# INFSCI 2725 Data Analytics

## Assignment 5 – Causal Discovery

Date: 3/20/2016
Student name: Tong Wei, Zhenyu Peng
pittID: TOW6, ZHP6

# Executive Summary

This report aims to find the possible cause of low retention by running PC algorithm on a database containing information on 170 U.S. colleges. It begins by analyzing exploring the assumption underlying PC algorithm in GeNIe. In the second part, the causal structure will be illustrated based on different significance level and temporal sequence knowledge. Based on the same temporal sequence as in the paper of Druzdzel & Glymour, we found out the main causal structure under p=0.001 among top10, tstsc and apret (Top10 -> tstsc -> apret) are similar with the finding in the paper. Moreover, as is the same with the result in the paper, the connection for data in 1993 between apret and variables like spend, strat, or salar are through tstsc or top 10. The main difference between the causal structure under p=0.05 stems from the factor that there is nocausal relationship between salar and apret, whereas strat has causal relationship with apret.

# 1.Assumption Check

As PC algorithm assumes all the variables follow a normal distribution, all the data distribution for these eight variables are presented in the graphs below. Druzdzel & Glymour stated that after some data preprocessing, all histograms for their variables are close to symmetric unimodal distributions with the exception of two positively skewed variables, spend and strat. However, in our dataset for 1993 used without any preprocessing, it seems that rejr, spend and top10 all have comparatively skewed distribution, which could possibly contribute to the different casual results in subsequent analysis. As for another implicit assumption of PC algorithm, linearly dependent data, we can see from Figure 9, the top three variables that have greatest correlation with apret are tstsc, top10 and salar (in descending order). Compared with the correlation matrix in the paper, we can observe a decreased correlation between tstsc and apret and increased correlation between top 10, salar and apret. Besides, there is a decrease in correlation between tstsc and top 10. This correlation matrix illustrates the second difference between data in 1993 and 1992.

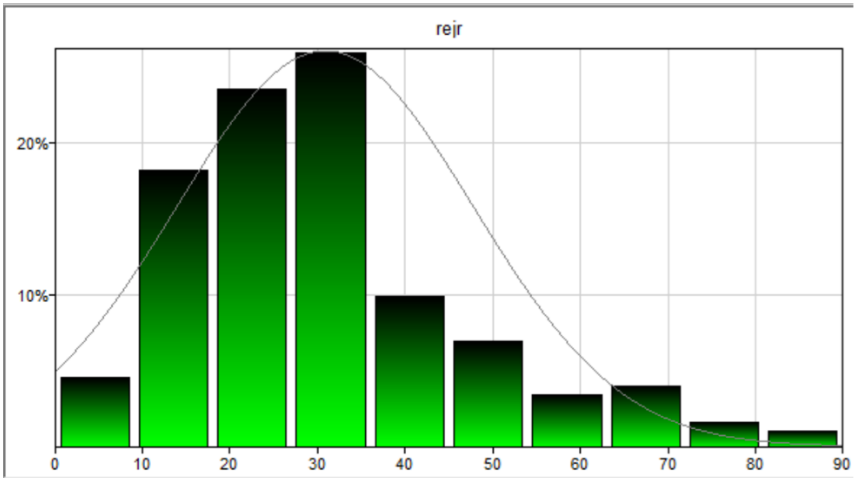


Figure 1: Histogram of rejection rate rejr for the 170 data points.

