



DSPACE

<https://dspace.org/>

A chow-liu score-based structure learning approach to refining course curricula in higher education

Naicker, Leantha

2024

<https://hdl.handle.net/10539/37528>

A Chow-Liu Score-Based Structure Learning
Approach to Refining Course Curricula in Higher
Education
Research Report

Leantha Naicker

788753

Supervised by Dr. Ritesh Ajoodha

June 6, 2023

Abstract

High failure and dropout rates have challenged higher education learning institutions to support students and keep them motivated throughout their undergraduate and postgraduate learning. This is not only beneficial to tertiary institutions but also to South Africa as a whole.

This research adds to the field of curriculum learning using structure learning graphical modelling to refine course curricula at a tertiary level. This will assist faculties in ensuring that students are provided sufficient knowledge in their respective fields through the co-requisite and prerequisite subjects of each program. This problem is constantly being looked at in literature, with most solutions embedded in manual methods; however, in recent years, there has been a shift to looking at how artificial intelligence and machine learning can be used to produce unbiased solutions.

Chow-Liu's score-based structure learning method in conjunction with K2 scoring is used in this research, due to its ability to handle large node spaces and reduce complexity with its tree-structured approach. The method is first validated using synthetically generated data before it is exposed to the real-world observational database.

The data set used for this study is obtained from a South African university after removing all socio-economic and demographic data. The results have two noted benefits; one is to help refine course curricula for undergraduate degrees by suggesting co-requisite and prerequisite courses to be added for various programs and the second is to help prescribe subject selection for postgraduate students.

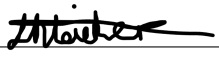
Keywords— Chow-Liu, Score-Based Structure Learning, Graphical Modelling, K2, Hill-Climbing

Declaration

I, Leantha Naicker, (Student number: 788753) am a student registered for the postgraduate research course CMS7009A, MSc(Computer Science) Research Report.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledgment of the original source) is wrong.
- I confirm that the work submitted for assessment is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand, Johannesburg may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  _____
Leantha Naicker (788753)
06 June 2023

Contents

1	Introduction	5
2	Related Work	7
2.1	Existing Curriculum Refinement Solutions	7
2.2	Student Academic Success Using Graphical Models	9
2.3	Probabilistic Graphical Modelling Background	12
3	Methodology	14
3.1	Data and Ethical Implications	14
3.2	Data Preprocessing	15
3.3	Models	17
3.3.1	Hill Climbing Method	17
3.3.2	Chow-Liu Method	18
3.4	Evaluation Techniques	19
4	Results and Discussion: Synthetic Data Model Evaluation	20
4.1	Ground Truth	20
4.2	Data Generation	20
4.3	Learned Models	21
4.4	Evaluation	23
5	Results and Discussion: Real World Data Model Evaluation	26
5.1	Undergraduate Course Refinement Results	26
5.2	Postgraduate Course Suggestions	30
6	Conclusion	34
6.1	Implications	34
6.2	Contributions	35
6.3	Future Work	35
	References	40

Chapter 1

Introduction

In South Africa, education is seen as the key to success in bettering futures and elevating families from poverty. Pursuing a higher education qualification is expensive and requires dedication, time, self-discipline, and support. Whilst getting accepted into a university is a great achievement that many work tirelessly for their whole schooling careers, every year universities experience significant failure rates [Nabil *et al.* 2021; Mti *et al.* 2021; Lauritzen 1996]. This results in debt and time wasted; as a failure at this level often leaves students disillusioned, causing them to drop out and give up the pursuit of degrees and diplomas to take up low-income jobs in the workforce.

A factor that gives rise to the failure rate can be attributed to a student's insufficient knowledge of certain examinable concepts; which could be caused by a student having missed the opportunity to grasp a fundamental outcome from a previous phase of education. This could also be caused by a failure in the proposed curriculum of the school to adequately prepare a student for concepts that are needed later on in their qualifications.

The purpose of this research is to identify these knowledge gaps and misalignments structured in the course curricula of higher education institutions, given the historic academic performances of students [Nabil *et al.* 2021]. The aim of this study is to use a structure learning graphical modelling approach to see if there are inefficiencies in the current curriculum for a given field of study and determine if additional prerequisite or co-requisite subjects are required to be added or mitigated in order to better support and equip students for success in their chosen field.

The assumption in this research is that each student's best effort mark/grade in a subject is directly correlated to that student's understanding of the content for the particular subject in question. The ultimate goal of this research is to help assist the youth of South Africa in completing their higher education qualifications; to put them in the most suitable positions in society for the success and growth of the country.

There have been many attempts to understand the high failure rates at tertiary education levels and how to mitigate it [Saranti *et al.* 2019; Nudelman *et al.* 2019; Lacave *et al.* 2018]. These studies often focus on predicting student success based on academic history, with the approach being a classification problem, whereby students are predicted to fall into one of two categories of 'Pass' or 'Fail' for an academic year or as the outcome of a specific course being taken. These studies also often rely on demographic and socio-economic factors to make their predictions. The research presented in this report aims to focus

purely on the academic history of students to produce its outcomes.

The problem of curriculum refinement is ongoing in literature today. Most approaches recommend manual processes which are left open to human error and biases [Purita P. Bilbao 2008; Rhodes 2018; Winter 2016]. In recent years, however, researchers have started looking to artificial intelligence and machine learning systems to find more robust and time-efficient methods of solving this problem [Ball *et al.* 2019; Rawatlal 2017]. Using probabilistic graphical modelling to address this issue will add to furthering the field of curriculum learning.

The research question is therefore; *Can the Chow-Liu method with the K2 scoring algorithm be used to provide a graphical ordering of subject matter into co-requisite and prerequisite relationships for tertiary-level courses, and be evaluated using KL-divergence against baseline models?*

This research produces graphical models demonstrating the relationships between subjects from the higher education data provided. The Chow-Liu method is selected to learn these structures from the observational data due to its ability to handle a large number of random variables time-efficiently [HC Peng and Ding 2012]. In order to validate the accuracy of the Chow-Liu algorithm, synthetic data is generated to represent a ground truth model and compare the learned model against it.

The relationships from the real-world observational data are verified by university course guidebooks and documentation, which contains courses that fall within each school at a university, with all their prerequisites and co-requisites. This verified model is then used to identify if relationships are missing in the current undergraduate standard curriculum that should be added as prerequisites or co-requisites to a course, or if there are subjects in a curriculum that serves little to no purpose for the chosen field of study.

Given that postgraduate degrees through coursework and research do not have prescribed course curricula, leaving students to sometimes select subject courses haphazardly to make up their credits. These graphical models are found to be useful for demonstrating how postgraduate courses relate to each other and help students choose subjects aligned better with their interests, strengths, or research topics.

The approach for building these structure learning graphical models were aimed to be as generalised as possible. Meaning that the systems do not rely on socioeconomic or demographic factors such as age, race, or gender. The data inputs to the models were purely based on academic results obtained for various subjects over a 6-year period (2012 to 2017).

This study contributes to the field of curriculum learning which is a topic in causation by providing a probabilistic graphical recommendation to help structure undergraduate course curricula at tertiary institutions and creating a postgraduate course selection guide to help better inform postgraduate hopefuls.

To that end, the structure of this research report is as follows: In this chapter, the problem area is introduced with a motivation for the research. In Chapter 2, relevant work already done in the areas of curriculum learning and graphical modelling are reviewed and discussed. The concepts and methods discussed in chapter 2 are used in chapter 3 to devise the methodology used to conduct and evaluate this experiment. Chapter 4 explores the results and discussion of the synthetic data model, to validate the legitimacy of the graphs that are obtained from real-world academic data. Chapter 5 then presents the results and findings from the observational data graphical models. Finally, chapter 6 discusses the implications and contributions of this research on the topic of curriculum learning.

Chapter 2

Related Work

In the previous chapter, the research area is presented, whilst introducing the problem that this study attempts to solve. In this chapter, an exploration of how curriculum refinement problems have been solved thus far is done in section 2.1. Section 2.2 looks deeper into the problems of using structure learning to improve student academic success and section 2.3 provides background information on probabilistic graphical modelling that is useful for understanding the research presented in this report.

2.1 Existing Curriculum Refinement Solutions

At all levels of education, institutions around the world strive to ensure that they have a comprehensive and robust curriculum to best support teaching and learning in the classroom. However, the approach to putting together such a curriculum varies according to country, institution and circumstances. This section explores some of the widely used propositions in the area of curriculum learning.

Purita P. Bilbao [2008] published an article on curriculum development, where she highlighted five main techniques used to solve this problem at primary, secondary and tertiary education levels. The approaches are as follows:

- *Behavioral Approach:* This involves specifying goals, objectives, contents and activities. The learning outcomes are evaluated against the defined objectives and are therefore perceived as a change in behaviour, which indicates the measure of success.
- *Managerial Approach:* Here a selected member such as the head of school, head of department, or principal becomes the curriculum leader. The leader sets the educational goals, plans the curriculum, designs each program and prepares instructional guides.
- *Systems Approach:* With influences from systems theory, parts of the educational systems are examined on how they relate to each other. For example, in this research one could look at how the stream of computer science relates to the stream of mathematics.
- *Humanistic Approach:* This approach is rooted in child development and progressive philosophy, by taking into account the formally planned curriculum and the student's total development. Here, the student is at the center of the curriculum.

- *Reconceptualist Approach*: Here the educator or instructor focuses on the needs and ability of each student rather than employing a standardised practice. This approach is certainly not feasible with a large volume of students.

Rhodes University published their guide to curriculum structure [Rhodes 2018] and highlighted that all curricula should be informed by the field of study, the philosophical beliefs of the school and the students themselves. The design of curricula at all levels (macro, meso and micro) needs to align with the academic needs of students and their social context. It is also important to provide learning objectives at the course, module and qualification levels. This helps structure course programs in accordance with the Higher Education Qualifications Sub Framework (HEQSF) and South African Qualifications (SAQA) requirements [Rhodes 2018].

The university [Rhodes 2018] also acknowledges the need to review the curriculum plan on a regular basis to accommodate new ideas, knowledge and practices. They aim to take into account the learning needs of students and the affordances of new technology to enhance the process of learning and teaching and redress issues that need to be accommodated for in the curriculum.

Winter [2016] explores a technique called Curriculum Mapping, which requires real-time electronic sharing of content and objectives. This collection of data can be used to align what is going on in a department, other departments and even other institutions. This could be of great value, to use other course curricula from various other tertiary institutions to refine your own. Ensuring that quality education is given to all students by matching the standards.

Ball *et al.* [2019] explored statistics and machine learning to develop a curriculum for the School of Computer Science using data from two universities. The authors recognised that there are socio-economic and personal factors that influence a student's success at university, they excluded such factors from the curriculum learning, using only information on academics, age and gender. They used machine learning to develop their curriculum with the academic history of 4266 students between the years 2005 and 2017.

Machine learning Identification Trees were used with correlation by Rawatlal [2017] to determine the relevance of a course to its requisites based on academic data and independent of human bias. Assessment results were compared against the existing co- and prerequisite structure to evaluate whether the current curriculum structure is sufficient for student progression. This experiment was performed on data from the school of engineering and resulted in a 72% accuracy. The author was able to make recommendations on courses that should be moved to a different year and identify useful prerequisite and co-requisite subjects for the school of engineering.

Most of the approaches mentioned above [Purita P. Bilbao 2008; Rhodes 2018; Winter 2016] are based on human assumption and perception. This means that there is room for human error and biases. Hence, for this research, a machine learning approach is used without the inclusion of biographical and socio-economic data factors.

2.2 Student Academic Success Using Graphical Models

Graphical models have been used in numerous fields such as medicine and biology for classification, biomarker selection and functional analysis using biomedical data [AJ *et al.* 2019] and in the field of computer science to model cyber systems [Sarkar *et al.* 2017; Riaz and Ali 2011] to help solve and model complex problems. Unfortunately, there is not much work in the field of graphically modelling course structures to study university curricula and student academic success; however graphical modelling studies in other Educational Data Mining subjects have helped refine and guide the approach of this study.

Scanagatta *et al.* [2019] provides a comprehensive review of the rapidly advancing field of Bayesian network structure learning over the past 30 years, that covers 24 constraint-based algorithms, 22 score-based algorithms and 15 hybrid learning algorithms. The review highlights the increasing reach of problems that can be solved by the algorithms, and it also goes on to highlight the discrepancies within the literature. The study finds and concludes that approaches such as hill climbing, Greedy Equivalence Search (GES) and Peter and Clark (PC) remain competitive over newly proposed algorithms, which conclude their own superior performance. The study further concludes that approximate score-based algorithms tackle high-dimensional problems better than other alternative approaches.

Table 2.1 gives a summary of some of the methods used for modelling student academic data and what corresponding evaluation metrics were used. This helps inform the approach used for this experiment. In the summary table MCMC is the Markov Chain Monte Carlo method, STEP is the Student-Teacher-Exam-Performance model which is a novel model by Shalem *et al.* [2014], BEM is Bridge Ensemble Measure and GDM is Global Direction Measure which are novel methods compared to an Extended Overlap Method (EOM) by Alsaad *et al.* [2019] and COMMAND is a novel method that for Combined Student Modelling and prerequisite Discovery. The evaluations; acc stands for accuracy as a percentage, r is correlation and AUC is the Area Under the Curve of the Receiver Operating Characteristic (ROC) which is a curve metric calculated from 10-fold cross-validation.

Table 2.1: A summary of authors' contributions using different Bayesian network graphical models in the field of educational data mining.

