Georgia Tech Enrollment Project
Team #39: Daniel Lyczak

Enrollment Patterns in Higher Education: A Micro and Macro Approach

Introduction

Georgia Tech does not operate with a centralized student model. Course enrollment comprehension is static at best, and often, non-existent. There is no dynamic tool to identify nor aid student enrollment progression to track flow, forecast demand, and understand the interconnectivity of courses and students. Through this project we hope to gain insight while propelling innovation in high impact areas.

Our aim is to develop an interactive tool for student progression and course enrollment / participation at Georgia Tech. The purpose of the tool is to quickly understand student enrollment, first to conceptualize how students flow now, then working to optimize that flow. The existing tools, and any attempts to visualize them, are strictly linear, and predicated on intuition.

On a macro (Institute wide) level, we built a graph to identify how courses are connected to one another based on enrollment patterns. By identifying commonalities in courses taken from one semester to the next, we hope to improve course offerings, work towards improving academic advising, and help the institute with its faculty / staffing allocation. We communicate these findings through a graph which emphasizes these findings with visualizations that tell a story.

On the student (micro) level, we built a tool which analyzes a student's past enrollment schedule, finds other students with similar enrollment patterns, and recommends to the student what courses they should enroll in next. This is communicated through an easy-to-understand chart, which when generated, shows what the student has done, and what they are recommended to do next. The purpose of this approach is twofold; to boost student satisfaction (by increasing retention and the 4-year graduation rate) and begin creating advisor tools that work.

Problem Definition

Despite its name, Georgia Tech has a technology problem. Administratively, there is little attempt to formalize and understand student enrollment beyond the anecdotal. While exists software to help (if purchased) fill this gap, there arise three problems: 1) high cost, 2) high risk sharing Federally protected data, 3) These tools take a one-size fits all approach, which is not optimal for a specialized school.

Georgia Tech, and higher education in general, need to gain more understanding in student behavior (reflected in course participation) and to turn that insight into action. How a student behaves in their courses is not only quantifiable in data that each university has access to, it also directly helps measure their likelihood of being retained (staying in college), graduating on time (by the end of their fourth year), and their overall satisfaction. Without understanding how their students interact and proceed through courses, a university cannot take informed action, and any changes implemented cannot be quantified to measure success / failure. The problem, which we aim to address, is to gain quantifiable insight into student course participation which uses free technologies, is housed locally, and which objectively improves the student experience. We further want to convey these findings visually, in ways which make complex relations and predictions easy to understand, and adaptable to each user's needs.

Background & Literature Survey

Course progression impacts a university's reputation due to its role on student / faculty ratios and time to graduation, key metrics for families choosing a college and national rankings. Qualitatively, course

offerings influence overall student satisfaction, as course offerings have been shown to directly impact retention, often a quantitative measure for a student's happiness with their university life (Dietz-Uhler, 2013). Dietz-Uhler highlights the impact of successful course offerings on qualitative measures of satisfaction. Research (Fraysier, 2020) further confirms the role of engagement, retention, and satisfaction. Though this work emphasizes pre-enrollment characteristics, its incorporation of data on Georgia students (60% of Georgia Tech's undergraduate population is in-state) serves as an aid, as a growing number of students are enrolling at Tech with incoming credits, which impacts student flow. Therefore, these pre-college characteristics must be considered when analyzing course progression.

The common approach to predicting and forecasting course enrollment for four-year universities is a linear regression analysis and classification (Luna, 2009). While this approach has been effective when allocating finances and resources, it minimally considers student course demand (especially across multiple disciplines) into its model. By only focusing on the monetary components of enrollment, (Luna, 2009) best reinforces current methodologies while advocating for a more dynamic / holistic approach.

More recent research (Shao, 2022) reinforces that the standard method of modeling is linear. While emphasizing the need for such models, this paper is limited in its overall application as it only focuses on a single class. Shao does take demographic data and incorporates it with past enrollment numbers, which could be beneficial later to help identify additional bottlenecks in subpopulations. However, it does not consider the interconnectedness of enrollment, which our approach aims to address.

There have been attempts to implement Markov Decision Processes (MDP) in higher education (Slim, 2021), however, this approach was focused on a Monte Carlo simulation aimed at improving curriculum issues based on course complexity. We will build on this (Slim, 2021) theoretical framework with real world data, and with applications beyond course complexity. This paper does add credence to our planned approach to modeling course flow, as their simulation was successful. This paper provides invaluable theory to support future progress in the development of a dynamic and interconnected model, capable of evolving over time.

