Journal of the Statistical and Social Inquiry Society of Ireland
Vol. XXXVI

BARRINGTON LECTURE 2006/07

ASSOCIATION RULE ANALYSIS OF CAO DATA

P.D. McNicholas

Trinity College Dublin

(read before the Society, 30 November 2006)

This lecture is delivered under the auspices of the Barrington Trust (founded by the bequest of John Barrington, Esq.) with the collaboration of the Journal of the Statistical and Social Inquiry Society of Ireland.

***Abstract*:** Central Applications Office (CAO) application data is analysed using a data mining technique, association rule mining, to investigate relationships between course choices across applicants. The role of gender as a factor in course selection is examined as well as a larger question around the functionality of the application system – what attracts students to a course; is it a topic of interest or is it the perceived status of the course associated with high entry points? The expected gender imbalances in areas like primary teaching and engineering appear, along with some others. Association rules generated suggest that students select courses based primarily on topic but sometimes with geographical location in mind. No evidence is found to suggest that students are selecting courses based on points status. Further in-depth analysis was carried out on two subgroups of students – those who applied for at least one medicine course and those who applied for at least one law course. Once again, the resulting association rules give little or no evidence that applicants are selecting courses based on points status.

Keywords: Association rules, college application, Central Applications Office, CAO, points race, college entry
JEL Classifications: C10, I21

1. INTRODUCTION

1.1 Background

On Monday 21st August 2006 the Irish Times carried an article entitled *'Points race' may be over as CAO requirements tumble*. The author, Seán Flynn, explained that the points race may be coming to an end; a claim supported in the article by John McGinnity, deputy registrar at NUI Maynooth (NUIM), who was quoted as stating that "this year has seen a rebalancing between the supply and demand for places."

The term 'points race' is used to describe the struggle to attain points; the metric by which school leavers have come to judge their Leaving Certificates. In past generations the vital statistic in assessing the strength of one's Leaving Certificate was the number of passes and honours attained.

Today however, the assessment is usually given in cold hard points, so much so that the attainment of points has become, for many, an end as well as a means.

Points are awarded by the Central Applications Office (CAO) on the basis of grades attained at the Leaving Certificate examination. The intended purpose of these points is to allow colleges to decide which students should get into which courses. In this sense, it is understandable that students should want to get as many points as possible; the more points attained, the less likely a student is to be pipped at the finishing line and lose out on their preferred college course.

However, it would be a disturbing prospect if students were choosing their courses based on points value rather than on topic; if the status of taking a college course with high points outweighed the wisdom of taking one that the student might enjoy or even excel at. This would be a points race in a different sense: points for points-sake. Considering that the points requirement for a course is dictated by the demand for entry to the course versus the amount of places available, it is not necessarily true that courses requiring higher points are better than those requiring lower points. Furthermore, in a situation where this type of points race existed, our prospects of succeeding as a knowledge-based society in the future would have to viewed as questionable.

1.2 Literature

Tuohy (1998)[1] completed the first research paper for the Commission on the Points System, which was established in 1997 by then Minister for Education and Science, Micheál Martin. Amongst his conclusions, the author found that gender may have been an issue in course choice. Moreover, he maintained that the answer to the question over whether or not the points system directly affected students' choices was unclear.

Lynch *et al.* (1999) completed the Points Commission's fourth research paper. Amongst their conclusions, the authors stated that "while there is a relationship between LCGPA [Leaving Certificate grade point average] and performance, students with identical Leaving Certificate grades display very different performance outcomes in higher education."

The final report of the commission, including recommendations (Hyland, 1999)[2] found that the Leaving Certificate was a valid method of predicting higher education performance, pointing to Lynch *et al.* (1999) for substantiation. In fact, the report largely vindicated the Leaving Certificate together with the CAO system as a college entrance system. The report highlighted application for medicine courses as an exceptional case and recommended further investigation by the relevant state and healthcare bodies.

Ideas around allocation of bonus points for subjects especially relevant to third-level course of interest, for an applicant's preferred courses, or for consistency of choice in completing the CAO form were considered but none were thought favourable. Furthermore, it was recommended that the practice of awarding bonus points for higher level mathematics in some institutions should end; as should the awarding of extra points to students who answer through Irish.

Clancy (2001) completed an extensive study of access to higher education. This analysis was conducted via questionnaires and raised some important questions around the availability of third level education to would-be applicants from various socioeconomic backgrounds. The effect of gender was also examined. An updated version of this study was carried out, by O'Connell *et al.*

[1] Tuohy's analysis included the utilisation of repeated factor analyses on binary data plus another variant on principal components analysis offered in SPSS.

[2] Áine Hyland was the chairperson. A full list of authors is given on page 175 of the report.

(2006), based on applicants in 2004. Like Clancy (2001), O'Connell *et al.* (2006) found that socioeconomic background was having a notable effect on the profile of entrants to various third-level courses and that more females than males were entering third-level education, with the stereotypical gender biases still present in areas like engineering and primary teaching.

O'Connell *et al.* (2006) also reported that the principal factor influencing a student's choice of college was the reputation of the institution (37% of students' chosen "main factor"), followed by the proximity of the institution to their home (21%) and the fact that the chosen course was not available elsewhere (17%). The two major factors influencing the choice of course within institution were the students' interest in the topic (57%) and the quality of qualification on offer (27%). These figures are positive as a reflection on the application process but they do vary across the different types of higher level institution.

Gormley & Murphy (2006) analysed CAO data from the year 2000 by assuming the existence of an underlying mixture model, with a Plackett-Luce model (Plackett, 1975) used for each mixture component. This analysis yielded a model with 22 mixture components, which is essentially claiming that there are 22 different 'groups' of students filling in CAO forms; 21 subject-based groups plus a noise group. These results support the CAO system since no group emerged that chose courses that were linked only by high points value and not by topic.

They went further and looked at the subset of students that their model assigned to the health sciences component, listing the 30 courses with the greatest probability of appearing on their CAO forms. The medicine courses across the five institutions offering medicine – University College Dublin (UCD), Trinity College Dublin (TCD), National University of Ireland Galway (NUIG), University College Cork (UCC) and the Royal College of Surgeons in Ireland (RCSI) – topped this list with probabilities of 0.47, 0.24, 0.20, 0.12 and 0.06 respectively. Two law courses and an engineering course were amongst these 30, with probabilities in the range 0.0083–0.0091; the authors ascertain that this could be regarded as adding "some weight" to the argument that a points race does exist. This however must be taken in context with the fact that the arts degree offered at UCD also makes this top thirty, coming in with probability 0.0092.

Furthermore, Gormley & Murphy (2006) then considered the courses that, if present on an applicant's CAO form, gave the highest probabilities of belonging to the health sciences component. Of the 25 courses listed, there were two courses, both of which had high entry points, that seemed out of place. However, both of these courses, which are part of the two-subject moderatorship offered at TCD, had relatively few applicants (three and 35 respectively). The authors concluded that the points race is more prevalent amongst the health sciences courses than elsewhere.

Lynch *et al.* (2006) looked at the CAO system as an admission system for dentistry courses. They concluded that while there was a statistically significant relationship between Leaving Certificate performance and first year dental examination results, there was "no association between the number of points achieved by students in their Leaving Certificate examination and their performance in the final dental examination."

Interestingly, Moran & Crowley (1979),[3] even though they analysed data based on a slightly different application system over 25 years ago, also found that there was a relationship between Leaving Certificate performance and first year college exam performance - this relationship continuing in a less definite fashion through the remainder of the university years.

Notably, Moran & Crowley (1979) concluded at the time that science-based subjects, in particular

[3] Moran and Crowley used a variety of statistical techniques in their analysis, including multiple regressions.

mathematics, had better predictive value than non-science-based subjects. They also vindicated, to an extent, the Leaving Certificate as a college entrance exam, concluding that "the prospects for other selection systems which can equal or improve on the Leaving Certificate seem poor."

1.3 Methodology

CAO degree applications data from the year 2000 were analysed using association rule mining. Whereas similar data have been analysed before using fairly complicated statistical models – Moran & Crowley (1979), Tuohy (1998) and Gormley & Murphy (2006) – the analysis herein is conducted at a very intuitive level. In particular, no underlying statistical model is assumed and no hypotheses are proposed. Therefore, the analysis essentially represents a convenient way of looking at the data.

1.4 Data

The data represents CAO degree applications from the year 2000; the same data that was analysed by Gormley & Murphy (2006). In the year 2000, 53,757 applicants – 24,419 male, 29,337 female and one with unspecified, misspecified or incorrectly recorded gender – chose up to ten of 533 degree courses; note that the two-subject moderatorship offered at TCD was counted as separate courses since, unlike the arts degrees offered at the NUI colleges, each applicant explicitly selected their course options at the application stage.

The most popular course was Arts at UCD, which was chosen by 11,007 (20.5%) applicants. The ten most popular courses are shown in Figure 1; these are all either arts or business courses, with the exception of Science at UCD. The top four courses in Figure 1 are the common entry arts courses offered at NUI colleges in Dublin, Galway, Maynooth and Cork.

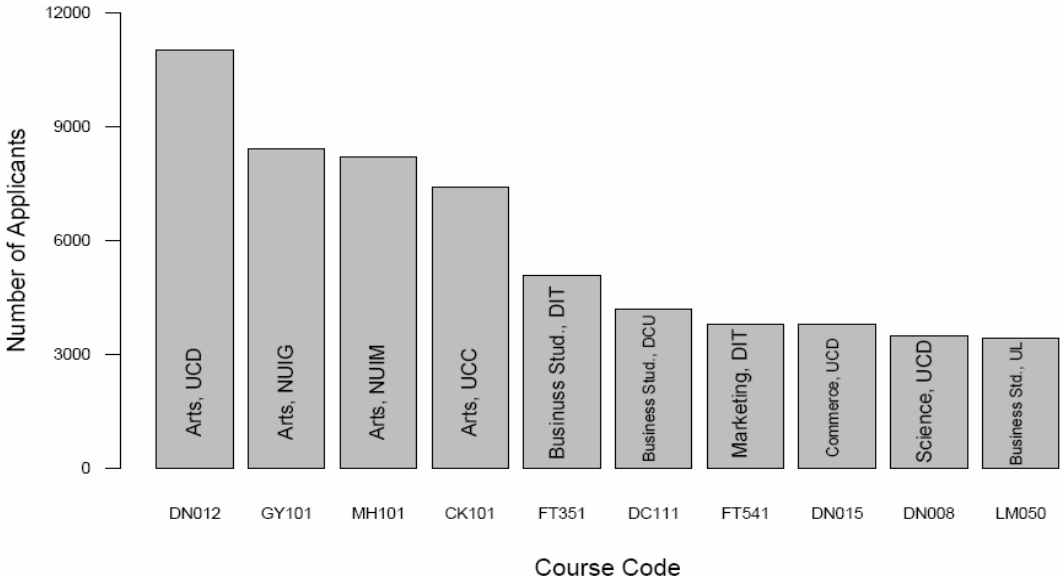


Figure 1: The ten most popular CAO degree course choices in the year 2000

Note that the abbreviation DIT is used herein to denote 'Dublin Institute of Technology' and IT is used in general to denote 'Institute of Technology'.

