Intelligent Information Systems IV
Proceedings of the Workshop held in
Augustów, Poland, 5-9 June, 1995

pages 1- 10

What Do College Ranking Data Tell Us About Student Retention: Causal Discovery In Action

Marek J. Druzdzel[1] and Clark Glymour[2]

[1] University of Pittsburgh
Department of Information Science
and Intelligent Systems Program
Pittsburgh, PA 15260
marek@lis.pitt.edu

[2] Carnegie Mellon University
Department of Philosophy
Pittsburgh, PA 15213
cg09+@andrew.cmu.edu

Abstract. We describe an application of the TETRAD II causal discovery program to the problem of search for causes of low student retention in U.S. universities. TETRAD II discovers a class of possible causal structures of a system from a data set containing measurements of the system variables. The significance of learning the causal structure of a system is that it allows for predicting the effect of interventions into the system, crucial in policy making.
Our data sets contained information on 204 U.S. national universities, collected by the *US News and World Report* magazine for the purpose of college ranking in 1992 and 1993. One apparently robust finding of our study is that student retention is directly related to the average standardized test scores of the incoming freshmen. When test scores of incoming students are controlled for, factors such as student faculty ratio, faculty salary, and university's educational expenses per student are all independent of graduation rates, and, therefore, do not seem to directly influence student retention. As the test scores are indicators of the overall quality of the incoming students, we predict that one of the most effective ways of improving student retention in an individual university is increasing the university's selectivity.

Keywords: TETRAD II, causal discovery, knowledge discovery in databases

1 Introduction

TETRAD II[1] [4] is a computer program embedding recently developed methods for causal discovery from observation. These methods, described in [5], consist of search procedures that have as goal identifying the causal structure of a system under study, i.e., the class of causal graphs that are compatible with the observed values of the system variables. The significance of learning the causal structure of a system is that it allows for predicting the effect of interventions into the system, crucial in policy making. The methods employed by TETRAD are closely related to the methods of induction of probabilistic models from data (e.g., [1]). They come with theoretical proofs of correctness and reliability and we believe that they are a significant improvement on the methods such as regression searches, used in standard practice.

The application of TETRAD that we describe in this paper is a search for the causes of low student retention in U.S. universities. Low student retention is a major source of worries for many U.S. universities. Even though some American universities achieve a student retention rate of over 90%, the mean retention rate tends to be close to 55% and in some universities fewer than 20% of the incoming students graduate (see Figure 1 for the distribution of graduation rates across a set of 200 U.S. national universities). It should be noted here that the data include both academic and non-academic dropout (e.g., students who dropped out because of financial reasons or those who transferred to other schools). Low student retention usually means a waste in effort, money, and

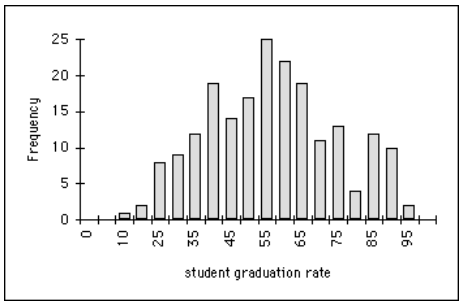

Fig. 1. Histogram of the 1993 graduation rates for 200 U.S. national universities (Source: *U.S. News and World Report, 1994 College Guide*).

human potential. Retention rate is often thought to indicate student satisfaction with their university program and, hence, indirectly, the quality of the university. Indeed, a significant correlation can be observed between university ranking and retention rate — universities close to the top of ranking lists tend to have high retention rates. Is a university's low student retention rate an indication of shortcomings in the quality of

[1] We will abbreviate the name of the program to TETRAD in the remainder of the paper.

education, facilities available to students, tuition costs, university's location, or perhaps wrong admission policies? More importantly, what action can the university take to improve the student retention rate? Can such actions as higher spending on student facilities, increasing the student/faculty ratio, increasing quality standards for teaching faculty, or modifications to admission policies make a difference?