The MDP approach (Slim, 2021) builds on previous work (Kelley, 2008) which mapped curriculum for a pharmaceutical program. The methodology from this effort is useful in that it focuses on a specific, niche university setting (pharmaceutical studies) which is analogous to the more restricted general course dynamics found at a STEM school like Georgia Tech. Kelley provides an outline on how to map courses while offering best practices in doing so. However, its goal is on program accreditation, not on optimizing the student experience. The systematic nature of their approach will help guide the focus and future development of our work, with the more technical support from other papers (i.e. Slim, 2021).

Attempts to model and understand why a student is not retained at college finds course interaction is the primary factor (Ishitani, 2003). Additionally, Ishitani concludes that GPA (course outcomes) is the strongest indicator to predict a student not being retained. By working to understand how each student individually flows through courses, observing enrollment patterns and grade trends, a classification system could be made to identify at risk students to take intervening action before it is too late.

Proposed Method: Intuition

The current methodologies behind course enrollment and prediction of coursework at Georgia Tech is largely predicated on the intuition of its students. **Appendix I** shows an example of the current tools used. This static worksheet does not take into consideration an individual student's needs or past

performance. Additionally, it embodies the current state of Georgia Tech, and higher education in general, unadaptable. Our approach and intuition are simple; form a connection to the data and convert it into multi-purpose data structures to allows for easy manipulation to calculate, convert, and model the data, and to convey any findings as visually and dynamically as possible. With data structured intentionally, we can build customized and personalized analytical and predictive tools which can operate on the macro (course and college) level, or the micro (student) level. The tools developed as part of this project are already growing in demand. The Quality Enhancement Plan (QEP), which impacts the University's accreditation and helps measure its success in its mission statement, announced in April of 2023 that academic advising at Georgia Tech is one of the three finalists for the QEP topic consideration. The emphasis of the topic is on the lack of refinement and usefulness of the current advising environment. This project's intent is already being validated.

This project's use of dynamic data personalizes the student enrollment process; predicated on past data but generated so that it never becomes static. It moves the advising and predicting methodologies from intuitive and anecdotal to quantitative and analytical. Our graph approach to identify patterns in course enrollment through analysis of a student progress through coursework will enable a predictive model for future course enrollment based on past class pairings to anticipate course demand, positively impacting course offerings, visualize the data such that it is interpretable to a broad audience, and create a personalized advising experience for students.

For individual students and their advisors, the work done in this project can lead to advances in course recommendations that are quantitative instead of anecdotal, and through starting the work to classify students by finding commonalities among their peers, we work to build classifiers to ensure students are retained (intervention models) and graduate on time (course optimization). By working to improve student life, we acknowledge the interconnectedness of how that leads to bettering the university.

Proposed Method: Approaches

Data was pulled from the staging tables of Banner (an Enterprise Resource Planner prevalent in higher education). The query was written in SQL with the limitations on course enrollment history filtered to only include undergraduate courses and students, to include only courses with a lecture component, include only students assigned to the Main Campus, and only for terms since Summer of 2018. The export fields were 1) Student unique identifier, 2) the term the student took a course (example "201808" for Fall 2018), 3) the course subject (example "CS" for Computer Science), and 4) the course number (example "1301"). The extracted data fields were limited due to the nature of this project, which involves sharing the data for confirmation of the project outcomes. Inclusion of additional datapoints would require special FERPA (Family Educational Rights & Privacy Act) clearance. The resulting table was extracted as a CSV file, and the unique student identifiers were randomized to comply with the existing FERPA access granted.

To read, parse, clean, and store the data, base Python (3.10.7) was used. For data prep and calculations, no libraries were imported. This was done to help aid in future applications and adoption of any tool from this project. Multiple variations of Python nested dictionaries were created to mimic an API. This approach was taken because the current quantity and type of data is limited due to FERPA but will be able to be built on if Georgia Tech chooses to adopt any of these tools.