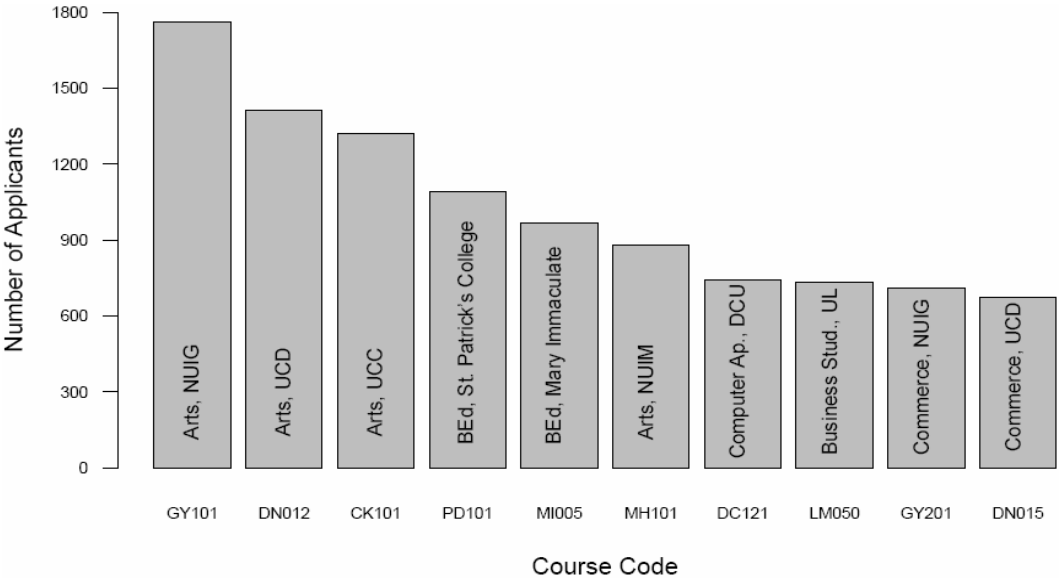


Figure 2: The ten most popular first preference CAO degree choices in the year 2000

Interestingly, the profile of the ten most popular first choices (Figure 2) is different; Arts at UCD and Arts at NUIG switch places as the first and second most popular courses respectively. Furthermore, in addition to arts and business courses, primary teaching courses also feature this time along with the computer applications course at DCU.

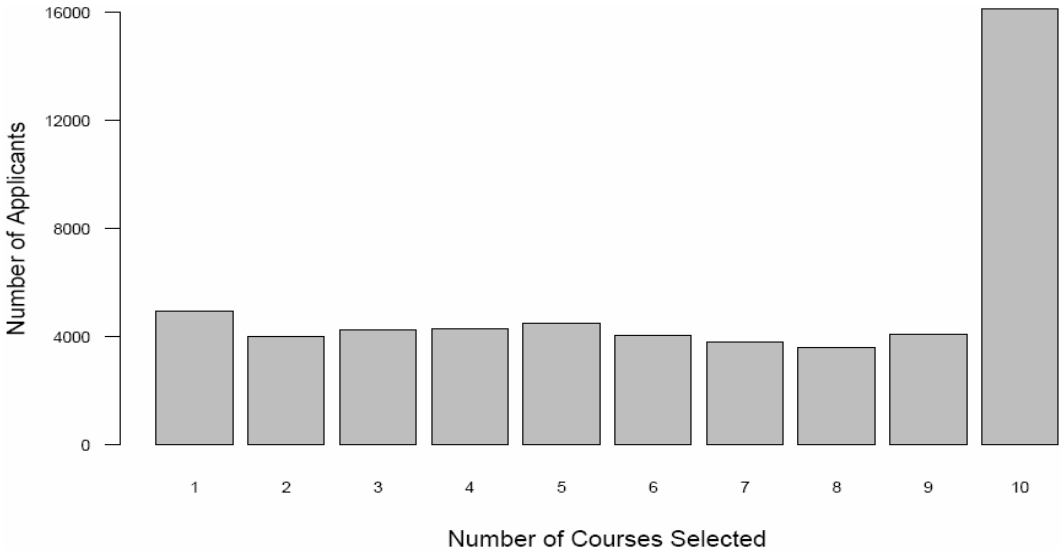


Figure 3: The number of degree courses selected by applicants in the year 2000

Figure 3 represents the distribution of the number of courses chosen by applicants on the degree section of their respective CAO forms in the year 2000. The form was completed in full by 16,138 (30%) students, which means that 70% of applicants did not select the maximum number of degree

courses. One was the next most popular number of courses selected, with 4,952 (9%) students opting to gamble on getting their first and sole choice. The number of applicants choosing 2–9 courses was roughly the same, taking values from 3,618 to 4,518.

2. ASSOCIATION RULES

2.1 Background

Association rules (Agrawal *et al.,* 1993) are used to discover relationships between variables in transaction databases. A transaction database is one that consists solely of binary variables; an item can either be purchased during a transaction or not. The CAO applications data can be viewed as a transaction database – each applicant chose up to ten of 533 courses.

Although formally introduced by Agrawal *et al.* (1993), many of the ideas behind association rules can be seen in the literature at least as far back as Yule (1903). Applicable to a myriad of spheres, such as convenience store data, credit card data, voting data and healthcare data; association rule analysis is one of the most versatile data mining techniques available to the analyst.

Association rule mining is particularly suited to the analysis of large datasets. The apriori algorithm (Agrawal & Srikant, 1994; Borgelt & Kruse, 2002; Borgelt, 2003) presents an easy-to-implement method of generating, or mining, association rules and is used herein.

2.2 Definition of an Association Rule

Given a non-empty set I, an association rule is a statement of the form $A \Rightarrow$ B ('A implies B'), where $A, B \subset I$ such that $A \neq \varnothing$, $B \neq \varnothing$, and $A \cap B = \varnothing$. The set A is called the antecedent of the rule, the set B is called the consequent of the rule, and I is called the itemset. Association rules are generated over a large set of transactions, denoted Ψ.

An association rule is deemed interesting if the items involved occur together often and there is evidence to suggest that one of the sets might in some sense lead to the presence of the other set. Association rules were originally characterised by mathematical notions called 'support', 'confidence' and 'lift'; more recently, a variety of other functions have been introduced, one of which, Gray and Orlowska's 'interestingness' (Gray and Orlowska, 1998) is utilised herein.

2.3 Support, Confidence & Lift

The notation $P(A)$ is introduced to represent the proportion of times that the set A appears in the transaction set Ψ. Similarly, $P(A, B)$ represents the proportion of times that the sets A and B coincide in transactions in Ψ. It is also necessary to define

$$P(B \mid A) = \frac{P(A, B)}{P(A)}.$$

That is, the proportion of times that the set B appears in all of the transactions involving the presence of the set A. Table 1 gives the functions by which association rules are traditionally characterised.

Table 1: Functions of association rules

Function	Definition
Support	$s(A \Rightarrow B) = P(A, B)$
Confidence	$c(A \Rightarrow B) = P(B \mid A)$
Lift	$L(A \Rightarrow B) = c(A \Rightarrow B)/P(B)$

The lift of an association rule can be viewed as some measure of the distance between $P(B \mid A)$ and $P(B)$ or, equivalently, as a function giving the extent to which A and B are dependent on one another.

3. ANALYSIS

3.1 Rule Generation

Association rules were generated using the apriori algorithm of the arules package (Hahsler *et al*., 2005) in R (R Development Core Team, 2005). The apriori algorithm requires that minimum thresholds are preset for support and confidence, the maximum length of a rule can also be preset; the length of a rule is the number of items (courses) within the rule. The minimum support of a rule was set at 0.5% and the minimum confidence was set at 80%. The maximum length of a rule was set at ten. Practically, support of 0.5% means that at least 269 students must have selected a particular combination of courses for a rule comprised of that combination to feature within the mined rules. In analysing these rules, two subsets of rules were of particular interest; those interrelating courses and those relating courses and gender.

3.2 Rules Interrelating Courses

3.2.1 Pruning Method

Discovery of interesting rules that interrelate courses was facilitated by using only the course choices in the input data, thereby omitting the gender variable; 145 association rules were generated using R. In order to view courses at the highest level of grouping, the following pruning technique was devised:

1. Consider the items in an association rule.
2. If a larger rule contains those items then delete the smaller association rule.

This method was applied and more than halved the number of association rules, leaving 72 rules, some of which comprised the same courses in different order.

For example, a rule comprised of four medicine courses, such as

{Medicine at NUIG, Medicine at TCD, Medicine at RCSI} $\Rightarrow$ {Medicine at UCC},

was pruned in favor of a rule comprised of these four, plus one additional course, such as

{Medicine at NUIG, Medicine at UCC, Medicine at TCD, Medicine at RCSI} $\Rightarrow$ {Medicine at

UCD}.

Note that when this method is faced with two or more rules of the same length, containing the same items (courses), all rules are retained. The option of a third step where ties of this sort can be broken, using confidence say, was not availed of in this case. It can be seen in Section 3.2.2 that the order of such rules, when ranked by confidence, can provide information about their constituent courses.

3.2.2 Top Twenty Rules

Table 2 gives the top twenty rules, ranked by confidence, relating distinct items. Following inspection of these rules, they can each be grouped into one of a few categories so that a non-model-specific 'clustering' of the courses emerges. An explanation of all course codes appearing in this work is given in Table 9 (Appendix B).