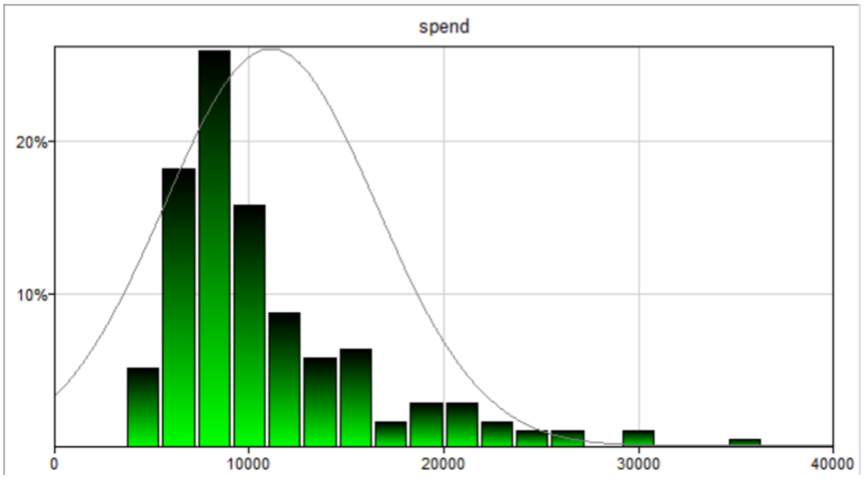


Figure 2: Histogram of the average spending per student spend for the 170 data points.

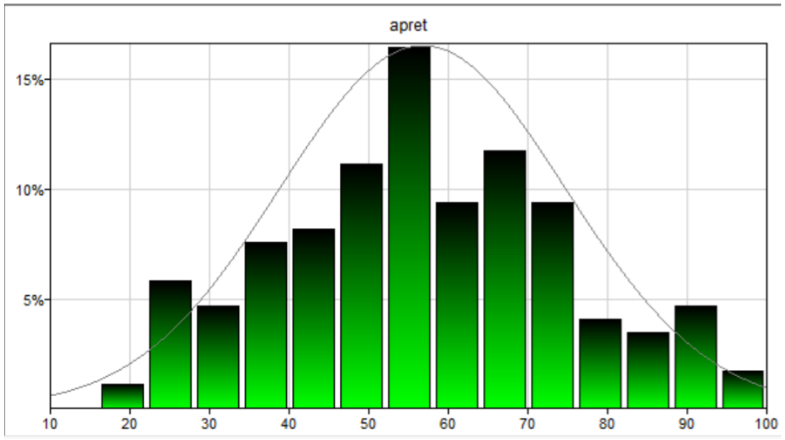


Figure 3: Histogram of the average freshmen retention rate apret for the 170 data points.

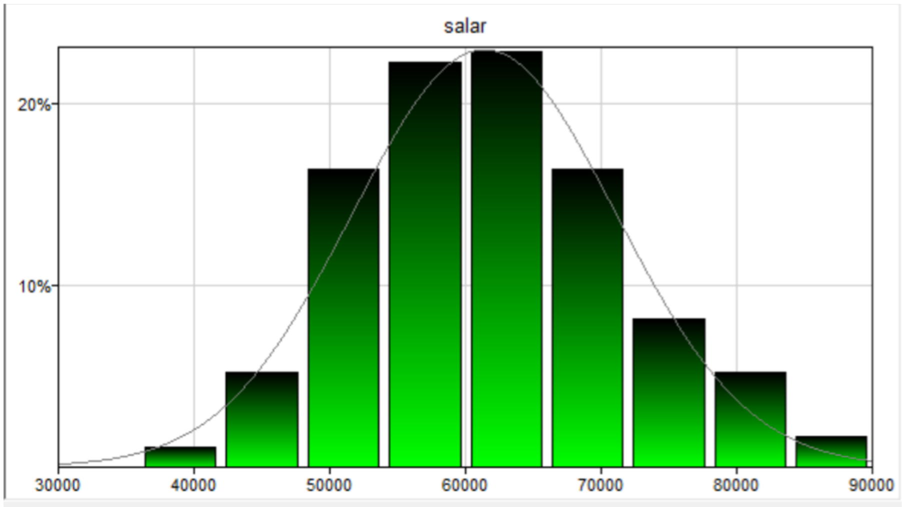


Figure 4: Histogram of faculty salary salar for the 170 data points.