The significance of learning the causes of low student retention, and the significance of learning the causal structure of a system in general, is that this allows for predicting the effect of interventions into the system. While applying, for example, simple regressions to the data would allow to make predictions about the value of a variable of interest given the values of other variables, this would not be sufficient for the purpose of policy making. What the leadership of a university wants is to predict the effects of external manipulations of the system by means of new policies aimed at improving the retention rate. For this, we need information about the underlying causal structure of the system. We therefore believe that determining the interactions among different relevant variables, including the direction of these interactions, is the necessary first step in addressing the problem. As large-scale experiments may be too expensive, ethically suspect, or otherwise impractical, research on this problem needs to rely mainly on observations. The analysis has to be practically limited to extracting patterns from large collections of measurements of relevant variables.

This paper describes a preliminary effort to see what, if anything, aggregate data for many U.S. universities can tell us about the problem. Our analysis involved data concerning 204 U.S. universities, collected annually by the *U.S. News and World Report* magazine for the purpose of their college ranking.[2] While we are far from giving clear cut answers to the questions posed above, we believe that our analysis provides some interesting insight into the problem. The available data suggests that the main factor in student retention among the studied variables is the average test scores (or other measures of academic ability) of incoming students. The test scores of matriculating students are a function of the quality of the applicants and the university's selectivity. High selectivity leads to high average test scores of the incoming students and effectively to higher student retention rates. Factors such as student faculty ratio, faculty salary, and university's educational expenses per student do not seem to be directly causally related to student retention. This hypothesis should be checked using data internal to any particular university, especially since the national data are aggregated to include both academic and non-academic dropout. If the national pattern is confirmed locally, we would suggest that, wherever possible, steps aimed at making the university more selective be taken. Improving the comparative image of the school, and therefore increasing the number of applicants, increasing the selectivity of the admission process, increasing the chance that good applicants will accept admission offer rather than choosing another university, should improve student retention in the long run.

The remainder of the paper is structured as follows. We describe the analyzed data sets in Section 2. Section 3 summarizes our assumptions about this data and prior information about the problem. The results of our analysis are presented in Section 4. Section 4.1

[2] The data available to us is for the years 1992 and 1993. We report our preliminary findings from the 1992 data in [2].

presents the results of TETRAD's search for possible causal structures that generated the data and Section 4.2 reports the results of applying simple regression to selected interactions identified by TETRAD. Section 5 contains a discussion of these results and policy suggestions.

2 The Data

The data used in our study consists of a set of statistics concerning 204 U.S. national universities[3] collected by the *U.S. News and World Report* magazine for the purpose of college ranking. To prepare the data for its annual ranking of colleges, *U.S. News and World Report* each year goes through a laborious process of data collection from over a thousand U.S. colleges. The data is collected from various university offices, such as admissions or business office, by means of surveys prepared by outside companies. The information obtained from each of the colleges is subsequently verified by the schools representatives. The process of collecting the data and combining them into the final college ranking is described in [3].

The data for national universities provided by *U.S. News and World Report*, contains over 100 variables measured for each of the 204 universities. There are compelling reasons for reducing the number of variables studied. Firstly, the power of statistical tests and the reliability of TETRAD's search depend on the ratio of the number of sample points to the number of variables: the higher the ratio, the better. Secondly, the complete data sets contained considerable redundancy, as many of the variables are analytical derivatives of other variables (e.g., retention rate was simply the ratio of graduating seniors to incoming freshmen, both numbers included separately in the data).

A related issue is that of missing values. Including variables with missing values and calculating covariances by skipping a particular unit for a particular variable would undermine the theoretical reliability of statistical tests. Testing partial correlations involves multiple correlations from the correlation matrix and, since these would not be based on a fixed sample size, the sample size used in the tests would be indeterminate.

We selected the following eight variables as the most, if not only, relevant to our analysis: average percentage of graduation (*apgra*), rejection rate (*rejr*), average test scores of the incoming students (*tstsc*),[4] class standing of the incoming freshmen (*top10*), which is percentage of the incoming freshmen who were in top 10% of their high school graduating class, percentage of admitted students who accept university's offer (*pacc*), total educational and general expenses per student (*spend*), which is the sum spent on instruction, student services, academic support, including libraries and computing services, student teacher ratio (*strat*), and average faculty salary (*salar*).