Author	Data	Features	Models	Evaluation
Culbertson [2014]	1,215 students who enrolled for Introduction to Classical Mechanics course at University of Illinois	Sub-domain knowledge areas of final examinations	MCMC	No measure of study success
Bhattacharyya <i>et al.</i> [2021]	Large synthetic distribution data	Numerical Values	Chow-Liu	KL Divergence and Conditional Independence Testing
Suzuki [2012]	Discrete and Continuous Data Distributions of a large nature	Numerical Values	Chow-Liu	KL Divergence
Selman and Gomes [2001]	N-Queens Problem Data (Large data set)	Placement of Queens, Moves, Board size	Hill-Climbing Search	Learning Score
Nudelman <i>et al.</i> [2019]	783 students enrolled for Computer Science at UCT between 2007 and 2016	Year, Matric Results, NBT Results, Registered Courses, Progression Results, Financial Aid, Secondary Schooling Info	Bayesian Network Modelling	72%
Saranti <i>et al.</i> [2019]	Data collection from application with 1164786 game instances from 9050 unique users	Answers to multiplication question	Learning Competence model	84.2%
Shalem <i>et al.</i> [2014]	Large data set crowd-sourced from Amazon's Mechanical Turk (AMT)	Student, teacher, topic, question, and correctness	STEP	$r = 0.999$
Alsaad <i>et al.</i> [2019]	Lecture transcripts	Analysis of lecture transcripts, student academic and demographic data	BEM, GDM and EOM	AUC:0.8, 0.8 and 0.74 respectively
Chen <i>et al.</i> [2016]	Data collected from a commercial non-adaptive tutoring system	Student performance data from Mathematics tests in Chapters 2 and 3	COMMAND and Fully Connected Model	AUC:0.803 ± 0.008 and 0.791 ± 0.007 respectively
Lacave <i>et al.</i> [2018]	383 students enrolled for Computer Science four courses	Academic and social data (gender, enrolment, age, first enrolment year, grade, scholarship)	K2 algorithm	K2 Scores

From the literature the Hill Climbing and Chow-Liu methods seem to be well-versed in handling large and complex problems in Bayesian graphical modelling [Suzuki 2012; Selman and Gomes 2001], such as the problem aimed to be solved with this research. It is noted that the Chow-Liu algorithm is hailed for its fast run time and remains one of the few structure learning procedures which does not require assumptions to be made as an input to the model [Bhattacharyya *et al.* 2021].

Whilst Culbertson [2014] compared various models against each other of different network levels; he along with Conati *et al.* [2004] noted that there is no consensus as yet on evaluating the quality of a graph based on the proposed techniques for comparing models. The following evaluations can be considered:

- *Mutual Information*: shows how two variables are associated with each other. Mutual Information is useful in comparing structures since Bayesian networks rely on conditional independence [Culbertson 2014; Liu 2009].
- *AUC*: is the Area Under the Curve of the Receiver Operating Characteristic (ROC) which is a curve metric calculated from 10-fold cross-validation [Alsaad *et al.* 2019; Chen *et al.* 2016]
- *Model Fit Indices*: Williamson *et al.* [2000] investigated the performance of the Ranked Probability Score [Epstein 1969], Weaver’s Surprise Index [Warren 1966], and Good’s Logarithmic Score [Good 1952] under several different model error conditions and found that Ranked Probability Score and Weaver’s Surprise Index performed well in detecting Bayesian network graphical errors.
- *Diagnostic Plots*: provides information about whether a given network fits the provided data [Sinharay *et al.* 2004]. Curves or plots can be used to assess discrepancies from the predicted data to the proposed model [Culbertson 2014].

Whilst there isn’t a substantial body of work in using Bayesian graphical models in the field of Educational Data Mining (EDM) to structure course curricula, there has been much research done relating to student academic performance within tertiary-level institutions. EDM exploits statistics, machine learning and data mining to analyse education data and resolve educational research issues. Predicting student success at the higher education level is often a classification problem in the field of Machine Learning (ML) [Hlomuka *et al.* 2021], where students are predicted to fall into one of two categories of ‘Pass’ or ‘Fail’ as the outcome of a specific course within the course curriculum for their qualification.

Culbertson [2014], as part of his PhD of Philosophy in Educational Psychology, studied graphical models for student knowledge. He used his research to motivate efficient testing of students based on the prerequisite and hierarchical structural relationships. It is noted that whilst Bayesian networks have not been used much in the educational measurement field, it has been used extensively in the field of AI for intelligent tutoring systems (ITS) [Mashiloane and Mchunu 2013].

Predicting students at risk of academic failure, at an early stage of a semester, is a useful tool to allow for early intervention to take place. Taking actions early enough to enhance knowledge and performance can save entire academic careers [Nabil *et al.* 2021; Hlomuka *et al.* 2021]. Nabil *et al.* [2021] approached this topic using Deep Neural Networks (DNN) to attempt to find students at risk of failure early in a semester to provide additional support in order to reduce failure rates at the end of each semester. They noted the challenge in predicting student academic performance due to the way in which demographic features,

economic background and student behavioural patterns can affect the biases of machine learning models.

Most studies take into account the socio-economic and demographic factors of students along with academic history to make these kinds of predictions [Nabil *et al.* 2021; Dukhan 2020; Thiele *et al.* 2016; Dukhan *et al.* 2012]. Machine learning models and artificial intelligence does have the ability to become biased based on socio-economic and demographic factors. Given this knowledge, it brings into question how these features are used and how much weight they carry on the predictive outcomes in research [Nabil *et al.* 2021; Dukhan *et al.* 2012]. This study, therefore, aims to focus purely on the academic history of students to produce its outcomes [Hlomuka *et al.* 2021], therefore removing any potential biases.

The question of using probabilistic graphical modelling to verify curriculum structure at a tertiary education level does not have an ample answer based on the literature assessed. There are some allusions to the idea of using graphical models to solve other EDM-related topics, as highlighted in this section but no study has deeply honed in on this exact research topic.

2.3 Probabilistic Graphical Modelling Background

For this research, the aim is to use a graphical model to find relationships between existing subjects offered at a university in order to aid in curriculum refinement. Graphical modelling, probabilistic graphical modelling (PGM) and structured probabilistic modelling; are all terms used to describe a probabilistic model which uses a graph to express conditional dependence structures between random variables [Anandkumar *et al.* 2011; Koller and Friedman 2009; Lauritzen 1996]. The random variables are presented in a graph by nodes and the relationships between the nodes are expressed by edges to represent relationships in a system.

There are two branches of graphical models commonly used; one is Bayesian networks which combines probability theory with graph theory and causality [Scanagatta *et al.* 2019; Koller and Friedman 2009] and the other is Markov random fields which is a random field that satisfies Markov properties described by an undirected graph. Both branches encompass the properties of factorization and independence. They do however differ in the set of independence that they can encode and the factorization of the distribution they induce [Koller and Friedman 2009]. Both branches of graphical modelling are commonly used in probability theory, statistics and machine learning.

Bayesian networks are the most popular modelling tool used for complex systems involving uncertainty. A Bayesian network is a Directed Acyclic Graph (DAG) used for modelling joint probability distributions over a set of variables [Onisko and Druzdzal 2014]. A Bayesian network B is represented by $B = (G, \Theta)$, which is defined by a tuple consisting of a DAG G and a set of parameters Θ . DAG G , consists of a set of nodes or vertices denoted by $\mathbf{X}$; each of which corresponds to one of the n variables in the following, $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$; and a set of directed edges $\mathbf{E}$ or arcs such that $G = (\mathbf{X}, \mathbf{E})$ [Scanagatta *et al.* 2019].

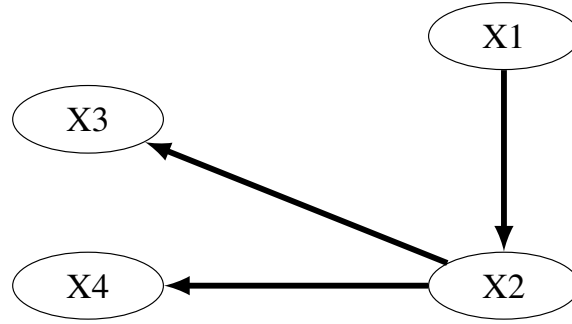


Figure 2.1: Example DAG with four random variables X_1 , X_2 , X_3 and X_4 making up the nodes in $\mathbf{X}$ with set of directed edges $\mathbf{E}$; $X_1 \rightarrow X_2$, $X_2 \rightarrow X_3$ and $X_2 \rightarrow X_4$.

Learning a DAG requires a structure learning approach. Structure Learning is the process of using data to learn the links between nodes of a Bayesian network, which is pertinent to the object of this research. There are two types of structure learning as seen in figure 2.2: *Constraint-based Methods* which prune edges based multiple conditional independence tests; and *Score-based Methods* which follow an approach to maximise an objective function over different graphs [Scanagatta *et al.* 2019; Koller and Friedman 2009].

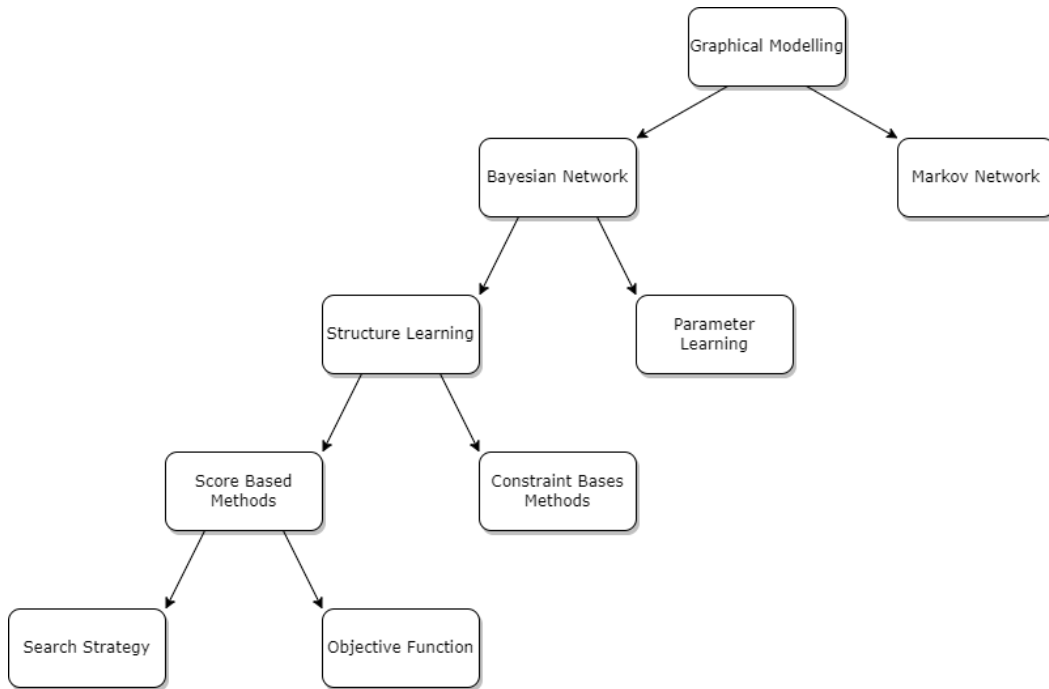


Figure 2.2: Hierarchy diagram showing the various branches that exist in the field of probabilistic graphical modelling and where the score-based methods that will be used in this research fit in.

Chapter 3

Methodology

This research aims to find strong relationships between the different subjects offered at a university using a structure learning method to obtain a probabilistic graphical model, in the hopes of improving university curricula. Having well-structured and robust curricula that ensure a solid foundation of outcome knowledge requirements for students will increase pass rates at tertiary institutions. The approach used here compares the results of a score-based structure learning graphical model called the Chow-Liu method.

To validate the results of the models obtained with the observational data, synthesised structures are generated and used to create the various graphical models. These models are then evaluated using Kullback–Leibler Divergence (KL-divergence) graphs to conclude the validity of the Chow-Liu method. Once this is successfully proven, the models obtained from the real-world data can be trusted to draw conclusions on the subject relationships in the course curriculum.

The data used in this research is from a South African university, the ethical considerations for which are discussed in section 3.2. Required data preprocessing to reduce noise and produce better results are further discussed in section 3.3. Data processing focused on two main areas: feature selection and noise filtering. After being processed, the data is run on the probabilistic graphical modelling algorithm to find relationships and correlations between the different subjects and courses in the data set. Details on the learning methodology can be found in section 3.4. Lastly, section 3.5 looks at the evaluation techniques of the learning structures and how to measure success.

3.1 Data and Ethical Implications

The data set used for this experiment is obtained from a South African university’s faculty of science. The data shows each student with the corresponding subjects they have taken and their resultant grade for each of those subjects. Ethical clearance was required from the University’s Human Research Ethics Committee for the data to be used in this research as it contains the private information of real students attending the university between the years 2008 and 2018.

Due to the fact that this experiment is neglecting the contributions of socio-economic and demographic information, and that all student numbers were fully encrypted to ensure anonymity; the University’s Human Research Ethics Committee has approved the study’s ethics application to use the data purely

for its academic history record. The protocol number for the clearance certificate is CSAM-2022-02.

Table 3.2 in section 3.3 on data preprocessing shows a sample of the observational data with all the socio-economic and demographic fields removed. A good point to take note of when talking about the data is that in order to confirm the validity of the score-based structure learning algorithms, synthetic data was generated using a forward sampling method. This process and its use are discussed further in results Chapter 4.

3.2 Data Preprocessing

Data preprocessing is required for graphical modelling and performance prediction models in higher education because such high-volume data sets can suffer from noise and other issues like class imbalance. Therefore, feature selection and noise filtering are necessary steps to ensure that data is ready for modelling [Hassan *et al.* 2020].

One of the most vital steps in machine learning is that of data preprocessing; whereby raw data collected from the real world is cleaned and converted to a suitable format to achieve better results in machine learning models [Nabil *et al.* 2021]. This is achieved with the following techniques:

- *Preliminary Feature Selection:* This is a method whereby features are filtered out upon first sight, due to their irrelevance or the criteria of the study. In this instance, demographic data such as age, race and gender are features that are discarded from the data set at the onset for the purposes of this research.

Features can also be filtered out manually due to large amounts of missing data. Using a missing statistics function, it was possible to find which columns of data had large volumes of missing information. The columns missing a large percentage of data were related mostly to secondary schooling information and residency, which are not needed for this study anyway. A full assessment of the missing values in the data set can be found in Appendix A of this report.