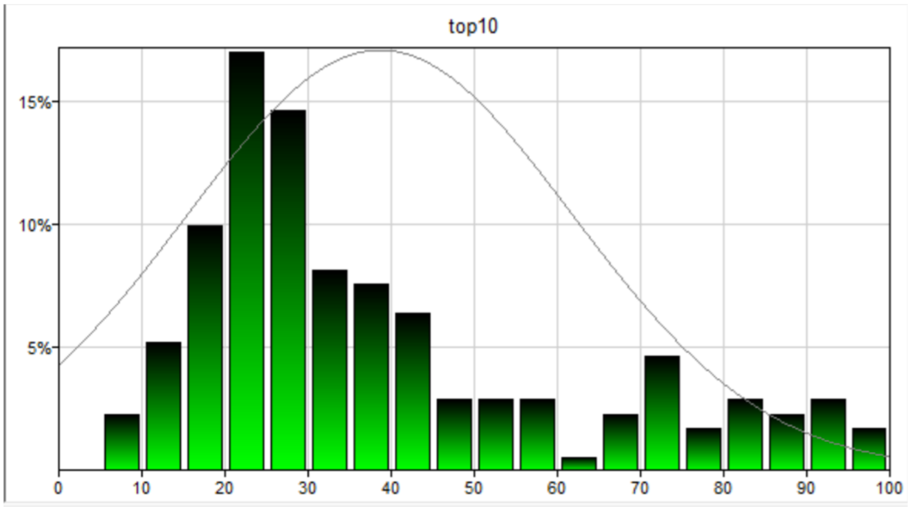


Figure 5: Histogram of class standing top10 for the 170 data points.

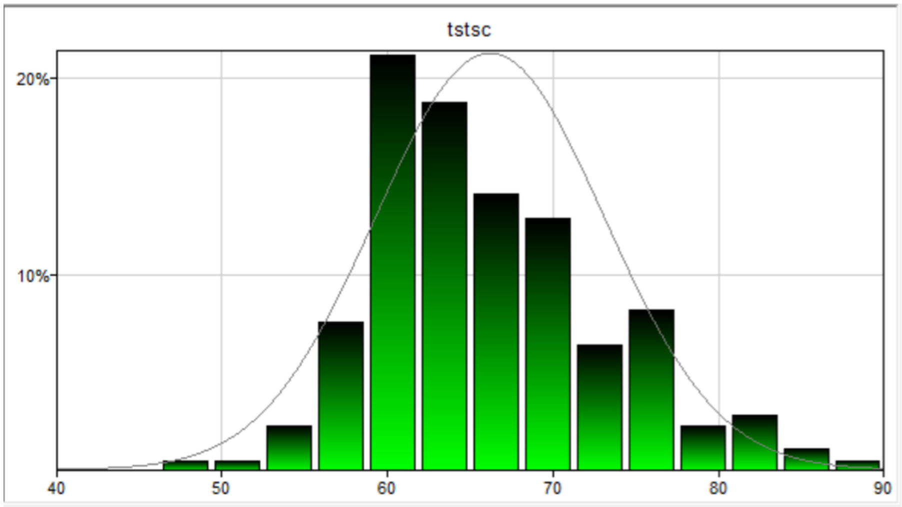


Figure 6: Histogram of average test scores of incoming freshmen for the 170 data points.

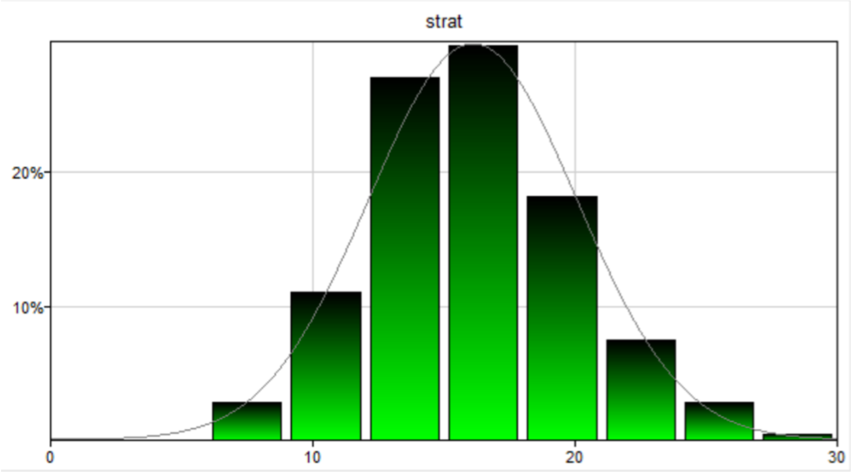


Figure 7: Histogram of student-teacher ratio for the 170 data points.

