Mohammed Alharbi **Spring 2016** MAA271@PITT.EDU
Yixuan Edison Wang YIW72@PITT.EDU

INFSCI 2725: Data Analytics
Assignment 5: Causal Discovery

Introduction

In this assignment, we are going to investigate potential causes of low freshmen retention in the 1993 data of the US universities and comparing the findings with Druzdzel & Glymour's conclusions in their paper, titled “Application of the TETRAD II Program to the Study of Student Retention in U.S. Colleges”.

The analysis has three parts. The first part includes details about the data variables. The second part compares the correlation matrix in both data sets, 1992 and 1993. The last and important part talks about the causal graphs (or patterns) that are suggested by GeNIe.

Note our answer to the required questions will be implicit the two parts bellow. That includes a question such as What causes student retention.

The data features

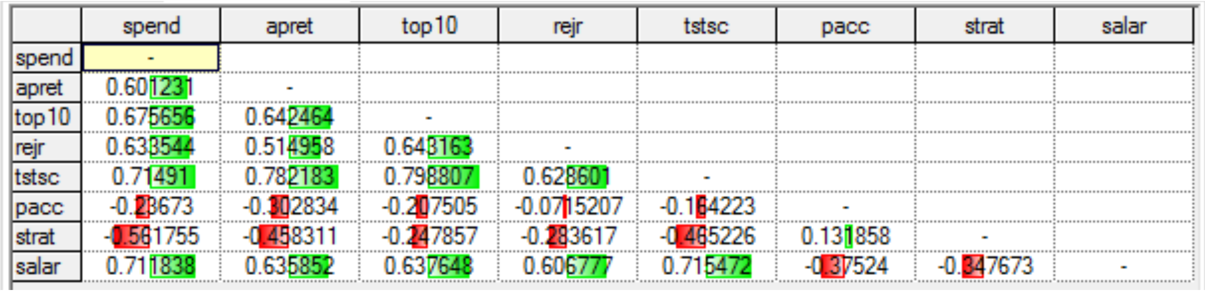
The data has the following nine variables for our analysis:

- apret – average percentage of freshmen retention.
- rejr – rejection rate.
- tstsc – average test scores of the incoming students.
- top10 – class standing of the incoming freshmen, which is percentage of the incoming freshmen who were in top 10% of their high school graduating class.
- pacc – percentage of admitted students who accept university’s offer.
- spend – total educational and general expenses per student, which is the sum spent on instruction, student services, academic support, including libraries and computing services.
- strat – student teacher ratio.
- salar – average faculty salary.

Correlation

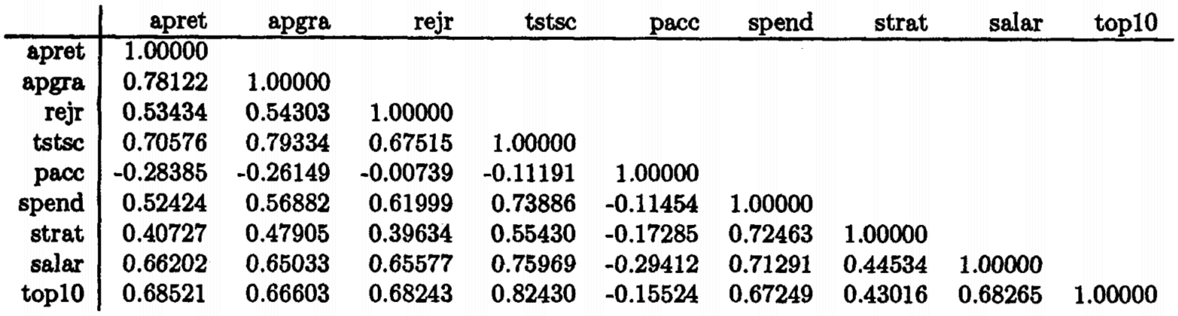
To examine the dependence between all data variables, a correlation matrix has been used. The correlation matrix results the coefficient, ranging between -1 to 1. Zero to One signifies the two variables tend to increase or decrease together. -1 to 0 signifies one variable increases as the other decreases. The most correlation is represented by 1, whereas the perfect negative (or opposite correlation) is represented by -1. The correlation matrix for data of 1993 and 1992 is produced below. There are quite a few differences between this correlation matrix and 1992 correlation matrix.