Table 3.1: Table showing the percentage of missing data for a few select sets of data columns between the years 2008 and 2018 for the faculty of science at a South African higher-education institution

Column Name	% Missing
Program Title	0.00
Encrypted Student No	0.00
Course Code	0.00
Course Title	0.01
Final Mark	0.01
Final Grade	0.01

- *Fill in missing values:* There are three methods to deal with missing data values for critical fields: For continuous variables, the data can be extrapolated via a mathematical formula or using the k nearest neighbors averaging method to find an estimate for the missing value. For discrete variables, a median option of the discrete options can be taken as Onisko and Druzdzal [2014] did for their medical diagnosis study with graphical modelling. Lastly, one can delete rows containing missing important data from the data set if the number of entries with missing values is small

enough to not deeply impact the value of the data set.

For this data set the missing values in the 'Final Mark' column, as seen in table 3.1, is averaged using the k nearest neighbor method, where $k = 5$ and the marks are mapped to the 'Final Grade' column using the grading system of the institute.

- *Categorical to Numerical transformation*: This allows categorical data features to be mapped to numerical values via a key to make the data more suitable for machine modelling. This method was needed for the string values in the 'Final Grade' column.
- *Table Transformation and Mapping*: For the algorithms to consume data in a fashion that will produce the desired outcomes, the encrypted student numbers and academic data needed to be formatted to a table whereby each column represented a unique course code and each row a unique encrypted student number.

The end result would be data in table 3.2, mapped to a representation like that in table 3.3.

Table 3.2: Table showing how the real-world data is presented after removing all fields related to the student's socio-economic and demographic traits

Year of Study	Course Code	Final Mark	Encrypted Student No	...
2017	CMS1015	58	E0D140BC16941C9D78E0D3A3547F08EB	...
2018	CMS1016	70	E0D140BC16941C9D78E0D3A3547F08EB	...
2018	CMS1017	68	E0D140BC16941C9D78E0D3A3547F08EB	...
2016	CMS1016	66	E0D141BC16961C9D78F0D3A3578F08EV	...
2016	CMS1017	61	E0D141BC16961C9D78F0D3A3578F08EV	...

Table 3.3: Table showing how the data needs to be presented in order to be consumed by the Chow-Liu structure learning procedure

Encrypted Student No	CMS1015	CMS1016	CMS1017	...
E0D140BC16941C9D78E0D3A3547F08EB	58	70	68	...
E0D141BC16961C9D78F0D3A3578F08EV	-	66	61	...

Whilst working with the data after following the pre-processing steps, two further discoveries were made that required further data clean-up. First, it was found that the data set contained postgraduate courses for honors and masters, despite being labeled as 'undergraduate'. In order to refine the study, it was decided to focus on undergraduate curricula to help students achieve that first milestone and separately look at the postgraduate data.

Furthermore, postgraduate degrees (such as honors or masters by course work and research) do not have prescribed curricula structures to evaluate against. Students are free to choose whichever subjects they want to do and in whichever order they would like to do them in. This study is extended to use the learned graphical model to help find strong relationships among the courses for postgraduate students available to help them pick subjects that align with their interests, strengths or research topics.

Secondly, with further knowledge from the university which the data was extracted from, it was found that there was a subject refresh that occurred during the years in which the information was collected. This resulted in a duplication of courses being reflected in the 'unique' course name extraction. Further cleaning had to be done on the data to eliminate these duplications, and thus at the end of the exercise, only data between the years 2012 and 2017 could be used.

3.3 Models

As discussed in chapter 2 of this report, probabilistic graphical modelling has two branches, namely Bayesian networks and Markov random fields. For this research Bayesian networking was used for its combination of probability theory, graph theory and causality [Scanagatta *et al.* 2019]. Learning a dependency graph from a Bayesian network requires a structure learning approach. In this study, score-based learning approaches were selected and examined more closely. Score-based methods consist of a search strategy and an objective function.

The challenge with score-based approaches is to find the highest, or near to the highest, scoring graph from several feasible graphs. Finding the exact highest scoring graph is an exhaustive, complex and time-consuming effort that only becomes more inefficient as the problem size grows [Scanagatta *et al.* 2019; Koller and Friedman 2009]. Therefore, for this study approximate algorithms were utilized. These typically produce a single Directed Acyclic Graph (DAG) with a locally optimal score.

Further research into Hill Climbing (HC) and the Chow-Liu algorithms were conducted as methods of interest. A summary of the research for each method can be found in the subsections that follow.

3.3.1 Hill Climbing Method

Hill climbing is understood to be an iterative mathematical optimization technique in the area of score-based learning; whereby the algorithm starts with an initial solution to a problem and makes incremental changes to the solution to establish a better solution; until a better solution can no longer be found [Chinnasamy *et al.* 2022; Bouckaert 1994].

To put this into a mathematical description, hill climbing tries to maximise a target function which is termed $f(\mathbf{x})$; in which $\mathbf{x}$ is a vector of values. For each iteration of the algorithm, hill climbing will adjust a single element in $\mathbf{x}$ to determine if the adjustment improved the value $f(\mathbf{x})$. This process continues until there is no longer an improvement in $f(\mathbf{x})$ and at this point, it can be agreed that $\mathbf{x}$ is ‘locally optimal’. It is important to note the algorithm stops when it reaches an optimal solution or if it runs out of time or resources.

Hill climbing and greedy algorithms are both heuristics used for optimization purposes to find a good localized solution. Hill climbing does run into problems when it enters one of the following regions: plateau, ridge or a local maximum [Chinnasamy *et al.* 2022].

Plateau: A point in the graph where all neighbors have the same value. This can be overcome by making a big jump with the chance that you will land in a non-plateau region.

Ridge: Any point on a ridge may appear as a peak because all possible directions move downwards. This can be overcome by moving in several (two or more) directions at once.

Local Maximum: At a local max, all neighboring states have values worse than the current state. Applying a greedy approach means that the algorithm will choose to terminate itself at this stage, even though a better solution may exist. This can be overcome by using a backtracking technique that maintains a list of visited states. Backtracking will allow the algorithm to return to a previous state and explore a new path.

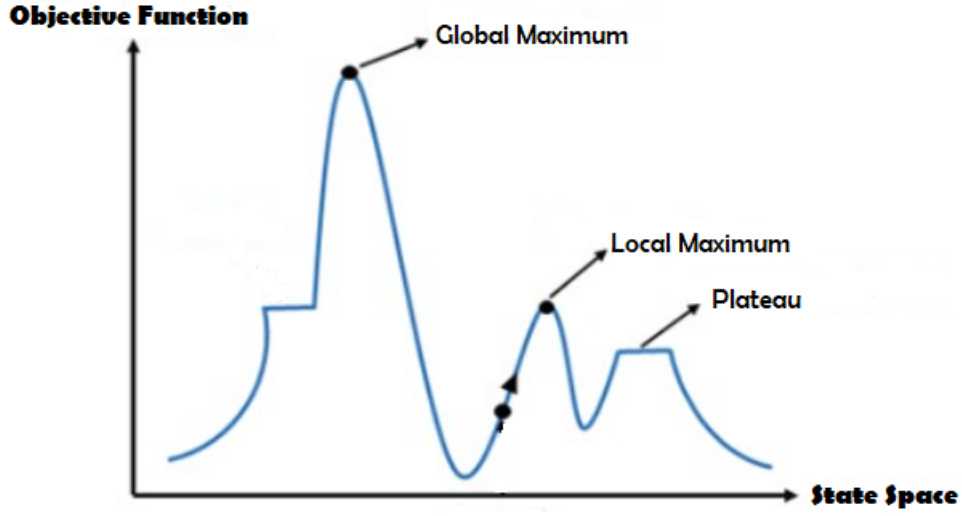


Figure 3.1: Adapted State Space Diagram demonstrating the regions that the hill climbing method can fall into [Chinnasamy *et al.* 2022].

3.3.2 Chow-Liu Method

The Chow-Liu algorithm was developed in 1968 and uses a maximum likelihood tree structure search approach to optimise its score. The tree structure of the algorithm limits each node to have at most one parent; but has no limit to the number of child nodes it may bare. The advantage of this approach is that the complexity is limited by the parent policy, making it faster to run on data sets with a large number of random variables, as is the case with this research.

Chow-Liu works by making an approximation on a joint distribution with a tree structure. The algorithm can effectively be boiled down to three fundamental steps:

1. Computing the weight $I(X,Y)$ of each possible edge between all nodes in the data set. This is achieved with formula 3.1; where X and Y are two nodes with an edge drawn between them.

$$I(X,Y) = \sum_X \sum_Y P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right) \quad (3.1)$$

2. Find the maximal weight-spanning tree that connects all nodes in the graph using the Kruskal algorithm. In each step, one edge is greedily added and a check must be done to ensure that the structure is still a tree before proceeding to add the next edge. The tree with the greatest total weight becomes the final model.

$$\sum_{(X,Y) \in E} I(X,Y) \quad (3.2)$$

Kruskal's algorithm as represented by equation 3.2 is derived from graph theory and acts greedily to discover the shortest path between two points in a connected weighted graph. The algorithm changes a graph into a forest and then views every node as a separate tree to perform its function.

3. Select a node to be the root node from which all other node-directed edges will radiate out.

The work of Chow and Liu (1968) shows that the maximum-likelihood tree-structured network can be found in quadratic time [Koller and Friedman 2009]. If distribution P is tree-structured, then Chow-Liu finds the correct structure in polynomial time $O(n^2[m + \log n])$. If P is not tree-structured, Chow-Liu finds a tree structure Q that has minimal KL-Divergence of $\argmin_Q KL(P, Q)$ [HC Peng and Ding 2012].

The caveat to this approach is that due to the parental limitation, the method is unable to capture the complete structure of the distribution. For example, based on the research topic of this report, first-year courses may not appear aligned as co-requisites to one another at first glance, but further interpretation will highlight the causal relationship.

3.4 Evaluation Techniques

It is important to evaluate each of the selected structure learning models to quantify how well they each perform with the data given. In order to validate how accurate the models are, first an experiment must be run on a synthetic data set that reflects the true data set concerning the number of random variables which in this case are the various undergraduate course codes in the field of science offered at this higher-education institution.

The experiment on the ground truth is run as follows; first, a ground truth random structure is created, then the data is generated using a forward sampling method, the next step involves learning the model and lastly the ground truth distribution is measured against the learned distribution using Kullback–Leibler Divergence or KL-Divergence. These methods are discussed further in Chapter 4.

Once the synthetic data is used to validate the outputs of the structure learning algorithms, the real-world data is then applied to produce models from the Chow-Liu method. Each model is then further evaluated with the university course guidebooks. Looking for accuracy in the models such as a second-year course depending on a first-year course, similar course codes for each school being grouped together, etc.

Once this evaluation is completed, to depict how accurately the models generated compare to the current curricula at the university’s science faculty; it then becomes possible to look for other strong inferences from the models such as potential subjects that should be added to the various programs offered to better the course curricula and strengthen a students knowledge base for success in other co-requisite or pre-requisite subjects.

The study also evaluates the postgraduate data set to make recommendations to students on which subjects to take for their honors or masters coursework in order to help guide them in achieving the right learning outcomes and achieve success. More information on the finding from the observational dataset can be found in chapter 5.

Chapter 4

Results and Discussion: Synthetic Data Model Evaluation

Before one can go ahead and make inferences on the observational data set, it is important to first validate the reliability of the resultant models obtained from the algorithm. This chapter explores how to do that and discusses the findings of the model outputs using synthetically generated data which reflect our real-world data format, in the sense that every column represents a course and every row an individual student as shown in table 3.3. The experimental setup, results and discussion will be laid out in sections 4.1 to 4.4.

4.1 Ground Truth

The first step in this experiment is to generate a random structure with the same number of variables as the real-world undergraduate data set. This is done to ensure that the synthetic data results and real-world data results are comparable. For course names in the synthetic set, numbers ranging from 1 to 486 are used, this is reflective of the number of undergraduate courses that are available in the faculty of science for this research-intensive higher-education institution.

This structure is generated 10 times and randomly each time to avoid biases, to get an average of the results and ensure reliable feedback. Once the course structure is generated, the matrix table can be filled with data that can be fed into the models to produce a DAG. The ground truth model is to be used in this experiment as the target or validating model.

The ground truth models for a five-node example can be seen in figures 4.1 and 4.2 for the hill climbing and Chow-Liu methods respectively.

4.2 Data Generation

A forward sampling method is used over likelihood-weighted sampling and rejection sampling to populate the data set for each random structure that is created. Each generation produces 100 samples for the

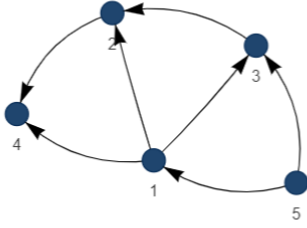


Figure 4.1: Five-node ground truth DAG using the hill climbing score-based method

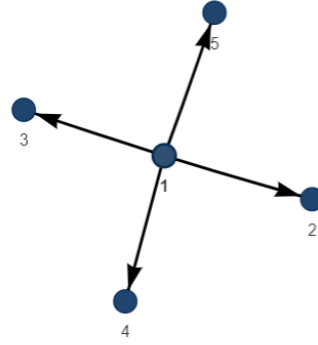


Figure 4.2: Five-node ground truth DAG using the Chow-Liu score-based method

KL-Divergence graph in figure 4.5 and a variable n number of samples in figure 4.6.

Forward sampling is a method used to produce a sample $\xi[1], \dots, \xi[M]$ of every variable X in a Bayesian network so that each sample is generated in proportion to its ground truth probability distribution $P(X)$ [Koller and Friedman 2009]. This allows the estimation of the prior probability of any variable to be possible. This is what is known as topological order, where values for the parents of a variable are generated first.

Meaning that if first X_1 is sampled, followed by each of the variables according to the ordering of variables, by the time the algorithm reaches X_M it already has values for the parents of X_M . It can now sample a value for X_M from the distribution of X_M given the values assigned to its parents. This way it always knows exactly which distribution to use. The algorithm continues this generation process for every single variable.