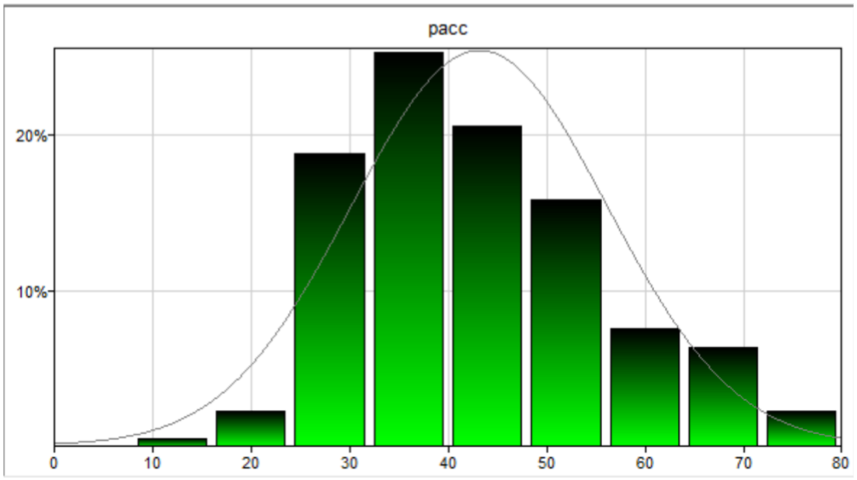
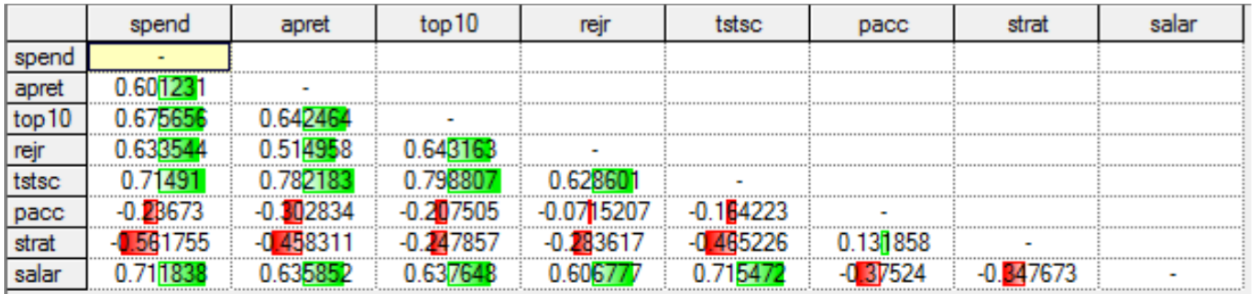


Figure 8: Histogram of percent of admitted applicants who accept university's offer for the 170 data points.



|  | 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 9: Matrix of correlations among the analyzed variables (170 data points).

## 2. Results

In the second part, PC algorithm in Genie is run on the dataset. We will first use the same temporal sequence as the paper (illustrate in Figure 10) to identify the causal structure.

