**UNIVERSIDAD AUTÓNOMA DE QUERÉTARO**

**FACULTAD DE INGENIERÍA**
**DOCTORADO EN INGENIERÍA**

# “Modelado lineal y algorítmico aplicado a la evaluación de procesos educativos en ingeniería.”

TESIS

QUE PARA OBTENER EL TÍTULO DE

**Doctor en Ingeniería**

PRESENTA

M. en I. Eric Leonardo Huerta Manzanilla

DIRIGIDO POR

**Dra. Rebeca del Rocío Peniche Vera**

SANTIAGO DE QUERÉTARO, QUERÉTARO, 2021.

# Universidad Autónoma de Querétaro

## Facultad de Ingeniería

Doctorado en
Ingeniería

**“Modelado lineal y algorítmico aplicado a la evaluación de procesos educativos en ingeniería.”**

TESIS

Que como parte de los requisitos para obtener el grado de
Doctor en Ingeniería

Presenta:

**Eric Leonardo Huerta Manzanilla**

Dirigido por:

**Dra. Rebeca del Rocío Peniche Vera**

SINODALES

| | | |
|---|---|---|
| Dra. Rebeca del Rocío Peniche Vera | | |
| Presidente | | Firma |
| Dr. Manuel Toledano Ayala | | |
| Secretario | | Firma |
| Dr. Juan Carlos Antonio Jáuregui Correa | | |
| Vocal | | Firma |
| Dr. Juvenal Rodríguez Reséndiz | | |
| Sinodal | | Firma |
| Dr. Avatar Flores Gutiérrez | | |
| Sinodal | | Firma |

Centro Universitario
Querétaro, QRO
México.
Junio 2021

*This work is dedicated to my family—Aurora, my strength, Paola, my pride and joy—also, to my friends that make the degree possible; their generosity was limitless.*

# Acknowledgments

This work was supported by the Consejo Nacional de Ciencia y Tecnología (Mexico's National Science and Technology Council) under Grant No. 32033, the National Science Foundation (NSF Award No. 1545667) and The Fulbright Commission of the US Department of State's Fulbright Visiting Scholar Program.

# Abstract

College retention has been studied for more than four decades, but it remains a concern for educational institutions, and it is still an important research subject. It is of particular interest in engineering for reasons including the potential workforce shortages, its impact on competitiveness, and socioeconomic equity. The analysis of college networks presents an interesting new perspective that may assist in discovering structural aspects of the interaction of students with the college systems that may, in turn, offer predictors for educational outcomes, like retention. Co-enrollment density is a novel metric estimated with enrollment records related to the probability of the graduation logit. Its algorithms and metanalytic models applied to retention are introduced in this work.

# Resumen

La eficiencia terminal en programas de licenciatura se ha estudiado por más de cuatro décadas, sin embargo continúa siendo un tema de investigación importante. El tema es de particular interés en ingeniería por razones que incluyen la posible falta de profesionales, su impacto en la competitividad y en la equidad social. El análisis de redes en los programas de licenciatura presenta una interesante y novedosa perspectiva que puede ayudar a descubrir aspectos estructurales de la interacción de estudiantes con los sistemas de la universidad, que pueden ser predictores de resultados educativos, como la eficiencia terminal. La densidad de co-matriculación es un nuevo índice estimado con registros académicos que está relacionado con la probabilidad logística de titulación. En este trabajo se introducen los algoritmos y modelos meta-analíticos aplicados al estudio de densidad de co-matriculación y su relación con la eficiencia terminal.

# Contents

# List of Figures

# List of Tables

CHAPTER 1

# Introduction

Persistence is a concern for engineering colleges, first, due to its impact on the efficient and effective use of university funding, second, due its consequences on potential workforce shortages and the impacts on economic development (Becker, 2010; Belser, Shillingford, Daire, Prescod, & Dagley, 2018; M. H. Johnson, 2013; W. Johnson & Jones, 2006).

The available data published show that one out of two engineering students will never graduate; the inquiry on the subject also shows that the differences in retention rates between institutions and countries are not significant (Aljohani, 2016b; Braxton et al., 2013; Braxton, Milem, & Sullivan, 2000). Literature reviews on the subject report that the persistence rates in countries like Canada, the United States, Great Britain, and Australia, among others, are similar and are close to one of two students ever graduating (Tight, 2020). Research reports on the persistence of engineering students in Latinamerica are scarce; but, they also confirm similar retention and persistence rates (Carales, 2020; Lucena, Downey, Jesiek, & Elber, 2008).

The research on retention has a very long tradition; the first works were reported as early as the sixties and the seventies. The research on retention has more than four decades and has been carried mainly in the American Educational System (Aljohani, 2016b). Other countries that have devoted funds and talents to the problem are England, New Zealand, Australia, and some Asian countries (Hodges et al., 2013; Krause & Armitage, 2014; Willcoxson, Cotter, & Joy, 2011). These studies have approached the problem with

the paradigmatic longitudinal model, that implements analysis that follows the paths of student departure, particularly the work after Tinto (1975, 1988, 1993) and Tinto and Cullen (1973).

At the time of this work, it was not known if the approximation of sociometric indexes was possible using academic records only; therefore, that was one of the hypotheses explored by implementing an algorithm to estimate sociometric indexes. The social network index mutuality (Rao & Bandyopadhyay, 1987) was approximated. Thus, one of the main goals for this work was to demonstrate the potential of academic records as sources for relational data and its use to analyze persistence. Mutuality, also called reciprocity, measures the frequency in which two nodes in a network reciprocate choices; for example, students enrolling in the same course. It allows the evaluation of clustering and centrality, other important properties of social networks and their neighborhoods (Wang & McCready, 2013).

## 1.1 Problem definition

The problem approached in this work is how to identify the engineering students' risk of not graduate as early and as efficiently as possible, by designing and testing network indexes using academic records.

## 1.2 Motivation

The traditional longitudinal analysis for the inquiry of persistence and retention requires data that is not available as part of the normal operations of educational institutions. Therefore, their use requires the implementation of special projects. The relational data for network analysis also requires special studies, usually self-report data obtained with questionnaires. The works after Biancani and McFarland (2013); Grunspan, Wiggins, and Goodreau (2014); Israel (2020); Israel, Koester, and McKay (2020); Israel et al. (2020) and Tudor (2008) are examples of these projects. The resources allocated to implement the longitudinal analysis or the network analysis are considerable; therefore, they are, by their nature, special studies with special funding, which makes them not useful for the continuous monitoring of academic outcomes.

The access to the database MIDFIELD, maintained by Ohland and Long (2016) was instrumental, be-

cause it offered the opportunity to implement the algorithm with robust data that was also well-curated, saving time and money for data cleaning. Actually, without the access to this resource, the project would be unfeasible.

## 1.3 Main goal

The project's goal was to develop, implement, and test algorithms to derive relational indexes from academic records to estimate linear models to understand, explain, and predict educational outcomes, particularly those related to the risk of not graduating.

### 1.3.1 Specific goals

- To develop algorithms for the assessment of students' relational patterns.
- To fit linear models based on novel relational indexes to explain, analyze and predict educational outcomes.
- To build statistical models for the evaluation of persistence and retention in engineering programs.

## 1.4 Dissertation structure

The thesis is organized as follows:

- Chapter 2 presents a brief review of the literature.
- Chapter 3 explain the methods used and the data.
- Chapter 4 is about the results and their discussion.
- Chapter 5 includes the work's conclusions.

CHAPTER 2

# Literature Review

## 2.1 Early work on retention

Retention in college began to be an inquiry subject in the late sixties and early seventies. William Spady (1970) reviewed over 80 papers and proposed what he called an interdisciplinary review and synthesis on dropouts. He was greatly influenced by Durkheim's work on suicide to offer his conclusions on the subject. In his findings, he offered a preliminary model, that was later improved in a follow-up paper published one year after the literature review (Spady, 1971). Family background, academic potential, and normative congruence were the primary individual factors related to persistence, according with the findings of this author. Grade performance was the sole academic factor reported. Friendship support and social integration were two social factors that he found that may affect the dropout decision. The dynamic aspects reported were intellectual development, satisfaction, and commitment. Since this early work, we can see conclusions that show common trends in the literature related to success in college that are: Factors that may affect persistence could be individual characteristics, either psychological, demographic, financial, academic, and attitudinal; others are related to academic performance, like college GPA, and the last are social in nature, like friendship support, social integration, and commitment. These factors are stated to translate, dynamically, into intellectual development and satisfaction that may finally produce the decision to drop or stay. We argue that there should be patterns behind this complexity that may reveal the social dynamics for each student during the college passage. These social dynamics indexes were found to be related to persistence.

Tinto and Cullen (1973) made another influential review of literature on dropping out of college. After two years of Spady's work, these authors contributed further to understand the reasons for students not completing degrees. After Durkheim's work, Tinto proposed the same individual, academic and social factors found in Spady's. In this case, family background, individual attributes, and pre-college schooling were argued to produce the first level of commitment with graduation and with the institution. Academic integration was related to grade performance and intellectual development, and all influenced the commitment to graduation. Finally, social integration leads to institutional commitment based on peer-group interactions and faculty interactions. Tinto's work evolved over a couple of decades (see, Tinto, 1975, 1988, 1993, 1997). A later work arguing that the classroom is the center of the academic experience provides a strong foundation for our work, because it proposed that measuring aspects of the interaction at the classroom level, or at least the probability of such interactions, may reveal the construction of the social networks that students are involved with and that may lead to integration and eventually to persist or drop from college (Tinto, 1997). Terenzini and Pascarella (1980) summarized six studies validating Tinto's framework explaining students' departure from college. Three main findings deem interesting for our proposal: Background characteristics were not significant, social and academic integration were significant, and also their interactions with background characteristics, and the frequency of informal interactions with faculty members accounts for the third part of the variation, mainly those interactions were academic. There was found that academic integration compensates for social integration. Academic and social integration may compensate in between. A metric based on the frequency of informal interactions with faculty encourages an approach based on frequencies of interactions, but in our case, we studied peer-to-peer relations.

Bean (1980) proposed a model for student attrition in college. The work after Bean, as Spady did, attempted to provide empirical evidence for the theoretical relationships that the model included. Interestingly, John Bean challenged the premise of attrition models based on the suicide theory by Durkheim. Bean also pointed out a common issue with quantitative analysis of attrition/persistence analysis: their lack of theoretical foundations. The "analytical variables" (predictor variables) in Tinto and Spady's models do not allow path analysis or causality. Instead, this model departs from Durkheim's work and derives from the

theory of turnover in business. The model explores background variables, organizational determinants, intervening variables, and their impact on dropout instead of the previous categories for explanatory variables found in Spady and Tinto's theoretical models. The results were reported disaggregated by gender, another significant difference with other models on the subject. According to Bean, there is a common feature in students leaving college; either male or female, they lack social integration. Bean did not address social integration as a variable; but, he reported that male students leaving may be characterized as living with parents and not knowing the social and academic rules well. Female students who drop as not belonging to campus organizations, not satisfied with being students at the institution, feel that they are not treated fairly. The adjusted correlation coefficient for the model's total explanatory power was .21 for women and .12 for males. Spady's model reported .31 for males and .39 for females, unadjusted correlations in both cases.

## 2.2 Student's integration and retention

Student involvement and integration are known to improve multiple colleges' outcomes, particularly retention; however, the standard sociometric methods for assessing integration are challenging to implement and not always convenient. The method proposed may be an alternative because higher education institutions have detailed enrollment records already available.

Student involvement is critical for retention, according to Astin's theory of student involvement and Tinto's interactionalist model of student departure (Astin, 1999; Tinto & Cullen, 1973). These theories are similar in dynamics, as pointed by Pascarella (1980). Thus, evaluating student's involvement is instrumental for higher educational institutions. Tinto's model proposed that integration impacts attrition. It has related the attrition phenomenon and the process of leaving college with the pioneering work on suicide by Durkheim, a similarity initially explored by Spady (1970). The analogy is thus that the student who leaves college is understood to have committed "academic suicide"; the individual that does not fit within the college social or educational systems (Spady, 1970, 1971; Tinto, 1975). Consequently, the individual with weaker ties within the social system of the college is more prone to leave.