It should be noted that forward sampling is a randomized process, so each run can generate a different result each time and the estimated distribution from different runs will usually be different as well. Therefore the sampled distribution is only an estimate of the true distribution. This highlights two important points about forward sampling:

- One particular sample distribution will be different from the true distribution, but on an average of multiple sample distributions, they would be equal to the true distribution. This makes the sampling process unbiased.
- The more samples used, the closer one gets to the true distribution. This means that the variance of the sampling process decreases with more samples.

4.3 Learned Models

Now that synthetic data tables are generated; one can advance to using the hill climbing and Chow-Liu structure learning procedures to learn a model for each procedure. To create the data model, the Maximum Likelihood Estimator (MLE) is used to fit the data. The estimated model or sample is scored using

the K2 scoring method for each learning procedure.

In statistics, the Maximum Likelihood Estimator is a framework for estimating the assumed probability distribution and parameters that best explain the observational data. When estimating parameters for Bayesian networks, MLE has a problem of overfitting the data when there is insufficient data available. Thankfully, the data sets used for the synthetic model and the real-world data are well-informed and thus MLE can be used [Brownlee 2019a].

Equation 4.1 defines the likelihood function for a given choice of parameters θ , where $P(\xi[m] : \theta)$ is the parametric model for which we wish to estimate parameters for and ξ an instance of X , which depicts the random variables or nodes. It is required for each choice of parameters θ that $P(\xi[m] : \theta)$ is a legal distribution, is not negative and that $\sum_{\xi} P(\xi[m] : \theta) = 1$ [Koller and Friedman 2009].

$$L(\theta : D) = \prod_m P(\xi[m] : \theta) \quad (4.1)$$

Once the likelihood function is defined, one can use the MLE to choose the parameter values $\hat{\theta}$ for a given data set D that satisfies equation 4.2 [Koller and Friedman 2009]:

$$L(\hat{\theta} : D) = \max_{\theta \in \Theta} L(\theta : D) \quad (4.2)$$

So MLE is used for this experiment to fit the synthetic data using equations 4.1 and 4.2 . The models are then scored using the K2 scoring method. The K2 algorithm was published by Cooper, G. and Herskovits, E. in 1992 as a score-based greedy search algorithm for learning Bayesian networks from observational data. K2 uses a Bayesian score $P(Bs, D)$ to rank different structures and then greedily maximises $P(Bs, D)$. Bs is the score and D is the observational data.

K2 relies on the following assumptions to functionally score Bayesian networks; there are little to no missing values present in the data set, there is parameter independence and modularity, and there is uniformity in the prior distribution of the parameters in the given network structure. K2 can be expressed with equation 4.3 [de Campos 2006]; where $P(G)$ is the structure prior probability that is a constant for each graph G . n is the number of nodes in the graph, q_i the state of the parent of the i^{th} node, r_i the mutually exclusive states of the i^{th} node and N_{ijk} the instance of the i^{th} node being in the k^{th} state when its parents are in their j^{th} configuration.

$$K2(G : D) = \log(p(G)) + \sum_{i=1}^n \left[\sum_{j=1}^{q_i} \left[\log \left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right] \right] \quad (4.3)$$

There are a multitude of scoring methods one can use to achieve score-based networks such as; the Bayesian Dirichlet or BD scoring function, the likelihood-equivalence Bayesian Dirichlet or BDe scoring function, log-likelihood (LL) scoring function, Minimum Description Length (MDL) scoring function, Normalized Minimum Likelihood (NML) scoring function, Mutual Information Tests (MIT) scoring function or even the Bayesian Information Criterion (BIC) scoring function. These are amongst the most popular methods used in the field today. Carvalho [2021] conducted an experiment amongst all these methods and concluded that the K2 method is best used for larger data sets and NML for smaller ones [Carvalho 2021].

Figures 4.3 and 4.4 were produced using the hill climbing and Chow-Liu learning algorithms respectively, using a K2 scoring system. Below are examples of DAG graphs produced from running the algorithms with a five-random node (course) structure.

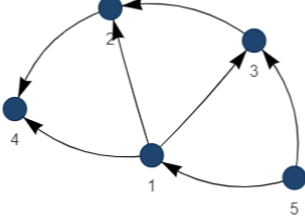


Figure 4.3: Five-node DAG example generated using the Hill Climbing method

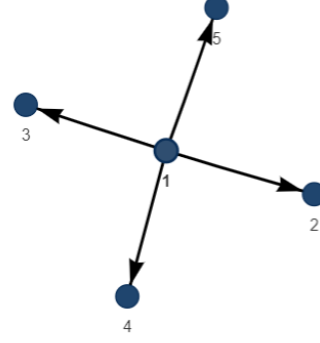


Figure 4.4: Five-node DAG example generated using the Chow-Liu method

As seen in figures 4.3 and 4.4; both outputs agree on the relationships between nodes labeled (1,2), (1,4), (1,3), (1,5); however, the hill climbing method finds further relationships between nodes (2,4), (3,2) and (5,3), which the Chow-Liu method does not represent. This leads to the conclusion that whilst the Chow-Liu method does find meaningful relationships and associations, it can not give a fully connected graph like the hill climbing method, where all relationships are present.

4.4 Evaluation

Kullback-Leibler Divergence or KL Divergence is named after the two authors of the method Solomon Kullback and Richard Leibler. It is an evaluation metric commonly used to measure the error of how a ground truth probability distribution P diverges from a second approximate probability distribution Q [Bu *et al.* 2018; Brownlee 2019b]. Note that the ‘||’ operator indicates divergence as illustrated by equation 4.4.

$$KL(P||Q) \quad (4.4)$$

KL Divergence is commonly calculated as the positive sum of the probability of each event in distribution P multiplied by the log of the probability in P over the probability of the event in Q , as seen in equation 4.4. In order to get the negative sum of the probability of each event in P by the event in Q , the terms of the fraction are swapped, demonstrated by equation 4.5.

$$KL(P||Q) = - \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \quad (4.5)$$

$$KL(P||Q) = \sum_{x \in X} P(x) \log \left(\frac{Q(x)}{P(x)} \right) \quad (4.6)$$

From equations 4.5 and 4.6, the intuition is that when the probability for an event P is large and the probability for the same event in Q is small, there will be a large KL Divergence score. Similarly, when

there is a small probability of the event in P but the probability of the event in Q is large, there will also be a large KL Divergence score. KL Divergence is sometimes known as ‘relative entropy’ and it is important to note that it is not symmetrical, $KL(P||Q) \neq KL(Q||P)$ [Brownlee 2019b].

During this experiment a hardware limitation on the hill climbing method was discovered which made it impossible to evaluate a 400+ node model. The hill climbing score-based algorithm has a time complexity of $O(\infty)$ and a space complexity of $O(b)$, where b is the branching factor [Weld 2003]. In order to process the full set of courses, a supercomputer with sufficient memory and RAM is needed; therefore the Hill Climbing experiment is left incomplete. Only the Chow-Liu method will be used on real-world data.

To fully evaluate the Chow-Liu algorithm, the KL Divergence metric values are averaged over the 10 runs to get the mean value and standard deviations and are used to produce the plots that are seen in figures 4.5 and 4.6.

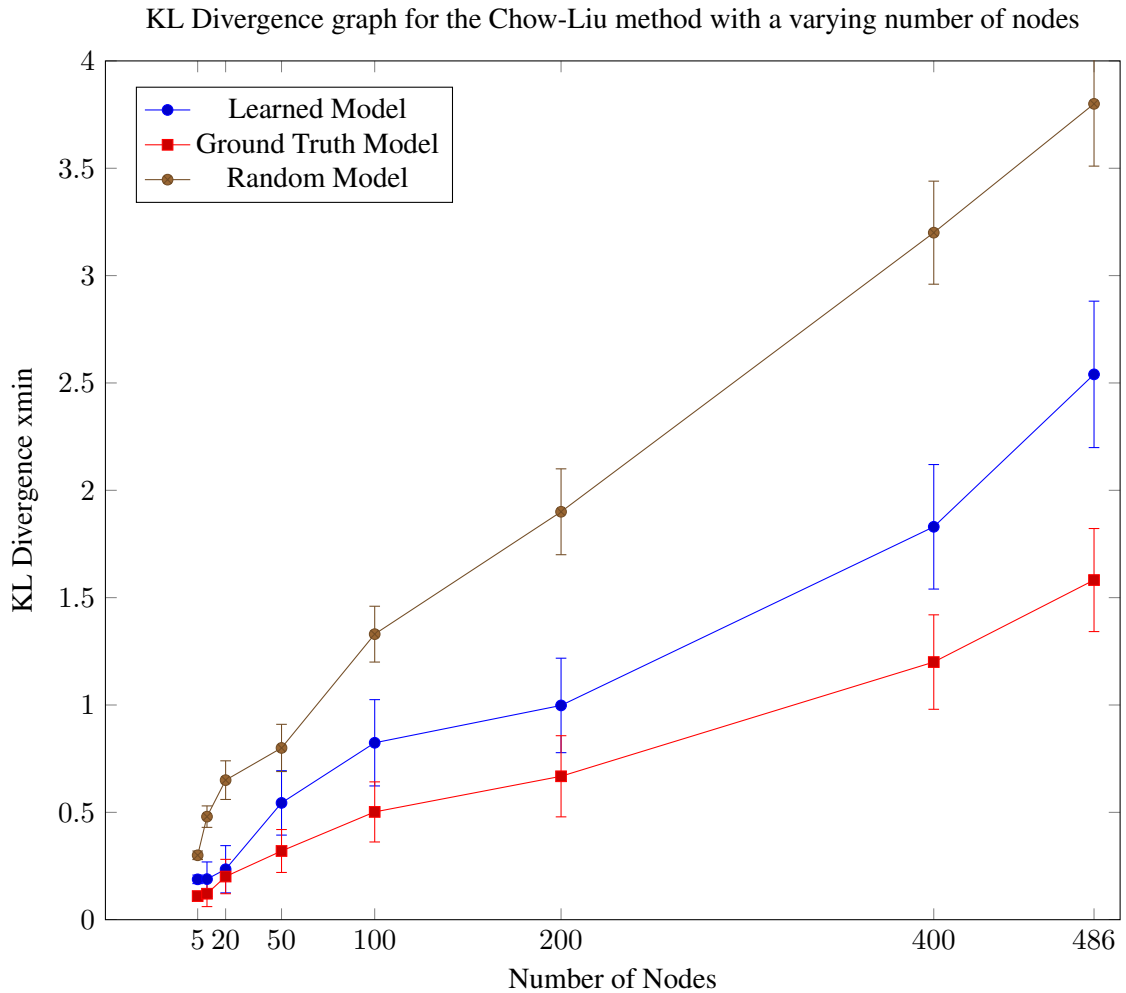


Figure 4.5: KL Divergence graph with varying number of nodes and 100 samples for the Chow-Liu algorithm

From the KL divergence graphs in figure 4.5, it can be deduced that as the numbers of nodes increase

the less reliable the results become and our error bars of standard deviation also increase. It is also possible to deduce that the learned model performs quite well in comparison to the ground truth model and significantly better than the random model.

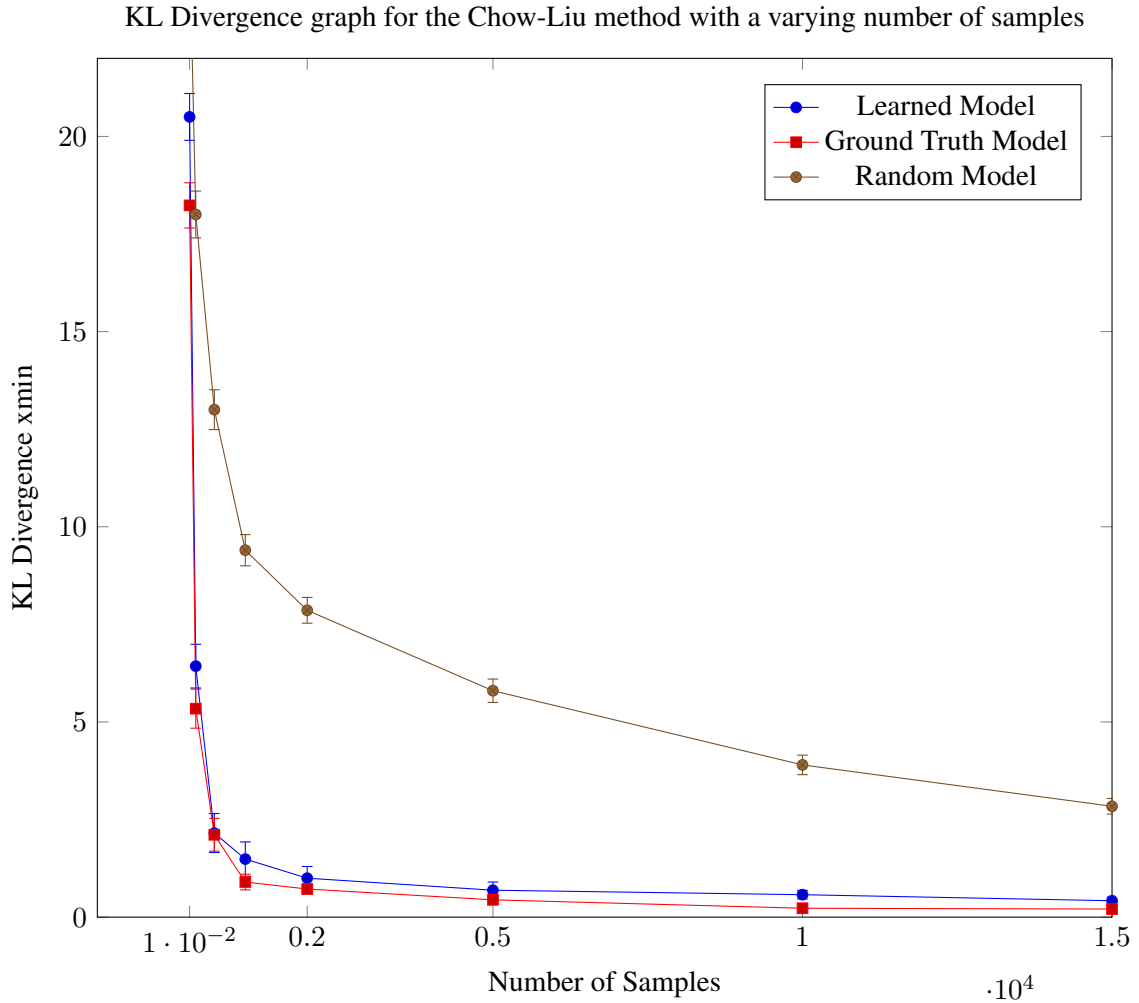


Figure 4.6: KL Divergence graph with varying samples sizes and 486 nodes for the Chow-Liu algorithm

Figure 4.6 on the other hand verifies that the more samples you have in the data set, the more reliable your results become and the more likely one is to find the true model formation. It can be concluded from these graphs that a reasonable number of nodes and a large sample size is required for convergence [Suzuki 2012]. The ground truth and learned model start to become very reliable at about 2000 (0.2×10^4) samples, whilst the random model needs many more than 15 000 samples to achieve the same performance.