Table 2: The top twenty rules, ranked by confidence, interrelating courses

	Rule	Support	Confidence	Lift
1	{GY501, CK701, TR051, RC001} ⇒ {DN002}	0.53%	97.92%	33.76
2	{MI005, DN012, CM001, FR001} ⇒ {PD101}	0.55%	97.67%	21.22
3	{PD101, FR001, LM047} ⇒ {MI005}	0.55%	96.71%	21.17
4	{CM001, FR001, PD103} ⇒ {PD101}	0.60%	96.41%	20.95
5	{MI005, PD103} ⇒ {PD101}	0.51%	95.82%	20.82
6	{CM001, LM047} ⇒ {MI005}	0.56%	95.54%	20.91
7	{GY101, MI005, CM001, FR001} ⇒ {PD101}	0.58%	95.44%	20.74
8	{DN012, CM001, MH101, FR001} ⇒ {PD101}	0.64%	95.04%	20.65
9	{MI005, CM001, MH101, FR001} ⇒ {PD101}	0.58%	94.24%	20.48
10	{CK701, TR051, DN002, RC001} ⇒ {GY501}	0.53%	92.76%	38.84
11	{CK101, CM001} ⇒ {MI005}	0.54%	92.33%	20.21
12	{TR084, FT472} ⇒ {FT471}	0.59%	92.11%	17.38
13	{GY101, DN012, FR001} ⇒ {PD101}	0.52%	92.11%	20.01
14	{DN008, TR051} ⇒ {DN002}	0.54%	91.25%	31.47
15	{GY501, CK701, DN002, RC001} ⇒ {TR051}	0.53%	89.52%	41.31
16	{DC111, FT542, DN015} ⇒ {FT351}	0.66%	89.44%	9.47
17	{GY101, MH101, FR001} ⇒ {PD101}	0.57%	89.24%	19.39
18	{CK101, FR001} ⇒ {MI005}	0.56%	89.05%	19.49
19	{MI005, DN012, PD101, CM001} ⇒ {FR001}	0.55%	88.82%	29.33
20	{TR004, GY251} ⇒ {DN009}	0.50%	88.67%	27.93

Medicine

Rules 1, 10 and 15 in Table 2 relate the medicine courses across the five institutions offering medicine: NUIG, UCC, TCD, RCSI and UCD. In this case, the fact that the UCD course is in the consequent of Rule 1 suggests that it was the most popular of the five medicine courses; in rules 10 and 15 and elsewhere amongst the 72 remaining rules, rules relating the medicine courses appeared with different courses in the consequent (Medicine at NUIG was the second most popular medicine course and Medicine at TCD was the third).

Rule 14 associates three courses within Dublin; Science at UCD, Medicine at TCD and Medicine at UCD. This is possibly an example of a geographical effect on course selection.

Teaching & Arts

Rules 2–9, 11, 13 and 17–19 relate arts (BA) and education (BEd) courses across nine institutions throughout Ireland. The course PD101, BEd at St. Patrick's College, is the consequent part of the majority of these rules, suggesting that it was a significant draw to such-minded people and the most popular of the BEd courses.

Social Care

Rule 12 suggests that people who applied for Social Studies (Social Work) at TCD and Early Childhood Care and Education at DIT also selected Social Care at DIT. In fact less than 8% of applicants choosing the former pair failed to choose the latter course. This may be indicative of a geographical effect – applicants wishing to study social care in Dublin.

Business

Rule 16 relates four business courses in Dublin: Business Studies at DIT, Business Studies at DCU, Management and Marketing at DIT and Commerce at UCD. There are further rules within the 72 that related business courses in Dublin.

Law

Rule 20 associates the law degrees at TCD, NUIG and UCD; the UCD degree being the most popular choice. This rule can be used to illustrate the lift of an association rule; given that it is known that an applicant selected law at both TCD and NUIG, they are 27.93 times more likely to have selected Law at UCD than if no information was available regarding their other course choices. Moreover, since lift is symmetric (see Appendix A.1) it is also true that if it is known that an applicant selected Law at UCD, they are 27.93 times more likely to have selected Law at TCD and Law at NUIG than if no information regarding their other course choices was available.

3.2.3 Remaining Rules

All 72 rules are given in Table 10 (Appendix B). When the antecedent and consequent of these rules are taken into account, almost all of them can be considered as belonging to one of the following categories, or 'clusters'.

Arts	Primary Teaching & Arts
Business[4]	Psychology & Arts
Business & Communications	Science
Engineering	Social Care[5]
Law	Theology & Arts
Medicine	

All of the remaining rules fall under some combination of these categories – typically, but not

[4] Includes Business Studies, Commerce, Finance and Marketing.

[5] Includes Social Care, Social Science, Early Childhood Care and Early Childhood Studies.

exclusively, a combination involving arts. Most importantly, there were no combinations, like medicine and law perhaps that associated two very different disciplines that are regarded as prestigious in terms of CAO points. Therefore, it can be said that all of the 145 mined rules were grouped by topic of interest and not by perceived prestige associated with high points value.

Looking within the categories into which these rules fall, there is evidence of further grouping by geographical area. For example, Rule 12 in Table 2 relates social care courses within Dublin whereas Rule 38 in Table 10 relates the same type of courses without this geographical constraint.

Moreover, Rules 29 and 49 (Table 10) relate science courses in the Dublin area, while Rule 51 relates business courses within Dublin. The geographical effect is not, however, restricted to the Dublin area. Rule 57 relates three courses in Cork: Commerce, Government and Public Policy, and Arts, all at UCC. This may in fact be an example of a rule where topic was second to geographical location as the main motivator for course selection.

3.2.4 Remarks

By inspection of the top 20 rules, five 'clusters' emerge. These 'clusters' are consistent with the findings of Tuohy (1998) and Gormley & Murphy (2006), yet come about from the output of a very simple analysis that assumed no underlying model. Furthermore, inspection of the whole 72 rules led to the discovery of six additional 'clusters'.

The presence of these 'clusters', especially arrived at in such a natural fashion, is evidence of the effectiveness of the points system; this analysis yielded no evidence to suggest that students choose courses with the highest points, instead the data suggest that they choose courses by topic, with geography sometimes a factor. Moreover, due to the pruning method that was employed, it can be inferred that amongst all 145 mined rules, there was no evidence to suggest that course choices are made based on prestige associated with high CAO points.

3.3 Further Exploratory Analysis

3.3.1 The Idea

Further analysis was carried out to search for any weak evidence that might suggest that a points race exists amongst a smaller cohort of students. The support threshold was reduced and pruning was affected. Then subsets of prestige (high-point) courses within the remaining rules were examined.

3.3.2 Methodology

The rules were mined once again, this time the support threshold was set at just 0.1%, meaning that a rule would be mined so long as it was supported by at least 54 applicants. The confidence threshold remained at 80% and the pruning method employed in Section 3.2.1 was applied to the 2,537 rules that were generated; this reduced the number of rules to 1,020.

3.3.3 Results

Medicine

Ninety-nine of these 1,020 rules contained at least one medicine course; most of these rules involved courses in areas like medicine, physiotherapy and science. However, there were three rules involving law courses; they are listed in Table 3.

Table 3: Rules involving law courses within the medicine subset

	Rule	Support	Conf.	Lift
1	{Law at UCD, Medicine at TCD} ⇒ {Medicine at UCD}	0.12%	86.67%	29.88
2	{Law at UCD, Medicine at NUIG} ⇒ {Medicine at UCD}	0.13%	89.61%	30.90
3	{Law at TCD, Medicine at NUIG} ⇒ {Medicine at UCD}	0.11%	81.43%	28.08

Law

Looking at the cohort who chose law, the rules in Table 3 surface once again. There are also rules with commerce and business present, however, it could be argued that law, business and commerce are not very different disciplines. However, the rule in Table 4 seemed out of place; associating Civil Law at NUIG with Psychology at NUIM. Psychology at NUIM was a new course and so had no previous points-related prestige but entry points for the year 2000 were relatively high at 465.

Table 4: Rule involving law and psychology

Rule	Support	Conf.	Lift
{Psychology at NUIM, Law at NUIG} ⇒ {Arts at NUIG}	0.12%	81.48%	5.21

It should be noted that the rules in Table 3 and Table 4 are each supported by less than 70 applicants. Further exploratory analysis, with lower confidence thresholds, is carried out in Appendix C. The rules involving commerce and law are also mentioned in Appendix C.

3.4 Gender Related Choices – Female

3.4.1 Initial Analysis

The gender variable was included in the mining process, the support threshold was reset to 0.5% and 536 association rules were generated using the arules package in R. The resulting rules were pruned, as in Section 3.2.1, so that only supersets survived. The remaining rules were syntactically constrained to have 'female' (denoted *F* herein) in the consequent and the top ten such rules are given in Table 5.

Table 5: The top ten rules, ranked by confidence, with consequent {*F*}

	Antecedent	Support	Conf.	Lift
1	{BEd at Froebel, Early Childhood Care & Education at DIT}	0.54%	99.32%	1.82
2	{Arts at NUIG, Early Childhood Care & Education at DIT}	0.86%	99.14%	1.82
3	{BEd (Home Economics) at St. Catherine's}	1.26%	99.12%	1.82
4	{Early Childhood Care & Education at DIT, Tourism Marketing at DIT}	0.59%	99.07%	1.82
5	{Arts at UCC, Early Childhood Care & Education at DIT}	0.54%	98.97%	1.81
6	{Early Childhood Care & Education at DIT, Social Care at DIT, Early Childhood Studies at UCC}	0.84%	98.90%	1.81
7	{BEd at St. Patrick's, Early Childhood Care & Education at DIT}	0.66%	98.89%	1.81
8	{BEd at Mary Immaculate, Early Childhood Studies at UCC}	0.60%	98.80%	1.81
9	{Early Childhood Care & Education at DIT, Hospitality (Hotel & Catering) Management at DIT}	0.58%	98.42%	1.80
10	{Social Science at UCC, Early Childhood Care & Education at DIT}	0.51%	98.21%	1.80

From Table 5 it is apparent that the Early Childhood Care and Education course at DIT is in the antecedent of eight of the ten rules. The aim of this section was to find relationships between course choice and gender (female), and the rules in Table 5 largely failed in this aim. Therefore, an alternative approach was necessary.

Note that rules were chosen with $\{F\}$ in the consequent rather than the antecedent because the proportion of applicants to a particular course that were female was viewed as more interesting than the proportion of females that applied to a particular course. In any event, since lift is symmetric (see Appendix A.1), it applies identically to the converse of each rule.

3.4.2 Interestingness

Since the method of pruning so that only supersets survive was not satisfactory in this case, an alternative approach was necessary. Gray and Orlowska (1998) define the 'interestingness' of an association rule $A \Rightarrow$ B, as follows;

$$\text{Int}(A \Rightarrow B;\, K,\, M) = \left[\left(\frac{P(A,B)}{P(A)P(B)} \right)^{K} - 1 \right] (P(A)P(B))^{M}.$$

This metric presents a compromise between the distance of lift from one and the respective magnitudes of $P(A)$ and $P(B)$. The choice of K and M is somewhat arbitrary, which can be regarded as either a strength or a weakness of Gray and Orlowska's interestingness.

The interestingness of all 536 mined rules was computed with $K = M = 2$. The top ten rules with consequent $\{F\}$, ranked in order of interestingness, are given in Table 6. Choosing $K = M = 2$ can be viewed as giving a balanced compromise between the distance of lift from one and the respective magnitudes of $P(A)$ and $P(B)$. Further discussion on the choice of $K = M = 2$, including an alternative formula for Gray and Orlowaska's interestingness, is given in Appendix A.2.