## 2.3 Classroom proximity as a proxy for student's involvement

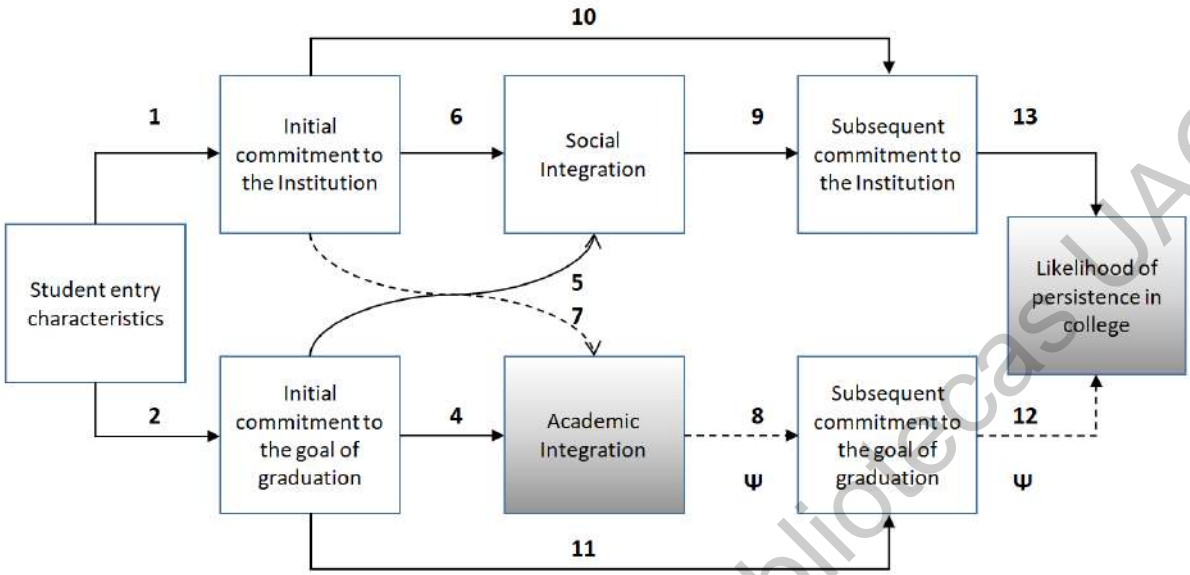
Tinto's Internationalist Theory describes thirteen propositions to explain the longitudinal process of student departure. The propositions are interrelated, as shown in Figure 2.1, it shows the propositions (arrows) linking the stages of student departure or staying (boxes) for explaining the decisions that lead the student to drop from, or stay in, a study program, according to Tinto and Cullen (1973) and Tinto (1988). At the institutional level, the college is the context in which, under the logic of Durkheim's work[1] individuals are more likely to drop out if they have weaker ties within the social and academic networks of the Institution. There are two main patterns for an individual to lack of integration into the college structure that may lead to dropout, insufficient interactions with others and insufficient identification of personal values with those of the college collectivity (Tinto, 1975). The two aspects of integration—the interaction between individuals and the identification of personal values with their collectivity—reinforce each other and are syntrophic. These processes are represented by propositions 5 and 7 that relate the stages of initial commitments with social and academic integration in Figure 2.1. Classroom proximity $(\psi)$ may be a proxy for propositions 8 and 12, subsequent commitment to the goal of graduation, and the likelihood of persistence in college, measured as graduation. Tinto's theory depicts pre-college students' entry characteristics to be instrumental in the college's departure process. Factors like family background, skills, abilities, and prior schooling. They influence the student's initial commitment to the Institution and their goals (graduating in engineering in this case). The Institution's structure defines the student's integration into the college's formal system. Therefore, a student's entry characteristics and the Institution's structure affect the student's fit and integration at the Institution (Tinto, 1988, 1993).

Tinto also stated that the classroom is at the center of the academic experience. It is the place and time where and when pivotal activities of the social and academic life happen (Tinto, 2006). The activities that the students have in the classroom are still a central part of the structure of the college. It is the only place and time for building relationships for those students that do not live on campus (Tinto, 2017). Thus, we propose that the sociometric index called classroom proximity as a proxy for academic integration at the classroom level, estimated using only academic records.

[1] Tinto's original work is partially based after Durkheim's theory of suicide.

Figure 2.1

*Tinto student's persistence model.*




*Note*. The arrows represent the thirteen propositions in Tinto's theory (Tinto, 1975), and the boxes are the stages leading to leave or stay in an educational program, according to him. $\psi$ is co-enrollment density (the index proposed in this work) and indicates that it may be related to Tinto's propositions 8 and 12.

### 2.3.1 Other challenges in engineering education

The interest for engineering and STEM careers has been declining worldwide (Becker, 2010; Belser et al., 2018; M. H. Johnson, 2013; W. Johnson & Jones, 2006; Sithole et al., 2017). The graduation rate in engineering is below the global rate in tertiary education of one in two (Aljohani, 2016a; Braxton et al., 2013). The literature on retention expands four decades in US (see Aljohani, 2016b, chap. 2), two decades in Europe, and a decade or more in other regions like Australia and Latin America. Also, there is an unbalanced participation of women in engineering, finally, the brightest students are choosing other careers than engineering, these phenomenons appear to be related with the public image of engineering as a field of work that has less prestige than Medical Doctors, Stock Market Agents and Lawyers, and it is culturally related with the image of "nerds", in developed countries (Borri & Maffioli, 2007). In contrast, under-developed countries shown better career prospects for engineers.

Other perspective on engineering education is that its curriculum is diverse, as it has been historically (Corlu et al., 2018). The roots for western engineering education, as known today, may be traced to European early technical schools, where continental Europe approached engineering as a public service, involving knowledge on advanced mathematics and science. the École Nationale des Ponts et Chaussés in France is an example of such schools. In contrast, Anglo-American engineers were trained on the job; England's early engineering schools represented such a model that evolved after World War I, when industries demanded from engineers higher levels of scientific knowledge (Corlu et al., 2018). However, even today, there is no standard set of skills and expertise to train engineers (Lucena et al., 2008; Passow & Passow, 2017). The diversity of the discipline's curriculum offers an additional layer of complexity to the inquiry on persistence and graduation in engineering, which is part of the emerging field of research in engineering education (Borrego & Bernhard, 2011), that is still ambiguous in its identity and status (Jesiek, Newswander, & Borrego, 2009). The study of the problems related to the education of engineers is a relatively recent field of inquiry (Borrego & Bernhard, 2011). Until the late nineties, the notion that engineering education had a research plan included low retention and persistence issues. There is no generally accepted set of terms for the study of students leaving college nor methods.

## 2.4 Definitions

The terms related to students leaving their studies are not standardized; therefore, the concepts applied in this work are defined in this section. Please refer to these definitions when a clarification for a term is required[2]

**Attrition** A reduction in a school's student population because of transfers or dropouts;

**Co-Enrollment** Students that voluntarily enroll in the same section of a course in the same term (CE);

**Dropout** The temporary or permanent voluntary withdrawal from an education or training program before completion. This term should not to be confused with academic dismissal;

**Dual-Enrollment** Enrollment of students in two schools at the same time;

**Cohort** Group members that share a common educational experience. In the context of the study, students that enroll in the same courses and terms due to compulsory institutional practices;

**Enrollment** The total number of individuals registered in a program accounts for a relationship between student and institution;

[2]These definitions were adapted from the Educational Resource Information Center's Thesaurus (ERIC, 2020), except as cited.

**Graduation** Receiving a diploma or degree for completing a phase of formal education. It is an institutional and an individual goal;

**Persistence** The continuance of a student's enrollment from the first to the second year—measured as the enrollment in one additional term after the first year. It is measured by the percentage of students who return to college for their second year (NCES, 2019). In this study, retention rate and persistence rate are the same;

**Retention** The ability of an educational institution to prevent student attrition and keep students enrolled until graduation. Its rate is measured as the percentage of students who return to the same institution (NCES, 2019). In this study, retention rate and persistence rate are the same;

**Transfer** Students who have transferred or intend to transfer from one higher education institution or program to another achieve more advanced or different educational goals (College transfer students).

CHAPTER 3

# Materials and methods

The methodology is described in Figure 3.1. The algorithm includes matrix operations to derive a classic social network analysis' adjacency matrix from enrollment records. Their total events for dyads per student are used to compute co-enrollment, and then, the density of co-enrollment. Finally, logistic regression models and their area under the receiver operating curve were computed to test the hypotheses.

## 3.1 The research questions

This work explores some items in the research agenda proposed by Tinto; first, it is an operationalization of a sociometric-index built with academic records that allows a comparative analysis between institutions using longitudinal rather than cross-sectional data. Second, co-enrollment density allows the construction of logistic regression models; that may lead to predictive analysis. Third, the analysis of race/ethnicity and gender are explored, along with the index in a disaggregated way (Tinto, 1975).

The main research question posed was whether, if it is possible to assess the relationship of academic integration to retention by estimating indices like co-enrollment density, using only academic records. This work has the following research questions:

- Could relational data be estimated with academic records?
- Are relational indexes estimated with academic records related to retention?
- Is co-enrollment density a predictor for graduation?

## 3.2 The data used in this study: Enrollment records

The Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) was the data that allows the implementation of this work (Ohland & Long, 2016). MIDFIELD holds information for more than two million undergraduates at twenty institutions from 1987 to 2012. We used the records for eight of the institutions in the sample that keep full enrollment data that include unique course's section. Those interested in replicate the results demonstrated here will need a database holding course records along with student records, both with unique identifiers. Records in MIDFIELD have information on students enrolling in particular sections of a class. Thus, the frequency of student pairs attending the same class could be calculated from the records. We demonstrated how enrollment data could be used to estimate social network indices. This work presents one of such indices, which was called co-enrollment density.

## 3.3 Institutions general profile

Institutions are the main unit of analysis and comparison in this study. The institutions provide a natural category due to the concurrency of institutional properties that may allow the emergence of common trends in the students that share the academic environment. The profile of the eight institutions that were included in the study may lead to interesting conclusions; therefore, we will use these properties as the theoretical background for the results section. The data set used for the analysis hold academic records for 702,532 students. Table 3.1 shows the students that graduated from the institutions. The total refers to the quantity of students whose records are hold in the database. The totals can also reveal the relative size of the institutions.

Table 3.1

*Student records per institution*

| Institution | No Grad. | Grad.(%) | Total |
|---|---|---|---|
| B | 29,422 | 14,137 (32.4) | 43,559 |
| C | 64,825 | 69,008 (51.5) | 133,833 |
| D | 20,920 | 28,670 (57.8) | 49,590 |
| E | 22,883 | 13,107 (36.4) | 35,990 |
| F | 38,029 | 45,594 (54.5) | 83,623 |
| H | 41,336 | 28,000 (40.3) | 69,336 |
| I | 61,900 | 76,628 (55.3) | 138,528 |
| J | 63,715 | 84,358 (56.9) | 148,073 |
| Total | 343,030 | 359,502(51.1) | 702,532 |

*Note*. The data covers a period of twenty years until 2013 at the eight engineering colleges included in the study.

Table 3.2

*Engineering student records per institution*

| Institution | No Grad. | Grad. | (%) | | Total | (%) |
|---|---|---|---|---|---|---|
| B | 1304 | 586 | (31.0) | | 1890 | (4.3) |
| C | 1529 | 617 | (28.7) | [-] | 2146 | (1.6) |
| D | 9486 | 11775 | (55.3) | | 21261 | (42.8) |
| E | 1348 | 1123 | (45.4) | [+] | 2471 | (6.8) |
| F | 7108 | 8227 | (53.6) | | 15335 | (18.3) |
| H | 1521 | 571 | (27.2) | [-] | 2092 | (3.0) |
| I | 3887 | 4636 | (54.4) | | 8523 | (6.1) |
| J | 4036 | 2832 | (41.2) | [-] | 6868 | (4.6) |
| Total | 30219 | 30367 | (50.1) | | 60586 | (8.6) |

Institution E is the smallest institution in the sample. It is a public, historically black college with enrollment over 10 thousand undergraduate students. Institution D also shows another interesting property for its engineering students that graduate in greater percentage than the overall graduation rate. The largest one, H, is another public research university. It holds over 30 thousand enrolled students[1].

Table 3.2 shows the engineering students in the database. Figures are broken down by graduation and the total shows also the percentage of engineering students to the total students in record for each institution. The column for students that graduated shows also the percent of graduation. It also shows three institutions were the graduation rate is lower than the global graduation rate showed in Table 1. Institution C is a STEM oriented public research university, more than forty percent of the students on record are engineering students. This particular property allows the comparison of an engineering oriented college against other universities with general orientation.

Ethnicity is also another interesting category to keep consideration of for the interpretation of the results and to discover theoretical constructs of the findings. Table 3.3 shows the engineering students break down by ethnicity. The codes are as follow. A= Asian, B=African American, H=Hispanic, I=Native American, N=International, W=White and X=Other/Unknown. Ethnicity is other category that may inform the interpretation of the final results. One group is integrated by institutions A and D against the other six institutions. Six of the eight institutions have predominantly white engineering students with 60 to 84 percent of the total students on record. This provides a contrast against colleges A and D where white students are 0.9 and 12.8 percent, respectively. African American students are the largest minority. A and D institutions may provide a good reference for results that may include the effect where white students are a minority. Asian Americans are the second larger minority group in the sample. Institution C provides a contrast being the only institution where Ethnicity A is larger than B, and all the other minority groups. One limitation

[1]The curriculum of the engineering programs included in the sample are mostly leveled in their academic workload. This means that the chances of students co-enrolling is similar along the program of study.

of the sample is that all the colleges included are public; therefore, there is no evidence if the findings are applicable to private institutions. It remains for future analysis that may be derive from our proposal.

Table 3.3

*Ethnicity of engineering students in the sample.*

| Institution | A | % | B | % | H | % | I | % | N | % | W | % | X | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 8 | 0.4 | 1803 | 95.4 | 11 | 0.6 | 0 | 0 | 51 | 2.7 | 17 | 0.9 | 0 | 0 |
| B | 84 | 3.9 | 467 | 21.8 | 221 | 10.3 | 9 | 0.4 | 72 | 3.4 | 1288 | 60 | 5 | 0.2 |
| C | 2396 | 11.3 | 1456 | 6.8 | 683 | 3.2 | 38 | 0.2 | 765 | 3.6 | 15792 | 74.3 | 131 | 0.6 |
| D | 18 | 0.7 | 1995 | 80.7 | 10 | 0.4 | 12 | 0.5 | 113 | 4.6 | 316 | 12.8 | 7 | 0.3 |
| E | 712 | 4.6 | 1416 | 9.2 | 213 | 1.4 | 111 | 0.7 | 249 | 1.6 | 12634 | 82.4 | 0 | 0 |
| F | 59 | 2.8 | 114 | 5.4 | 24 | 1.1 | 9 | 0.4 | 92 | 4.4 | 1758 | 84 | 36 | 1.7 |
| G | 643 | 7.5 | 128 | 1.5 | 531 | 6.2 | 67 | 0.8 | 154 | 1.8 | 6695 | 78.6 | 305 | 3.6 |
| H | 607 | 8.8 | 499 | 7.3 | 846 | 12.3 | 27 | 0.4 | 94 | 1.4 | 4759 | 69.3 | 36 | 0.5 |
| Total | 4527 | 7.5 | 7878 | 13 | 2539 | 4.2 | 273 | 0.5 | 1590 | 2.6 | 43259 | 71.4 | 520 | 0.9 |

## 3.4 Social network theory and its application to measuring integration

The focus of this study is on operationalizing academic integration. We believe that co-enrollment density has the potential to measure students' integration within formal educational environments. Academic integration is the degree of congruence between the student's academic behavior and the practices and norms of the university's system. Integration affects the student's commitment to the Institution, and therefore, the student's commitment to persist and graduate. The student's commitment to the Institution and graduation are, in Tinto's model, directly proportional to the academic and social integration of the student, and these factors are presumed to be linked to persistence and graduation (Tinto, 1975).