Given the two KL Divergence graphs, for which the data point values can be found in Appendix B of this report, it is safe to say that the learned models strongly represent the ground truth models from which they are sampled. Therefore one can be assured that the Chow-Liu algorithm is producing consistent results and finding edges in accordance with the true model.

Chapter 5

Results and Discussion: Real World Data Model Evaluation

In Chapter 4, it is established that the learned models are fairly reliable and that the results seen from these models can be trusted within a certain degree of accuracy. In this chapter, the Chow-Liu method is used to learn a model from the real-world data set. The results of which will not only help refine course curricula at tertiary institutions for undergraduate programs; but also assist postgraduate students in making better-informed decisions about their coursework.

5.1 Undergraduate Course Refinement Results

The Chow-Liu method was able to produce a full graphical model with all 486 undergraduate subjects on offer in the faculty of science at this university from which the data is taken. These are subjects across all years of study between the years 2012 and 2017. It is important to note that only the best effort marks of each student are used from the data set, so if a student failed a subject previously but then went on to pass in the next year, the failing grade is ignored.

The results of this model are aimed to validate current course curriculum relationships and further find possible subjects that could aid students in their chosen field of study. Figure 5.1 depicts a small portion of the final learned model; and contains some interesting information that will be discussed in the next paragraph. For a high-level overview of what the complete graphical model looked like, please see Appendix C of this report.

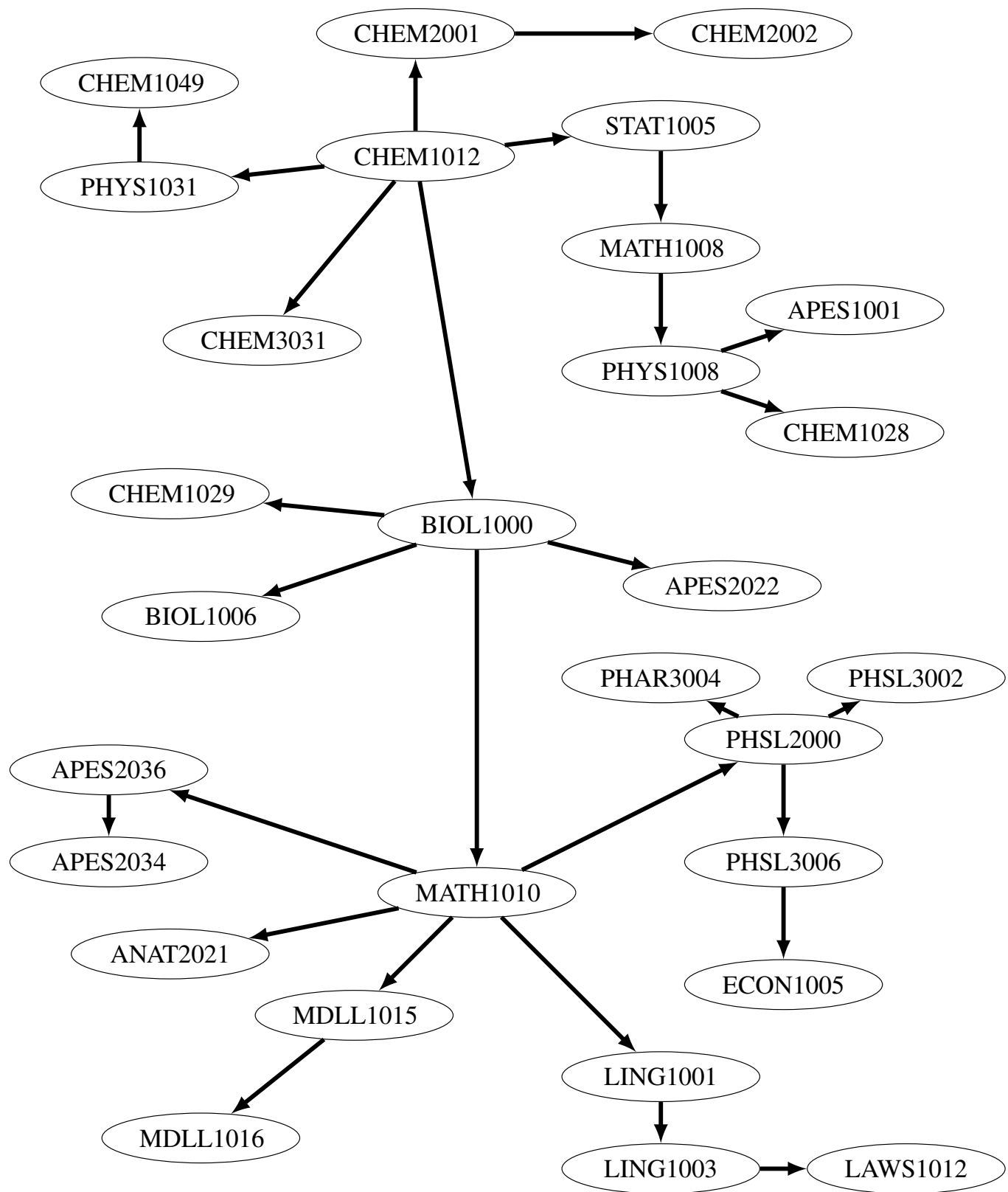


Figure 5.1: Part of the resultant undergraduate data model. The graphical model shows validation for courses that are already prescribed as co-requisites and prerequisites, and further finds relationships that exist between modules that are not yet prescribed.

Table 5.1: Table showing a key for the course codes to their respective course titles, which gives more insight into what each course is about in the faculty of science.

Course Code	Course Title
ANAT2021	Human and Comparative Biology II
APES1001	Introduction to Medical Sciences I
APES2022	Marine and Coastal Systems Fieldwork II
APES2034	Aquatic Ecology II
APES2036	Fundamentals of Ecology II
BIOL1000	Introductory Life Sciences
BIOL1006	Complementary Life Sciences I
CHEM1012	Chemistry I
CHEM1028	Chemistry (HC)
CHEM1029	Chemistry (HC)
CHEM1049	Chemistry (Auxiliary) PT
CHEM2001	Chemistry IIA (Major)
CHEM2002	Chemistry IIB (Major)
CHEM3031	Undergraduate Research in Applied Chemistry III
ECON1005	Economic Concepts IA
LING1001	Introduction to the Structure of Language I
LING1003	Language, Mind and Society
LAWS1012	Law of Persons
MATH1008	Ancillary Mathematics I
MATH1010	Ancillary Mathematics and Statistics I
MDLL1015	South African Sign Language: Basic IA
MDLL1016	South African Sign Language: Basic IB
PHAR3004	Pharmacology III
PHSL2000	Physiology II
PHSL3002	Applied and Experimental Physiology III
PHYS1008	Physics 1 (HC)
PHYS1031	Physics 1 (HC) Part-time
PHSL3006	Human Physiology III
STAT1005	Ancillary Statistics I

From figure 5.1 it can be seen that the learned graphical model picks up prerequisite relationships such as PHSL2002, Physiology II which is a second-year subject, being a foundational step to doing Applied and Experimental Physiology, PHSL3002, CHEM1012, Chemistry I, being foundational to CHEM2001 and CHEM2002, Chemistry IIA and Chemistry IIB, and so on. These relationships exist within the curriculum of the university already with proof in the course guideline for the faculty.

Also seen are co-requisite subjects that students benefit from taking in conjunction with one another, such as first-year subjects PHYS1008 (Physics I), APES1001 (Introduction to Medical Science I) and CHEM1028 (Chemistry (HC)). What is significant here is that the university course guidebook does not have PHYS1008 as a co-requisite to APES1001. Therefore it is worth investigating if this course should be added to the curriculum for students wishing to pursue a qualification in the medical science field.

As mentioned in section 4.3 of this report, not all relationships that exist between the courses can be demonstrated by the Chow-Liu method. Only strongly correlated subjects have edges between them. This means that there may be some subjects that have relationships with each other that are not fully

represented here. For example, APES2032 - Integrated Principles of Aquatic Ecology II is a co-requisite to APES2034 - Aquatic Ecology II, which is not depicted in this learned model.

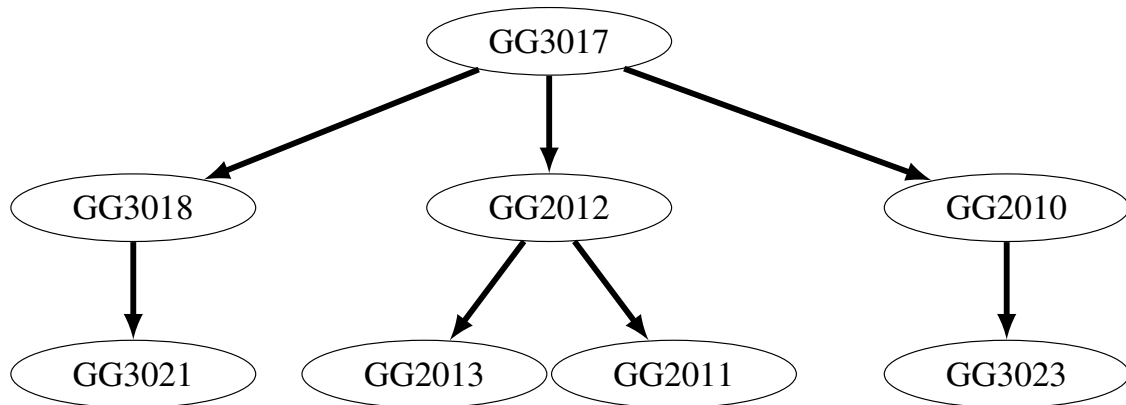


Figure 5.2: Part of the resultant Geographical Science cluster from the learned model

Table 5.2: Table showing a key for the course codes to their respective course titles, which gives more insight into what each course is about in the school of geographical science.

Course Code	Course Title
GG2010	Earth and Atmospheric Processes II
GG2011	An Introduction to Climate and Society II
GG2012	Environmental Governance: From Local to Global II
GG2013	Methods, Models and Geographic Information Systems II
GG3017	Geographic Information Systems and Remote Sensing III
GG3018	Fundamentals of Conservation Biogeography III
GG3021	Advanced Atmospheric Sciences III
GG3023	Theory and Practice in Sustainability Science and Sustainable Development III

In figure 5.2, it can be observed that GG3017, which is a third-year course in Geographic Information Systems and Remote Sensing, has arrows leading to second-year courses GG2012 and GG2010. Similarly, in figure 5.1 PHSY3006 has an arrow going to ECON1005, which is a fundamentals course in Economic Concepts. Thus it can be noted that correlation does not imply causation. Due to the difficulty in uncovering causal relationships in observational data, it is accepted that there is no way to distinguish the directionality of the arrows in a probabilistic model using observational data.

This experiment was conducted with uninformative priors. It was purely data-driven and yet still produced consistent results with each run of the algorithm. What is observed between runs is that sometimes the direction changes in the edge between two nodes. Due to the use of a score-based learning method, a class of networks called the I Equivalence Class is created with varying orientations of the edges that all have the same learning score. There is no statistical test available today to advise one on the correct orientation of these arrows [Binucci *et al.* 2022; Kurata 2007].

Whilst Bayesian Model Averaging can give you a probability mass of the preference of the orientation of the arrow, it is computationally impossible to achieve with a large number of nodes.

Curriculum learning is a topic in causation and the contributions made here to the field are of significant value. An important point to note, however, is that this approach of curriculum refinement may

not be possible for all tertiary institutions. Those institutions that have rigid course curricula with no course selection opportunities for undergraduate students will never find causation outside of that rigidly structured curriculum. This experiment worked well with the faculty of science at a research intensive higher-education institution due to the immense freedom given to students to pursue their credits in any field of interest outside of their basic primary science subjects.

5.2 Postgraduate Course Suggestions

Currently, at most tertiary-level institutions, postgraduate programs that include course selection have no prescribed curricula recommendations. Students pursuing honors or master's degrees by coursework are free to study any subject available within the school. Unless each student takes the time to delve into the offerings of each semester course, these accredited courses are chosen haphazardly according to criteria such as the course name, the lecturer taking the course and what previous students have said about the difficulty level of the course.

Using the real-world historic academic postgraduate data between the years 2012 and 2017, a learned model was produced using the score-based Chow-Liu structure learning method, with a K2 scoring approach, to produce some recommended pairings between the 430 postgraduate subjects. What this allows institutions to do is provide guidance to students on what subjects are closely related and help them to navigate a path to success in completing their postgraduate degrees via coursework.

Figure 5.3 shows a portion of the results from the postgraduate observational data. Based on the best effort marks of students between the years 2012 and 2017, it is seen that APES7019, a course on Energy and the Environment, is a good precursor to courses MINN7076, APES7018, APES7003 and APES7020, the course titles for which can be found in table 5.2. What is very interesting and adds to the validity of the model, is the fact that the algorithm only has information on the course code and marks for each student (as seen in table 3.3), yet the model learned to group and associate the relationships between environmental type subjects in the faculty of science. This means that the learned model shows a relationship between courses based solely on academic history, validating that this association is possible.