[3] Defined as major research universities and leading grantors of doctoral degrees.

[4] We owe an explanation for readers who are not familiar with the admission procedure to U.S. universities. Practically every U.S. university requires a prospective student to take a nationwide test, administered by a private educational testing institution. The most popular test is SAT (Scholastic Aptitude Test), but there are others required for specialty schools, such as law or management schools. The score on such a standardized test gives a reasonable measure of the preparation of the applicant and is an important factor in admissions.

Describing each of over 100 remaining variables and discussing why we have not considered them for our analysis would make this paper unacceptably long. We limit ourselves to a few remarks. The values of a large number of the variables were included indirectly in the eight chosen variables. Average test scores of incoming students (*tstsc*), for example, is a normalized compilation of values of 14 variables, including a breakdown of average results for various parts of SAT and ACT tests. Average percentage of graduation (*apgra*) express the essence of 14 variables concerning student retention, such as breakdown of dropout rates across all semesters of the freshman year. Rejection rate (*rejr*) and percentage of admitted students who accept university's offer (*pacc*) express, along with the average test scores (*tstsc*) and class standing (*top10*), selectivity of the school. We chose the total educational and general expenses per student (*spend*), student teacher ratio (*strat*), and average faculty salary (*salar*) as indicators of the quality of school's teaching and its financial resources.

From the complete set of 204 universities, we removed 26(31)[5] universities that had missing values for any of the eight variables of interest. This resulted in a set of 178(173) data points.

3 The Assumptions

Although TETRAD's algorithms are independent on the actual distribution of the variables, they rely on the outcomes of a series of statistical tests. The necessary tests are especially powerful if we can assume normally distributed, linearly related variables. We studied how reasonable this assumption was for the available data set by plotting histograms of each of the eight variables. By visual inspection of the histograms, we removed 8(14) additional data points that appeared to be outliers. The resulting data set, consisting of 170(159) data points, reasonably satisfied the normality assumptions. All histograms were close to symmetric unimodal distributions (see Figures 1 and 2 for example), with the exception of two positively skewed variables, *spend* and *strat*.

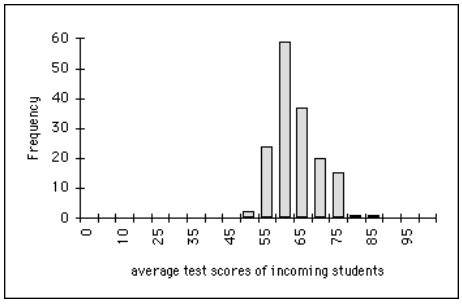

Fig. 2. Histogram of the test scores *tstsc* for the 170 data points (1993 data).

[5] In the sequel, we will report the counts for the 1992 data followed by the counts for the 1993 data in parentheses.

An important assumption made by TETRAD is that the causal structure that generated the data points is acyclic. This assumption was not necessarily true in our data set. For example, most of the variables considered influence the image of the university. The image, in turn, can be argued to influence all of the eight variables. We still think that the acyclicity assumption is reasonable in our data set, as all feedback processes that we can think of in this context are extremely slow acting (at least on the order of decades as opposed to the interaction of our interest between the measured factors and graduation rate), so that in the snapshot provided by our data they can be assumed negligible.

An assumption frequently made in causal modeling is causal sufficiency, which is an assumption that the analyzed variables form a self-contained structure — there are no latent common causes. (An equivalent of this assumption is the assumption that all error terms are independent.) TETRAD allows for search with both the causal sufficiency assumption and without it. As it is unlikely that the selected variables form a self-contained structure, we have run TETRAD without making the causal sufficiency assumption. Several control runs with causal sufficiency assumption did not reveal anything that would put our main conclusions in question.