The intention to co-enroll with particular others in classes, and the frequency with which these co-enrollments occur, may be related to higher levels of integration and, therefore, with higher retention levels. This result suggests that academic integration might be operationalized using the social network concept of mutuality or reciprocity, which is essential to the goal and intention to graduate, as predicted by the theories of Tinto, Astin, and Pascarella Astin (1999); Chapin (2019); Pascarella (1980); Tinto (1993). Co-enrollment density is based on the concept of reciprocity. It is an index that assesses the tendency for individuals in a group to reciprocate choices more frequently than would occur by chance. Reciprocity is one of many structural characteristics of a social network, that reflects the cohesiveness of a group. It is an indicator of social integration; in our case, we believe it is more related to academic integration than

to social integration. Co-enrollment density only measures the frequency of encounters in the classroom, and do not imply actual social interaction; but, it may reflect academic affinity and potential academic contacts. The records used in this study reflect only the frequency of mutual encounters of students with particular others. The proximity index is aggregated at the institution level. Institutions are the context where the students decide to stay or to leave, and they provide an environment where students interact with others, and with the structure. Institutions are the context where a student's level of integration can lead to retention or attrition.

## 3.5 Measuring interactions between peers

The analysis of social networks uses graph theory or statistics. Both methods translate the theoretical statements on the structure of the network and their relations into sets of graphs or statistical models, respectively. The statistical analysis approach for communication networks assessment uses sociometric algorithms and requires social-relational matrices. The data to build such matrices come from questionnaires or ethnographic methods. More recently, the internet is a source for data on social relations. This work extends those previous efforts to consider academic records, specifically class enrollment data, as inputs to estimate a relational index.

Sociometric questionnaires regularly address the intention to meet with particular others in a network. In the case of students, the classmates are the potential pairs that may be the subject of such intention. In this context, the data provide a probability of meetings between members of a social network. However, this probability is based on self-reported data only, with no way to verify if the frequency of meetings occurs. Wasserman and Faust (1994) explained that the statistical analysis approach to social network analysis tests the stochastic assumptions about relational data contained in the social network dataset, and this analysis can be local or global, the first at the graph level, and the second at the whole network level . This work uses the approach based on statistical analysis, evaluating proximity with a probabilistic algorithm per Institution, where institutions are independent networks and the final level of aggregation.

The dyad, or pair, is the fundamental structural element of a social network. The term dyad is two individuals who can interact because they are part of a group. Reciprocity is an index that describes the relationship of dyadic proximity, also known as reciprocity or mutuality. An adaptation of the algorithm proposed originally by Katz and Powell (see, O'Malley & Onnela, 2019), and the idea of standardization principle proposed by Rao and Bandyopadhyay (1987), were applied to student records of class enrollment; instead of standard social network data. This strategy aims to establish the feasibility of social network analysis using large existing databases rather than relying on data that is much more difficult to gather and less reliable. Equation 3.1 presents the estimation for reciprocity proposed by Katz

and Powell, for the expected value of mutual selections between two actors in a network.

$$t_\alpha = \frac{2(N-1)m - Nd^2}{Nd(N-1-d)} \tag{3.1}$$

In Equation 3.1, N is the number of individuals in the group, d is the number of choices the N individuals have expressed, and m is the frequency of reciprocity. The variable m is typically obtained with surveys or through ethnographic methods; however, for this study, m was obtained counting instances of pairs of the enrollment records. Equation 3.2 shows the expected value of reciprocal choices under these assumptions.

$$E(m) = \frac{Nd^2}{2(N-1)} \tag{3.2}$$

The expected value of mutual, reciprocal choices in a network is also the expected number of choices that may occur only by chance, i.e., random reciprocity. Katz and Powell's equation discounts the random reciprocity from the estimation of voluntary reciprocity.

At the time of the study, there were no methods to derive sociometric information from academic records. Thus, this work begins from the premise that because academic records can describe the frequencies with which students are co-enrolled in classes, those frequencies may be used to approximate the expected value for reciprocity. Therefore, classroom proximity may be understood as a proxy for the reciprocity of students with their classmates. It is a simple, dyadic sociometric index that measures the occurrence of reciprocal non-random meetings between a particular pair of individuals (Kadushin, 2012). None of the Institutions included in the data set had any large-scale cohorts of students that would create non-random reciprocity based on institutional policy; therefore, that should not produce bias to our results.

The algorithm that was implemented in this study is presented in Equation 3.3 It was partially based on Katz and Powell's algorithm. The variable is the count of dyads of the cohort J, in all the group-class items of the same Institution for a student i. Then, for each individual out of the N-1 persons in the J cohort. The cohort size in our model is dynamic and depends on the different courses the student enrolled in. The cohort size in the original reciprocity model should be fixed. Our metric is a more appropriated one for the case of students at college whose potential classmates are changing each academic term, unless institutional cohorts were implemented. Even then, the algorithm may work, but confirming this should be left for future work that has access to data from institutions with cohort practices.

## 3.6 The co-enrollment density algorithm

We provide a formula and a chart to explain our proposal for this novel relational metric that can be calculated using academic standard records. co-enrollment density is based on the social network concept of reciprocity. Reciprocity is a measure of the stability of a social network. It also reflects interdependence between pairs in the network (Rao & Bandyopadhyay, 1987). They emphasize the need for the index to be standardized to make its magnitude comparable across different networks. Thus, in our work, we standardized the results and reported the aggregated results for each Institution.

## 3.7 The co-enrollment density formula

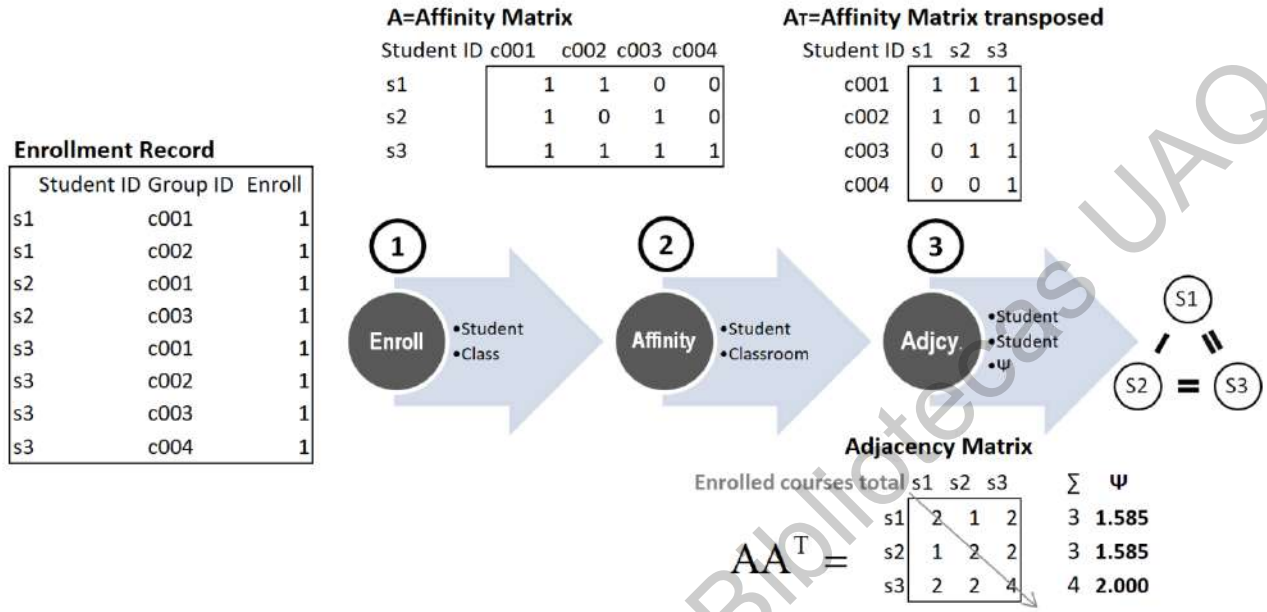
The analysis was based on the frequency of dyads. A dyad was defined as two students who were in the same class section. A dyad with the frequency of one was count when two students were in the same class section one time; a count of two out of the same dyad was reported if the same pair of students meet again in a different class section and so on. The count of mutual encounters was accumulated per dyad, and the log2 of the total was calculated to estimate the number of classmates a student joint with, during the terms enrolled in the program of study. Equation 3.3 defines co-enrollment density $\psi$, and Figure 3.1 explains it graphically. The term $AA^T$ is the adjacency matrix. It is the result of the affinity matrix times its transpose. It contains the frequency of dyads between students.

$$\psi = log_2 \sum (x_{ij}) - x_{i=J} \tag{3.3}$$

The student's co-enrollment density $\psi$ is the base two logarithm of the sum of $x_{ij}$ minus $x_{i=J}$ items, in the adjacency matrix $AA^T$. The $x_{ij}$ items are the times the student i co-enrolled with a classmate j. The terms $x_{i=J}$ are the total courses a student i was enrolled. Therefore, the sum equals the total co-enrollment events a student had in a program. The term $log_2$ is an approximation of how many classmates the student co-enrolled with during the program of study. Therefore $\psi$ is an approximation for social network's mutuality and reciprocity. However, it does not imply any social connection, only that two students decided to join the same course, and when that happens more frequently, it may be related with similarities in the students' progress in the curriculum and the pace of progress, that in turns, may be related with other educational outcomes, like persistence and the eventual graduation.

Figure 3.1

*Computing the relational matrix $AA^T$ from academic records.*




## 3.8 The co-enrollment algorithm chart

Figure 3.1 describes the co-enrollment algorithm, $\psi$. It shows an example that uses a tiny enrollment record, including three students and four courses only.

### 3.8.1 Enrollment record

The simplified enrollment record: The first step of the algorithm is to prepare an enrollment record with a single student I.D. column, a single section-course I.D. column, and the enrollment instance.

### 3.8.2 Affinity matrix

The second step is to derive the affinity matrix from the enrollment record by spreading the enrollment events for each student over the course sections. The affinity matrix will have as many columns as the total different individual sections for all the courses taken by a student's community in a time frame. The affinity matrix content is the enrollment record spread over all the courses reported in the data.

### 3.8.3 The adjacency matrix

The third step is to compute the adjacency matrix with the total dyads in the affinity matrix. It is the product of the affinity matrix times its inverse. The adjacency matrix is the square Matrix with the total co-enrollment events per student. The adjacency matrix diagonal is the total courses that each student has been enrolled in. The totals per row, or column, less the values in the adjacency matrix's diagonal, are the total co-enrollment events per student. The co-enrollment density is the log2 of the total co-enrollment events per student in the record. co-enrollment density, as calculated, represents an approximation to the total number of fellows that a particular student has joined with twice or more, based on the affinity matrix derived from the enrollment record.

## 3.9 Data analytics

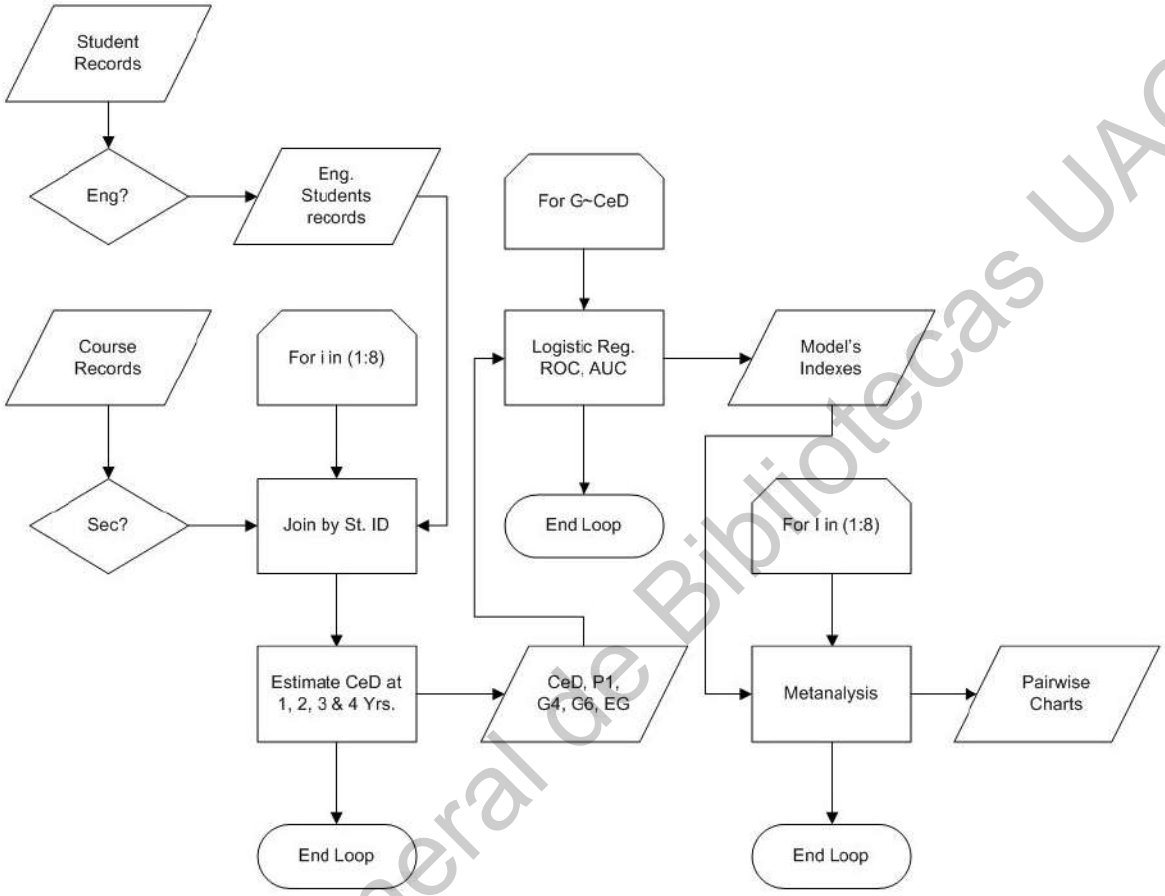
### 3.9.1 Preparing the data

The goal of data preparation is to end with clean enrollment records holding a single identifier for students associated with all courses for each student and an enrollment marker, one per course. In the Figure 3.1 such table is called the enrollment record. The original MIDFIELD data were kept in files in sas7bdat format. They were two tables, one for students and one for the courses[2]. Both tables have the same student identifier. The information is fully anonymized for the students and for the institutions. The first step was to filter the dataset for those institutions that have complete course's data to build a single course identifier up to the section of the course. In the American higher education system a course is identified by the period in which the course was offered, a course's code and a section. The section is a separated code for each session in which the course was offered. A particular course may be offered in different sections in the same term and with the same identifier, but each section represents a different group of students that meet in different schedules to the other sections. Originally, MIDFIELED data base had twelve institutions, but four of them do not hold section data in their courses' records; therefore, those institutions were not considered for the analysis. The single course identifier for each course was built using a term code, a course identifier and the sections offered. The course identifiers were kept associated with the corresponding student identification number to build the enrollment record per institution. The student records were filtered for those students that declare engineering at the time of enrollment at the institution and that confirm engineering when majored. Also, transfer students were not considered as the transfer altered the opportunity to co-enrolled with other students at the institution. The libraries sqldf and stats provide functions that were used for data preparation, including the filtering and joining of the original data into simplified enrollment records with single course I.D. and single student I.D. (Grothendieck, 2017). The

