Assignment 5: Causal Discovery

Jingyi Lin (jil173) & Yichao Chen (yic85)

I. Data

In this assignment, we will use GeNIe to verify the findings about the university retention data. The datasets provided for our exploration is of the year 1993, while the data used in Druzdzel & Glymour's study was of the year 1992. There may be some differences between these two data sets, which will influence the study's conclusions.

II. Assumption Verification

Firstly, in the study, two authors have mentioned that "*All histograms were close to symmetric unimodal distributions, with the exception of two positively skewed variables, spend and strat.*" To better understand the distribution of each data column in 1993's data, eight histograms will be generated as shown below.

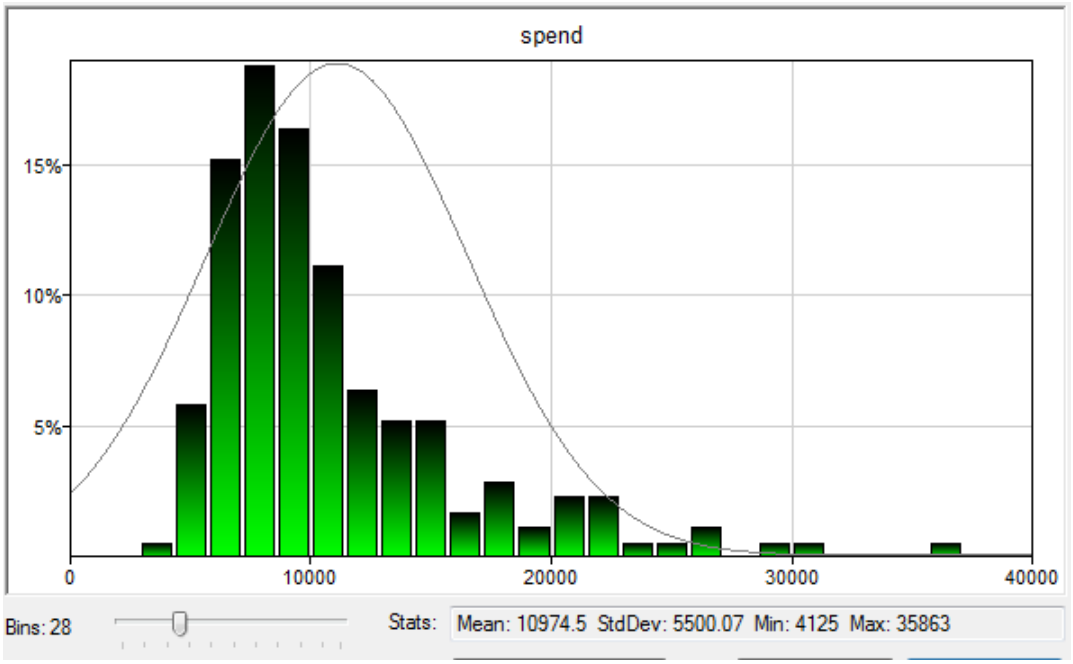


Figure 2.1 Histogram of "spend" – average spending per student

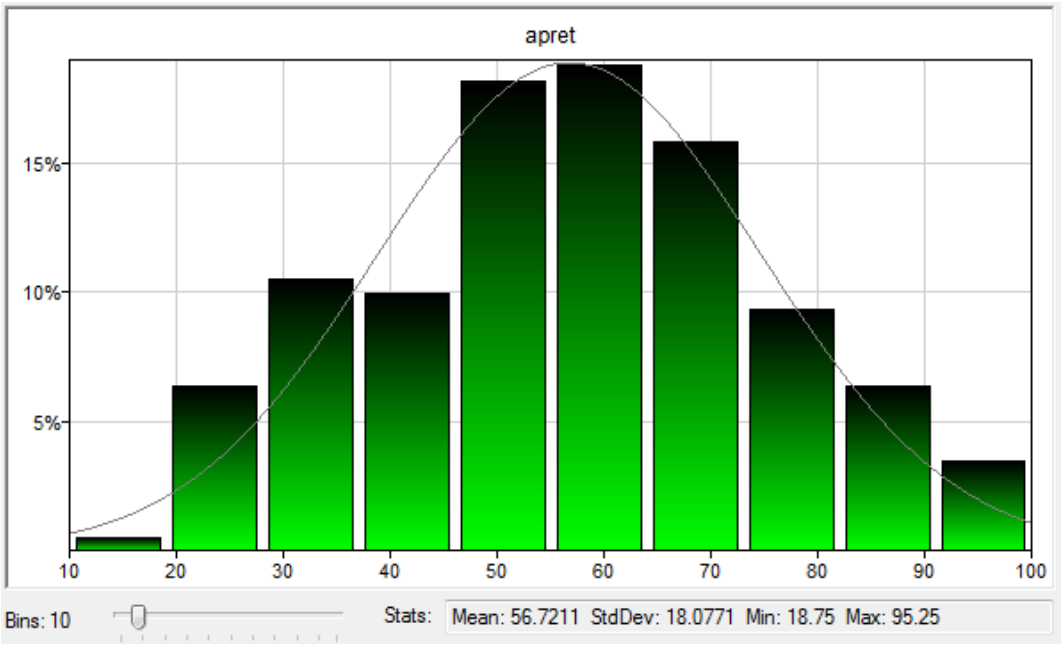


Figure 2.2 Histogram of "apret" – average retention rate

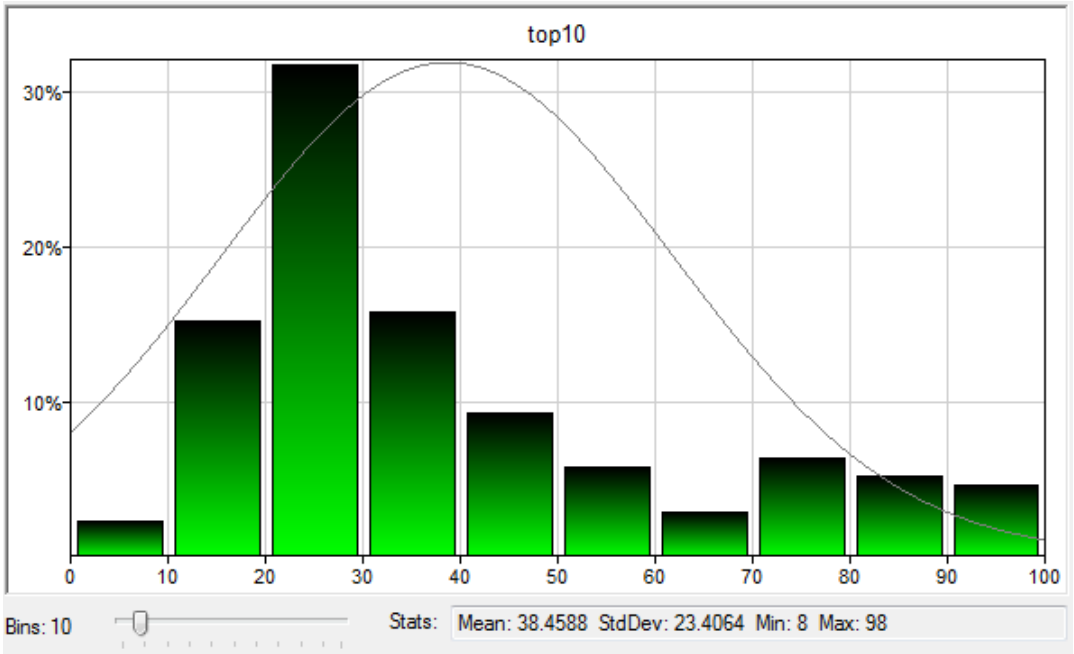


Figure 2.3 Histogram of "top10" – freshmen percentage of top 10% students