Table 6: The ten highest ranked, by interestingness, rules with consequent {*F*}

	Antecedent	Support	Conf.	Lift	Interest.
1	{Social Care at DIT}	4.99%	94.10%	1.72	0.00165
2	{Early Childhood Care & Education at DIT}	4.49%	98.13%	1.80	0.00140
3	{BEd at St. Patrick's}	4.01%	87.23%	1.60	0.00098
4	{BEd at Mary Immaculate}	3.89%	85.18%	1.56	0.00089
5	{Social Science at UCD}	3.47%	85.79%	1.57	0.00072
6	{Early Childhood Studies at UCC}	3.19%	96.30%	1.78	0.00069
7	{Social Science at UCC}	3.02%	87.88%	1.61	0.00056
8	{BEd at Froebel}	2.77%	91.34%	1.68	0.00049
9	{Early Childhood Care & Education at DIT, Social Care at DIT}	2.49%	98.10%	1.80	0.00042
10	{BEd at Coláiste Mhuire}	2.51%	88.27%	1.62	0.00039

The magnitude of the values in interestingness column (Interest.) of Table 6 is not of particular importance here; it is used for ranking purposes only.

3.4.3 Top Ten Rules

Primary Teaching

Rules 3, 4, 8 and 10 associated various types of primary teaching degrees (BEd) to female applicants. These rules all have confidence in the range 85.18%–91.34%, which is consistent with the proportion of primary school teachers in Ireland that are female. Drew (2006) reports UNESCO educational statistics that reveal that 86% of primary school teachers in Ireland in the year 2002 were female. Table 7 gives the corresponding figures for the years 1999, 2001 and 2004, as well as the year 2002.

Table 7: Gender breakdown of primary teaching staff in Ireland, 1999–2004

	1999	2000	2001	2002	2003	2004
Total	21,148	20,865	21,865	22,979	23,972	24,792
Female	17,924	–	18,680	19,772	–	20,671
Male	3,224	–	3,185	3,207	–	4,121
% Female	84.8	–	85.4	86.0	–	83.4
% Male	15.2	–	14.6	14.0	–	16.6

The raw data for Table 7 was sourced from the UNESCO Institute for Statistics website (www.uis.unesco.org), UIS Database, updated following the *Final release of data from the 2005 education survey*. The gender-specific data for the years 2000 and 2003 were listed as unavailable.

Over the four years for which data on the proportion of female primary teachers are available, this number is between 83.4% and 86.0%. Drew (2006) suggests a variety of reasons that may explain why men are not attracted to primary teaching courses. Moreover, this issue is not confined to applicants within the Republic of Ireland. The situation is similar north of the border. The

Stranmillis Annual Report 2003/04 stated that "male entrants to the Primary B.Ed. represent 17% of the Primary intake [in the year 2003], compared with 9% in 2002." However, their 2004/05 annual report revealed that this number had fallen to 10% in the year 2004.

Earlier this year it appeared that the Government recognised this imbalance as an issue and was planning to address it. In an article entitled *Competition Intense for Places on Teaching Degree Courses* in the Irish Times on 25th January 2006, Brian Mooney wrote that "some 90 per cent of primary school teachers under 40 are women and today, the Department of Education is launching a promotion campaign to attract men into primary teaching."

Social Care & Childhood Care

Rules 1, 5 and 7 link courses in social care and social science to female applicants. Rules 2 and 6 link courses in early childhood care and early childhood studies to female applicants; for example, over 96% of applicants to Early Childhood Studies at UCC were female. Rule 9 shows that over 98% of applicants who applied for both Social Care and Early Childhood Care and Education at DIT were female.

This area shows a gender imbalance on par with, or worse than, that shown in primary teaching. The reasons for such an imbalance are likely to be similar; such as the perception of such jobs as 'women's jobs'. However, there is far less literature around this topic; suggesting that it may be perceived as being of less importance to society.

3.4.4 Further Rules

Extending to the 20 most interesting rules, as listed in Table 11 (Appendix B), all courses that emerge are in the areas of primary teaching, social studies and arts. Therefore, the stereotype that primary teachers and social workers are predominantly female is, the data suggests, apparent from the application stage.

3.5 Gender Related Choices – Male

3.5.1 Rule Generation & Ranking

The 536 rules generated in Section 3.4.1 were ranked according to their interestingness, with the syntactic constraint that the consequent is 'male', denoted $\{M\}$.

3.5.2 Top Ten Rules

The top ten such rules are given in Table 8.

Table 8: The ten highest ranked, by interestingness, rules with consequent {*M*}

	Antecedent	Support	Conf.	Lift	Interest.
1	{Engineering at DIT}	2.91%	87.12%	1.92	0.00062
2	{Engineering at TCD}	2.56%	80.30%	1.77	0.00044
3	{Construction Management at Waterford IT}	2.32%	90.09%	1.98	0.00040
4	{Construction Management at GMIT[6]}	2.27%	91.59%	2.02	0.00039
5	{Construction Economics (Quantity Surveying) at Limerick IT}	2.27%	87.85%	1.93	0.00038
6	{Construction Management at Limerick IT}	2.17%	91.75%	2.02	0.00036
7	{Construction Economics & Management (Quantity Surveying) at DIT}	2.18%	88.74%	1.95	0.00035
8	{Electrical & Electronic Engineering at DIT}	2.01%	89.65%	1.97	0.00030
9	{Computer Engineering at DIT}	1.77%	84.07%	1.85	0.00022
10	{Electronic Engineering at DCU}	1.47%	88.55%	1.95	0.00016

Engineering

Rules 1, 2, 8, 9 and 10 of Table 8 relate Engineering courses to male applicants; for example, almost 90% of all applicants to Electrical and Electronic Engineering at DIT were male. It is generally accepted as a problem that there is such a gender-imbalance in engineering in Ireland. Attempts are currently being made from various quarters to address this problem; for example Science Foundation Ireland (SFI), in a press release on 9th May 2006, announced the *SFI Young Women in Engineering Scholarship* aimed at attracting school leavers into engineering sciences. In the same press release, SFI stated that "based on CAO statistics only 16.4% of students who accepted places on four-year engineering degree courses in 2002–2004 were female". This analysis reveals that the imbalance can be traced back even further, to the application stage.

Whereas the consequences of the under-representation of women in engineering may not have the same social consequences as that of men in primary teaching, it is not an ideal situation. At the recent launch of Trinity College Dublin's new *Centre for Women in Science & Engineering Research*, the Minister for Enterprise, Trade and Employment, Micheál Martin, stated that "there is a clear and unequivocal case for increasing the participation of women in science, engineering and technology in Ireland. The under representation of women threatens our global competitiveness and requires a coordinated set of interventions."[7]

Construction Management and Construction Economics

The remaining rules from Table 8 associate courses in construction management and construction economics to male applicants; for example, over 91% of applicants to Construction Management at Limerick IT were male. Like the gender imbalances in social care and early childhood care, the gender imbalance in the construction sector does not seem to be an active issue.

[6] Galway-Mayo Institute of Technology.

[7] Source: press release, Trinity College Dublin, 10th November 2006.

3.5.3 Further Rules

The top 20 rules with consequent {*M*} are given in Table 12; the later ten rules all associated courses in engineering, construction and manufacturing. The data suggest, as in the female case, that the usual stereotypes hold in the male case and are visible at the application stage.

4. CONCLUSIONS

Through the application of a simple analysis to CAO application data from the year 2000 – which consisted only of a useful way to look at the data without imposing any statistical model, testing any hypotheses or computing any confidence intervals – some important questions regarding the CAO application system have been answered, while others may have been raised.

Definite patterns of course choice amongst applicants emerge, based primarily on course topic, with geography also a factor for some applicants. In no instance, save for very small cohorts of students, did points alone dictate course choice. Moreover, no evidence to link courses purely based on points was found. This is a very important result, supporting the CAO application system as an effective third level application channel.

There is strong gender linkage amongst the data. The vast majority of applicants to courses in social care, early childhood care and primary teaching are female; whereas the majority of applicants to courses in engineering and construction management are male.

Whereas the gender imbalances in primary teaching and engineering are active issues, comparatively little is being done to address the imbalances that exist in the social care and construction sectors. Whether or not these imbalances are detrimental to our society is debatable, however, it is clear from this analysis that they can all be traced back to the application stage.

A Word of Warning

The conclusions of this report should not be taken as a vindication of the CAO college application system. Whereas it is true that there is no evidence herein to suggest that a ‘points race’, in the negative sense, is prevalent, it is not necessarily true that this application system puts the best applicants into each college course.

Tuohy (1998) gives the analogy of a bank using a round of golf to decide which applicant should get the management post; it is a fair, consistent, system but the best golfer will not necessarily be the best branch manager.

5. ACKNOWLEDGEMENTS

I would like to record a special note of thanks to Dr. Brendan Murphy and Dr. Myra O’Regan for their advice and suggestions. Thanks to Dr. Shane Whelan for his numerous and valuable suggestions at the refereeing stage.

I also owe thanks to Prof. Eileen Drew for kindly giving me a copy of her book (Drew, 2006) before its publication and to Ms. Sharon King and Ms. Deirdre Toher for thoroughly proof-reading this work.

I would like to thank the Statistical and Social Inquiry Society of Ireland and the Barrington Trust for giving me the honour of being the one-hundred-and-twentieth Barrington Lecturer.

The support of an SFI Basic Research Grant (04/BR/M0057) is gratefully acknowledged.

BIBLIOGRAPHY

Agrawal, R. & Srikant, R. (1994), Fast algorithms for mining association rules, *in* 'Conference on Very Large Data Bases'.

Agrawal, R., Imielinski, T. & Swami, A. (1993), Mining association rules between sets of items in large databases, *in* 'ACM SIGMOID Conference'.

Borgelt, C. (2003), Efficient implementations of apriori and eclat, *in* 'Workshop on Frequent Itemset Mining Implementations', Melbourne, FL.

Borgelt, C. & Kruse, R. (2002), Induction of association rules: Apriori implementation, *in* '15th COMPSTAT Conference', Berlin, Germany.

Clancy, P. (2001), *College entry in focus: a fourth national survey of access to higher education*, The Higher Education Authority.

Drew, E. (2006), *Facing Extinction? Why Men are not Attracted to Primary Teaching*, The Liffey Press, Dublin.

Gormley, I.C. & Murphy, T.B. (2006), 'Analysis of Irish third-level college applications data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**(2), 361–379.

Gray, B. & Orlowska, M. (1998), CCAIIA: Clustering categorical attributes into interesting association rules, *in* 'Pacific-Asia Conference on Knowledge Discovery and Data Mining'.

Hahsler, M., Gruen, B. & Hornik, K. (2005), *arules: mining association rules and frequent itemsets*. R package version 0.2–4.

Hyland, A. (1999), *Commission on the Points System. Final Report and Recommendations*, The Stationary Office, Dublin.

Lynch, C.D., McConnell, R.J. & Hannigan, A. (2006), 'Dental school admissions in Ireland: can current selection criteria predict success?', *European Journal of Dental Education* **10**, 73–79.

Lynch, K., Brannick, T., Clancy, P. & Drudy, S. (1999), *Commission on the points system. Research paper no. 4. Points and performance in higher education: A study of the predictive validity of the points system.*, The Stationary Office, Dublin.