[2]The dictionaries of both databases are included in the apendix.

Figure 3.2

*Data processing flow diagram.*




student records in MIDFIELD hold information on the career's majors in the form of a six digits CIP codes. These refer to the taxonomy of educational programs defined by the U.S. Department of Education's National Center for Education Statistics (*CIP user site*, 2021).

### 3.9.2 Data flow

#### Estimating co-enrollment density

The following steps are shown in the two first columns in the flow chart of Figure 3.2.

**Process:** Estimating co-enrollment density

**Input:** Students and course records in MIDFIELD.

**Output:** Table with CeD at 1 to 4 years paired with binomial data for P1, G4, G6, and EG.

### Computing the logistic regression models and their tests

These steps are depicted in the third column in Figure 3.2. The logistic regression models and their corresponding tests were calculated from the tables produced in the previous steps. Logistic regression models' coefficients, receiver operating characteristic curves, area under the receiver operating curves, cut-off points for CeD, and a table with coefficients per model were computed. The files obtained as an output of this step are included in the article.

**Process:** Logistic regression and tests

**Input:** Table with co-enrollment density (at one to four years) paired with binomial data (1 or 0 for persistence and graduation at four to six years).

**Output:** The logit models, AUROC curves and coefficients in OR.

## 3.10 The adjacency matrix

The main part of the computing process is to get the adjacency matrix from the enrollment records. To do that, each course identifier is a column header of a new table called affinity matrix where each row is for the records of one student. The row will have 0 for the courses that the student did not take and 1 for the courses where the student was enrolled. The affinity matrix has as many columns as different courses were offered for the entire sample of students, therefore it could grow quickly as more student's records are included in the analysis. We prepare samples of one thousand students per institution to make the computation feasible with a standard laptop computer with a conventional CPU, 16 GB of RAM and 1 TB NVME disk.

The affinity matrix is the key to get relational data from the enrollment records. Functions available in the libraries Matrix and tidyr allow computing of the affinity and adjacency matrices from the previously formatted enrollment records Wickham et al. (2019); Wickham, Chang, et al. (2020); Wickham, code), and RStudio (2020). The Matrix library facilitated the processing of the large affinity matrix and the transpose and the cross matrix product to obtain the adjacency matrix from the affinity matrix. We used dense matrices to overcome the limitation of the R Language that requires all the data to be available in RAM, in order to be processed.

## 3.11 The regressions and the area under the receiver operating curve

Predicting graduation requires the identification of a function $f(s)$ to map the change in probability of Y's odds ratio from negative to positive $OR = p(Y)/(1 - p(Y))$, with a potential predictor ($\psi$ in this case). One solution is called the logit function, see Equation 3.4 (Dobson & Barnett, 2008).

$$p(Y) = \int f(S)ds = \frac{exp(\beta_0 + \beta_1\psi}{1 + exp(\beta_0 + \beta_1\psi)} \tag{3.4}$$

Where $f(S) \in X : \omega$ and integrates to the link function in Equation 3.5—the logistic regression model (see, Dobson & Barnett, 2008, p. 126). Co-enrollment density $(\psi)$ is the predictor for the logit of Y (graduation/persistence).

$$log[\frac{p(Y)}{1 - p(Y)}] = \beta_0 + \beta_1\psi + \epsilon \tag{3.5}$$

An example of the seventy two logistic regression models (see the full models' set in, E. L. Huerta-Manzanilla, Ohland, Toledano-Ayala, & Jáuregui-Correa, 2021, Supplementary data) obtained with the methodology is shown in the Figure 3.3 that presents the logit of $Y \approx \psi_4$ mapped to co-enrollment density at 4-years predicting graduation at 6-years (G6) for Institution D. Its coefficient is shown in Listing 3.1, see the coefficient value and its 95%C.I. at Line 3.

Listing 3.1

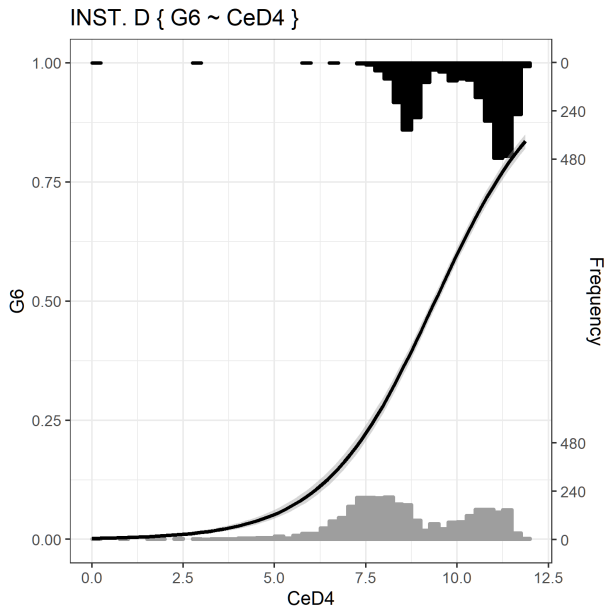
*Logistic regression model's summary for Institution D: G6 ≈ 4-year co-enrollment.*

```
1    Logistic regression predicting y
2  OR(95%CI)          P(Wald's test) P(LR-test)
3  x (cont. var.) 1.93 (1.86,2.01)   < 0.001
4  Log-likelihood = -3469.5124
5  No. of observations = 5999
6  AIC value = 6943.0248
```

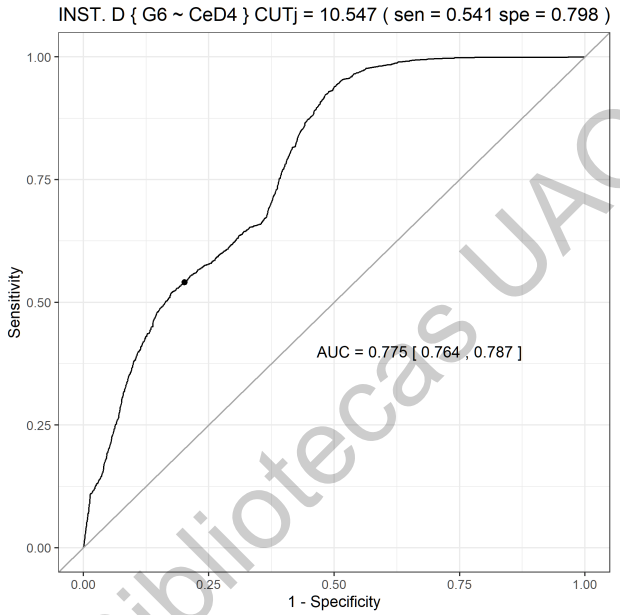
We prepare the logistic regression models and their corresponding receiver operating characteristic curves per institution. The {stats::glm} function included in the base R's library allows the computation of the logistic regression models (Team, 2020). We use the {epidisplay} library to prepare the logistic regression model's displays (Chongsuvivatwong, 2018). The library {proc} produces the receiver operating characteristic analysis and the estimation of the area under the ROC curve (Robin et al., 2011). We reported the best model based on AIC and the corresponding AUC values. Co-enrollment density was computed for college networks filtered for courses offered at the first, second, third and fourth years of enrollment. co-enrollment density1 was tested to predict P1, G4 and G6. co-enrollment density at

Figure 3.3

*Institution D: (a) Logit Model; (b) AUC Chart.*




(a) *Logistic regression plot.*




(b) *Receiver operating curve and its area.*

two years was fitted to G4 and G6. co-enrollment density at three years and co-enrollment density at four years were fitted to G6, and EG. Nine models were obtained per Institution for a total of seventy two logit models.

CHAPTER 4

# Results and discussion

## 4.1 Odds ratios for the logistic regression models

Table 4.1 shows the summary for the logistic regression models, in odds ratio ranges. The institutions are the rows, the second column are the labels for three rows per institutions, that are: OR – odds of graduation, AUC – area under the receiver operating curve and CX – the classroom proximity index for 50% odds of graduation. There are four columns for models that were estimated after the first, second, third and fourth year of enrollment of engineering students. The last column has information for the total engineering students in record, the percent of graduation and a sign indicating if the engineering college's graduation rate is less or greater than the institution's graduation rate.

Table 4.1

*95% C.I. for the odds ratio for the logistic regression models with co-enrollment computed for one to four years.*

| | Year 1 | | Year 2 | | Year 3 | | Year 4 | |
|---|---|---|---|---|---|---|---|---|
| Institutions | .025 | .975 | .025 | .975 | .025 | .975 | .025 | .975 |
| B | 1.17 | 1.39 | 2.22 | 3.02 | 3.83 | 5.63 | 5.53 | 8.57 |
| C | 1.01 | 1.13 | 1.22 | 1.42 | 1.67 | 2.03 | 2.77 | 3.67 |
| D | 1.37 | 1.44 | 1.61 | 1.72 | 1.86 | 2.00 | 2.18 | 2.36 |
| E | 0.84 | 0.92 | 1.04 | 1.16 | 1.24 | 1.40 | 1.40 | 1.58 |
| F | 1.32 | 1.39 | 1.45 | 1.54 | 1.57 | 1.67 | 1.71 | 1.83 |
| H | 0.67 | 0.79 | 1.00 | 1.18 | 1.44 | 1.75 | 2.20 | 2.79 |
| I | 0.72 | 0.80 | 1.27 | 1.45 | 2.30 | 2.64 | 3.56 | 4.15 |
| J | 1.09 | 1.21 | 2.04 | 2.33 | 2.49 | 2.82 | 2.90 | 3.30 |

## 4.2 Area under the receiver operating curves

Table 4.2 shows the models' results for the area under the receiver operating curves of the models. One more time, the rows are the institutions and the colums are the 95% confidence intervals for the area under the ROC curves.

Table 4.2

*95% C.I. for the area under the receiver operating curves (AUROC) for the logistic regression models estimated at one, two, three and four years.*

| | Year 1 | | Year 2 | | Year 3 | | Year 4 | |
|---|---|---|---|---|---|---|---|---|
| Institutions | .025 | .975 | .025 | .975 | .025 | .975 | .025 | .975 |
| B | 52.8 | 58.2 | 69.7 | 74.3 | 77.5 | 81.5 | 82.4 | 85.9 |
| C | 50.0 | 55.3 | 56.4 | 61.3 | 66.8 | 71.3 | 77.7 | 81.6 |
| D | 67.3 | 69.8 | 73.4 | 75.7 | 78.2 | 80.3 | 81.7 | 83.6 |
| E | 56.1 | 60.6 | 49.2 | 53.8 | 56.7 | 61.2 | 61.4 | 65.7 |
| F | 66.3 | 68.8 | 69.1 | 71.5 | 71.7 | 74.0 | 74.5 | 76.7 |
| H | 60.3 | 65.0 | 48.6 | 53.4 | 58.8 | 63.4 | 68.6 | 72.9 |
| I | 59.5 | 61.9 | 51.5 | 54.0 | 62.6 | 65.1 | 70.9 | 73.2 |
| J | 47.8 | 50.6 | 65.2 | 67.8 | 71.7 | 74.1 | 76.7 | 78.9 |

*Note*. The top row holds the number of years for which co-enrollment was calculated (1–4). For each year there are the two limits for a 95% C.I. [Limit at .025, Limit at .975]

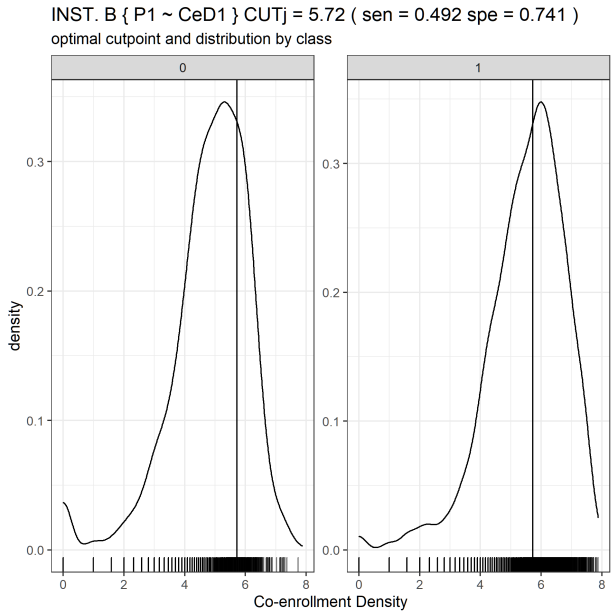
## 4.3 Discussion

Institution A shows that the Classroom proximity Logistic Regression (LR) model's prediction power is random after the first year (AUC is close to fifty percent). After the second year the Log-R's AUC increases above sixty percent and the OR shows that students that co-enrolled with more than 8 students are two to three times more likely to graduate than students co-enrolling with less class-mates. After the third year the students that co-enrolled with more than 8 classmates are three to five times more likely to graduate and the AUC is above seventy percent. The same trend can be observed in all the institutions, AUC increases each year and at the fourth year the predictive power of Log-R models is around 70%. This finding suggests that the proposed Log-R may be a useful tool to forecast graduation. Students that show classroom proximity under eight may be associated with low odds of graduation; thus, it may be a warning index to further investigate particular cases.