One way that TETRAD can be aided in its search for the set of causal structures that could have generated the data is by an explicit encoding of prior knowledge about causal relations. TETRAD allows for specifying temporal precedence among variables, information about presence or absence of direct causal connections between pairs of variables, and information about absence of common causes between pairs of variables. With respect to the available data set, we believe that the average spending per student (*spend*), student teacher ratio (*strat*), and faculty salary (*salar*) are determined based on budget considerations and are not influenced by any of the five remaining variables. Rejection rate (*rejr*) and percentage of students who are offered admission (*rejr*) and who accept the university's offer (*pacc*) precede the average test scores (*tstsc*) and class standing (*top10*) of incoming freshmen. Finally, our only assumption about graduation rate, *apgra*, was that it does not cause any of the other variables. These assumptions are reflected in the temporal ordering in the following table, which was the only prior knowledge given to TETRAD.

fte, *spend*, *strat*, *salar*
rejr, *pacc*, *pdoct*
tstsc, *top10*
apgra

4 The Results

4.1 TETRAD

When TETRAD is run on normally distributed data with the linearity assumption, it converts the raw data into a correlation matrix. The values of the elements of this matrix is all that matters in discovery and are all that is needed to reproduce our results whether using TETRAD search or any other approach. The correlation matrix for the 159 data points of the 1993 data set is reproduced in Figure 3.

	spend	apgra	top10	rejr	tstsc	pacc	strat	salar
spend	1.0000							
apgra	0.5455	1.0000						
top10	0.6381	0.5879	1.0000					
rejr	0.4766	0.4720	0.5674	1.0000				
tstsc	0.6732	0.7403	0.7655	0.5813	1.0000			
pacc	-0.3807	-0.4237	-0.2498	-0.0810	-0.2985	1.0000		
strat	-0.7713	-0.3867	-0.3099	-0.2721	-0.4688	0.1909	1.0000	
salar	0.6954	0.6328	0.6025	0.4885	0.6515	-0.5159	-0.3737	1.0000

Fig. 3. Matrix of correlations among the analyzed variables (1993 data set, 159 data points).

When making decisions about independence of a pair of variables conditional on a subset of the remaining variables, TETRAD uses statistical tests (in the normal–linear case, standard **z**–test for conditional independence). The search begins with a complete undirected graph. Edges in this graph are removed by testing for appropriate conditional independence relations. If two variables a and b become independent when conditioned on a subset $\mathcal{S}$ of the remaining variables, there is no direct causal connection between them — all interactions between a and b take place through intermediate variables included in $\mathcal{S}$. This is a simple consequence of two assumptions known as *causal Markov condition* and the *faithfulness condition* [5]. Orientation of the remaining edges is based on a theorem proven in [5]. For example, suppose that two variables a and b are not directly connected (i.e., there exists a subset of the remaining variables $\mathcal{S}$ that makes a and b conditionally independent) and there is an edge between a and c and an edge between b and c. If a and b are independent conditional on $\mathcal{S}$ and dependent conditional on $\mathcal{S} \cup c$, then a and b are both direct causal predecessors of c. In other words, the edges can be oriented from a to c and from b to c. Both, the process of removing edges and the process of orienting edges, can be aided by prior information about the underlying graph. TETRAD allows for specifying presence or absence of direct connections between pairs of variables and also temporal precedence among the variables. Knowledge of temporal precedence allows for limiting the number of tests for conditional independence and, under certain circumstances, aids in orienting the edges of the graph. The details of TETRAD's search algorithm are given in [5].