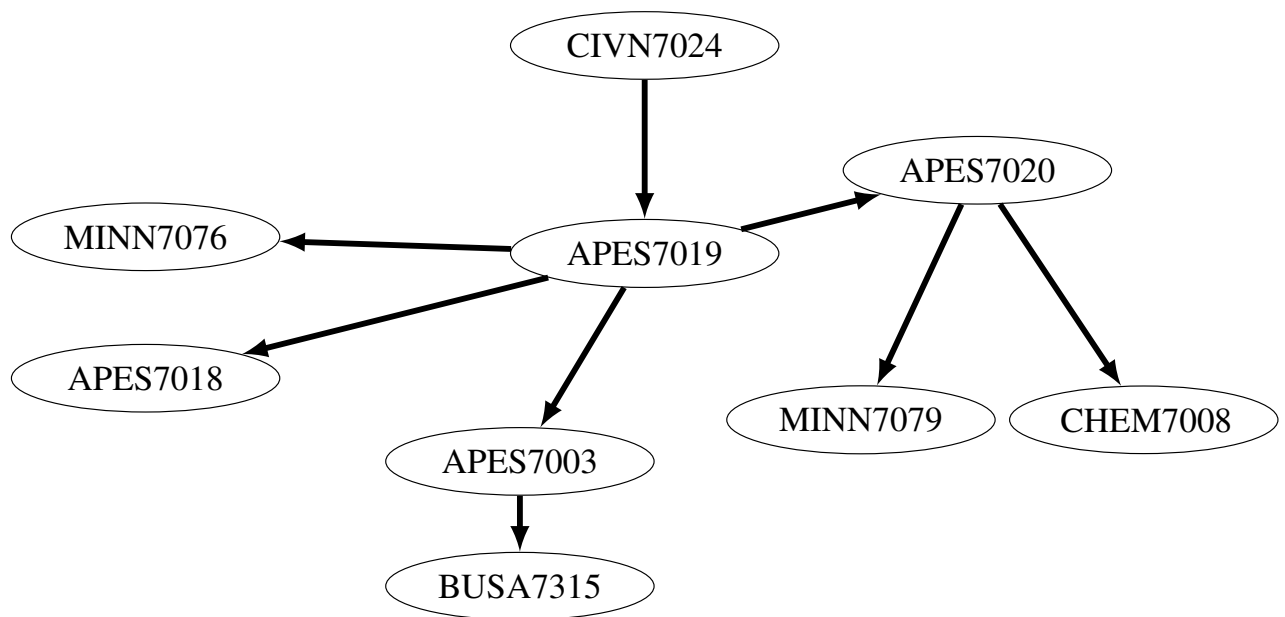


Figure 5.3: Part of the resultant model, clustered on masters subject offerings around the course code APES7019, for subject 'Energy and the Environment'

Course Code	Course Title
APES7003	Sustaining Populations and Resources: Frontiers
APES7018	Adaptive Management in Conservation and Catchment Management
APES7019	Energy and the Environment
APES7020	Environmental Impact Assessment - Concepts and Critical Review
BUSA7315	Business of the Environment
CHEM7008	Environmental Chemistry
CIVN7024	Environmental Management
MINN7076	Sustainable Development in Mining and Industry
MINN7079	Safety, Health, Environment and Community Systems

Table 5.3: Table showing a key for the course codes to their respective course titles, which gives more insight into what the course is about.

Whilst using these models may not stop students from choosing postgraduate subjects haphazardly based on the personal preference criteria mentioned earlier, it can certainly help interested students to align their subjects with a specific field of interest or on subject matter pertinent to their research topics. This will help improve postgraduate success rates at institutions that do not have prescribed postgraduate curricula by keeping students focused on their key areas.

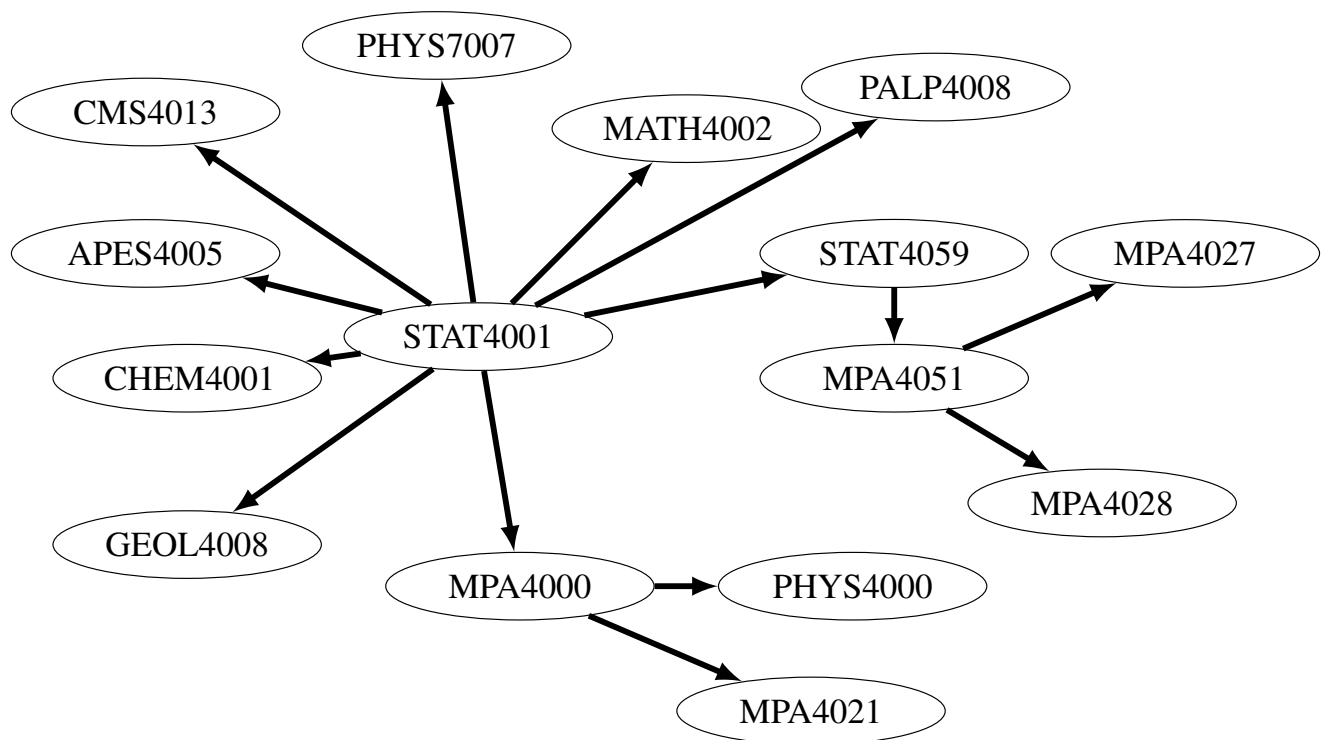


Figure 5.4: Part of the resultant model on honors subjects centered around the course code STAT4001, for subject 'Actuarial Science Honours FT'. Upon closer inspection, it can be seen that courses with very few data points (under 10 students attempted the course during the 6 year period) converged here.

Course Code	Course Title
APES4005	Zoology Honours PT2
MPA4000	Applied Mathematics Honours
MPA4021	Advanced Mathematics of Finance Honours FT
MPA4027	Applications of Stochastic Processes
MPA4028	Portfolio Optimisation
MPA4051	Introduction to Derivative Securities
CHEM4001	Applied Chemistry Honours
CMS4013	Computer Science Honours PT2
GEOL4008	Geology Honours Ext
MATH4002	Mathematics Honours PT2
PALP4008	Palaeontology Honours Part II - Jan to June
PHYS4000	Physics (Exp and Theory) Honours FT
PHYS7007	PHYS MSc Research Report
STAT4001	Actuarial Science Honours FT
STAT4059	Actuarial Science Honours Ext

Table 5.4: Table showing a key for the course codes to their respective course titles, which gives more insight into what each course is about; where FT indicates 'Full Time' studies and PT indicates 'Part Time' studies.

Figure 5.4 demonstrates that relationships may be found for other statistical reasons, such as in this case where the data points are low in number, the courses were grouped around STAT4001, which is a

full-time actuarial science honors course. This means that the model is not necessarily recommending that these courses aid each other in conceptual understanding but rather that they share the characteristic of not having enough data to make meaningful conclusions on their causality. It is therefore important to validate these results with information from the respective schools and historic data before making changes to any course curricula.

Chapter 6

Conclusion

In this research, the ultimate goal was to use probabilistic graphical modelling to help tertiary institutions better their course curricula using only historic academic data. This is done to aid improvement in student success rates in South Africa and move the country forward through higher education. Section 6.1 discusses the findings and implications of this report. Section 6.2 highlights the contributions made to the fields of curriculum learning and EDM. Based on the outcomes of this research, section 6.3 provides ideas on possible future work.

6.1 Implications

From the results in chapter 4, it was found the learned model attained using the Chow-Liu algorithm performed relatively better than the randomly generated model and generally well by the ground truth model in the KL-Divergence graphs. This proved the validity of the learned models and allows the research to proceed using the real-world observational data obtained from a South African university. What the graphs also demonstrated was that as the number of nodes increases, the performance of the algorithm becomes slightly less reliable; however, as the number of samples increases with a large number of nodes, the output model becomes more reliable.

Chapter 5 on the results of the real-world data indicates two key findings; one is to help validate and refine course curriculum for different undergraduate programs in the field of science, and the second is to provide guidance to postgraduate students on how to select courses that align to their strengths, field of interest or on their research.

The Chow-Liu algorithm was able to find existing relationships between subject courses offered that are implemented in the school already according to the course guidebook for undergraduate degrees. It is noted that not all existing relationships were indicated in the learned model; this could be due to the limitations of the Chow-Liu method not finding a fully connected graph or a suggestion by the data that maybe these subjects do not need each other for student success. Due to the tree structure of the method, every node has at most one parent, meaning that co-requisite subjects are often of different levels on the tree but still highlights the significance of the need for one subject to aid in the completion of the other.

What the model also demonstrated was there exist relationships between offered courses that are not part of the prescribed curriculum. This is powerful in that one can now use the outcome model to make

suggestions to better inform the current curriculum and thereby improve student success in undergraduate programs by providing them with all required foundational knowledge in their field of study.

Another powerful finding is that the Chow-Liu method, using postgraduate data, is able to find relationships between postgraduate courses given only historic academic marks and course codes. This can greatly aid students in making beneficial subject selections for their coursework accreditation.

6.2 Contributions

The contributions of this research to fields of study in EDM, probabilistic graphical modelling and curriculum learning are as follows:

- Providing a probabilistic graphical recommendation to help structure undergraduate course curriculum at tertiary institutions based on historic academic data from students of that institution. Proving that unbiased graphical models can be created without the influence of socioeconomic and demographic factors.
- Creating a postgraduate guide on course selection for students based on historic postgraduate data for the research-intensive higher-education institution that the data is collected from.
- Contribution to the field of curriculum learning, which is a topic in causation, on how intelligent systems can inform curriculum structuring without human bias or error.

6.3 Future Work

This study merely scratches the surface of the possibilities of how probabilistic graphical modelling can be used in curriculum learning at higher education institutions. However, it must be remembered that these methods can only work at institutions or departments that do not have rigidly fixed curricula and allow some degree of subject selection by students.

A furthering of this study would be to extend the research by using other high-performance score-based learning methods and comparing them to the Chow-Liu method presented in this report. It would be of great value to see the limitations of the hill climbing method overcome by creating performance enhancement for the algorithm through the use of multi-threading and running the algorithm on a super-computer or even in the cloud.

It would also be interesting to gather a larger range of data; over a few decades, up to the present, to see if and how trends change in course recommendations with different generations of students. For example, a student attending university now in the faculty of science would be much more well-versed with computer technology and coding practices than a student that did not grow up with computers from the 1980s. This is bound to affect course prerequisite subject recommendations in the field.

Lastly, this study could be furthered to secondary schooling to aid high school students in subject selections by using nationwide data. This would work well for schools that do not have predefined subject selection courses available and where students are free to do any combination of electives that they want. The graphical model would be able to inform students on which selections prove most successful

together; for example, taking pure Mathematics improves your outcome in Accounting.

There are boundless directions in which to take this study forward and make great impacts in societies that need more highly qualified workers to better the living conditions of the society as a whole. Bettering educational systems and setting students up for every opportunity to succeed has no downside and using advancements in structural graphical learning to do so is a wonderful way to ensure that human error and bias are left out of the equation.

References

- [AJ *et al.* 2019] Sedgewick AJ, Buschur K, Shi I, Ramsey JD, Raghu VK, Manatakis DV, Zhang Y, Bon J, Chandra D, Karoleski C, Sciurba FC, Spirtes P, Glymour C, and Benos PV. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. In *Oxford Journals, Bioinformatics (Oxford, England)*, pages 1204–1212. Oxford University Press, April 2019.
- [Alsaad *et al.* 2019] Fareedah Alsaad, Assma Boughoula, Hari Sundaram, and Chengxiang Zhai. Mining mooc lecture transcripts to construct concept dependency graphs. 08 2019.
- [Anandkumar *et al.* 2011] Animashree Anandkumar, Vincent Y. F. Tan, and Alan S. Willsky. High-dimensional graphical model selection: Tractable graph families and necessary conditions. In *NIPS*, pages 1863–1871, 2011.
- [Ball *et al.* 2019] Robert Ball, Linda DuHadway, Kyle Feuz, Joshua Jensen, Brian Rague, and Drew Weidman. Applying machine learning to improve curriculum design. pages 787–793, 02 2019.
- [Bhattacharyya *et al.* 2021] Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and N. V. Vinodchandran. Near-optimal learning of tree-structured distributions by chow-liu. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2021*, page 147–160, New York, NY, USA, 2021. Association for Computing Machinery.
- [Binucci *et al.* 2022] Carla Binucci, Walter Didimo, Michael Kaufmann, Giuseppe Liotta, and Fabrizio Montecchiani. Placing arrows in directed graph layouts: Algorithms and experiments. *Computer Graphics Forum*, 41:364–376, 01 2022.
- [Bouckaert 1994] Remco R. Bouckaert. Properties of bayesian belief network learning algorithms. In *UAI*, 1994.
- [Brownlee 2019a] Jason Brownlee. *A Gentle Introduction to Maximum Likelihood Estimation for Machine Learning*, Oct 2019.
- [Brownlee 2019b] Jason Brownlee. *How to Calculate the KL Divergence for Machine Learning*, Oct 2019.
- [Bu *et al.* 2018] Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V. Veeravalli. Estimation of kl divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4):2648–2674, 2018.
- [Carvalho 2021] Alexandra M. Carvalho. *Scoring functions for learning Bayesian networks*, 2021.
- [Chen *et al.* 2016] Yetian Chen, José P. González-Brenes, and Jin Tian. Joint discovery of skill prerequisite graphs and student models. In *EDM*, 2016.