Once compiled, node relations representing course interconnectivity were built. Previous work in this field only looked at a single course, or a specific sequence of courses (Calculus I to Calculus II to Calculus III). To build an agnostic graph for relations among all classes, pairs were made for every course a student took. **Appendix II** provides a simple visualization of this process. We did not want to only look at course progression for a given subject or school (example CS1301 -> CS1331). Additionally, we did not want the chain to break if a student did not enroll in a particular semester. As most Georgia Tech undergraduates will at one point either: 1) not enroll in summer courses or 2) participate in an internship and coop (and therefore not be registered for lectures in that term), the localized term order was used. If, for example, a student was enrolled in Fall 2021 (term 0) did not enroll in Spring 2022 (term 1) but did enroll in Summer 2022 (term 2), their courses for Fall 2021 were linked to Summer 2022. This ensures all course enrollment is captured, and accounts for other random anomalies.

Concurrently, an enrollment count for each course and its representation as a node was calculated. When counted as an origin node, the last term in the dataset was not considered, and as a terminal node, the first term in the dataset was not considered. This was done to ensure that only the terms for which a course can participate in a directed graph is considered in strength calculations. Further, although each student is represented multiple times, our method ensured that it is only the relative strength of the pair, meaning each student can only be counted an equal number of times to the count of how many times they took the class as in the eligible direction of their course flow. These results were stored in dictionaries of weights, representing the conditional probability of a student enrolling in a particular course, conditional on them taking the paired course their previous term.

To use these weights to predict enrollment, initially a class (as node) was inputted, the node dictionary was used and the enrollment of the most recent semester for all corresponding courses was taken and multiplied. This was unsuccessful, and often resulted in overestimation. A potential student pool was then used to prevent students from being calculated multiple times. While this decreased the estimation, it did not fix it. Therefore, a student-by-student basis was implemented, and the conditional probability of any course for Student A was captured as a list value. Additionally, the Level of the course (i.e. 1XXX for intro, 4XXXX for upper level) was considered. This was done due to the decreasingly homogenous enrollment patterns of students as they progress through college.

A damping factor was implemented which was higher for lower-level courses. This damping factor first acted as a filter, wherein if a pair weight was under the damping factor, it was not considered. If, for example, *Course J* has a 0.05 weight to *Course K*, if the damping factor was 0.06, *Course J* would no longer be considered as a feeder course for *Course K*. Once the list of courses and potential students were assembled, each student's weights were summed, the corresponding value was put into the logistic distribution to ensure that its value could not exceed one, and the damping factor was subtracted. This was summed across all students to give a predicted enrollment of a given course the next time it was offered. **Appendix III** provides a simple illustrative example of this process.

To predict / aid individual student enrollment, a function was written that looks at the target student's most recent course enrollments and converts it to a set. Concurrently, it builds a historic bank of courses which the student already took. It takes the set of most recent courses, and iterates through other students, finding all who have taken the same set of courses in a term other than their most recent. If another student's set matches, it looks at the courses they took in their subsequent semester, if the target student has not already taken that course, it is added to a bank. Once all students are compared,

a strength index, wherein the number of enrollments in each course is divided by the matched students determined the strength of a given course's recommendation. **Appendix IV** visualizes this process.

The final user interface chosen was a Python (Jupyter) notebook. While it sacrifices some style and interactivity, the benefits of the format better align with the real-world applications of the tools created. The included data is comprised of all colleges. However, if this is used by Georgia Tech, when sent to colleges / schools, the data would have to be truncated, as university wide data is rarely shared openly to all university units. Additionally, a notebook allows for various calculations with the data, while allowing localized personalization. Finally, using a Jupyter notebook ensures that the data never need be made public, as Georgia Tech would not authorize the use of open-source tools with a web connection. Finally, Jupyter notebooks and Python code are powerful enough to handle all current tasks while also being simple enough to train and troubleshoot with units once they have their own instance of that data and determine their particular use cases. As the current version functions on multiple levels, the notebook format allows for systematic walkthrough of the calculations represented visually.

Experiments / Evaluation

As our data was a long form CSV file, the first issue was ensuring it was correctly parsed and structured. Initial debugging consisted of artificial data with predetermined answers. Once this passed, truncated data was used to verify the parsing was preformed correctly. The initial data is temporal, a result of a one-time pull from a database query. However, our approach was intended to accommodate data if / when it is implemented. Thus, artificial corner cases were used to test functions could handle all anticipated cases. The initial pull resulted in 524,000 records, of which 24,000 were from the most recent semester, Spring 2023, isolated and used as a test group.

Once parsed, the same process of simulation, to truncation was used for calculating functions. However, additional whiteboarding was used to ensure graph pairings were calculating correctly. Cumulatively, the two pairing lists produced 2.7 million tuples. Corner cases were again used, and speed tests were performed. While averaging 10 seconds, ultimately it was decided that this process should not be run every time, so a Pythonic class structure was implemented to ensure data was run once, and all its generated output stored for future use.