Depending on the significance level used in independence tests, TETRAD's individual statistical decisions regarding independence may be different and a different class of causal structures may result. It is, therefore, a good practice to run the program at several significance levels. We ran TETRAD with the following four significance levels: $p = 0.2$, 0.1, 0.05, and 0.01. Our earlier simulation studies have indicated that this range is appropriate for data sets of the size available for our problem. The danger of making p too small is that TETRAD will reject weak dependences as insignificant and, in effect, delete arcs that represent weak but genuine causal influences. Classes of the graphs proposed by TETRAD for significance levels $p = 0.05$ and $p = 0.01$ are presented in Figure 4. The edges of the graphs have the following meaning: A single arrow ($\longrightarrow$) denotes a direct causal influence. A double headed arrow ($\longleftrightarrow$) between two variables

denotes presence of a latent common cause of these two variables. An single arrow with a circle at one end (o⟶) expresses TETRAD's inability to deduce whether there is a direct influence between the two variables (⟶) or a latent common cause between them (⟷). An edge with circles at both ends (o—o) expresses TETRAD's inability to deduce whether there is a direct influence between the two variables and, if so, what is its direction (⟶ or ⟵) or a latent common cause between them (⟷). The core of the structure, i.e., how *apgra* is related to the remaining variables, was insensitive to changes in significance. Variations in the interactions among the remaining variables can be attributed to the relatively small size of our data set.

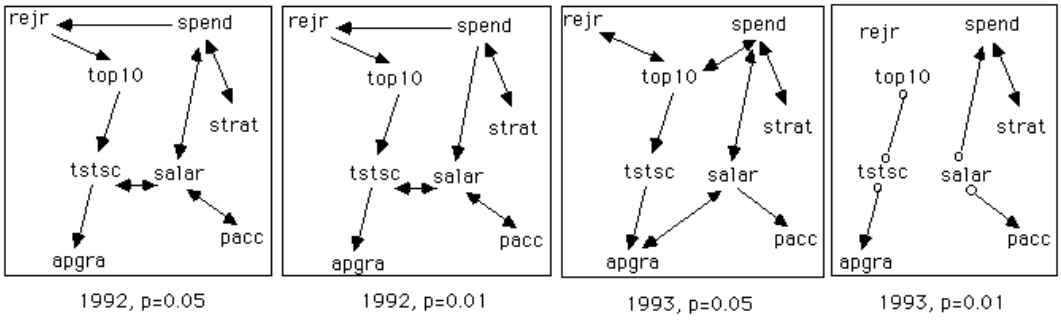

Fig. 4. Causal graphs proposed by TETRAD (significance levels p=0.05 and p=0.01).

In the graphs in Figure 4, as well as in all other graphs suggested by TETRAD, any connection between *apgra* and variables like *spend*, *strat*, or *salar* is through *tstsc*. The "latent common cause" connection between *salar* and *apgra*, shown in Figure 4 for $p = 0.05$, disappears at $p < 0.01$. Virtually all graphs contained a direct causal connection between the average test scores and student graduation rate.

TETRAD's algorithms are much more reliable in determining existence of direct causal links than in determining their orientation. Therefore, prior knowledge supplied to TETRAD may be critical for the orientation of edges of the graph. We used the temporal sequence described in Section 3, but we also checked the robustness of our result to temporal ordering by running TETRAD with no assumptions about temporal precedence. Although TETRAD proposed different orderings of variables, all direct links, and the direct link between test scores and graduation rate in particular, were the same in both cases.

To check whether the causal structure is the same for the top research universities we prepared additional data sets for TETRAD with universities that were in the top 50 universities with respect to their academic ranking. Our results were essentially similar to those of the complete data sets. Any differences between graphs concerned influences among variables different than *apgra* and can be partially attributed to a small number of data points and, hence, susceptibility to chance variations.

4.2 Linear Regression

We applied linear regression to the relation between the main indicator of the quality of incoming freshmen, *tstsc* (average test scores), and *apgra* (graduation rate) to obtain a quantitative measure of the strength of these interactions. We emphasize that we used regression only to estimate the coefficients in a linear model obtained by the TETRAD search. If regression were used instead to search for the variables influencing retention and graduation, it would include variables that TETRAD says have no direct influence on the outcome, and that are conditionally independent of the outcome variables.[6]

In the full data set of 170(159) data points, linear regression applied to *apgra* on *tstsc* resulted in the following equations:

```
1992: apgra = -77.4 + 2.03 tstsc, R-sq(adj) = 60.9%
1993: apgra = -64.9 + 1.89 tstsc, R-sq(adj) = 54.5%
```

In the restricted set of 50 top ranked research universities, the regression equations were:

```
1992: apgra = -84.6 + 2.13 tstsc, R-sq(adj) = 70.0%
1993: apgra = -63.6 + 1.88 tstsc, R-sq(adj) = 64.5%
```

In the group of top ranking universities, the average test scores of incoming freshmen explain as much as 70%(64.5%) of the variance in graduation rates.

