnson. 2018. Data Mining and Machine Learning in Education with Focus in Undergraduate CS Student Success. In *Proceedings of 2018 International Computing Educational Research Conference (ICER '18).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3230977.3231012

ICER '18, August 13–15, 2018, Espoo, Finland

ACM ISBN 978-1-4503-5628-2/18/08.
https://doi.org/10.1145/3230977.3231012

1 PROGRAM CONTEXT

CS education is the center focus of my career. After a Bachelor's in CS and more than a decade in practice with large US corporations and the US Federal government, I returned to academe, completed a Master's in CS from GSU in 2004. I began teaching online CS courses and later, teaching CS courses as a full time faculty at a local community college. Being a first generation college student motivated me to work in a community college with strong emphasis on education excellence. The merger of my institution and my Master's alma mater created an opportunity to enter the PhD program in CS at GSU. The Summer semester of 2016 initiated my full time status as a PhD student. Passing the qualifying exams in the Fall of 2017, leaves my remaining course requirements of a CS foundation course and one external departmental course. Fulfillment of one additional dissertation committee person is required before a thesis proposal submission. Completing time-line for these is Fall of 2019, and a target defense date is sometime in Fall 2020 or Spring 2021. My current project is analyzing the approved data of my institutional review board (IRB) for a study that compares performance of CS students in CS curriculum foundation courses based on a prerequisite fulfillment by either the discrete mathematics (MATH2420) or theoretical foundations of computer science (CSC2510). This work will be submitted to the ACM SIGCSE in August 2018, and will establish a basis for several related projects combining additional datasets within the university.

2 CONTEXT AND MOTIVATION

A preliminary data analysis shows an increase in population of transfer CS students at GSU and a trend to enroll in more CS courses per semester than native students. It also shows that transfer student performance is consistently different than that of native students, namely higher CS course fail rates and lower CS GPA scores. These detrimental findings [4], motivate me to investigate and identify other predictors of retention or discovery of performance attributes in the CS undergraduate program. My doctoral project will use graph data mining to find communities where behavioral outcome strategies like student-faculty relationships, cognitive and non-cognitive gains, and social network analysis can be applied and evaluated. Using these data-driven findings in my research, I can develop experimental models of predictive course sequencing, adaptive and targeted tutoring, and a student facing learning analytics reporting system to better understand the CS student perceptions and perceived effects of the models. These models along with improving the CS student experience, lowering attrition, and

decreasing loss of time and resources are the motivating goals of my doctoral work.

3 BACKGROUND & RELATED WORK

Much research has focused on undergraduate student success factors resulting from flipped classrooms, hybrid learning, technology usage, and intuition-driven designs. Educational data mining (EDM) focuses on exploring this unique and large-scale research challenge with data-driven analysis to better understand students and their learning environments. [3] EDM is well positioned to analyze components of GSU's CS student digital avatar because they are mixed among sources of the university's learning management systems (LMS), CS courses, grades, demographics, socio-economic indicators, network usage data, and social networking interactions. According to New [5], achieving data-driven CS education systems are a result of four goals: 1) personalization, 2) evidence-based learning, 3) school efficiency, and 4) continuous innovation, all the while protecting the individual student privacy. The inclusion of data-driven education in CS starts with leveraging data to provide a more effective education system. An analysis of 240 EDM works over three years (2010 to 2013) by Penã-Ayla [6], shows that most common approaches to gage performance are student modeling and assessment function. Barker et al., found most higher education retention studies focus on the combining of STEM disciplines, giving an unclear picture of computer science. [2] My research will be completely focused in the CS program.

4 STATEMENT OF THESIS/PROBLEM

Producing bachelor degreed CS students to work in practice and academe is an increasing demand and is driving higher enrollments in colleges and universities in CS programs. Using EDM and learning analytics, a discovery of knowledge relating from student learning and behavior, demographics, academic advising, CS course sequencing, network usage/interactions will produce evidence based data, enabling higher retention and graduation pathways for these students. Using data mining and modeling to detect communities in the evidence based data, CS student prediction models can be made, tested, and improved through empirical research. Creating a learning analytics reporting model where CS students can see predictors of achievement, identification of skill deficits, and receive targeted and personalized intervention modalities, beyond the face-to-face or online interactions is the challenge in my research.

5 RESEARCH GOALS & METHODS

My research goals are:

- Analyze CS student data that fulfill a prerequisite from either MATH2420 or CSC2510 and:
  - Compare effective results on remaining core CS courses (CS grades, course failures, population behaviors)
  - Compare performance and job placement resulting from fulfillment of the prerequisite course (IT vs CS)
  - Currently under research with IRB-H18480
- Represent the CS student in/with graphical model(s) and use social network analysis of the data (clickstream) from our LMS and the internal GSU network to discover relationships based on graph data mining and graph theory principles
- Compare transfer vs native CS student data in CS course loads per semester within course levels and student classifications
- Develop a position paper on transfer CS students addressing:
  - CS student's desire for lower cost of core courses at a 2-year college
  - Conflict of 2-year college's desire for their students to complete an AS degree, resulting in more classes at 2-year, leaving mostly CS courses with a 4-year university
  - And the desire for students to finish a BS degree within three years, post AS degree, with CS course overloading (>= 3 per semester) resulting in performance impacts

6 DISSERTATION STATUS

I have analyzed a cumulative data set of CS student data that compares transfer to native student performance impacts with results published at ACM SIGCSE. [4] In the summer of 2018, my plans include investigate the math versus CS prerequisite effect on remaining requirements of CS curriculum courses. GSU's managing group for our LMS has agreed to meet and discuss a protocol to incorporate data elements for a learning analytics study in the CSC4350, Software Engineering course, I will teach in Fall 2018. The thesis is currently in outline and my third year (Fall 2018, Spring 2019) will be used to complete the thesis proposal and defend it. The following years involve research and preparation for my dissertation defense.

7 EXPECTED CONTRIBUTIONS

The design of a framework or model for a student and faculty facing learning analytics reporting system to deliver advanced analysis of evidence data, LMS interactions, and student performance outcomes. Also, identify and compare communities of CS students through analysis of LMS data for improving the usage of and deeper integration into the undergraduate CS curriculum. Finally, I would like to design and present a common coding scheme or foundation grammar for evidence based education data to open new analysis and measurement instruments. These along with research data will be made available to other undergraduate STEM programs for education research.

REFERENCES

[1] *Assessing and Responding to the Growth of Computer Science Undergraduate Enrollments*. 2017. URL: https://www.nap.edu/catalog/24926, doi:10.17226/24926.
[2] L. Barker, C. L. Hovey, and L. D. Thompson. Results of a large-scale, multi-institutional study of undergraduate retention in computing. In *Frontiers in Education Conference (FIE), 2014 IEEE*, pages 1–8. IEEE, 2014.
[3] EDM Society. International Educational Data Mining Society. URL: http://jedm.educationaldatamining.org/index.php/JEDM.
[4] W. G. Johnson, R. Sunderraman, and A. G. Bourgeois. Performance Impact of Computer Science Course Load and Transfer Status. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, page 1076. ACM, 2018.
[5] J. New. Building a data-driven education system in the United States. *Center for Data Innovation, November*, 25, 2016.
[6] A. Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.