To check predictive capabilities, enrollment data from Spring 2023 (202302) was used to test our model's expected enrollment for a sample of courses and compare that with the existing prediction of enrollment being equal to previous term. The initial findings consistently resulted in over estimation, so weights / dampers were used. These were eventually found to need stratification, as upper-level courses were consistently outperforming lower-level courses in a point (exact count) prediction. While this made the model perform well in predicting course enrollment increasing or decreasing, lower-level courses continued to struggle with point estimation. Upon further testing, the cause for this is twofold: 1) Summer terms are inherently different, 2) lower-level courses require additional student information.

Student course recommendation pairing followed this approach, implementing the simulated, truncated, and corner case sequencing. With such a large dataset, every tested student found a match. The full process is detailed above and illustrated in **Appendix IV.** Each tested student experiment was developed to provide some output. Working with trial users, they wanted to know the strength of the recommendation, so that information was added, with a final example in **Appendix V**.

The deliverable was more straightforward. The restrictive parameters of this project, and the hope of having this work expand soon, meant any tools with an internet connection which were not authorized by the Office of Information Technology (OIT) could not be used. This made Jupyter notebooks an attractive option, and during implementation it was learned that Georgia Tech is planning to migrate towards Cloud storage, meaning using tools like this could help serve as a training method for the future. To aid in the user experience, all code was offshored outside the workbook. Verbiage and color schemes were deliberately chosen to comply with Institute Communications. While there is some user input required, future expansions could lead to simpler interactions, and with this in mind, the prospect of modification in data entered or output requested was built into the existing product's current iteration. While the result is not optimized for interaction, it serves as a necessary transition to more complex tools beyond Excel, while empowering users and accommodating various levels of competency. These assumptions were validated via user trials with existing staff in various colleges / units.

The efficacy and success of these intentions in visuals and tools. Having confirmed that the data was correctly parsed, the calculations were done correctly, and the output was as expected structurally, the final question of the project was, is it effective? To answer this, Georgia Tech trial users were leveraged, deliberately choosing technical non-technical audiences. Most of the helpful feedback came from the non-technical, as the need to convey more complex information succinctly and with as little words as possible, was constantly emphasized. This resulted in an iterative process to help determine what information was essential. **Appendix VI** illustrates this best, where the user feedback resulted in a better story, conveyed through graphics but maximizing data-ink ratio. Additional comments emphasized the less is more approach, and help inform and constrain, all with the intent of optimizing the message.

Conclusion and Discussion

Course enrollment and outcomes are linked to every metric important in higher education: student success, university reputation, and finances. Through this project, we established methods and tools to freely and locally parse, clean, analyze, interpret, and visualize course enrollment trends. While working on this project, and what resonated throughout the literature, it became increasingly clear that without course outcomes (grades), any modeling attempts were severely handicapped. Future work in this project should include outcomes, as well as certain course related metrics, such as pre-enrollment incoming credits. Further, Summer term enrollment behaviors are clearly unique, meaning two models may be needed and ideal. This project's work lend itself to creating a classification system to lead to early intervention if needed, with the aim of creating a clustering model and visual to easily identify at risk students. Also, by visually representing how course enrollment is connected across the university, how students flow through courses is easy to interpret, and lends itself to course optimization.

A tool is only as good as it is used. This project confirmed that the information Georgia Tech already stores in its data warehouses, and the capabilities of free-to-use technology, allows for the creation of powerful tools to help inform and drive purposeful action, while conveying complex information in simple visual mediums. By better understanding how students are, rather than the anecdotal or hypothetical, a university can and should work to make each student's life quantifiably better. This project is a start, and we sincerely hope it aids in reaching that goal.

Work Distribution

All team members contributed a similar / equal amount of effort.

References

Dietz-Uhler, B. and Hurn, J.E., (2013). Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective. *Journal of Interactive Online Learning*, *V12 n1* (pp. 17-26). http://www.ncolr.org/jiol/issues/pdf/12.1.2.pdf

Fraysier, K. Reschly, A. and Appleton, J. (2020). Predicting Postsecondary Enrollment with Secondary Student Engagement Data. *Journal of Psychoeducational Assessment V38 n7* (pp. 882-899).