As it shows in figure 5 Page 426 in the paper and copied bellow, the biggest difference is the relationship between strat variable and other variables. In 1992 only pacc variable has a negative correlation with strat, while correlations of all variables are negative in 1993.



	spend	apret	top10	rejr	tstsc	pacc	strat	salar
spend	-							
apret	0.601231	-						
top10	0.675656	0.642464	-					
rejr	0.633544	0.514958	0.643163	-				
tstsc	0.71491	0.782183	0.798807	0.628601	-			
pacc	-0.23673	-0.302834	-0.207505	-0.0715207	-0.164223	-		
strat	-0.561755	-0.458311	-0.247857	-0.283617	-0.465226	0.131858	-	
salar	0.711838	0.635852	0.637648	0.606777	0.715472	-0.37524	-0.347673	-

Figure 1: Matrix of correlations for 1993 data.



	apret	apgra	rejr	tstsc	pacc	spend	strat	salar	top10
apret	**1.00000**								
apgra	**0.78122**	**1.00000**							
rejr	**0.53434**	**0.54303**	**1.00000**						
tstsc	**0.70576**	**0.79334**	**0.67515**	**1.00000**					
pacc	**-0.28385**	**-0.26149**	**-0.00739**	**-0.11191**	**1.00000**				
spend	**0.52424**	**0.56882**	**0.61999**	**0.73886**	**-0.11454**	**1.00000**			
strat	**0.40727**	**0.47905**	**0.39634**	**0.55430**	**-0.17285**	**0.72463**	**1.00000**		
salar	**0.66202**	**0.65033**	**0.65577**	**0.75969**	**-0.29412**	**0.71291**	**0.44534**	**1.00000**	
top10	**0.68521**	**0.66603**	**0.68243**	**0.82430**	**-0.15524**	**0.67249**	**0.43016**	**0.68265**	**1.00000**

Figure 2: Matrix of correlations for 1992 data.

Note only a few correlations could be presented in patterns, which are generated by PC algorithm. It is an evidence for “**correlation does not mean causation**.”

Patterns

As the main topic in Druzdzel and Glymour's paper, our analysis focused on potential causes of low freshmen retention in the US universities, based on the US universities' data of 1993. We used GeNIe program, applying the PC algorithm on that data with a variety of significance levels, including 10%, 5%, 1%, and 0.1%. The PC algorithm could be run by either providing prior (background) knowledge or without. In both situation, slightly different changes may occur.

The following are testing PC algorithm on the data with the variety of significance levels:

Applying the PC algorithm with a **significance level of 10% or 5%**, which is the default significance level, displays the pattern bellow. Obviously, it shows the average freshmen retention rate (apret) is directly caused by the average test scores (tstsc). There is also a latent common cause between the average freshmen retention rate (apret) and and student teacher ratio (strat).

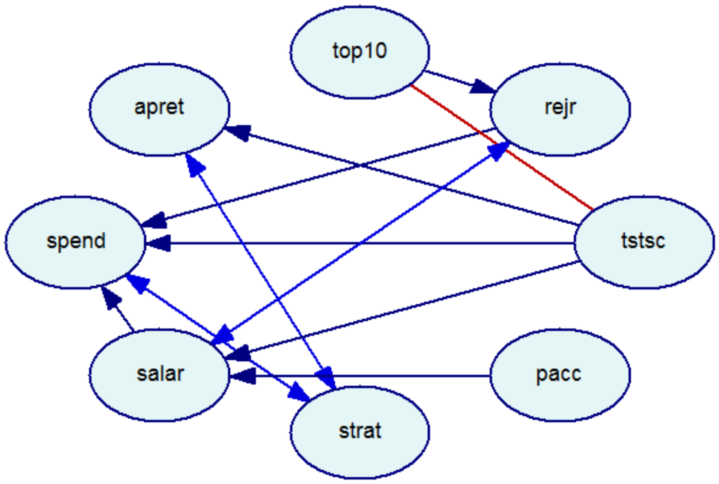


Figure 3: A causal graph (pattern) proposed by PC algorithm for a 1993 data (significance levels p=0.05).

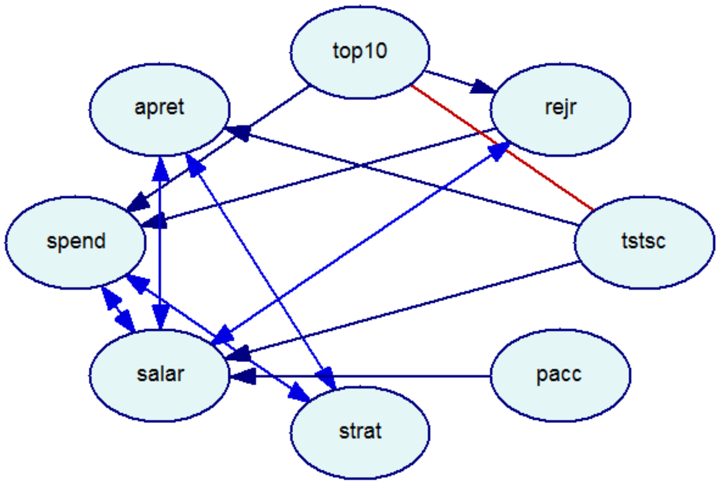


Figure 4: A causal graph (pattern) proposed by PC algorithm for a 1993 data (significance levels p=0.1).

Comparing with Druzdzel & Glymour's findings (figure 5) and for significance level p=0.05, the causal effect of apret is salar, tstsc, top10, and apgra in their paper, while the causal graphs to 1993 data with the same significance level (figure 3) proposed by PC algorithm expresses that the causal effect of apret is only tstsc and strat.