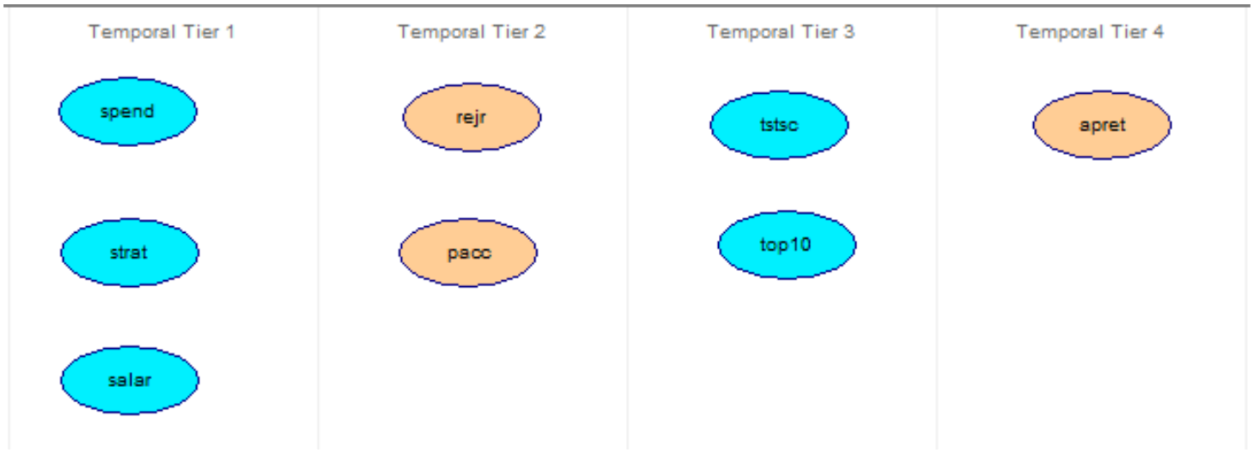
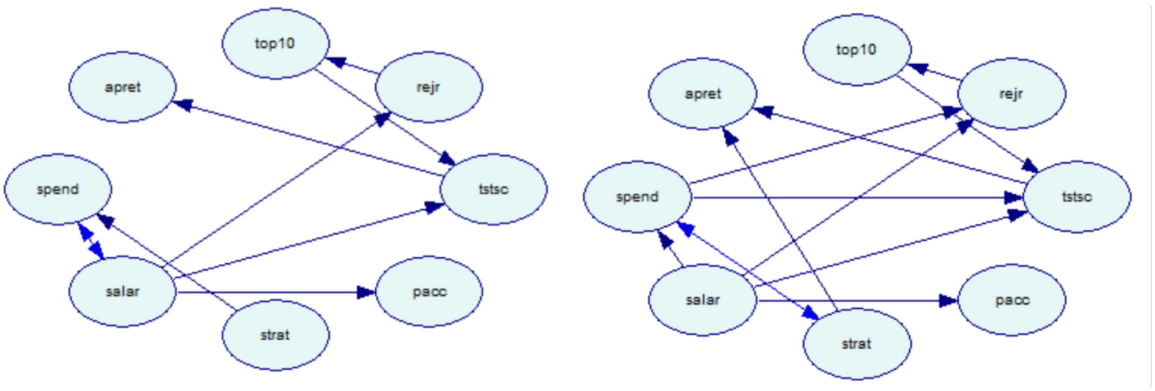