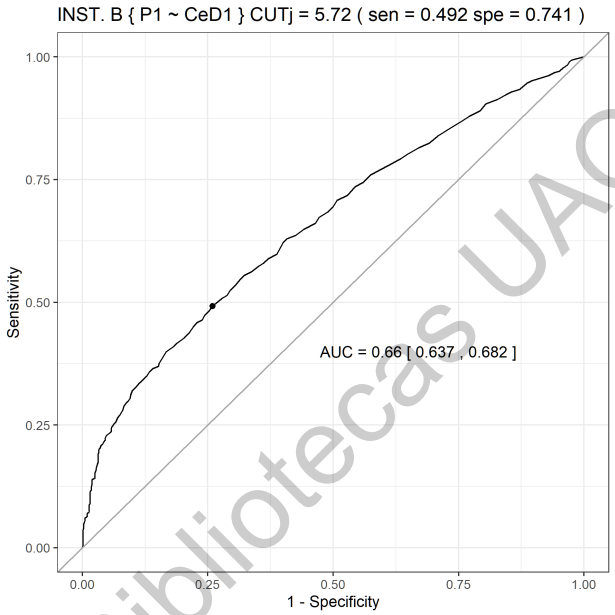
Each Institution, and each academic unit, may calculate its best cut-off points per period to have a balance between predictive power and opportunity to make interventions for mitigating low odds of graduation. Figure 4.1 shows the

Figure 4.1

*Institution B: (a) Cut-off chart; (b) AUC Chart.*




(a) *Cut-off point chart.*




(b) *AUROC chart.*

cut-off point estimation charts for Institution B. The curves shown in Figure 4.1a are density plots for the probability related with those students that may drop off their studies (0) and those that will persist (0). The vertical lines in both density plots shown the cut-off value with best predictive power $CUT_j = 5.72$ in this case. Figure 4.1b shows the area under the receiver operating curve, that is the plot for the probability estimated by the corresponding logistic regression model for $Sensitivity$ and $1 - Specificity$ and the position of the $CUT_j$. It shows the displays for ROC curves and AUC confidence intervals for Institution A. We further discuss the two cases that produced the best models of the sample, as well as the two that were the worst results. It is encouraging that even the worst models, those for M and C institutions, have better than guessing AUROC curves and that classroom proximity was significant in their logistic regression models. Also, the student's odds of graduation were 47% and 92% better per proximity unit. Figure 5 shows the logistic regression for Institutions A and F. The charts' x-axis shows classroom proximity. The y-axis is the scale for the probability of graduation. The logistic regression curve shows the confidence interval in light gray and the model prediction in a black line. There are histograms for students that graduated, at the top, and in light gray at the bottom for students that did not graduate. The figure also shows the ROC curves. The charts have sensitivity in the Y-axis and specificity in the X-axis. The ROC curve shows its 95% confidence intervals. The tables for the model's coefficients are shown along with the charts for convenience. Institutions A and F data sets produced the best results of the study. They have the highest AUCs and the highest odds ratios for classroom proximity as

a predictor for graduation. Students at Institution A show three times higher odds to graduate per unit of classroom proximity. In the case of institution F, students present twice the odds of graduation per each classroom proximity unit.

Institution F is one of the few where Ethnicity and Gender are significant, along with proximity. According to the LRM obtained, International students have 118% higher odds of graduation than Asian students, given that all other factors remain the same. Black students have a 33% reduction in their odds of graduation compared with Asians. Male students present 20% lower odds of graduation than their female counterparts. Considering that there are no specifics on the characteristics of the Institution, no further elaboration on these findings was intended. However, we learned from this model that classroom proximity may still have a share of the prediction power for graduation, even when other factors are also significant in an LRM model. Figure 6 shows the Institutions M and C models. It presents the logistic regression models at the top, the ROC curves, and the models' coefficients at the bottom. These are the models with the lowest AUCROCs. We found encouraging that even the models with the weaker prediction power were better than guessing. It remains the possibility to make a study with larger samples at these institutions, or even with the population data sets. We leave for future studies the elaboration on these findings with further details on the academic and social characteristics of these institutions.

## 4.4 Potential impact

The method proposed may help institutions to expand their efforts to identify students who could benefit from retention initatives by including co-enrollment density, or similar indexes. The logistic regression models that relate co-enrollment at the first year with persistence provides a parsimonious index and its cut-off threshold values may identify potential dropouts. It may be added to indicators known to affect persistence in the first year. The results may add a way to link the theory on persistence to the practice of engineering education using network analysis and empirical models based on course enrollment records. It is not known if co-enrollment density is related to institutional cohorting. Research has shown informal mentoring to be more effective than formal mentoring (Inzer & Crawford, 2005), which suggests that institutional cohorting (formal efforts to group students) may not have the same benefits or predictive value as found in this work.

The methodology and the co-enrollment density index add to the literature on retention. It is a system parameter that reveals patterns that the students tend to exhibit if they are more prone to stay in an educational program. Co-enrollment density may be the first of other college network metrics that may support policies in improving college outcomes, other researchers may use enrollment data to analyze college networks that may reveal patterns of student

activities related to academic outcomes. As Tinto (1997) suggested, network and data analysis may complement traditional longitudinal studies. Co-enrollment density appears to be a novel and parsimonious metric that may predict retention for students in 4-year engineering degrees.

## 4.5 Publications

This section refers the main published work that was derived from the ideas in this thesis. The first derived published work by E. Huerta-Manzanilla, Ohland, and Long (2013), was presented in the American Society for Engineering Education's conference in 2013. The following is an abstract for this paper derived after this dissertation:

The Impact of Social Integration on First Time in College Engineering Students Persistence, Longitudinal, Interinstitutional Database Analysis. Persistence of engineering students was 51.5% from 1987 to 2010, based on a large multi-institution dataset. Many approaches have been proposed to assess factors affecting persistence. The main models on persistence are Tinto's Theory of Student Departure, Astin's Theory of Involvement in Higher Education and Pascarella's General Model for Assessing Change. Tinto proposed that academic and social integration reinforce students' commitment totheir institution and educational goals. Sociometric techniques from Social Network Theory are being adapted to develop measures of social integration among undergraduate students using a large, multi-institution longitudinal dataset. This paper will introduce this approach and, in particular, discuss the social network parameter "mutuality" and study its relationship to persistence in engineering. Mutuality is an index that assesses the tendency for individuals in a group to reciprocate choices more frequently than would occur by chance. Subsets of students with the same major and starting years were sampled by institution. Unique institution-class-course identity codes were defined for section groups to establish which students took classes in each other's presence, and the mutuality index was evaluated for each student cohort in a section group. Mutuality reflects reciprocity beyond random grouping, due to students having free selection of groups. A matrix of section groups and cohorts was built as abridge data structure to assess mutuality. A simplified mutuality algorithm was evaluated per each cell in the matrix. A linear model for mutuality as predictor and persistence rate per cohort as the response was fit to subsets results. Two institutions with persistence rates of 73% and 44% were compared. Mutuality rate per group was 0.73 2.26% and .44 0.79% , respectively. Results suggest mutuality maybe related with persistence. Results for other institutions and subpopulations will be consideredin the final paper. Mutuality Distributions for Mechanical Engineers at two institutions. Institution A: Persistence rate 73%. Institution B: Persistence rate 44%.

This first work published was implemented using an stochastic approach. Probability density functions were calculated for an index that approximated mutuality with the enrollment records. There were apparent differences between students that graduate against those that did not graduate; however, there were no testing methods implemented to evaluate the significance of the difference nor to implement the index in institutional settings. The work was encouraging but it did not offer any further application.

The second work by E. L. Huerta-Manzanilla, Ohland, and Peniche-Vera (2021) implemented a more advanced algorithm with a more efficient matrix manipulation numerical methods. While the first algorithm read and compute the index in several hours, the algorithm included here runs in minutes.

College retention is a concern for educational institutions and researchers. This concern is particularly acute in engineering for reasons including workforce shortages, economic competitiveness, social justice, and socioeconomic equity. This study presents the evaluation of co-enrollment density (CeD) for engineering students at eight medium and large American public universities over 24 years. CeD is a novel metric estimated using enrollment records that may predict retention in 4-year bachelor of science programs in engineering. Graduation and persistence were fitted to CeD with logistic regression. Students in denser co-enrollment clusters—high CeD—tend to graduate more than their classmates in less dense neighborhoods—low CeD. The regression models predict graduation with odds ratio intervals 95 % CIs [3.24, 4.81] and area under the receiver operating curve [0.76, 0.80]. CeD is more sensitive to students who do not persist, particularly after the first year, so CeD's cut-off points may be indicators for dropouts' risk.

A third paper that is related with this work is by E. L. Huerta-Manzanilla, Ohland, Toledano-Ayala, and Jáuregui-Correa (2021). This article describes in detailed the processed data produced by the co-enrollment density algorithm that was previously published (see, E. Huerta-Manzanilla et al., 2013). The abstract for this work is:

This article describes the data related to co-enrollment density (CeD), a new network clustering index, that can predict persistence and graduation. The data hold the raw results and charts obtained with the algorithm for CeD introduced in "Co-Enrollment Density Predicts Engineering Students' Persistence and Graduation: College Networks and Logistic Regression Analysis." There are data for eight institutions that show CeD as a predictor for graduation at four years, graduation at six years, and ever graduated. The files were processed using R to estimate CeD at one, two, three, and four years. Logistic regression models, receiver operating characteristic curves, specificity, sensibility, and cut-off points were estimated for each model. The R code to reproduce the metanalysis for the summary data is included. The displays for the logistic regression models, receiver operating characteristic curves, density curves for classes, models, and parameters are included.

## 4.6 Future work

This work provides a framework to ask various research questions of value. As an extension of this work, does co-enrollment density have similar performance in other disciplines, and in engineering programs with different designs? Since this study used aggregate data, its results are most applicable to the majority group in U.S. engineering programs, which is White males. It would be valuable to ask how co-enrollment density and its usefulness vary based on an intersectional combination of race/ethnicity and sex. It would be useful to study whether curriculum frameworks such as CDIO or PBL alter co-enrollment density and its predictive value. As suggested above, does institutional cohorting produce the same effect as spontaneous co-enrollment? How does co-enrollment density perform in private colleges? Investigating the outliers and exceptional cases of institutions with atypical retention levels, such as Institution H, has the potential to address the most problematic cases with respect to retention. Exploring different methodologies to estimate cut-off points is also an interesting research direction, particularly for its practical use to improve college retention.

## 4.7 Meta-analysis

This section elaborates on the potential for generalization of the results obtained in this work. It provides a prospective view for its use and application in preventing students to leave their engineering programs for academic administrators.

### 4.7.1 Co-enrollment density as a predictor for graduation

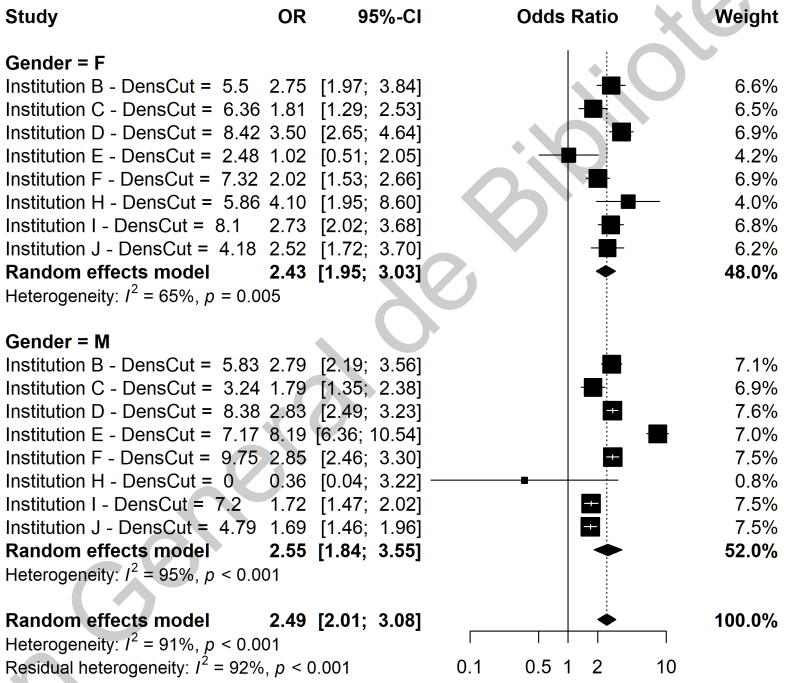
The results suggest that co-enrollment density may be a parsimonious predictor for first-year persistence (at the first year of studies) and graduation at 4-year engineering programs (after four or six years), that may be easier to implement and operate than multivariate models that explains retention with many contributing factors.