- [Chickering and Meek 2015] David Chickering and Christopher Meek. Selective greedy equivalence search: Finding optimal bayesian networks using a polynomial number of score evaluations. 06 2015.
- [Chickering 2003] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, mar 2003.
- [Chickering 2020] Max Chickering. Statistically efficient greedy equivalence search. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 241–249. PMLR, 03-06 Aug 2020.
- [Chinnnasamy *et al.* 2022] Sathiyaraj Chinnnasamy, M. Ramachandran, M. Amudha, and Kurinjimalar Ramu. *A Review on Hill Climbing Optimization Methodology*, Jan 2022.
- [Conati *et al.* 2004] Cristina Conati, Abigail S. Gertner, and Kurt VanLehn. Using bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, 12:371–417, 2004.
- [Culbertson 2014] Michael J. Culbertson. Graphical models for student knowledge: Networks, parameters, and item selection. 2014.
- [de Campos 2006] Luis M. de Campos. Ascoring function for learning bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7:2149–2187, 10 2006.
- [Desana and Schnörr 2020] Mattia Desana and Christoph Schnörr. Sum-product graphical models. *Machine Learning*, 109(1):135–173, 2020.
- [Dukhan *et al.* 2012] Shalini Dukhan, Ann Cameron, and Elisabeth Brenner. The influence of differences in social and cultural capital on students’ expectations of achievement, on their performance, and on their learning practices in the first year at university. *International Journal of Learning*, 18:337, 01 2012.
- [Dukhan 2020] Shalini Dukhan. Value for learning during this time of transformation: the first-year students’ perspective. *Higher Education Research & Development*, 39(1):39–52, 2020.
- [Epstein 1969] Edward S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982)*, 8(6):985–987, 1969.
- [Good 1952] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14:107–114, 1952.
- [Hassan *et al.* 2020] Hasniza Hassan, Nor Bahiah Ahmad, and Syahid Anuar. Improved students’ performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. *Journal of Physics: Conference Series*, 1529(5):052041, may 2020.
- [HC Peng and Ding 2012] F Long HC Peng and C Ding. Feature selection based on mutual information : criteria of max-dependency, max relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2012.
- [Hlomuka *et al.* 2021] Siyabnga Hlomuka, Shalini Dukhan, and Ritesh Ajoodha. To what extent do mark expectations of first years align with their academic achievement within biology based degrees at a south african university? In *2021 3rd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–5, 2021.

- [Ilieva and Yankova 2020] Galina Ilieva and Tania Yankova. Early multi-criteria detection of students at risk of failure. *TEM Journal*, 1:344–350, 2020.
- [Karger and Srebro 2001] David Karger and Nathan Srebro. Learning markov networks: Maximum bounded tree-width graphs. 08 2001.
- [Koller and Friedman 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 01 2009.
- [Kurata 2007] Yohei Kurata. Arrow symbols: Theory for interpretation. 2007.
- [Lacave *et al.* 2018] Carmen Lacave, Ana Isabel Molina, and José A. Cruz-Lemus. Learning analytics to identify dropout factors of computer science studies through bayesian networks. *Behaviour & Information Technology*, 37:1007 – 993, 2018.
- [Lauritzen 1996] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.
- [Liu 2009] Chao-Lin Liu. Selecting bayesian-network models based on simulated expectation. *Behaviormetrika*, 36:1–25, 2009.
- [Mashiloane and Mchunu 2013] Lebogang Mashiloane and Mike Mchunu. Mining for marks: A comparison of classification algorithms when predicting academic performance to identify “students at risk”. In Rajendra Prasath and T. Kathirvalavakumar, editors, *Mining Intelligence and Knowledge Exploration*, pages 541–552, Cham, 2013. Springer International Publishing.
- [Mti *et al.* 2021] Xolani Mti, Shalini Dukhan, and Ritesh Ajoodha. The influence of clarity on career choice and academic success in first year. In *2021 3rd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–6, 2021.
- [Nabil *et al.* 2021] Aya Nabil, Mohammed Seyam, and Ahmed Abou-Elfetouh. Prediction of students’ academic performance based on courses’ grades using deep neural networks. *IEEE Access*, 9:140731–140746, 2021.
- [Nudelman *et al.* 2019] Zachary Nudelman, Deshendran Moodley, and Sonia Berman. Using bayesian networks and machine learning to predict computer science success. In *47th Annual Conference of the Southern African Computer Lecturers’ Association, SACLA 2018*, pages 207–222, 01 2019.
- [Onisko and Druzdzel 2014] Agnieszka Onisko and Marek J Druzdzel. Impact of bayesian network model structure on the accuracy of medical diagnostic systems. In *ICAISC*, 2014.
- [Philippou *et al.* 2020] Nastassja Philippou, Ritesh Ajoodha, and Ashwini Jadhav. Using machine learning techniques and matric grades to predict the success of first year university students. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–5, 2020.
- [Purita P. Bilbao 2008] et. al Purita P. Bilbao. Curriculum development. 2008.
- [Rawatlal 2017] Randhir Rawatlal. Application of machine learning to curriculum design analysis. *Computing Conference 2017*, 07 2017.
- [Rhodes 2018] Rhodes. Brief guide for designing a curriculum and constructing a course document. 2018.
- [Riaz and Ali 2011] Ferozuddin Riaz and Khidir M. Ali. Applications of graph theory in computer science. In *2011 Third International Conference on Computational Intelligence, Communication Systems and Networks*, pages 142–145, 2011.

- [Rubin 1984] Donald B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–1172, 1984.
- [Saranti *et al.* 2019] Anna Saranti, Behnam Taraghi, Martin Ebner, and Andreas Holzinger. *Insights into Learning Competence Through Probabilistic Graphical Models*, pages 250–271. 08 2019.
- [Sarkar *et al.* 2017] S. Sarkar, Z. Jiang, A. Akintayo, S. Krishnamurthy, and A. Tewari. Chapter 18 - probabilistic graphical modeling of distributed cyber-physical systems. In Houbing Song, Danda B. Rawat, Sabina Jeschke, and Christian Brecher, editors, *Cyber-Physical Systems, Intelligent Data-Centric Systems*, pages 265–285. Academic Press, Boston, 2017.
- [Scanagatta *et al.* 2019] Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. A survey on bayesian network structure learning from data. *Progress in Artificial Intelligence*, pages 1–15, 2019.
- [Selman and Gomes 2001] Bart Selman and Carla P Gomes. Hill-climbing search. pages 333–336, 2001.
- [Shalem *et al.* 2014] Bar Shalem, Yoram Bachrach, John Guiver, and Christopher M. Bishop. Students, teachers, exams and moocs: Predicting and optimizing attainment in web-based education using a probabilistic graphical model. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 82–97, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.
- [Sinharay *et al.* 2004] Sandip Sinharay, Russell Almond, and Duanli Yan. Assessing fit of models with discrete proficiency variables in educational assessment. *ETS Research Report Series*, 2004, 06 2004.
- [Suzuki 2012] Joe Suzuki. The bayesian chow-liu algorithm. *IEEE Transactions on Information Theory*, pages 315–322, 2012.
- [Thiele *et al.* 2016] Tamara Thiele, Alex Singleton, Daniel Pope, and Debbi Stanistreet. Predicting students’ academic performance based on school and socio-demographic characteristics. *Studies in Higher Education*, 41:1424 – 1446, 2016.
- [Warren 1966] Weaver Warren. PROBABILITY, RARITY, INTEREST, AND SURPRISE. *Pediatrics*, 38(4):667–670, 10 1966.
- [Weld 2003] Dan Weld. Notes on the complexity of search. 09 2003.
- [West *et al.* 2012] Patti West, Daisy Rutstein, Robert Mislevy, Junhui Liu, Roy Levy, Kristen Dicerbo, Aaron Crawford, Younyoung Choi, Kristina Chapple, and John Behrens. *A Bayesian Network Approach To Modeling Learning Progressions*, pages 257–292. 01 2012.
- [Williamson *et al.* 2000] David M. Williamson, Russell Almond, and Robert Mislevy. Model criticism of bayesian networks with latent variables. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI ’00, page 634–643, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- [Winter 2016] Chantal Winter. Refining and revising curriculum. 2016.

Appendix A

Table 6.1: Table showing the percentage of missing data from 149 978 entries for each data column in the undergraduate real-world data set between the years 2008 and 2018 for the faculty of science at a South African University

Column Name	% Missing
Program Faculty Name	0.00
Academic Career	0.00
Program Code	0.00
Program Title	0.00
Plan Code	0.00
Plan Description	0.00
Year of Study	0.00
Registration Status	0.00
Progress Outcome Type	0.00
Progress Outcome Type Description	0.00
New to Program	0.00
New to University	0.00
Course Code	0.00
Course Title	0.00
Final Mark	0.01
Final Grade	0.01
Course Attempt Status	0.00
Encrypted Student No	0.00
Gender	0.00
Race Description	0.00
Age	0.21
Marital Description	0.46
Home Language Code	0.71
Home Language Description	15.64
Religion Description	100.00
Citizenship Status	100.00
Nationality Short Name	100.00
Country Of Birth	100.00
Permit Type Description	100.00
Sims Urban Rural	100.00
Sims Quintile	100.00
Secondary School Quintile	100.00
Secondary School Code	100.00

Secondary School Name	100.00
HR Address1	100.00
HR Address2	100.00
HR Address3	100.00
HR Address4	100.00
HR City	100.00
HR Postal Code	100.00
HR Province	100.00
HR Country	100.00
HR Country Name	100.00
HP Address1	100.00
HP Address2	100.00
HP Address3	100.00
HP Address4	100.00
HP City	100.00
HP Postal Code	100.00
HP Province	100.00
HP Country	100.00

Appendix B

Figure 6.1 shows the 10 node, 100 sample experiment with a the ground truth depiction and the 10 sample learned models.