5 Discussion

It seems that none of the variables in the data set were directly causally related to student retention except for standardized test scores. This result, following directly from the fact that graduation rate is, given average test scores, conditionally independent of all remaining variables, seems to be robust across varying significance levels, availability of prior knowledge, and data set size. The average test scores seem to have a high predictive power for student graduation rate. For the top 50 ranking research universities, average test scores explain as much as 70%(64.5%) of the variance in graduation rate.

Average test scores of incoming students can be viewed as indicators of the quality of incoming students. It seems that retention rate in an individual university can be improved by increasing the quality of the incoming students. This, in turn, can be improved by increasing the number and the quality of applicants. The better the pool of applicants from which an admission committee can select, the better the accepted students and, hopefully, the better the matriculating students are likely to be. Changing factors such as faculty salary, student teacher ratio, or spending per student should, according to our result, have no direct short-term effect on student retention.

One limitation in our study is that the available *U.S. News* data do not disaggregate academic from non-academic dropout. We predict that internal data will show a difference between average test scores of dropout (academic and non-academic) and graduates.

[6] We regressed *apgra* on the remaining seven variables purely as an academic exercise. For the 1992 data set, regression indicated three predictors to be significant: *tstsc* ($p < 0.001$), *pacc* ($p < 0.003$), and *strat* ($p < 0.023$). For the 1993 data set, four predictors were significant: *tstsc* ($p < 0.001$), *pacc* ($p < 0.002$), *salar* ($p < 0.012$), and *spend* ($p < 0.059$).

Another limitation is that our data do not disaggregate between different departments. Some departments may have many academic dropout, others few. Also, the available data set did not include other variables that may have been relevant, as geographical location (climate, urban/rural, etc.), available academic support, financial situation of the students, prominence of athletics on campus, etc.

Finally, it is possible to apply alternative prior models of interaction of the variables in our data set. One alternative, suggested to us by Steven Klepper, might involve one latent variable influencing all eight variables studied. This model, however, would not account for the strong conditional independences observed in the data, and is in fact rejected in the 1992 data set, for which we checked it, by the standard f ratio test (Chi square of 356 with 27 degrees of freedom).

6 Acknowledgments

Considerable data collection effort and generosity in making the collected data available on the part of *U.S. News and World Report* made this study possible. Steven Klepper, Chris Meek, Richard Scheines and Peter Spirtes contributed to our work with valuable suggestions. We thank Felicia Ferko, Kevin Lamb, and Jeffrey Bolton from Carnegie Mellon University's Office of Planning and Budget for enabling us to access the data files and providing insightful background information. Anonymous reviewers for the Workshop on Knowledge Discovery in Databases 1994, in which we presented the preliminary results of our analysis of the 1992 data [2], prompted us for more details in our presentation. Support for this work has been provided by ONR and NPRDC under grant N00014–93–1–0568 to Carnegie Mellon University.

References

1. Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(NUMBER):309–347, MONTH 1992.
2. Marek J. Druzdzel and Clark Glymour. Application of the TETRAD II program to the study of student retention in U.S. colleges. In *Proceedings of the AAAI–94 Workshop on Knowledge Discovery in Databases (KDD–94)*, pages 419–430, Seattle, WA, 1994.
3. Robert J. Morse, Senior Editor. U.S. News & World Report's America's Best Colleges Rankings: How it's done. Technical report, U.S. News and World Report, Washington, DC, May 8, 1992.
4. Richard Scheines, Peter Spirtes, Clark Glymour, and Christopher Meek. *TETRAD II: Tools for Discovery*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1994.
5. Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, 1993.