The summary in Figure 4.2 shows the result for the metanalysis of the studies per institution. The first column shows the Institution's references and the estimated cut-off values for $\psi_1$ to optimize the identification of potential persisters against those that may be in risk of not persist after the first year. The second column are the odds-ratios, the third column are the 95% CI for the odds-ratios. The forest chart follows and it has one as the reference for no-significance, because the effect should be greater than one to have predictive power. The last column at the right is the weight for each study to estimate the general random effect for the model. The chart is grouped by gender; at the top are the results for the models for females and at the bottom the results for male students. This grouping was applied only to estimate the effects of the logit models, but the computation for $\psi$ included male and females as potential peers

for the undirected graphs count. Institution E's logit model has not predictive power for female students. A similar phenomenon happened for male students at Institution H. This is a notable difference and accounts for a different networking pattern of males compared with females in the first-year of studies. Institution E had 86.4% of African American students at the time the database was accessed, therefore these two anomalies may be explained by biased co-enrollment patterns for these two communities.

Figure 4.2

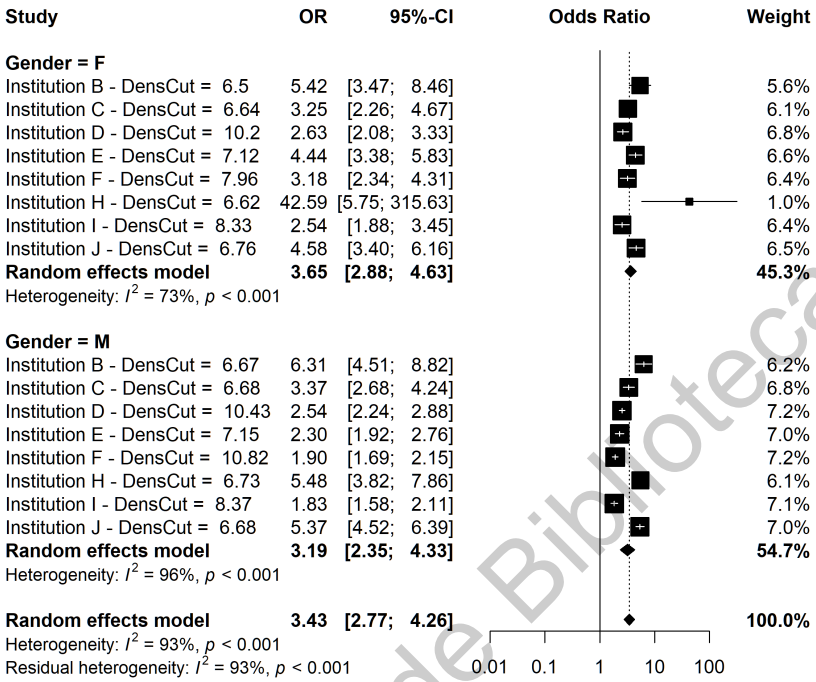
*Meta-analysis summary for co-enrollment density estimated at the first year of studies as predictor for retention.*




A similar forest chart depicts the meta-analysis for the odds ratios for the logistic regression models for graduation at 4-years predicted by co-enrollment density estimated at the second year of studies. The Figure 4.3 shows that meta-analysis and this time the predictive power is 95%[2.77; 4.26], overall. The most important finding is that after the second year co-enrollment was a potential predictor for persistence and graduation at all the institutions. However, Institution H shows a biased result that worth further investigation, that may be a subject of a future inquiry.

Figure 4.3

*Meta-analysis summary for co-enrollment density estimated at the second year of studies as predictor for graduation at four years.*




### 4.7.2 Is there another database to contrast the results?

The MIDFIELD database has been shown to be representative of a more comprehensive American database of engineering programs; however, it is impossible to determine if MIDFIELD is representative in this study, because we did not find other database capable of studying this phenomenon, to have it as a control unit. Indeed, the United States Department of Education concluded that it was infeasible to create such a database (Cunningham & Milam, 2005). Further studies with similar academic records may enlighten the question on the repeatibility of the results reported in this study.

CHAPTER 5

# Conclusion

We found that academic records were useful for the estimation of a relational index. Also, classroom proximity was a good predictor for graduation, and it was found that co-enrollment density may predict retention along with gender and ethnicity. The findings answer our research questions. This work proposes a method to estimate classroom proximity using academic records, an index that may be related to academic integration; it adds an alternative perspective to the growing literature on social network analysis applied to higher education. Classroom proximity was estimated with data on enrollment found in official academic records that may be an alternative source for the assessment of integration indexes. The evidence obtained suggests that the classroom proximity index is relevant as a metric of retention and a potential proxy for academic integration. This result opens the possibility to implement other integration metrics using academic records. Relational indexes requiring only academic records that are already available may be used when studies that generate sociometric data are not available.

In general, the results showed that persisters have higher levels of proximity, or in this case, more frequent proximity in enrollment than that of non-persisters. Two institutions showed lower differences between persisters and non-persisters, which deserves to be reviewed in future analysis, and maybe the source of even more exciting findings that may add to this study, as well to the research on retention—particularly if other institutions added to MIDFIELD that show this same tendency. An expansion of MIDFIELD is underway that intends to facilitate such questions at the institution level of analysis (Ohland et al., 2011, 2008)

An important finding was that retention is related to higher levels of proximity. There was evidence supporting our original hypothesis that persisters generally have higher proximity and a broader range of mutual encounters with others when compared with students that do not graduate in the same Institution. The results demonstrate that aca-

demic records may be used to estimate sociometric indexes within formal academic settings. Classroom proximity, for example, builds upon accepted theories. If other sociometric indexes can be calculated from existing institutional course records, social network parameters analysis becomes more accessible to institutions. Academic records may be used more as a source of data that may approximate some aspects of social integration. However, this type of data does not include social data as they are commonly understood. The primary source of information, in this case, is the frequency of encounters of individuals in the same classroom that may or may not result in actual interaction and may or may not be intentional. Nevertheless, the results suggest that the probability of interactions, even after these considerations, may reflect the intentionality to some degree—enough to show consistent differences between those students that decided to leave the discipline, non-persisters, when compared with those that stay, persisters.

The database used for this study has information on public universities, only; therefore, it remains unknown if the observed results follow the same patterns in other colleges systems that do not share the same structure due to funding. Also, the database reflects only American style colleges, and further analysis would be required to evaluate the potential implications for university systems in other countries. While some limitations described earlier can be addressed in future work, it is essential to acknowledge them when interpreting the results of this study. The academic records used for the analysis do not have sociometric information as it is commonly understood, and this was the major constraint for the study proposed – we can only infer that intentional proximity is evidence of social integration. This fact limits the extension of the arguments that can be inferred from the results. Also, it separates this study and its results from the traditional sociometric analysis. It cannot be stated that the reported ”selections” were intentional, due to the lack of actual social preferences in the dataset. We know that the students were in the same place during most of the meetings for a particular course in a term. We believe that classroom proximity is sensitive to the longitudinal process as it depends on the number of courses taken by the student to improve its AUC. As we previously stated, co-enrollment density may be a link between academic integration, the commitment to the goal of graduation, and the likelihood of persistence in college. We found the results encouraging for looking for other algorithms that allow the approximation of other well-known social network integration parameters.

The results suggest that co-enrollment density may be a robust and parsimonious predictor for first-year persistence and graduation at 4-year engineering programs. While MIDFIELD has been shown to be representative of a more comprehensive national database of engineering programs, it is impossible to determine if MIDFIELD is representative in this study, because no other database exists capable of studying this phenomenon. Indeed, the United States Department of Education concluded that it was infeasible to create such a database (Cunningham & Milam, 2005). In particular, our findings suggest that institutions with extremely low persistence and graduation rates may be

exceptions. Co-enrollment density may replace multivariate models, when formal enrollment records are available, and it allows the estimation of cut-off points that may help identify students at risk of not persisting after the first year, or not graduating later. It shows higher specificity than sensibility when estimated at one and two years. Therefore, co-enrollment evaluated at the first and second years seem to be more sensitive to students at the risk of leaving; while, the index estimated at the third and the fourth years are better to identify students that show a positive trend to graduation.

# References


Aljohani, O. (2016a, February). A Comprehensive Review of the Major Studies and Theoretical Models of Student Retention in Higher Education. *Higher Education Studies*, *6*(2), 1. Retrieved 2020-07-21, from http://www.ccsenet.org/journal/index.php/hes/article/view/57103 (ZSCC: 0000177 tex.ids: aljohaniComprehensiveReviewMajor2016a, aljohaniComprehensiveReviewMajor2016b publisher: Canadian Center of Science and Education) doi: 10.5539/hes.v6n2p1

9

Aljohani, O. (2016b, January). A Review of the Contemporary International Literature on Student Retention in Higher Education. *International Journal of Education and Literacy Studies*, *4*(1), 40–52. Retrieved 2020-10-30, from https://eric.ed.gov/?q=management+students+retention+review&pg=3&id=EJ1149286 (ZSCC: 0000048 Publisher: Australian International Academic Centre PTY, LTD) doi: doi:10.7575/aiac.ijels.v.4n.1p.40

1, 9

Astin, A. W. (1999). Student Involvement: A Developmental Theory for Higher Education. *Journal of College Student Development*, *40*(5), 12. (ZSCC: 0003902)

7, 16

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*(2), 155–187. Retrieved 2020-08-20, from http://link.springer.com/10.1007/BF00976194 (ZSCC: 0002913) doi: 10.1007/BF00976194

6

Becker, F. S. (2010, August). Why don't young people want to become engineers? Rational

reasons for disappointing decisions. *European Journal of Engineering Education*, *35*(4), 349–366. Retrieved 2021-01-17, from http://www.tandfonline.com/doi/abs/10.1080/03043797.2010.489941 (ZSCC: 0000168) doi: 10.1080/03043797.2010.489941

1, 9

Belser, C. T., Shillingford, M. A., Daire, A. P., Prescod, D. J., & Dagley, M. A. (2018, September). Factors Influencing Undergraduate Student Retention in STEM Majors: Career Development, Math Ability, and Demographics. *The Professional Counselor*, *8*(3), 262–276. Retrieved 2021-01-24, from http://tpcjournal.nbcc.org/category/pdf-articles/volumes/volume-8/volume-8-issue-3/ (ZSCC: 0000012) doi: 10.15241/ctb.8.3.262

1, 9

Biancani, S., & McFarland, D. A. (2013). Social Networks Research in Higher Education. In M. B. Paulsen (Ed.), *Higher Education: Handbook of Theory and Research: Volume 28* (pp. 151–215). Dordrecht: Springer Netherlands. Retrieved 2021-01-25, from https://doi.org/10.1007/978-94-007-5836-0_4 (ZSCC: NoCitationData[s1]) doi: 10.1007/978-94-007-5836-0_4

2

Borrego, M., & Bernhard, J. (2011). The Emergence of Engineering Education Research as an Internationally Connected Field of Inquiry. *Journal of Engineering Education*, *100*(1), 14–47. Retrieved 2021-02-06, from https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2168-9830.2011.tb00003.x (ZSCC: 0000216 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2168-9830.2011.tb00003.x) doi: https://doi.org/10.1002/j.2168-9830.2011.tb00003.x

10

Borri, C., & Maffioli, F. (Eds.). (2007). *TREE: Teaching and research in engineering in Europe ; re-engineering engineering education in Europe ; SOCRATES Erasmus thematic network* (No. 43). Firenze: Firenze University Press. (OCLC: 254994910)

9

Braxton, J. M., Doyle, W. R., III, H. V. H., Hirschy, A. S., Jones, W. A., & McLendon, M. K. (2013). *Rethinking College Student Retention*. John Wiley & Sons. (ZSCC: 0000388 Google-Books-ID:

sTukAQAAQBAJ)

1, 9

Braxton, J. M., Milem, J. F., & Sullivan, A. S. (2000, September). The Influence of Active Learning on the College Student Departure Process: Toward a Revision of Tinto's Theory. *The Journal of Higher Education*, *71*(5), 569–590. Retrieved 2020-09-08, from https://www.tandfonline.com/doi/full/10.1080/00221546.2000.11778853 (ZSCC: 0001314) doi: 10.1080/00221546.2000.11778853

1

Carales, V. D. (2020, April). Examining Educational Attainment Outcomes: A Focus on Latina/o Community College Students. *Community College Review*, *48*(2), 195–219. (Publisher: SAGE Publications) doi: 10.1177/0091552120903087

1

Chapin, L. A. (2019, March). Longitudinal predictors for Mexican Americans' high school and college graduation: Individual and ecodevelopmental factors. *Journal of Latinos and Education*, 1–16. Retrieved 2020-07-16, from https://www.tandfonline.com/doi/full/10.1080/15348431.2019.1588733 doi: 10.1080/15348431.2019.1588733

16

Chongsuvivatwong, V. (2018). epidisplay: Epidemiological data display package [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=epiDisplay (R package version 3.5.0.1)

24

*CIP user site.* (2021). Retrieved 2021-05-27, from https://nces.ed.gov/ipeds/cipcode/Default.aspx?y=56

22

Corlu, M. S., Svidt, K., Gnaur, D., Lavi, R., Borat, O., & Corlu, M. A. (2018, January). Engineering Education in Higher Education in Europe. In Y. J. Dori, Z. Mevarech, D. Baker, & K. C. Cohen (Eds.), *Cognition, Metacognition, and Culture in STEM Education.* (pp. 89–116). Springer. (ZSCC: 0000011[s0]) doi: 10.1007/978-3-319-66659-4_5

10

Cunningham, A., & Milam, J. (2005, March). *Feasibility of a Student Unit Record System Within the Integrated Postsecondary Education Data System.* Retrieved 2021-02-18, from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005160 (Publisher: National Center for Education Statistics)