Ishitani, T. DesJardins, S. (2002). A Longitudinal Investigation of Dropout From College in the United States. J College Student Retention, Vol.4(2), 2002-2003 (pp. 173-201).

Kelley, K. McAuley, J. Wallace, L. and Frank, S. (2008). Curricular Mapping: Process and Product. *American Journal of Pharmaceutical Education V72 n5 (Article 100).* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2630125/

Langston, R. Wyant, R. and Scheid, J. (2016). Strategic Enrollment Management for Chief Enrollment Officers: Practical Use of Statistical and Mathematical Data in Forecasting First Year and Transfer College Enrollment. *Strategic Enrollment Management Quarterly V4 n2* (pp.74-89). https://eric.ed.gov/?q=predict+enrollment&pr=on&id=EJ1109369

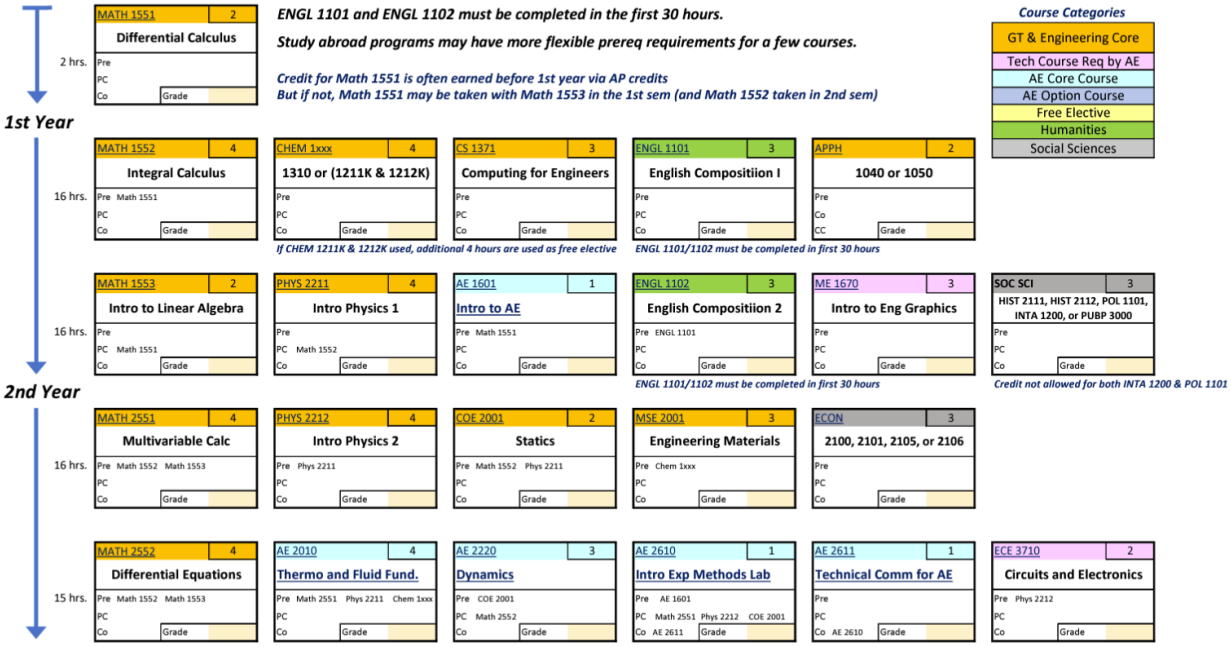
Luna, A. and Brennan, K. (2009). Using Regression Analysis in Departmental Budget Allocations. *IR Applications, Association for Institutional Research V24* (pp.1-14). https://eric.ed.gov/?id=ED508940

Shao, L. Leong, M. Levine, R. Stronach, J. and Fan, J. (2022). Machine Learning Methods for Course Enrollment Predictions*. Strategic Enrollment Management Quarterly, V10 n2* (pp. 11-29). https://asir.sdsu.edu/Documents/SMGDocs/SEMQ-1002-Shao.pdf