Moran, M.A. & Crowley, M.J. (1979), 'The Leaving Certificate and first year university performance', *Journal of the Statistical and Social Inquiry Society of Ireland* **14**(1), 231–266.

O'Connell, P.J., Clancy, P. & McCoy, S. (2006), *Who went to college in 2004? A national survey of new entrants to higher education*, The Higher Education Authority.

Plackett, R.L. (1975), 'The analysis of permutations', *Applied Statistics* **24**, 193–202.

R Development Core Team (2005), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Tuohy, D. (1998), *Commission on the points system. Research paper no.1. Demand for third-level places. Interests, fields of study and the effect of the points system in the application process for 1997*, The Stationary Office, Dublin.

Yule, G.U. (1903), 'Notes on the theory of association of attributes in statistics', *Biometrika* **2**(2), 121–134.

APPENDIX A

STATISTICAL DETAILS

A.1 Lift: A Symmetric Function

The lift of an association rule $A \Rightarrow B$ is the same as the lift of the rule $B \Rightarrow A$; this can be shown as follows;

$$L(A \Rightarrow B) = \frac{c(A \Rightarrow B)}{P(B)} = \frac{P(B \mid A)}{P(B)} = \frac{P(A,B)}{P(A)P(B)} = \frac{P(A \mid B)}{P(A)} = \frac{c(B \Rightarrow A)}{P(A)} = L(B \Rightarrow A).$$

This result is extremely useful when it is not clear which of the rules $A \Rightarrow B$ and $B \Rightarrow A$ is of more interest, or when both may be of interest. Equivalently, lift can be useful when it is not clear whether $P(A \mid B)$ or $P(B \mid A)$ is of more interest, or when both may be of interest. For example, the rule

{Social Care at DIT} $\Rightarrow$ {F}

was mentioned in Table 6. It is possible that the rule,

{F} $\Rightarrow$ {Social Care at DIT},

may also be of interest. The lift of both rules is the same (1.72), so that if it is known that an applicant is female then they are 1.72 times more likely to select Social Care at DIT than if no information about their gender is known. Furthermore, if it is known that an applicant has selected Social Care at DIT then they are 1.72 times more likely to be female than if no information about their course choices is available.

Moreover, lift can also be used in any situation, even outside of the area of association rule mining, when it is not clear whether $P(A \mid B)$ or $P(B \mid A)$ is of interest, or when both may be of interest. A further consequence of the symmetry of lift is that Gray and Orlowska's interestingness is also symmetric since

$$(P(A).P(B))^M = (P(B).P(A))^M .$$

Therefore, Gray and Orlowska's interestingness could also be used in any situation, even outside of the area of association rule mining, where it is unclear whether $P(A \mid B)$ or $P(B \mid A)$ is of interest, or when both may be of interest.

A.2 Interestingness: Alternative Definition & Setting K = M

Gray and Orlowska's interestingness can be written as

$$\text{Int}(A \Rightarrow B; K, M) = \left[c(A \Rightarrow B)^K - P(B)^K\right] P(A)^M P(B)^{M-K}. \qquad (1)$$

To see why this is so, notice that

$$\text{Int}(A \Rightarrow B; K, M) = \left[\left(\frac{P(A,B)}{P(A)P(B)}\right)^K - 1\right] \left(P(A).P(B)\right)^M$$

$$= \left[\frac{\left(P(A,B)/P(A)\right)^K - P(B)^K}{P(B)^K} \right] \left(P(A).P(B)\right)^M$$
$$= \left[P(B \mid A)^K - P(B)^K\right] P(A)^M \, P(B)^{M-K}$$
$$= \left[c(A \Rightarrow B)^K - P(B)^K\right] P(A)^M \, P(B)^{M-K}.$$

From Equation 1, an argument for choosing $K = M$ becomes apparent; since all of the rules are to be syntactically constrained to have a particular consequent (F or later, M), $P(B) = P(F)$ will be equal across all rules and choosing $K = M$ gives

$$P(B)^{M-K} = 1.$$

Therefore, setting $K = M = 2$ achieves a suitable compromise between measuring the distance of $c(A \Rightarrow B)$ from $P(B)$ and the magnitude of $P(A)$, giving

$$\text{Int}(A \Rightarrow B; 2, 2) = \left[c(A \Rightarrow B)^2 - P(B)^2\right] P(A)^2.$$

APPENDIX B

TABLES

B.1 Explanation of Course Codes Appearing Herein

Table 9: Course codes appearing in this work

CR	**Cork Institute of Technology**
CR107	Electronic Engineering
CR108	Mechanical Engineering
CK	**University College Cork (NUI)**
CK101	Arts
CK102	Social Science
CK111	Early Childhood Studies
CK201	Commerce
CK204	Finance
CK210	Government and Public Policy
CK301	Law
CK402	Biological and Chemical Sciences
CK701	Medicine
CM	**Coláste Mhuire**
CM001	B.Ed
DC	**Dublin City University**
DC111	Business Studies
DC121	Computer Applications
DC126	Financial and Actuarial Mathematics
DC131	Communication Studies
DC132	Journalism
DC181	Biotechnology
DC191	Electronic Engineering
DC192	Telecommunications Engineering
DC201	Common Entry into Science (Undenominated Entry)
FT	**Dublin Institute of Technology**
FT101	Architecture
FT111	Construction Economics and Management (Quantity Surveying)
FT125	Engineering
FT221	Electrical and Electronic Engineering
FT222	Applied Sciences
FT224	Optometry
FT281	Computer Engineering
FT351	Business Studies
FT352	Media Arts

FT353	Communications – Journalism
FT354	Information Systems Development
FT401	Hospitality (Hotel and Catering) Management
FT402	Tourism Marketing
FT471	Social Care
FT472	Early Childhood Care and Education
FT541	Marketing
FT542	Management and Marketing
FT543	Retail and Services Management
FR	**Froebel College of Education**
FR001	B.Ed
NC	**National College of Ireland**
NC001	Accounting and Human Resource Management
RC	**Royal College of Surgeons in Ireland**
RC001	Medicine
CS	**St. Catherine's College of Education**
CS001	B.Ed (Home Economics)
PD	**St. Patrick's College of Education**
PD101	B.Ed
PD103	BA in Humanities
TR	**Trinity College Dublin**
TR004	Law
TR032	Engineering
TR051	Medicine
TR071	Science
TR084	Social Studies (Social Work)
DN	**University College Dublin (NUI)**
DN001	Architecture
DN002	Medicine
DN003	Engineering
DN005	Veterinary Medicine
DN007	Social Science
DN008	Science
DN009	Law (BCL)
DN012	Arts
DN015	Commerce
DN020	Actuarial and Financial Studies
DN021	Business and Legal Studies
DN054	Psychology plus 2 permissible subjects under DN012 [Arts] for first year only

LC	**Limerick Institute of Technology**
LC017	Construction Economics (Quantity Surveying)
LC019	Construction Management
MI	**Mary Immaculate College**
MI005	B.Ed
LM	**University of Limerick**
LM047	Arts
LM050	Business Studies
LM069	Computer Engineering
LM081	Manufacturing Technology
GA	**Galway-Mayo Institute of Technology**
GA042	Construction Management
GY	**National University of Ireland, Galway**
GY101	Arts
GY103	Arts (Public and Social Policy)
GY201	Commerce
GY251	Bachelor of Civil Law (BCL)
GY301	Science
GY401	Engineering (Undenominated)
GY402	Civil Engineering
GY501	Medicine
MH	**National University of Ireland, Maynooth**
MH101	Arts
MH102	Finance
MH106	Psychology
MH201	Science
MU	**Pontifical University, Maynooth**
MU001	Theology and Arts
WD	**Waterford Institute of Technology**
WD025	Construction Management
WD026	Electronics

B.2 Rules Interrelating Courses

Table 10: The 72 rules mentioned in Section 3.2.3, ranked by confidence

	Rule	Support	Confidence	Lift
1	{GY501, CK701, TR051, RC001} ⇒ {DN002}	0.525%	97.92%	33.76
2	{MI005, DN012, CM001, FR001} ⇒ {PD101}	0.547%	97.67%	21.22
3	{PD101, FR001, LM047} ⇒ {MI005}	0.547%	96.71%	21.17
4	{CM001, FR001, PD103} ⇒ {PD101}	0.599%	96.41%	20.95

5	{MI005, PD103} ⇒ {PD101}	0.512%	95.82%	20.82
6	{CM001, LM047} ⇒ {MI005}	0.558%	95.54%	20.91
7	{GY101, MI005, CM001, FR001} ⇒ {PD101}	0.584%	95.44%	20.74
8	{DN012, CM001, MH101, FR001} ⇒ {PD101}	0.642%	95.04%	20.65
9	{MI005, CM001, MH101, FR001} ⇒ {PD101}	0.579%	94.24%	20.48
10	{CK701, TR051, DN002, RC001} ⇒ {GY501}	0.525%	92.76%	38.84
11	{CK101, CM001} ⇒ {MI005}	0.538%	92.33%	20.21
12	{TR084, FT472} ⇒ {FT471}	0.586%	92.11%	17.38
13	{GY101, DN012, FR001} ⇒ {PD101}	0.521%	92.11%	20.01
14	{DN008, TR051} ⇒ {DN002}	0.543%	91.25%	31.47
15	{GY501, CK701, DN002, RC001} ⇒ {TR051}	0.525%	89.52%	41.31
16	{DC111, FT542, DN015} ⇒ {FT351}	0.662%	89.44%	9.47
17	{GY101, MH101, FR001} ⇒ {PD101}	0.571%	89.24%	19.39
18	{CK101, FR001} ⇒ {MI005}	0.560%	89.05%	19.49
19	{MI005, DN012, PD101, CM001} ⇒ {FR001}	0.547%	88.82%	29.33
20	{TR004, GY251} ⇒ {DN009}	0.501%	88.67%	27.93
21	{MI005, PD101, CM001, MH101} ⇒ {FR001}	0.579%	88.60%	29.26
22	{TR071, MH201} ⇒ {DN008}	1.049%	88.13%	13.62
23	{CK101, PD101} ⇒ {MI005}	0.766%	88.03%	19.27
24	{GY101, CM001, MH101} ⇒ {PD101}	0.526%	87.89%	19.10
25	{DN021, TR004} ⇒ {DN009}	0.698%	87.41%	27.53
26	{TR071, GY301} ⇒ {DN008}	1.019%	87.12%	13.47
27	{DC111, FT541, DN015} ⇒ {FT351}	0.841%	86.92%	9.20
28	{CK402, TR071} ⇒ {DN008}	0.655%	86.91%	13.43
29	{TR071, FT222} ⇒ {DN008}	0.536%	86.23%	13.33
30	{MI005, FR001, LM047} ⇒ {PD101}	0.547%	85.96%	18.68
31	{DN012, MU001} ⇒{MH101}	0.737%	85.90%	5.64
32	{DC131, DC132, FT352} ⇒ {FT353}	0.549%	85.51%	26.42
33	{DN054, MH106, MH101} ⇒ {DN012}	0.809%	85.29%	4.17
34	{DC111, FT541, FT542} ⇒ {FT351}	0.969%	85.27%	9.03
35	{GY501, TR051, DN002, RC001} ⇒ {CK701}	0.525%	85.20%	46.54
36	{PD101, FR001, PD103} ⇒ {CM001}	0.599%	85.19%	30.01
37	{GY101, MI005, PD101, CM001} ⇒ {FR001}	0.584%	85.09%	28.10
38	{FT471, CK111} ⇒ {FT472}	0.845%	84.86%	18.53
39	{TR032, GY401} ⇒ {DN003}	0.541%	84.84%	22.35
40	{MH101, GY103} ⇒ {GY101}	0.590%	84.76%	5.42
41	{DC111, NC001} ⇒ {FT351}	0.783%	84.54%	8.95
42	{CK101, DN012, MH101, LM047} ⇒ {GY101}	0.939%	84.17%	5.38
43	{DN012, PD101, CM001, MH101} ⇒	0.642%	84.15%	27.79