35, 38

Dobson, A. J., & Barnett, A. G. (2008). *An Introduction to Generalized Lienar Models Third Edition* (3rd ed.). Boca Ratón, FL, US: CRC Taylor & Francis Group. (ZSCC: NoCitationData[s0])

24

ERIC. (2020). *Educational Resources Information Center Thesaurus of ERIC descriptors.* Retrieved 2020-09-15, from https://eric.ed.gov/ (ZSCC: NoCitationData[s0])

10

Grothendieck, G. (2017). sqldf: Manipulate r data frames using sql [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=sqldf (R package version 0.4-11)

21

Grunspan, D. Z., Wiggins, B. L., & Goodreau, S. M. (2014, June). Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research. *CBE—Life Sciences Education*, *13*(2), 167–178. Retrieved 2020-10-25, from https://www.lifescied.org/doi/10.1187/cbe.13-08-0162 (ZSCC: NoCitationData[s1]) doi: 10.1187/cbe.13-08-0162

2

Hodges, B., Bedford, T., Hartley, J., Klinger, C., Murray, N., O'Rourke, J., & Schofield, N. (2013). *Enabling retention: Processes and strategies for improving student retention in university-based enabling programs: Final report 2013*. Australian Government Office for Learning and Teaching. (ZSCC: 0000099)

1

Huerta-Manzanilla, E., Ohland, M., & Long, R. (2013). *The impact of social integration on engineering students' persistence, longitudinal, interinstitutional database analysis*. CQUni-

versity. Retrieved 2021-05-25, from /articles/conference_contribution/The_impact_of_social_integration_on_engineering_students_persistence_longitudinal_interinstitutional_database_analysis/13431182/1
31, 32

Huerta-Manzanilla, E. L., Ohland, M. W., & Peniche-Vera, R. d. R. (2021, September). Co-enrollment density predicts engineering students' persistence and graduation: College networks and logistic regression analysis. *Studies in Educational Evaluation*, *70*, 101025. Retrieved 2021-05-25, from https://linkinghub.elsevier.com/retrieve/pii/S0191491X21000511 doi: 10.1016/j.stueduc.2021.101025
32

Huerta-Manzanilla, E. L., Ohland, M. W., Toledano-Ayala, M., & Jáuregui-Correa, J. C. A. (2021, September). Logit models, the area under receiver characteristic curves, sensitivity, and specificity for co-enrollment density in college networks dataset. *Data in Brief*.
24, 32

Inzer, L. D., & Crawford, C. B. (2005). A Review of Formal and Informal Mentoring: Processes, Problems, and Design. *Journal of Leadership Education*, *4*(1), 31–50. (Publisher: Association of Leadership Educators)
30

Israel, U. (2020). *Networks, Community Detection, and Robustness: Statistical Inference on Student Enrollment Data* (PhD Thesis). (ZSCC: 0000000)
2

Israel, U., Koester, B. P., & McKay, T. A. (2020, April). Campus Connections: Student and Course Networks in Higher Education. *Innovative Higher Education*, *45*(2), 135–151. Retrieved 2021-01-25, from http://link.springer.com/10.1007/s10755-019-09497-3 (ZSCC: 0000002) doi: 10.1007/s10755-019-09497-3
2

Jesiek, B. K., Newswander, L. K., & Borrego, M. (2009). Engineering Education Research: Discipline, Community, or Field? *Journal of Engineering Education*, *98*(1), 39–52. Retrieved 2021-02-06, from

https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2168-9830.2009.tb01004.x (ZSCC: 0000094 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2168-9830.2009.tb01004.x) doi: https://doi.org/10.1002/j.2168-9830.2009.tb01004.x

10

Johnson, M. H. (2013). *An Analysis of Retention Factors In Undergraduate Degree Programs in Science, Technology, Engineering, and Mathematics* (Unpublished doctoral dissertation).

1, 9

Johnson, W., & Jones, R. (2006). Declining Interest in Engineering Studies at the Time of Increased Business Need. In *Universities and Business: Partnering for the Knowledge Society, by Luc E. Weber and James J. Duderstadt.* (pp. 243–252). London: Economica.

1, 9

Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings*. Oup Usa.

18

Krause, K.-L., & Armitage, L. (2014). Australian student engagement, belonging, retention and success: A synthesis of the literature. *The Higher Education Academy*, 1–45. (ZSCC: 0000031)

1

Lucena, J., Downey, G., Jesiek, B., & Elber, S. (2008). Competencies Beyond Countries: The Re-Organization of Engineering Education in the United States, Europe, and Latin America. *Journal of Engineering Education*, *97*(4), 433–447. Retrieved 2021-02-06, from https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2168-9830.2008.tb00991.x (ZSCC: 0000196 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2168-9830.2008.tb00991.x) doi: https://doi.org/10.1002/j.2168-9830.2008.tb00991.x

1, 10

Ohland, M. W., Brawner, C. E., Camacho, M. M., Layton, R. A., Long, R. A., Lord, S. M., & Wasburn, M. H. (2011). Race, gender, and measures of success in engineering education. *Journal of engineering education*, *100*(2), 225–252. (Publisher: Wiley Online Library)

37

Ohland, M. W., & Long, R. A. (2016). The Multiple-Institution Database for Investigating Engineering

Longitudinal Development: an Experiential Case Study of Data Sharing and Reuse. , 25. (ZSCC: NoCitationData[s0])

2, 14

Ohland, M. W., Sheppard, S. D., Lichtenstein, G., Eris, O., Chachra, D., & Layton, R. A. (2008). Persistence, engagement, and migration in engineering programs. *Journal of Engineering Education*, *97*(3), 259–278. (Publisher: Wiley Online Library)

37

O'Malley, A. J., & Onnela, J.-P. (2019). Introduction to Social Network Analysis. In A. Levy, S. Goring, C. Gatsonis, B. Sobolev, E. van Ginneken, & R. Busse (Eds.), *Health Services Evaluation* (pp. 617–660). New York, NY: Springer US. Retrieved 2021-05-27, from https://doi.org/10.1007/978-1-4939-8715-3_37 doi: 10.1007/978-1-4939-8715-3_37

17

Pascarella, E. T. (1980). Student-Faculty Informal Contact and College Outcomes. , 51.

7, 16

Passow, H. J., & Passow, C. H. (2017). What Competencies Should Undergraduate Engineering Programs Emphasize? A Systematic Review. *Journal of Engineering Education*, *106*(3), 475–526. Retrieved 2021-02-06, from https://onlinelibrary.wiley.com/doi/abs/10.1002/jee.20171 (ZSCC: 0000141 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jee.20171) doi: https://doi.org/10.1002/jee.20171

10

Rao, A. R., & Bandyopadhyay, S. (1987). Measures of reciprocity in a social network. *Sankhyā: The Indian Journal of Statistics, Series A*, 141–188. (ZSCC: 0000048 Publisher: JSTOR)

2, 17, 19

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, *12*, 77.

24

Sithole, A., Chiyaka, E. T., McCarthy, P., Mupinga, D. M., Bucklein, B. K., & Kibirige, J. (2017, January). Student Attraction, Persistence and Retention in STEM Programs: Successes and Continu-

ing Challenges. *Higher Education Studies*, *7*(1), 46. Retrieved 2021-01-24, from http://www.ccsenet.org/journal/index.php/hes/article/view/65810 (ZSCC: 0000108) doi: 10.5539/hes.v7n1p46

9

Spady, W. G. (1970, April). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, *1*(1), 64–85. Retrieved 2020-08-19, from http://link.springer.com/10.1007/BF02214313 (tex.ids: spadyDropoutsHigherEducation1970a) doi: 10.1007/BF02214313

5, 7

Spady, W. G. (1971, September). Dropouts from higher education: Toward an empirical model. *Interchange*, *2*(3), 38–62. Retrieved 2020-08-19, from http://link.springer.com/10.1007/BF02282469 doi: 10.1007/BF02282469

5, 7

Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

24

Terenzini, P. T., & Pascarella, E. T. (1980). Toward the Validation of Tinto's Model of College Student Attrition: A Review of Recent Studies. *Research in Higher Education*, *12*(3), 271–282. Retrieved from http://www.jstor.org/stable/40195370

6

Tight, M. (2020). Student Retention and Engagement in Higher Education. *Journal of Further and Higher Education*, *44*(5), 689–704. (ZSCC: 0000020 Publisher: Routledge) doi: 10.1080/0309877X.2019.1576860

1

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 37.

2, 6, 7, 8, 9, 13, 16

Tinto, V. (1988). Stages of Student Departure: Reflections on the Longitudinal Character of Student Leaving. *The Journal of Higher Education*, *59*(4), 438–455. Retrieved 2021-06-02, from https://

www.jstor.org/stable/1981920 (Publisher: Ohio State University Press) doi: 10.2307/1981920

2, 6, 8

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition. Second Edition*. University of Chicago Press, 5801 South Ellis Avenue, Chicago, IL 60637.

2, 6, 8, 16

Tinto, V. (1997, November). Classrooms as Communities: Exploring the Educational Character of Student Persistence. *The Journal of Higher Education*, *68*(6), 599. Retrieved 2020-08-20, from http://www.jstor.org/stable/2959965?origin=crossref doi: 10.2307/2959965

6, 31

Tinto, V. (2006). Research and Practice of Student Retention: What Next? , 19. (ZSCC: 0003326) doi: https://doi.org/10.2190/4YNU-4TMB-22DJ-AN4W

8

Tinto, V. (2017, November). Through the Eyes of Students. *Journal of College Student Retention: Research, Theory & Practice*, *19*(3), 254–269. Retrieved 2020-10-08, from http://journals.sagepub.com/doi/10.1177/1521025115621917 (ZSCC: 0000308) doi: 10.1177/1521025115621917

8

Tinto, V., & Cullen, J. (1973). Dropout in Higher Education: A Review and Theoretical Synthesis of Recent Research. *Teachers College Columbia Univesity*, 100.

2, 6, 7, 8

Tudor, G. M. (2008). Mapping the Tutoring Referral Network: Exploring the Student-To-Tutoring Connection. , 157. (ZSCC: 0000000)

2

Wang, X., & McCready, B. (2013, October). The Effect of Postsecondary Coenrollment on College Success: Initial Evidence and Implications for Policy and Future Research. *Educational Researcher*, *42*(7), 392–402. (Publisher: SAGE Publications) doi: 10.3102/0013189X13505683

2

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York, NY, US: Cambridge University Press. (Pages: xxxi, 825) doi: 10.1017/CBO9780511815478

17

Wickham, H., Bryan, J., attribution), R. C. h. o. a. R. c. a. a. C. c. w. e. c., code), M. K. A. o. i. R., code), K. V. A. o. i. l., code), C. L. A. o. i. l., ... code), E. M. A. o. i. l. (2019, March). *readxl: Read Excel Files.* Retrieved 2020-10-08, from https://CRAN.R-project.org/package=readxl (ZSCC: NoCitationData[s0])

23

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., ... Dunnington, D. (2020, June). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* Retrieved 2020-10-08, from https://CRAN.R-project.org/package=ggplot2 (ZSCC: NoCitationData[s0])

23

Wickham, H., code), E. M. A. o. i. R., & RStudio. (2020, June). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files.* Retrieved 2020-09-04, from https://CRAN.R-project.org/package=haven

23

Willcoxson, L., Cotter, J., & Joy, S. (2011). Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities. *Studies in Higher Education*, *36*(3), 331–352. (ZSCC: 0000251 Publisher: Routledge)

1

# .1 R Code examples

This appendix includes the Listings 1, 2, 3, and 4 that are examples of the R Code implemented to process the records in MIDFIELD to produce the results reported with the first generation algorithm (E. Huerta-Manzanilla et al., 2013).

Listing 1

*R Code examples for the first generation algorithm*

```
1
2           # Files students and courses
3           # Libraries ------------------------------
4
5           library (sqldf)
6           library (stringr)
7           library (lattice)
8           library (latticeExtra)
9           library (RColorBrewer)
10
11          # All Students Mutuality ------------------
12
13          attach (courses)
14          attach (students)
15
16          # courserecords ← sqldf ("SELECT *, COUNT (MID) AS qty
17          FROM students JOIN courses USING (MID)
18          WHERE institution = 'Florida A&M' OR institution = 'Florida State'
19          AND lastgrp = 'ENG' GROUP BY CourseYYYYT, CourseCombo,
20          section, graduated")
21
22
23          # Main Table Preparation ------------------------------------
24
25          courserec ← sqldf ("SELECT *, COUNT (MID) AS qty
26          FROM students JOIN courses USING (MID) GROUP BY CourseYYYYT,
27          CourseCombo, section, graduated")
28
29          # Range of Response -----------------------------------------
30
31          attach(courserec)
32
33          range ← sqldf ("select * from courserec
34          where qty < 80 and qty > 2 AND Gender <> 'N' ")
```

## .2 R Code for the metanalysis

This appendix shows the R Code implemented for the metanalysis published in E. Huerta-Manzanilla et al. (2013). The Listing 5.

## .3 Dictionary for courses database

Table 1 shows the captions for the fields in the courses database of MIDFIELD as they were available at the time of the study.

Table 1

*Dictionary for Students Database*

| **Variable Name** | **Type** | **Length** | **Description** |
|---|---|---|---|
| APCredit | Char | 1 | Was credit for this course awarded as the result of advanced placement? |
| CourseAbbrev | Char | 5 | Course alpha identifier |
| CourseCombo | Char | 10 | Helper field, CourseAbbrev,CourseNumber |
| CourseCredits | Num | 8 | The number of credits awarded upon succesful completion of the course. |
| CourseGPEarned | NUM | 8 | The gradepoints earned for the course = coursecredits times gradepoint. |
| CourseYYYYT | Char | 5 | The second year of the MidfieldYear and term. |
| Grade | Char | 2 | Grade awarded for the course. |
| Method | Char | 1 | The means by which instruction is predominately delivered. |
| MID | Char | 10 | A unique MIDFIELD generated identification number for each student. |
| Section | Char | 4 | Course section identifier |

## .4 Dictionary for students database

Table 2 shows the dictionary for the fields in the Students database.