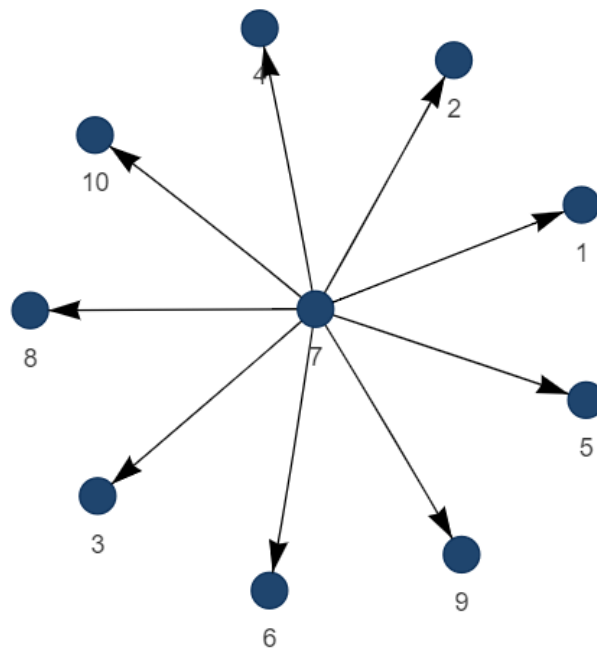


Figure 6.1: Ground Truth Chow-Liu Model

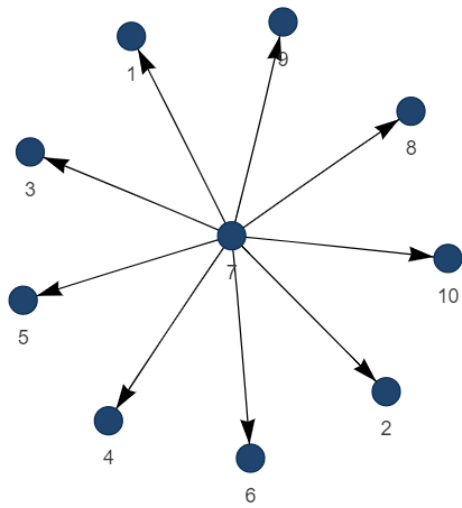


Figure 6.2: Learned Model 1

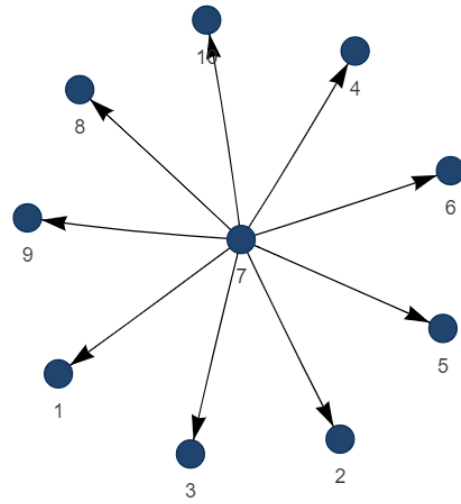


Figure 6.3: Learned Model 2

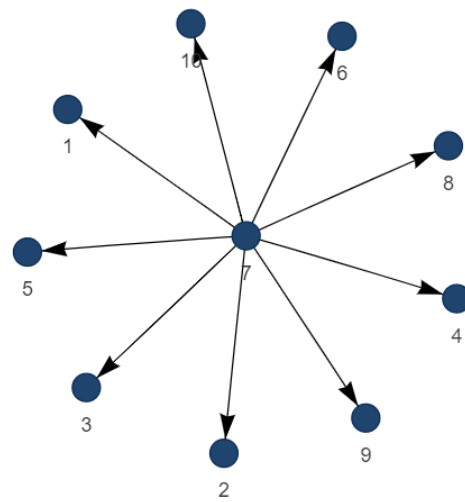


Figure 6.4: Learned Model 3

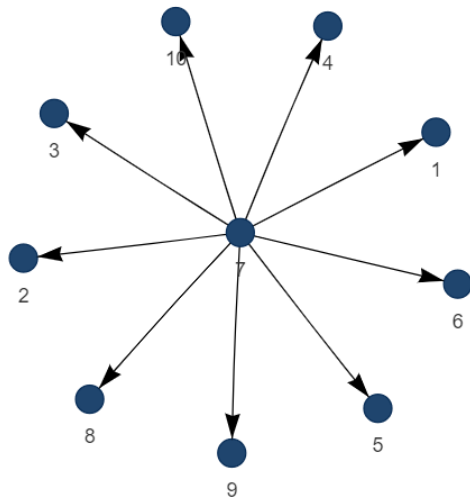


Figure 6.5: Learned Model 4

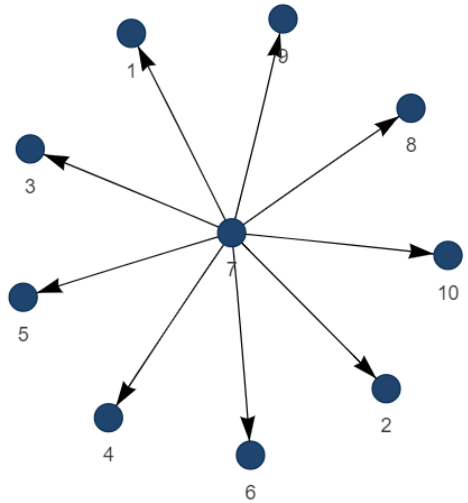


Figure 6.7: Learned Model 6

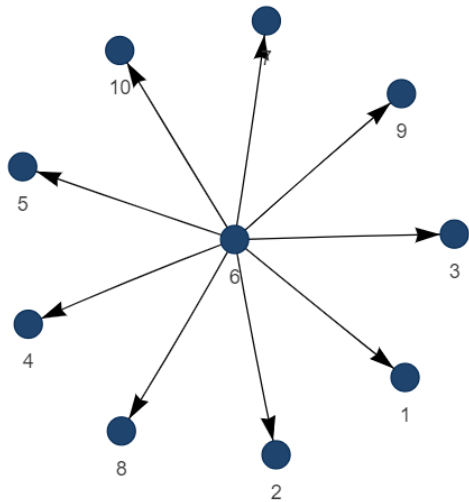


Figure 6.6: Learned Model 5

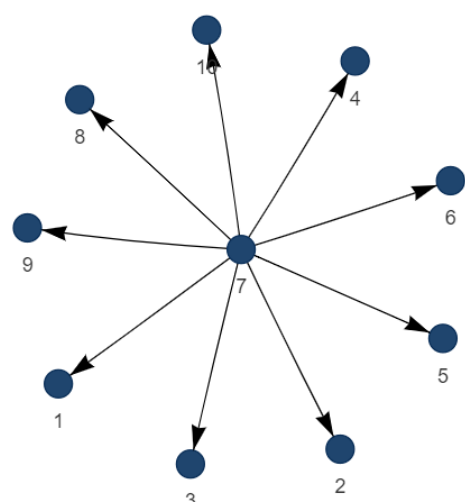


Figure 6.8: Learned Model 7

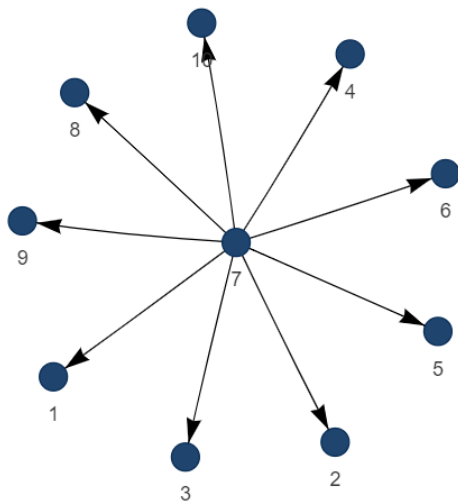


Figure 6.9: Learned Model 8

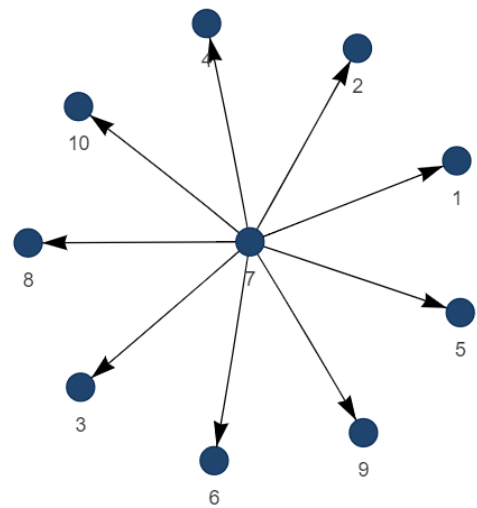


Figure 6.11: Learned Model 10

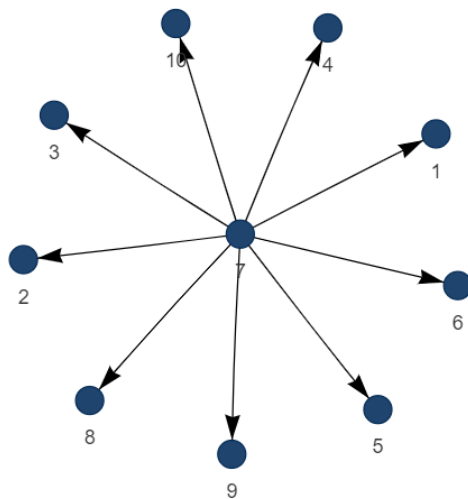


Figure 6.10: Learned Model 9

Table 6.3: Table showing the values used to produce figure 4.5 on the KL Divergence graph that varied the number of nodes in the network.

	Ground Truth		Learned Model		Random Model	
Nodes	Divergence	Std Dev	Divergence	Std Dev	Divergence	Std Dev
5	0.11	0.02	0.188	0.02	0.3	0.02
10	0.121	0.06	0.189	0.08	0.48	0.05
20	0.201	0.079	0.235	0.11	0.65	0.09
50	0.32	0.1	0.544	0.15	0.8	0.11
100	0.502	0.14	0.824	0.201	1.33	0.13
200	0.668	0.189	0.998	0.22	1.9	0.2
400	1.2	0.22	1.83	0.29	3.2	0.24
486	1.582	0.243	2.54	0.341	3.8	0.29

Table 6.5: Table showing the values used to produce figure 4.6 on the KL Divergence graph that varied the number of samples in the network.

	Ground Truth		Learned Model		Random Model	
Samples	Divergence	Std Dev	Divergence	Std Dev	Divergence	Std Dev
100	18.234	0.58	20.5	0.6	25	0.779
200	5.34	0.5	6.43	0.56	18.1	0.6
500	2.11	0.429	2.154	0.5	13.5	0.51
1000	0.9	0.2	1.487	0.44	9.4	0.4
2000	0.72	0.101	0.996	0.3	7.86	0.33
5000	0.443	0.04	0.69	0.21	5.81	0.3
10000	0.228	0.02	0.575	0.11	3.9	0.25
15000	0.207	0.003	0.42	0.09	2.84	0.2

Appendix C

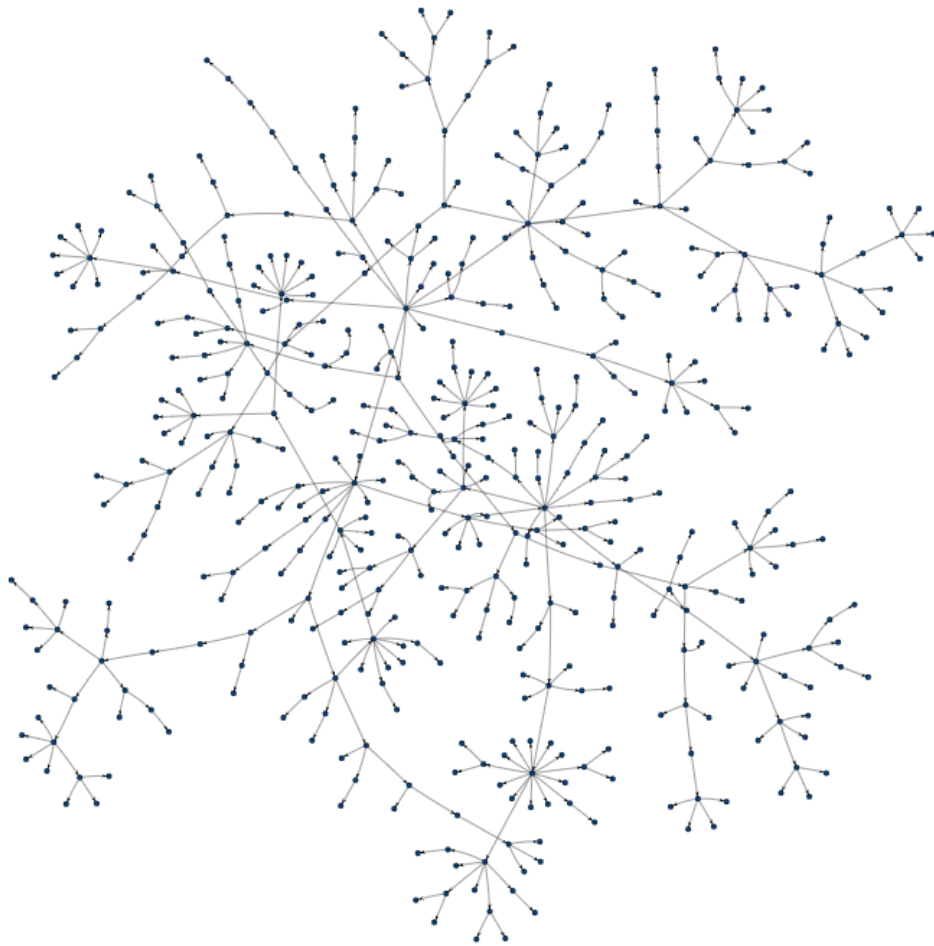


Figure 6.12: Complete Chow-Liu Undergraduate graphical model, including all courses offered in the faculty of science between the years 2012 and 2017

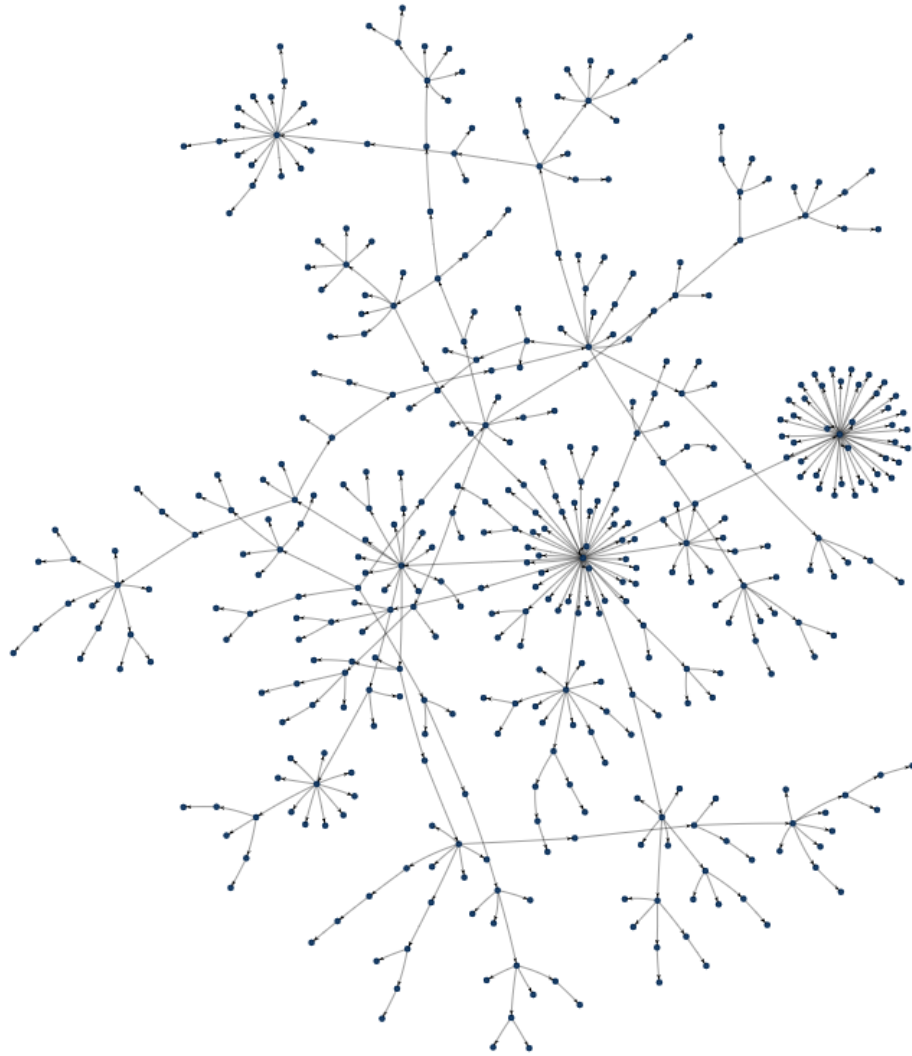


Figure 6.13: Complete Chow-Liu Postgraduate graphical model, including honors and masters courses offered in the faculty of science between the years 2012 and 2017