	{FR001}			
44	{GY101, MU001} ⇒ {MH101}	0.582%	84.14%	5.52
45	{GY201, MH101} ⇒ {GY101}	0.757%	83.92%	5.37
46	{GY101, DN012, PD103} ⇒ {MH101}	0.543%	83.91%	5.51
47	{MI005, DN012, MH101} ⇒ {PD101}	0.588%	83.60%	18.17
48	{MH106, GY101, DN012} ⇒ {MH101}	0.580%	83.42%	5.47
49	{TR071, DC201} ⇒ {DN008}	0.504%	83.38%	12.89
50	{CK102, LM047} ⇒ {CK101}	0.512%	83.08%	6.02
51	{FT541, FT542, DN015} ⇒ {FT351}	0.727%	83.01%	8.79
52	{CK101, GY101, DN012, LM047} ⇒ {MH101}	0.939%	82.92%	5.44
53	{DN012, TR004} ⇒ {DN009}	0.765%	82.86%	26.10
54	{CK101, FR001} ⇒ {PD101}	0.521%	82.84%	18.00
55	{CK301, TR004} ⇒ {DN009}	0.558%	82.64%	26.03
56	{LM050, FT351, DN015} ⇒ {GY201}	0.532%	82.18%	13.34
57	{CK201, CK210} ⇒ {CK101}	0.512%	82.09%	5.95
58	{GY101, PD101, CM001, FR001} ⇒ {MI005}	0.584%	81.98%	17.95
59	{MI005, DN012, PD101, FR001} ⇒ {CM001}	0.547%	81.89%	28.85
60	{GY101, MI005, DN012} ⇒ {PD101}	0.526%	81.56%	17.72
61	{DC111, FT354} ⇒ {FT351}	0.621%	81.46%	8.63
62	{FT351, MH102, DN015} ⇒ {DC111}	0.584%	81.35%	10.46
63	{FT351, FT542, FT543} ⇒ {FT541}	0.521%	80.92%	11.45
64	{TR071, DC181} ⇒ {DN008}	0.519%	80.87%	12.50
65	{MI005, PD101, MH101, FR001} ⇒ {CM001}	0.579%	80.78%	28.46
66	{PD101, CM001, PD103} ⇒ {FR001}	0.599%	80.70%	26.65
67	{DN007, MH101} ⇒ {DN012}	1.287%	80.65%	3.94
68	{CK204, CK101} ⇒ {CK201}	0.662%	80.54%	18.38
69	{DN054, MH106, DN012} ⇒ {MH101}	0.809%	80.41%	5.28
70	{DC111, LM050, DN015} ⇒ {GY201}	0.593%	80.35%	13.05
71	{DC111, FT541, DN012} ⇒ {FT351}	0.515%	80.29%	8.50
72	{DN012, PD101, MH101, FR001} ⇒ {CM001}	0.642%	80.23%	28.26

B.3 Rules with Consequent {F}

Table 11: The top twenty rules, ranked by interestingness, with consequent {*F*}

	Antecedent	Support	Confidence	Lift	Interest.
1	{FT471}	4.987%	94.10%	1.72	0.00165
2	{FT472}	4.494%	98.13%	1.8	0.00140
3	{PD101}	4.014%	87.23%	1.6	0.00098
4	{MI005}	3.892%	85.18%	1.56	0.00089
5	{DN007}	3.469%	85.79%	1.57	0.00072
6	{CK111}	3.192%	96.30%	1.78	0.00069
7	{CK102}	3.021%	87.88%	1.61	0.00056

8	{FR001}	2.766%	91.34%	1.68	0.00049
9	{FT472, FT471}	2.491%	98.10%	1.8	0.00042
10	{CM001}	2.506%	88.27%	1.62	0.00039
11	{PD101, FR001}	2.390%	91.46%	1.68	0.00037
12	{PD101, CM001}	2.245%	89.61%	1.64	0.00032
13	{DN007, DN012}	2.247%	85.07%	1.59	0.00030
14	{MI005, PD101}	2.175%	87.96%	1.61	0.00029
15	{TR084}	2.111%	92.13%	1.69	0.00029
16	{CM001, FR001}	1.961%	90.86%	1.66	0.00025
17	{PD103}	1.987%	84.03%	1.54	0.00023
18	{PD101, CM001, FR001}	1.823%	91.33%	1.67	0.00021
19	{DN012, PD101}	1.780%	89.77%	1.65	0.00020
20	{MI005, FR001}	1.717%	90.67%	1.66	0.00019

B.4 Rules with Consequent {M}

Table 12: The top twenty rules, ranked by interestingness, with consequent {*M*}

	Antecedent	Support	Confidence	Lift	Interest.
1	{FT125}	2.907%	87.12%	1.92	0.00062
2	{TR032}	2.556%	80.30%	1.77	0.00044
3	{WD025}	2.318%	90.09%	1.98	0.00040
4	{GA042}	2.269%	91.59%	2.02	0.00039
5	{LC017}	2.273%	87.85%	1.93	0.00038
6	{LC019}	2.17%	91.75%	2.02	0.00036
7	{FT111}	2.18%	88.74%	1.95	0.00035
8	{FT221}	2.015%	89.65%	1.97	0.00030
9	{FT281}	1.767%	84.07%	1.85	0.00022
10	{DC191}	1.468%	88.55%	1.95	0.00016
11	{CR107}	1.432%	88.71%	1.95	0.00015
12	{CR108}	1.397%	91.25%	2.01	0.00015
13	{DN003, FT125}	1.269%	83.99%	1.85	0.00011
14	{LC019, LC017}	1.205%	93.51%	2.06	0.00011
15	{LM069}	1.259%	81.86%	1.8	0.00011
16	{TR032, FT125}	1.161%	85.95%	1.89	0.000097
17	{WD026}	1.127%	90.99%	2	0.000095
18	{GY402}	1.168%	82.52%	1.82	0.000095
19	{DC192}	1.068%	84.16%	1.85	0.000081
20	{LM081}	1.027%	89.76%	1.98	0.000078

APPENDIX C

RESULTS OF ANALYSES FOR ALTERNATIVE

SUPPORT & CONFIDENCE THRESHOLDS

C.1 Reducing the Confidence

One of the criticisms sometimes levelled at association rule analysis is the sensitivity of the results to perturbation of the support or confidence thresholds. Therefore, in the interest of completeness, this section presents the results of analyses carried out with support threshold as low as 0.1% and confidence threshold as low as 50%.

Note that while a support threshold of 0.1% may be considered 'too' low, it is certainly the case hat a rule with confidence 50% is not useful. If a rule $A \Rightarrow B$ has confidence 50% then B is as likely to occur as not, given that A occurs.

The support threshold was set at 0.5% and rules were mined using a confidence threshold of 70%; these rules where then pruned so that only supersets remained, as explained in Section 3.2.1. This process was repeated for confidence thresholds of 60% and 50% respectively; the number of rules mined and remaining after pruning in each case is given in Table 13.

Table 13: Number of rules for a support threshold of 0.5%

Confidence Threshold	Rules Mined	Rules After Pruning
70%	291	143
60%	520	254
50%	728	366

C.2 Looking at Law & Medicine Subsets

C.2.1 Rules Containing Medicine Courses

From the 366 rules that came about using a support threshold of 0.5% and a confidence threshold of 50%, the subset of rules that contained at least one of the five medicine courses – one from each of the institutions offering medicine: UCD, TCD, RCSI, UCC and NUIG – was searched for the presence of any of the courses in Table 14.

Table 14: Courses for which the medicine subset was searched

Commerce at UCC	Law at UCC
Law and German at TCD	Law at TCD
Psychology at TCD	Law and French at TCD
Commerce at UCD	Law at UCD
Architecture at UCD	Law at NUIG
Actuarial and Financial Studies at UCD	Commerce at NUIG
Architecture at DIT	Law and Accountancy at UL
Financial and Actuarial Mathematics at DCU	

No occurrence of the presence of any of these courses within the medicine subset was observed.

C.2.2 Rules Containing Law Courses

The subset of the rules containing law courses – Law at UCC, Law at TCD, Law at UCD and Law at NUIG – was then examined for the presence of the courses in Table 15.

Table 15: Courses for which the law subset was searched

Commerce at UCC	Dentistry at UCC
Psychology at TCD	Dental Science at TCD
Pharmacy at TCD	Architecture at UCD
Actuarial and Financial Studies at UCD	Veterinary Medicine at UCD
Commerce at NUIG	Architecture at DIT
Financial and Actuarial Mathematics at DCU	Optometry at DIT
Commerce at UCD	

None of these courses were found to be present amongst any of the rules within the law subset. Note that the courses in Table 14 and Table 15 were chosen due to the high points requirement typically attached to them.

C.3 Reducing the Support & Confidence

The analysis of Appendix C.1 was repeated with the minimum support threshold reduced to 0.1% and the confidence threshold held at 50%. This resulted in the generation of 11,571 rules which reduced to 4,649 rules following the usual pruning method.

C.3.1 Rules Containing Medicine Courses

The subset of rules containing medical courses was searched for the courses given in Table 14. Six of these rules contained Law at TCD, with two of the six also containing Law at UCD, but none of these rules were supported by more than 73 students. Two further rules also contained Law at UCD, they were supported by 69 and 65 students respectively.

There were two rules that contained Psychology at TCD with Medicine at UCD and Medicine at TCD. However, these rules were supported by just 56 students and it is arguable that psychology and medicine are not very different disciplines. It is also possible that geography was a factor for these 56 students.

C.3.2 Rules Containing Law Courses

The subset of the 4,649 rules that remained after pruning containing at least one of the law courses offered at UCC, TCD, UCD and NUIG was searched for the presence of any of the rules in Table 15.