Table 2

*Dictionary for Courses Database*

| Variable Name | Type | Length | Description |
|---|---|---|---|
| ACT | Num | 8 | A two-digit composite raw score assigned to the applicant by the American College Testing Program |
| Age | Num | 8 | The age of the student at the time of first matriculation. |
| attendmonths | Num | 8 | Time attended in months. |
| attendyears | Num | 8 | Time attended in years. |
| CEEB | Char | 6 | A code by which the high school can be identified for those students who at their first entry to the institution were coming froma high school. The code is that used by ACT/ETS/CEEB. |
| citizenship | Char | 1 | Is the student a citizen of the United States? |
| ConMaj | Char | 3 | The 3 character major of the last major for students continuing as of last semester available |
| Coop | Char | 1 | Was the student ever a co-op student? |
| CoreGPA | Num | 8 | The grade point average for all attempted academic work that is required in the freshman and sophomore engineering curriculum. Non-engineering students will have a coregpa if they attempted any of the required core courses. |
| CumHoursEarned | Num | 8 | The total credit hours earned for all academic work. |
| CumHrsAttempted | Num | 8 | The total credit hours attempted for all academic work. |
| engineergpa | Num | 8 | The cumulative grade point average of engineering courses taken with an engineering prefix. |

*Continued on next page*

Table 2 – *Continued from previous page*

| Variable Name | Type | Length | Description |
|---|---|---|---|
| Ethnic | Char | 1 | Categories used to describe groups to which individuals belong, identify with, or belong in the eyes of the community. The categories do not denote scientific definitions of anthropological origins. A person may be counted in only one group. |
| EverAGR | Char | 1 | Was the student ever declared as an agriculture major? |
| EverAH | Char | 1 | Was the student ever declared as an arts and humanities major? |
| EverARE | Char | 1 | Was the student ever declared as an architectural engineering major? |
| EverASE | Char | 1 | Was the student declared as an aerospace engineering major? |
| EverBIE | Char | 1 | Was the student ever declared as an agricultural/biological engineering major? |
| EverBUS | Char | 1 | Was the student ever declared as a business major? |
| EverCHE | Char | 1 | Was the student ever declared as a chemical engineering major? |
| EverCPE | Char | 1 | Was the student ever declared as a computer engineering major? |
| EverCVE | Char | 1 | Was the student ever declared as a civil engineering major? |
| EverEGE | Char | 1 | Was the student ever declared as a general engineering major? |
| EverELE | Char | 1 | Was the student ever declared as an electrical engineering major? |
| EverENE | Char | 1 | Was the student ever declared as an environmental engineering major? |
| EverEngineer | Char | 1 | Was the student ever declared an an Engineering major? |
| EverEOE | Char | 1 | Was the student ever declared as an ”other” engineering major? |

*Continued on next page*

Table 2 – *Continued from previous page*

| Variable Name | Type | Length | Description |
|---|---|---|---|
| EverESE | Char | 1 | Was the student ever declared as an science and mechanics engineering major? |
| EverHSI | Char | 1 | Was the student ever declared an a history major? |
| EverMCE | Char | 1 | Was the student ever declared as a mechanical engineering major? |
| EverMTE | Char | 1 | Was the student ever declared as a materials engineering major? |
| EverNCE | Char | 1 | Was the student ever declared as a nuclear engineering major? |
| EverNMR | Char | 1 | Did the student ever not declare a major? |
| EverONS | Char | 1 | Was the student ever declared as an other non-STEM major? |
| EverSOC | Char | 1 | Was the student ever declared as a sociology major? |
| EverSTM | Char | 1 | Was the student ever declared ans a science and mathematics major? |
| EverTEC | Char | 1 | Was the student ever declared as a technology major? |
| EverTXE | Char | 1 | Was the student ever declared as a textile engineering major? |
| EverUND | Char | 1 | Was the student ever undeclared as a major? |
| Fall | Char | 3 | Did the student first matriculate in the fall term? IPEDS definition of fall entry includes students who strated in the summer term. |
| finalcumgpa | Num | 8 | The grade point average for all academic work at the institution. |
| FinalYear | Char | 7 | The year the student last attended the institution. |
| FirstGPA | Num | 8 | The cumulative GPA at the end of the first semester attended. |
| firstgrp | Char | 3 | The discipline group at first matriculation. |
| firstlevel | Char | 2 | The student level during the first semester of matriculation. |

*Continued on next page*

Table 2 – *Continued from previous page*

| Variable Name | Type | Length | Description |
|---|---|---|---|
| firstmaj | Char | 3 | The discipline major at first matriculation. |
| firstyyyyt | Char | 5 | The 4 digit MIDFIELD year and 1 digit MIDFIELD term of the first term attended |
| Gender | Char | 1 | Identifies the person by female or male classification. |
| grad6 | Char | 1 | Did the student graduate from the institution within 6 years? |
| gradgrp | Char | 3 | The discipline group at graduation. |
| gradmaj | Char | 3 | The discipline major at graduation. |
| graduated | Char | 1 | Did the student graduate from this institution? |
| gradyyyyt | Char | 5 | The 4 digit MIDFIELD year and 1 digit MIDFIELD term in which the student graduated. |
| groupath | Char | 203 | Declared discipline groups - 3 character groups |
| groupterm | Char | 200 | Time in semesters to declared major group - 3 characters groups. Leading X sets this variable to character when converting to Excel. |
| HomeZipCode | Char | 5 | The student's home Zip Code at time of admission. |
| hsgpa | Num | 8 | The high school grade point average upon which the student's application for admission is evaluated, based on a 4.0 grading system. A maximum of 5.0 is allowed for this variable since it is possible to obtain this ratio with extra weights on a 4.0 scale. |
| hsgparange | Char | 17 | High school grade categorized into 5 categories. |
| HSRank | Num | 8 | The ranked order of the student's high school class standing. |
| HSSize | Num | 8 | The size of the student's high school graduating class. |
| Instate | Char | 1 | Is the student a resident (based on home zipcode) of the state where the institution is located? |
| institution | Char | 13 | Institutional Name based on FICE code |

*Continued on next page*

Table 2 – *Continued from previous page*

| Variable Name | Type | Length | Description |
|---|---|---|---|
| LastGPA | Num | 8 | The cumulative GPA at the end of the last semester attended. |
| lastgrp | Char | 3 | The discipline group during the last term attended. |
| lastmaj | Char | 3 | The discipline major during the last term attended. |
| lastyyyyt | Char | 5 | The 4 digit MIDFIELD year and 1 digit MIDFIELD term of the last term attended |
| EverISE | Char | 1 | Was the student ever declared as an industrial and systems engineering major? |
| EverITD | Char | 1 | Was the student ever declared as a multi/interdisciplinary major? |
| MajorAfterFYE | Char | 6 | The first discipline major after the Freshman Engineering program. |
| majorpath | Char | 203 | Declared major groups - 3 character groups |
| majorterms | Char | 204 | A string of 3 character term numbers in which the student changed major. Leading X sets this variable to character when converting to Excel. |
| MID | Char | 10 | MIDFIELD created unique student identifier. |
| N | Char | 1 | A counting variable |
| othergpa | Num | 8 | The cumulative grade point average of all non science/-mathematics/engineering courses taken. |
| PES | Num | 8 | Percent enrollment in Free Lunch at the student's high school over the four years the student is expected to have attended high school. |
| PosGrad6 | Char | 1 | Could the student have graduated from this institution within 6 years of matriculation (time and a half of curriculum)? |
| PTFT | Char | 2 | Calculated from the average hours attempted per term. |
| SAT_M | Num | 8 | The three-digit, scaled score reported by the test publisher for the quantitative portion of the SAT. |

*Continued on next page*

Table 2 – *Continued from previous page*

| Variable Name | Type | Length | Description |
|---|---|---|---|
| SAT_V | Num | 8 | The three-digit, scaled score reported by the test publisher for the verbal portion of the SAT. |
| SAT | Num | 8 | The composite score. |
| sciencegpa | Num | 8 | The cumulative grade point average of all science and mathematics courses taken. |
| SecondMajTerm | Num | 8 | The term number in which the student first changes major. If a student never changes major this may be CON, TOL or GRD, |

Listing 2

*R Code examples for the first generation algorithm (Second part).*

```
1
2           # Charting by institution ALL MAJORS ---------------------
3
4           trellis.device(color = FALSE)
5           r1 ← densityplot (
6           ~ qty | institution,
7           groups = graduated,
8           key = simpleKey (text = (c("Non-Persisters", "Persisters")),
9           points = FALSE, lines = TRUE, columns = 2,  font=1),
10          layout = c(2,4),
11          data=range,
12          type = c("count"),
13          xlab= list("Dyads Frequency / Group-Class",font=1),
14          ylab= list("Probability Density of Dyads",font=1)
15          )
16          class (r1)
17          plot (r1)
18
19          # Majors All Institutions ----------------
20
21          trellis.device(color = FALSE)
22          majorallins ← densityplot (
23          ~ qty | lastgrp,
24          groups = graduated,
25          key = simpleKey (text = (c("Non-Persisters", "Persisters")),
26          points = FALSE, lines = TRUE, columns = 2,  font=1),
27          layout = c(3,4),
28          data=range,
29          type = c("count"),
30          xlab= list("Dyads Frequency / Group-Class",font=1),
31          ylab= list("Probability Density of Dyads",font=1)
32          )
33          class (majorallins)
34          plot (majorallins)
```

Listing 3

*R Code examples for the first generation algorithm (Third part).*

```
1
2
3           # Modelo 1 ----------------------------------
4
5           trellis.device(color = FALSE)
6           temp ← densityplot (
7           ~ qty | firstgrp * Gender,
8           groups = graduated,
9           key = simpleKey (text = (c("Non-Persisters", "Persisters")),
10          points = FALSE, lines = TRUE, columns = 2,  font=1),
11          layout = c(4,6),
12          data = range,
13          type = c("count"),
14          xlab= list("Dyads Frequency / Group-Class",font=1),
15          ylab= list("Probability Density of Dyads",font=1)
16          )
17          class (temp)
18          plot (temp)
19          rm (temp)
20
21
22          # Modelo General 2 ------------------------------
23
24
25          trellis.device(color = FALSE)
26          temp ← densityplot (
27          ~ qty | institution * firstgrp,
28          groups = graduated,
29          key = simpleKey (text = (c("Non-Persisters", "Persisters")),
30          points = FALSE, lines = TRUE, columns = 2,  font=1),
31          layout = c(2,4),
32          data = range,
33          type = c("count"),
34          xlab= list("Dyads Frequency / Group-Class",font=1),
35          ylab= list("Probability Density of Dyads",font=1)
36          )
37          class (temp)
38          plot (temp)
39          rm (temp)
```

Listing 4

*R Code examples for the first generation algorithm (Fourth and last part).*

```
1
2
3           # Testing ------------------------------------
4
5           range.ethnic ← sqldf ("select * from range where Ethnic <>
6           'X' and Ethnic <> 'A' ")
7
8           trellis.device(color = FALSE)
9           temp=densityplot ( institution ~ qty | Gender * Ethnic ,
10          groups = graduated,
11          key = simpleKey (text = (c("Non-Persisters",
12          "Persisters")), points = FALSE, lines = TRUE, columns = 2,
13          font=1),
14          layout = c(2,5),
15          data = range.ethnic,
16          type = c("count"),
17          xlab= list("Dyads Frequency / Group-Class",font=1),
18          ylab= list("Probability Density of Dyads",font=1)
19          )
20          class (temp)
21          plot (temp)
22          rm (temp)
23
24          # Example ------------------------------------
25
26          library (lattice)
27          trellis.device ( color = FALSE )
28          temp = densityplot ( institution ~ qty |
29          Gender * Ethnic, data = range.ethnic)
30          class (temp)
31          plot (temp)
32          rm (temp)
```

Listing 5

*R Code examples for the metanalysis of the second generation algorithm (First part).*

```
# Libraries
# The reader is encourage to use the {pacman} library for p_load()
# libraries instead of the more common library(), or require().

if (!require("pacman")) {
   install.packages("pacman")
   library(pacman)
} #Install {pacman} if it is not yet in the computer.

p_load(ggstatsplot, #Produce the charts: ggbetweenstats();
folderfun, #Set a folder function: setff();
ggplot2, #Allows to save the charts: ggsave();
ggthemes, #Select predefined themes: theme_hc();
readxl) #Read MS Excel files: read_excel();
```

Listing 6

*R Code examples for the metanalysis of the second generation algorithm (Second part).*

```
# folders
setff("io", "meta") #Set a subfolder meta to hold the files

# Data
meta ← read_excel(ffio("meta.xlsx"))
factores ← c("Ins", "Y", "X")
meta[,factores] ← lapply(meta[,factores] , factor)
met ← meta[(meta$CUT > 2),] #Screen out non significant models

# Function to save the charts to disk
sgr ← function(x,y){ggsave(filename = ffio(x),
   width = 10,
   height= 6,
   plot = y,
   device = "jpeg",
   dpi = 320,
   units = 'in')}
```

Listing 7

*R Code examples for the metanalysis of the second generation algorithm (Third and last part).*

```
# plot OR by X
or ← ggbetweenstats(
data = met,
x = X,
y = OR,
mean.ci = T,
type = "np",
bf.message = T,
results.subtitle = F,
outlier.tagging = T,
outlier.label = Ins,
ggtheme = theme_hc(base_size = 14),
xlab = "Predictor",
ylab = 'OR',
caption = FALSE,
pairwise.comparisons = T,
plot.type = "box",
messages = FALSE
); sgr("8-ORbyX.jpg",or)
```