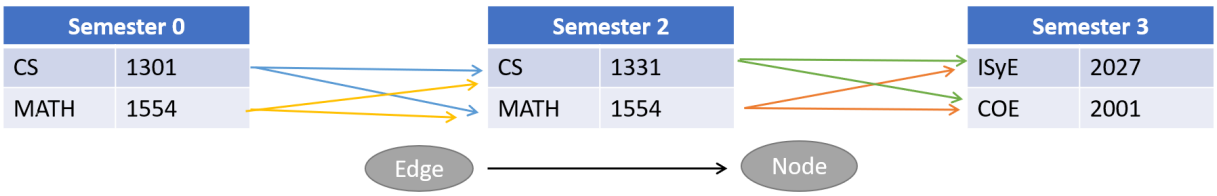
Slim, H. A. Yusuf, N. Abbas, C. T. Abdallah, G. L. Heileman and A. Slim. (2021). A Markov Decision Processes Modeling for Curricular Analytics. *20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, 2021* (pp. 415-421). https://ieeexplore.ieee.org/document/9680226

Appendix I: Example of Current Enrollment Tools / Aids (Worksheet)

Soure: https://d9.ae.gatech.edu/sites/default/files/file/2022/12/gt_ae_flow_chart_202109_0.pdf



Appendix II: Edge and Node Visualization

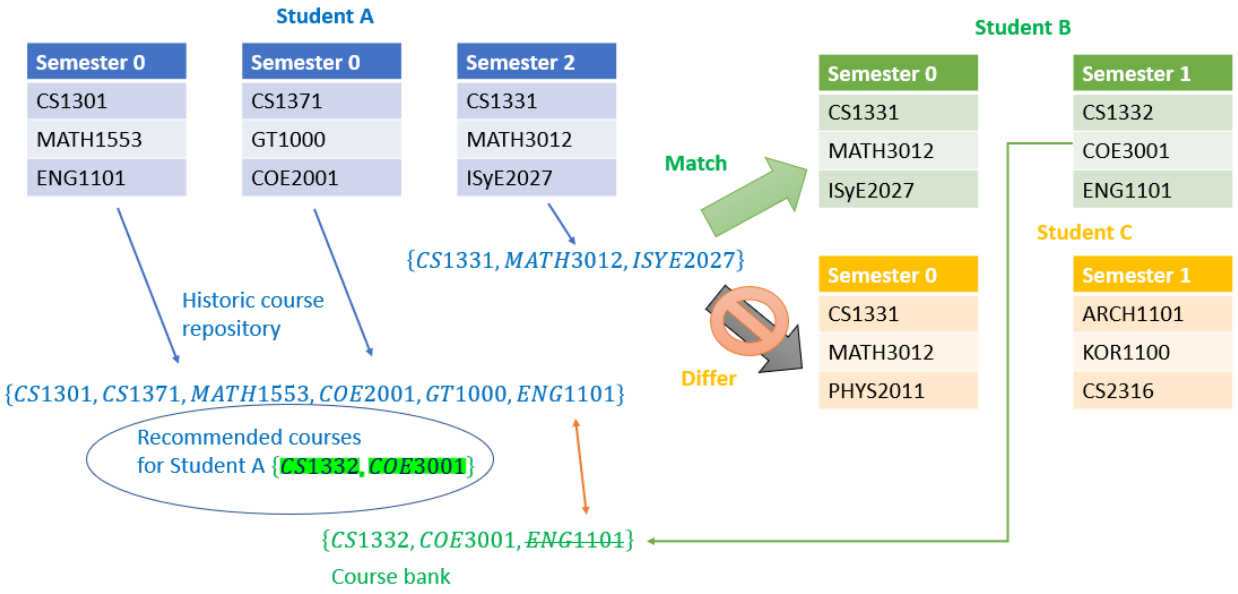


Appendix III: Estimated Likelihood a student enrolls in CS4641 Given past enrollment behavior

Course	Weight
CS1371	0.25
CSE4242	0.75
ISyE4032	0.32
MATH3012	0.27
CS4641	0.77

$$1 \times \frac{e^{\sum \mathcal{P}_{ij}}}{1 + e^{\sum \mathcal{P}_{ij}}} - weight = \mathcal{L}(student_i)$$

Appendix IV: Building a Student Course Matching System



Appendix V: Student Recommendation, Dynamic and Personalized (contrast with Appendix I)

	Spring 2022	Fall 2022	Spring 2023		Recommended Class	Recommendation Strength
0	MATH4317	CS4641	PSYC1101	-->	ENGL1102	13.56
1	MATH4107	MATH4541	-	-->	CS1331	8.96
2	CS2050	CS3600	-	-->	MATH1554	8.21
3	CS1332	CS3510	-	-->	PHYS2211	7.62
4	CS1100	CS2110	-	-->	CS1301	7.54
5	APPH1040	-	-	-->	MATH1553	7.13

Appendix VI: Course Graph