Figure 10: Temporal ordering of variables

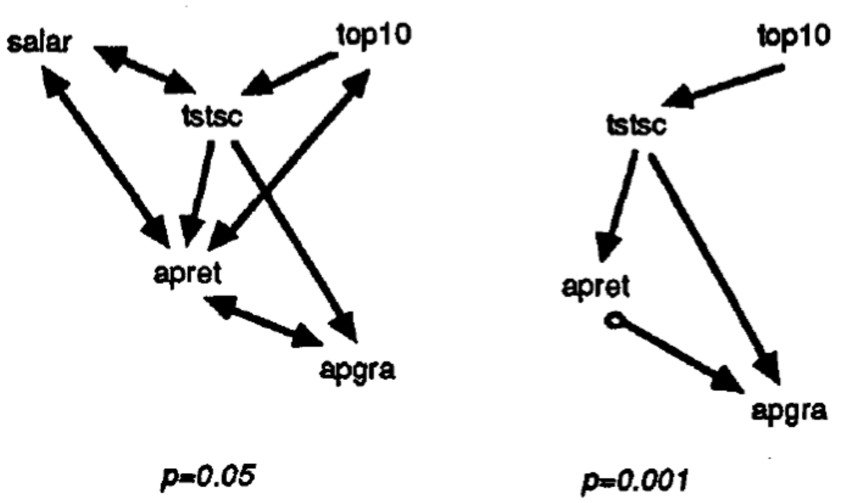


p=0.05

p=0.001

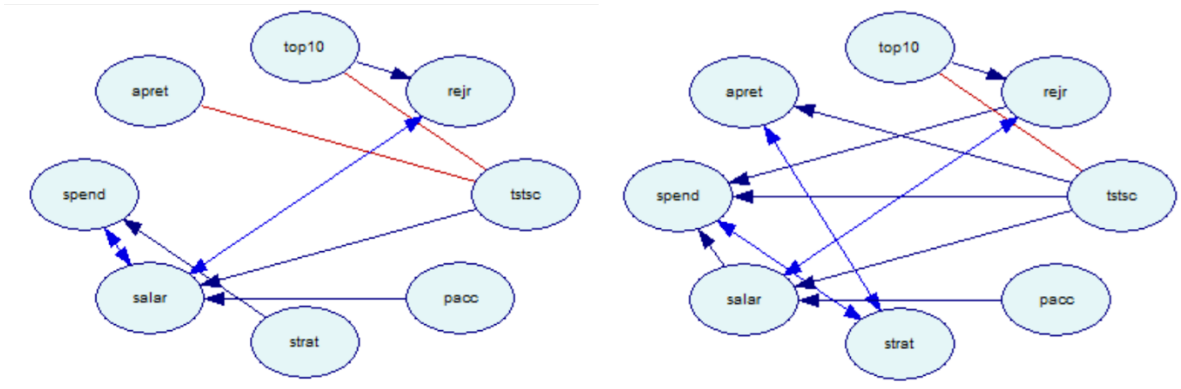
Figure 11: Two causal graphs proposed by PC algorithm for the completedata set of 170universities (significance levels p=0.05 and p=0.001).

Figure above illustrates the causal structure obtained. Changing significance level from 0.05 to 0.001, the latent common cause of spend and strat changes to the direct causal structure from strat to spend. The direct causal structure from salar to spend becomes latent common cause of these two variables. The causal relationship between spend and rejr disappeared. Moreover, we can see the direct causal relationship between strat and apret, spend and tstsc disappeared. Actually, we can see that the part of causal graphs (relationship among top10, tstsc and apret) under p=0.001 between data 1993 and 1992 (Figure 12) are the same. Class standing (top 10) in high school results in higher average test scores, which in turn leads to higher retention rate. Spend, strat and salar, the same as that stated in the paper, are connected with apret through tstsc. The figure for p=0.05 differ in that salar and apret for data in 1993 do not have causal relationship, while strat has causal relationship with apret. That is probably because correlation between apret and salar decrease from 0.65033 in 1992 to 0.63585.



*p=0.05*　*p=0.001*

Figure 12: Two relevant parts of causal graphs proposed by TETRAIDI for the complete data set of 175 universities (significance levels p=0.05 and p=0.001).



p=0.05　p=0.001

Figure 13: Two causal graphs proposed by PC algorithm for the complete data set of 170 universities without temporal sequence (significance levels p=0.05 and p=0.001).

In the paper of Druzdzel & Glymour, they checked the robustness of their result to temporal ordering by running TETRADII with no assumptions about temporal precedence. We also run PC algorithm in Genie with no assumptions about temporal precedence and the new causal graphs are demonstrated in Figure 12. It proves that prior knowledge supplied to PC algorithm is critical for the orientation of edges of the graph. Although the orientation of edges missed in both graphs, all direct links, the direct link between tstsc and apret in particular, were the same in both cases.

## 3. Linear Regression

We applied linear regression to variables (tstsc, strat, top10) on apret in order to get a quantitative measurement of these relationships. We use full data set of 170 colleges. The equations below are results. The first equation is about linear relation for tstsc(average test scores of incoming freshmen) and strat(student-teacher ratio) on apret(average retention rate). We can see that coefficient of tstsc is 1.882 which is much higher than coefficient of strat. Comparing to the first equation, the second equation is for tstsc alone on apret and tstsc alone can explain 60.9% of the variance of retention rate. This linear regression analysis also confirms the high relationship between average score of incoming students and retention rate, as illustrated in the paper. Thus, increasing colleges' selectivity can probably increase the retention rate of students.
Apret=-59.041+1.882 tstsc+-0.544 strat, R-sq(adj)=61.9%

Apret=-77.4+2.027 tstsc, R-sq(adj)=60.9%