Fifteen of these rules contained Commerce at UCC; all of these rules contained only one law course, that offered at UCC. Furthermore, 14 of these 15 rules were comprised exclusively of courses offered at UCC, with the other rule containing two UCC courses and a course at UL. This may be another example of geography, or perhaps university, being an important factor in course selection. It should be noted that, if the position of the '$\Rightarrow$' sign is ignored, there are really only eight distinct associations amongst these 15 rules and the best supported rule is supported by 83 applicants.

Sixteen of the rules within the law subset contain Commerce at UCD; this number reduces to nine

when the order of the items within the rules is ignored. All but three of these rules are comprised exclusively of Dublin courses and the best supported rule applies to 73 applicants.

Six of these rules involve Commerce at NUIG; which reduces to four when the order of the courses within the rules is ignored. Two of these rules (one, if order is ignored) are supported by 167 applicants and exclusively contain courses in NUIG; perhaps another example of geography being a factor in course selection. The four other rules were supported by no more than 63 applicants.

Four courses within the law subset contain Psychology at TCD, two of which comprise the same items in different order and all of which contain Dublin courses only; the best supported of these rules was supported by 84 applicants.

Interestingly, although not given in Table 15, it was noted that only one rule in this subset contained Law and Accountancy at UL and that this rule was supported by just 61 applicants. The other courses in this rule were Corporate Law at NUIG and Law at NUIG.

FIRST VOTE OF THANKS PROPOSED BY MS. CAITRÍONA RYAN, HEAD OF POLICY & PLANNING, HIGHER EDUCATION AUTHORITY.

Thanks to the author for a very interesting paper that highlighted some very important findings and I think findings which we would all welcome. Every year there is huge hype around the CAO; allegations of unfairness, help-lines, etc. There is also concern that sometimes students pick courses on the basis of the points that they think they will get rather than select the course that they would like to study. Claims are based on anecdotal evidence at best and are often made by onlookers not involved in the system. This paper addresses this issue and is very welcome.

It is very gratifying that the findings note that students pick courses on the basis of topic rather than points. Students anticipating getting high points are not necessarily feeling obliged to choose the courses that will require those points – this is good for society and creates confidence that those following particular career paths have a desire and interest to follow in that direction. Of course one thing the paper cannot get behind is the student's thought process prior to making an application, e.g. are there students interested in medicine, veterinary etc. dissuaded from even placing these on CAO application in the first place because they know that the points are well beyond their reach. For example, a student likely to get 400 or 450 points.

This paper puts back into focus what the CAO process is – merely a means to allocate a set of places among interested students. Not a way of valuing or ranking courses – this is very important. A dangerous aspect of the points system and one noted in the paper is the perception that courses are somehow ranked in value by the points they demand; this is obviously not the case and points only reflect supply and demand, number of places available. A recent example is the area of ICT and computer science; this area was critical to the economy, and while there has been some recovery in recent years, demand and hence points have fallen. There are concerns that this has led to a lowering of course quality and concerns that all people taking these courses today are of lower academic ability then in the past. However, there is evidence to show that this is not the case.

The proportion of acceptors gaining 450-plus points in their Leaving Certificate accepting Science courses at third level has in fact increased since the year 2000, as shown in the table below.

High Point (450+) Science Level 8 Acceptors 2000 – 2005.

Discipline	2000	2002	2003	2005
Science	12.5	12.6	12.2	12.8

There are some other interesting statistics that I would like to quote:

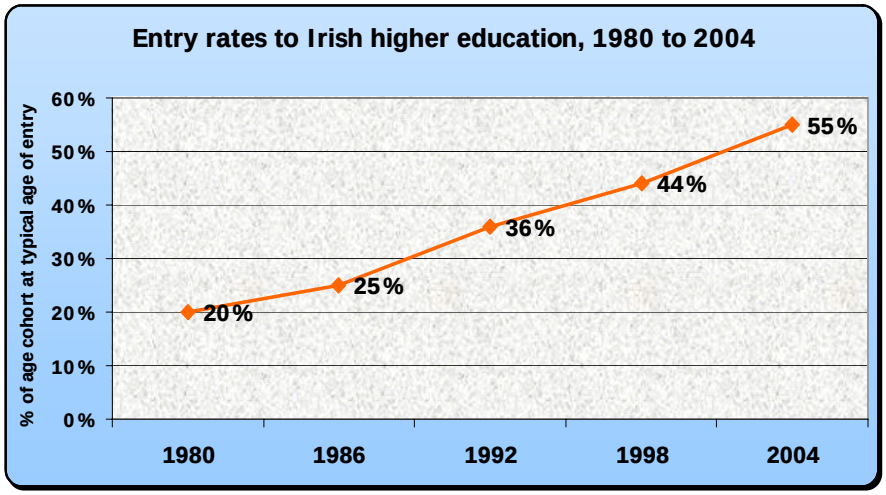
- 13.8% of people with more than 550 points go into Arts.
- 18.8% of people with more than 550 points go into Technology courses.
- 19.6% of people with between 500 and 545 points go into Arts.
- 27.4% of people with between 500 and 545 points go into Technology courses.

The HEA also had some concerns following the CAO process this year. The media coverage trumpeted the end of the points race and created an image of: declining student numbers due to declining school leaver cohort, reduced competition for courses (except for an elite subset) and

underutilised infrastructure.

Such inaccurate perceptions are negative for the further development of the HE sector; a sector so central to our social and economic development and to making the case for increased investment.

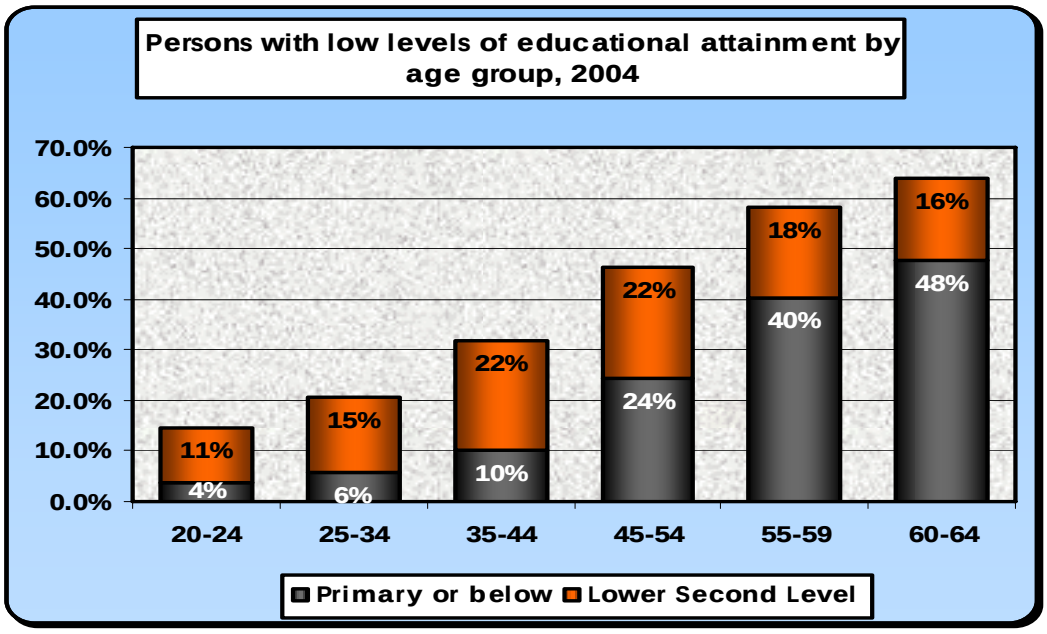
Media failed to contextualise some important points that are illustrated in the graphs and figures below.



Increasing participation rates

	Full-Time	Part-Time
1998/99	108,509	27,764
1999/00	115,696	31,469
2000/01	119,991	32,265
2001/02	124,589	34,965
2002/03	129,283	34,680
2003/04	133,887	34,000
2004/05	133,691	34,509
% Increase 1998/99 - 2004/05	23%	24%

Increasing enrolments over the period 1998/99 – 2004/05



Educational attainment levels by age for the year 2004

School leaving demographics

The overall picture is very good. There is increasing demand for higher education in Ireland, while other countries, such as the UK, struggle in this respect. Last year, CAO applications and acceptances, just under 39,000 (38,955), were the highest ever. Access to higher education is widening – individuals who get two honours or more have almost equal probability to go onto HE irrespective of socio-economic background (O'Connell study). The higher education sector is responsive to student demands and the needs of the economy in terms of supplying them with the graduates they need. Of course, this is not to say that there are not more challenges ahead.

This paper has legitimised CAO somewhat, but I think it is reasonable to say that it should not be only the entry route to higher education. The HEA conducted a study some years ago on courses with high entry points; the recommendation was that graduate entry for these courses should be considered as well as the traditional routes. This is going to happen for medicine courses; with graduate-entry to commence in autumn 2007.

There are issues around recognition of development and work experience, and around the appropriateness of the Leaving Certificate for measuring skills. What skills are being measured? Should qualities like independent thinking, the ability to work in a team, critical analysis, active citizenship, and other transferable skills be considered? However, what do we mean by aptitude, how can it be measured? We need to recognise diversity of careers and it would be unwise not to continue to interrogate our approach to assessment and entry.

Thanks for the time you have given me and thanks again to the author for the paper. Congratulations to the author on winning the Barrington medal.

SECOND VOTE OF THANKS PROPOSED BY MS. CLAIRE GORMLEY, SCHOOL OF MATHEMATICAL SCIENCES, UNIVERSITY COLLEGE DUBLIN.

It is a pleasure to second the vote of thanks expressed by Caitríona Ryan and congratulate the author on an excellent piece of statistical research and a valuable insight to the third level education applications procedure in Ireland. The set of applications from the year 2000 is a high dimensional data set; it is notoriously difficult to analyze and summarize such data but the methodology presented here achieves this seamlessly.

The paper presents an intuitive and straight forward method of examining high dimensional transaction data sets which has a wide range of applications. Association rules are now a part of every day life. Many retailers now record purchases and use association rules to both spatially organize their goods within store and also to provide 'special-offers' to customers suited to their purchasing habits. The theory behind association rules is intuitive and this is a large part of their attraction. The concepts of support and confidence for example are easily interpreted as in general people have an intuitive grasp of joint probability and conditional probability.

The area of third level applications is a topic which annually provides fodder for journalists, opposition parties and students. Clearly, the employment of an optimal third level applications procedure is vital to ensure that Ireland is moving forward towards a knowledge-based economy, which relies on high quality graduates. Hence ensuring the system works in the manner intended is paramount. The statistical approach presented provides a theoretically sound method of validating the system and complements previous investigations given in Gormley and Murphy (2006).

The use of association rules to analyse third level applications provides an intuitive, understandable method of summarizing the vast data set which arises. For example, the assertion

that applicants who select Law at both TCD and NUIG are 28 times more likely to select Law at UCD than if no information regarding their course choices is available, is a result that government, universities and applicants will feel at ease with. The strength of association rules lies in their clarity and interpretability.