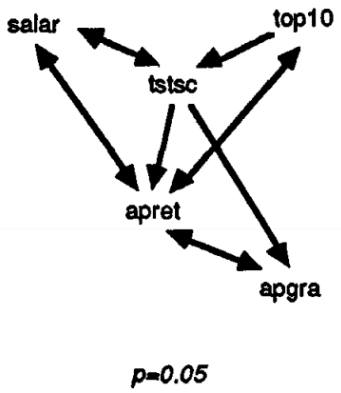


Figure 5: A causal graph (pattern) proposed by PC algorithm for a 1992 data (significance levels p=0.05).

Changing the significance level to a smaller number such as **1% and 0.1%** results the pattern bellow. Investigating this patterns, you can observe that the average freshmen retention rate (apret) does not be caused by student teacher ratio (strat). The edge between apret and strat variables is dropped. Moreover, it is unable to deduce whether there is a direct influence between the average freshmen retention rate (apret) and the test scores (tstsc). In terms of other variables, it is obviously observed that the causality between top10 and spend variables, which appears in the pattern of significance level of 10%. Also, the relation between spend and strat variables becomes more clear, so the average spending per student (spend) is affected by student teacher ratio (strat). Moreover, the connection between rejection rate (rejr) and the average spending per student (spend) is dropped in the pattern of the significance level of 0.1%. The connection between rejr and salar is also dropped.

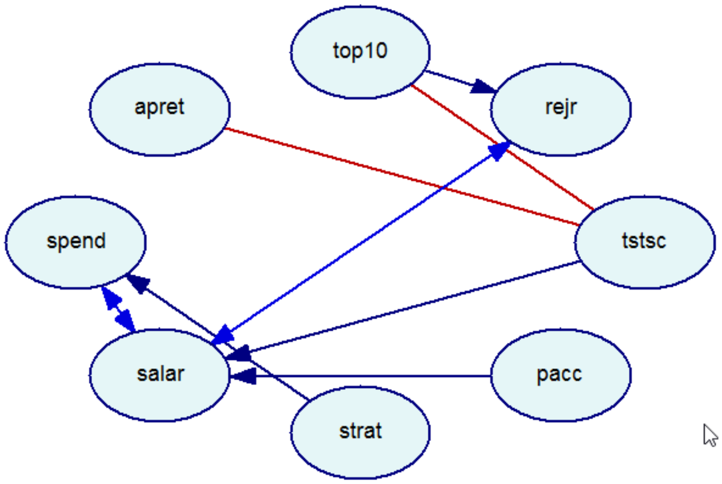


Figure 6: A causal graph (pattern) proposed by PC algorithm for a 1993 data (significance levels p=0.001).

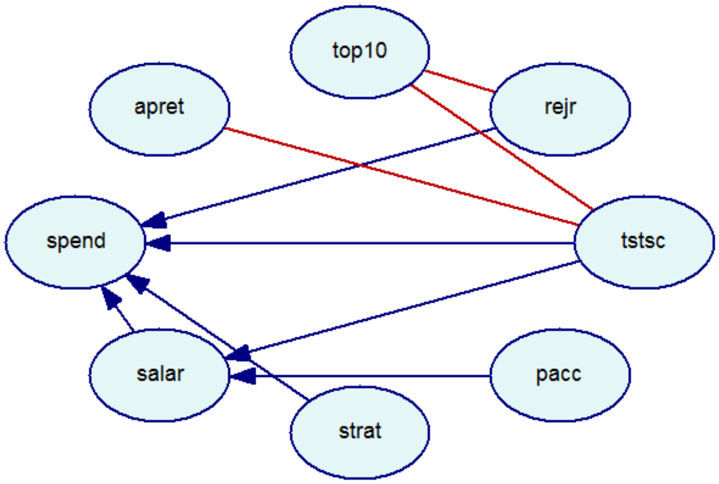


Figure 7: A causal graph (pattern) proposed by PC algorithm for a 1993 data (significance levels p=0.01).

Comparing with Druzdzel & Glymour's findings (figure 8) and for a smaller significance level p=0.001, the causal effect of apret is tstsc and top10 in their paper, while the causal graphs to 1993 data with the same significance level (figure 3) proposed by PC algorithm expresses that the causal effect of apret is tstsc variables, and it is unable to deduce whether there is a direct influence between them.

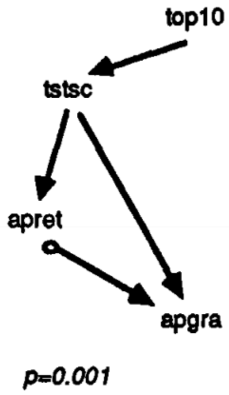


Figure 8: A causal graph (pattern) proposed by PC algorithm for a 1992 data (significance levels p=0.001).

Conclusion

Comparing the 1993 data findings with Druzdzel & Glymour's findings, there are apparently big different in the causality of the average percentage of freshmen retention as it is proposed by PC algorithm in GeNIe. However, the average of test scores (tstsc) remains the most (or strongest) causal effect of the average freshmen retention rate (apret) for both 1992 and 1993 data.