I wish to comment on a few statistical aspects of the paper. Quantifying uncertainty is a large part of what statisticians do. Quantifying uncertainty within the context of association rules is not however a straight forward concept. Bootstrapping the confidence of rules would be one method of providing a picture of the variation in the results although this would admittedly be computationally expensive with regard to this data set. Gormley and Murphy (2006) and Gormley (2006) employ a parametric modelling approach to this data set in which the estimation of standard errors is not problematic.

In a similar vein, the robustness of the reported rules to different mining algorithms is of interest. An alternative algorithm, the 'eclat' algorithm is available within the implemented R package. It would be interesting to examine correlations between the association rule clusters suggested by the 'eclat' algorithm and association rule clusters suggested by the 'apriori' algorithm.

There are many methods of calculating the number of 'clusters' present in a population within model-based approaches to clustering. Quantifying the number of clusters within a non-parametric approach however is not so intuitive. Estimating the number of association rule clusters using a more objective approach would be desirable. As pruning has already occurred perhaps rules with the same antecedent would be clustered together, providing a somewhat statistically based estimate of the number of clusters.

Overall these are minor points and the author should be congratulated for addressing a topic of both social and economic relevance in Ireland. The conclusions presented complement and further previous work done in this area. The author has dealt with computational issues and avoided complicated statistical modelling to provide insight to the Irish third level applications process. It is therefore my great pleasure to second the vote of thanks.

Bibliography

Gormley, I.C. & Murphy, T.B. (2006), 'Analysis of Irish third-level college applications data', *Journal of the Royal Statistical Society, Series A* 169(2), 361–379.

Gormley, I.C. (2006), Statistical models for rank data, PhD thesis, Department of Statistics, Trinity College Dublin.

DR. THOMAS BRENDAN MURPHY, DEPARTMENT OF STATISTICS, SCHOOL OF COMPUTER SCIENCE AND STATISTICS, TRINITY COLLEGE DUBLIN.

I'd like to congratulate the author on his excellent paper. The paper gives important insights into the Irish college applications system. In addition, his analysis provides an model-free alternative analysis to the model-based analysis completed in Gormley and Murphy (2006). It's comforting to see that two very different analyses found quite a few similar structures in the data.

The model-based analysis of Gormley and Murphy (2006) used a mixture of Plackett-Luce (Plackett, 1975) models to model the data; this modelling framework allowed for the ranking structure in the data to be exploited in the analysis. Can the association rule methodology be adapted to exploit the ranking data structure?

I would like to provide some extra information to support the conclusions of Gormley and Murphy (2006) that there is some evidence of a points race.

The Plackett-Luce mixture model provides estimates of the probability of an applicant belonging to each mixture component and also the probability of selecting courses in first place for applicants belonging to each mixture component; these probabilities can be used to calculate the probability of a course choice lower down the application than first place.

In our paper, we described the Health Science component in detail. In our analysis we found that UCD Arts, UCD Law, TCD Law and UCD Engineering were amongst the courses with the highest selection probability in the Health Science component; these probabilities are low for the probability of the course being chosen in first place on the application but are much higher for appearing further down the application. It is interesting that these courses had a higher selection probability than some health science courses. This shows some evidence of a points race.

You comment that UCD Arts has a higher selection probability than the Law degrees and offer this as evidence against our conclusions. It is worth noting that the UCD Arts course was the most frequently selected course in the 2002 CAO data; you show this in Figure 1 and 2. In our analysis, UCD Arts had a relatively high selection probability in most of the mixture components; the Cork-based, Limerick-based and "noise" components being the main exceptions. In fact, its position in the Health Science component is lower than in most other components.

In contrast, the UCD and TCD Law degrees only have a relatively high selection probability in a few components, Law and Health Science being the main examples.

Once again, congratulations on a thought provoking paper. It was very interesting to explore the similarities and differences in our results.

Bibliography

Gormley, I.C. & Murphy, T.B. (2006), 'Analysis of Irish Third Level College Applications Data', *Journal of the Royal Statistical Society, Series A*, 169, 361–379.

Plackett, R.L. (1975), 'The analysis of permutations', *Applied Statistics,* **24,** 193–202.

MS. DEIRDRE TOHER, DEPARTMENT OF STATISTICS, SCHOOL OF COMPUTER SCIENCE AND STATISTICS, TRINITY COLLEGE DUBLIN.

In reference to the findings that there is no evidence to support a points race within the applications of 2000, analysis of the subset of high achievers would perhaps be more appropriate. It is acknowledged that if such a points race exists, it is likely to do so within the group that assume, prior to sitting the Leaving Certificate, they are likely to achieve high points. However, the absolute numbers within such a group is likely to be dwarfed by the total numbers applying to the CAO, thus evidence of a points race would be difficult to find using only the data studied in this case.

By linking the total points scored by each applicant, the courses selected and the points assigned to those courses in the previous year, a more complete picture of whether applicants select courses solely by course content, or if there is an element of "playing the game" – choosing courses based on the points that the students expect to get.

To examine if there is a points race within high achieving students, I suggest filtering the students

by the actual points scored, either 480 (averaging 6 B2's) or 510 (averaging 6 B1's) points, as students who realistically expected to achieve extremely high points totals when filling out the CAO application form are unlikely to have obtained total points lower than these figures. Separately analysing the course selection of these individuals should be able to provide a definitive answer to the existence of a point race question.

Furthermore, by linking the points obtained by applicants and the points assigned to each course in the previous year (1999) to the course selections made, it could be determined if students are choosing courses, possibly within their field of interest, based on how they expect to perform. This may be of more interest within studies determining the ranking that students assign to courses - are they ranking courses based on the points value of the previous year, or solely on the desire to do that particular course?

REJOINDER

Thanks to Ms. Ryan and Ms. Gormley for proposing and seconding, respectively, the vote of thanks. I think that Ms. Ryan's comments are very positive. The data she presented suggest a very healthy third level education sector, as well as some challenges that the HEA seem to be addressing.

In the below sections, I reply directly to the questions and comments that were submitted by Ms. Gormley, Dr. Murphy and Ms. Toher. In addition to these colleagues, I would like to thank the following members for their interesting remarks about this paper: Dr. Hederman, Mr. Punch, Mr. Quill and Prof. Walsh.

Reply to Ms. Gormley

Bootstrapping

In the modelling environment used by Gormley & Murphy (2006), bootstrapping is a logical and useful approach to estimating the standard error of the model parameters. The association rule approach, however, as explained in this work, does not involve 'parameters' *per se* and no underlying statistical model is assumed. In all instances when the support, confidence, lift or interestingness of a rule is given, it is based upon the entire dataset of CAO degree course applicants from the year 2000. In a statistical sense, I regard these data as a population and not as a sample.

Furthermore, if these data were to be considered as a sample and bootstrapping was used, it may be better applied to lift rather than confidence.

Mining Algorithm

The apriori algorithm of the arules package in R was used in all instances to mine the association rules given in this work. This algorithm returns all association rules given predefined support and confidence thresholds – this algorithm works in such a way as to be entirely consistent, giving no variation in output for fixed support and confident thresholds. Therefore, I have no concerns regarding the robustness of this algorithm.

There are alternative algorithms available for association rule mining. I chose to use the apriori algorithm because it mines rules in such a transparent and efficient manner. The eclat algorithm, which is also available within the arules package, is a good alternative that outputs equivalence classes with support at least as great as some predefined threshold. These equivalence classes could then be pruned using the method of Section 3.2.1 and then grouped into clusters in the same

fashion as that used herein. Therefore, there would be no clear advantage in using eclat over apriori for this analysis and in fact, the nature of the output given by eclat even be could be considered a drawback. Further information about eclat is given by Zaki *et al*. (1997) and Borgelt (2003).

Method of Estimating Clusters

I accept that my method of 'clustering' the data was somewhat arbitrary and that if another analyst were to cluster the 72 rules given in Table 10, for example, then a different set of groupings may well emerge. However, I contend that under any alternative clustering of these 72 rules, none of the clusters would contain courses related by high entry points alone – and it is this that is of most importance.

I do not think that forming clusters based on common antecedent would provide meaningful clusters and I do not agree that such a method of clustering would be in any way statistically-based. Consider the four rules involving medicine that appear in Table 10, they all have different antecedents and yet clearly belong in the same cluster.

Furthermore, if clustering were performed based on identical antecedents then the 72 rules in Table 10 would have led to the formation of 71 clusters, since only two of them have the same antecedent.

Reply to Dr. Murphy

Association Rules for Rank Data

Methods of applying association rule mining techniques to categorical data are given by Tan *et* al. (2005, Chapter 7). However, these methods tend to involve treating membership of each category as a binary variable and so may not be entirely suitable to rank data. Furthermore, while I can envisage that the methodologies used herein could be modified to account for rank data, I think that the modifications required would take the resulting methods outside of the realm of association rule mining.

The best way to work ranking into the analysis of CAO data may well be to use the previous year's points requirements for courses to predict the ranking of courses on an applicant's CAO form. Such an analysis may well yield interesting results.

The Analysis of the Medicine Subset

I would like to begin by reiterating that I found that the analysis carried out by Gormley & Murphy (2006), on the same data as is analysed herein, provided great insight into the CAO application system. Specifically, their analysis of the health sciences component yielded very interesting results; in particular the placement of Arts at UCD, Law at UCD, Law at UCD and Engineering at UCD "amongst the courses with the highest selection probability in the Health Science component." It is also of interest that these courses had a higher probability of selection than some health science courses.

However, while I accept that this may be a reflection on the mentality of some students that are drawn to courses like medicine and law because of the status associated with them, I do not agree that this provides evidence of a points race. Although this difference may be semantic, I believe that it is important to distinguish between applicants in general and a small cohort of applicants who chose a health science course. In my opinion, the number of students involved is too small to assert that a points race does exist.

Various subsets of these data may contain a relatively high proportion of students who chose courses based primarily on points 'value'. However, as suggested by this analysis, the actual number of students in such cohorts will be small – too small, I contend, to assert the existence of a points race.

Reply to Ms. Toher

While I think that it would be interesting to investigate the psychology around the order in which students select their courses, I must emphasize my position that if a points race, in the negative sense, really existed then it would have revealed itself in the analyses contained herein.

Bibliography

Borgelt, C. (2003), Efficient implementations of apriori and eclat, in 'Workshop on Frequent Itemset Mining Implementations', Melbourne, FL.

Gormley, I.C. & Murphy, T.B. (2006), 'Analysis of Irish third-level college applications data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**(2), 361–379.

Tan, P.N., Steinbach. M. & Kumar, V. (2005), *Introduction to Data Mining*, Addison-Wesley, Boston, MA.

Zaki, M.J., Parthasarathy, S., Ogihara, M. & Li, W. (1997), 'New algorithms for fast discovery of association rules', Technical Report 651, Computer Science Department, University of Rochester, Rochester, NY 14627