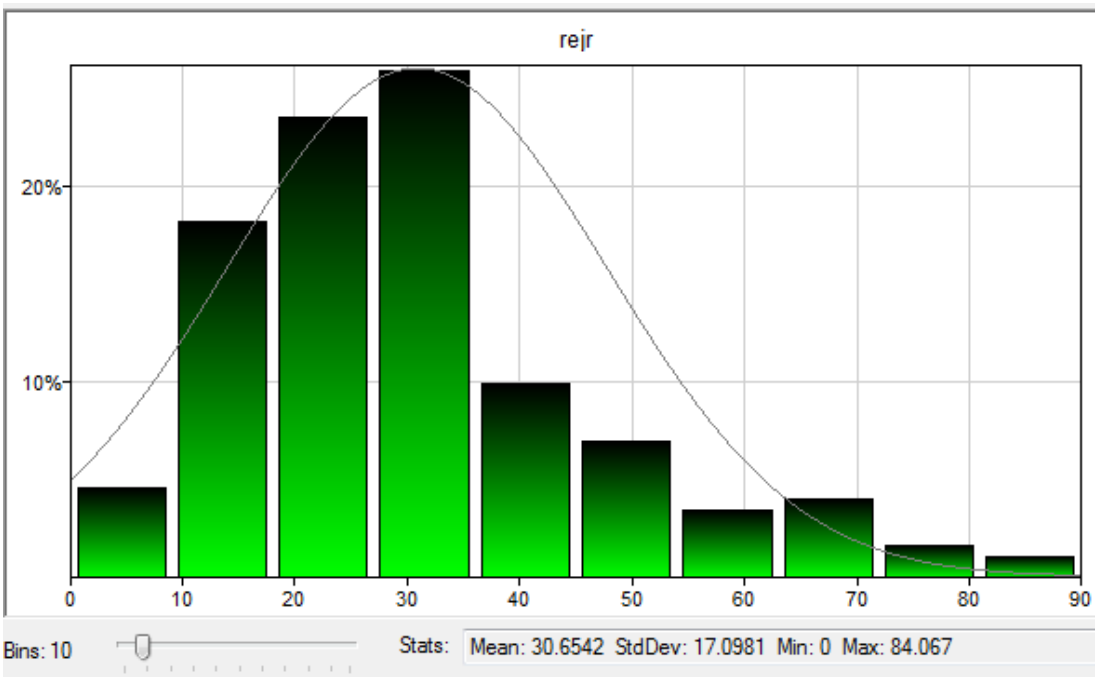


Figure 2.4 Histogram of "rejr" – rejection rate

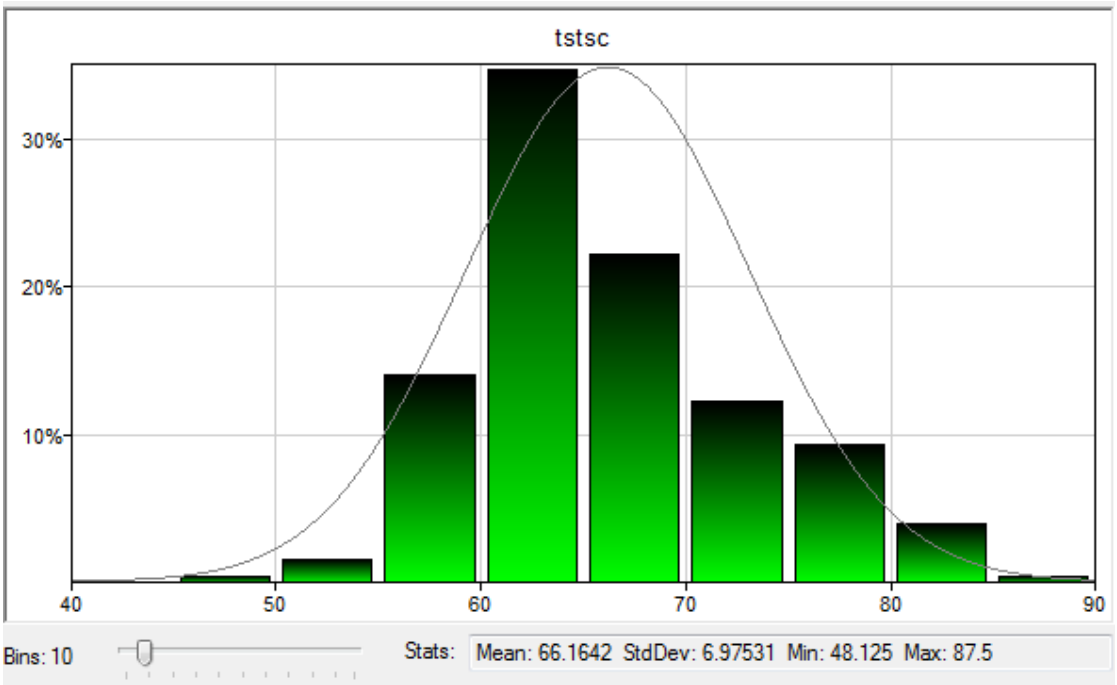


Figure 2.5 Histogram of "tstsc" – average test scores of freshmen

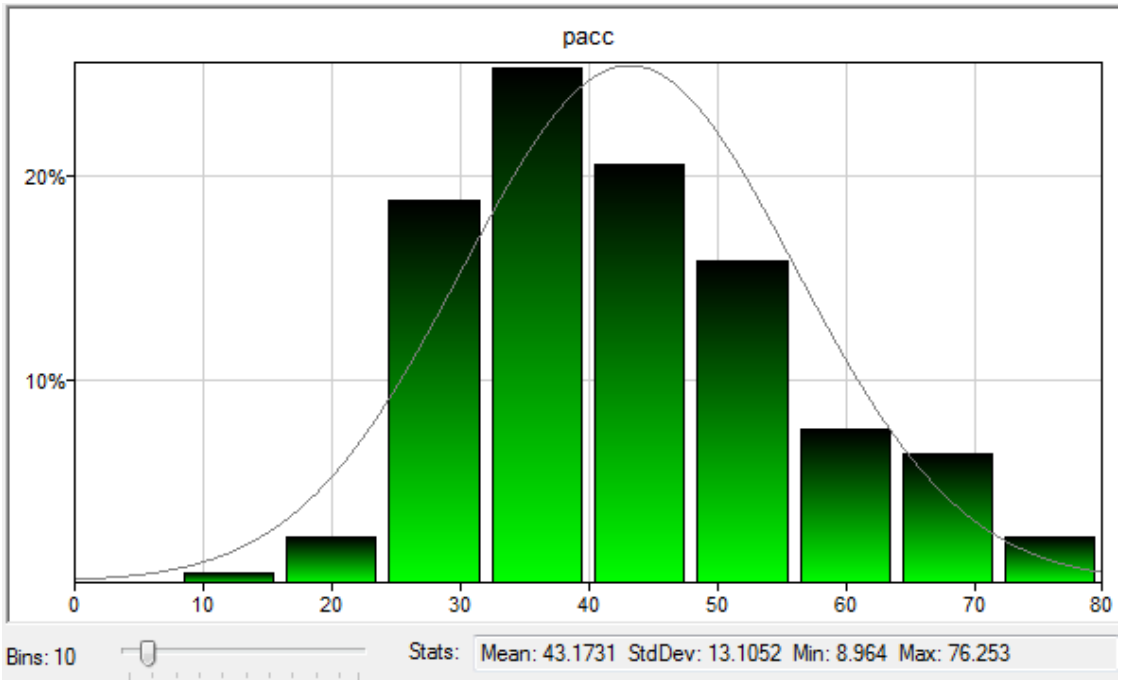


Figure 2.6 Histogram of "pacc" – percentage of acceptance of the offer

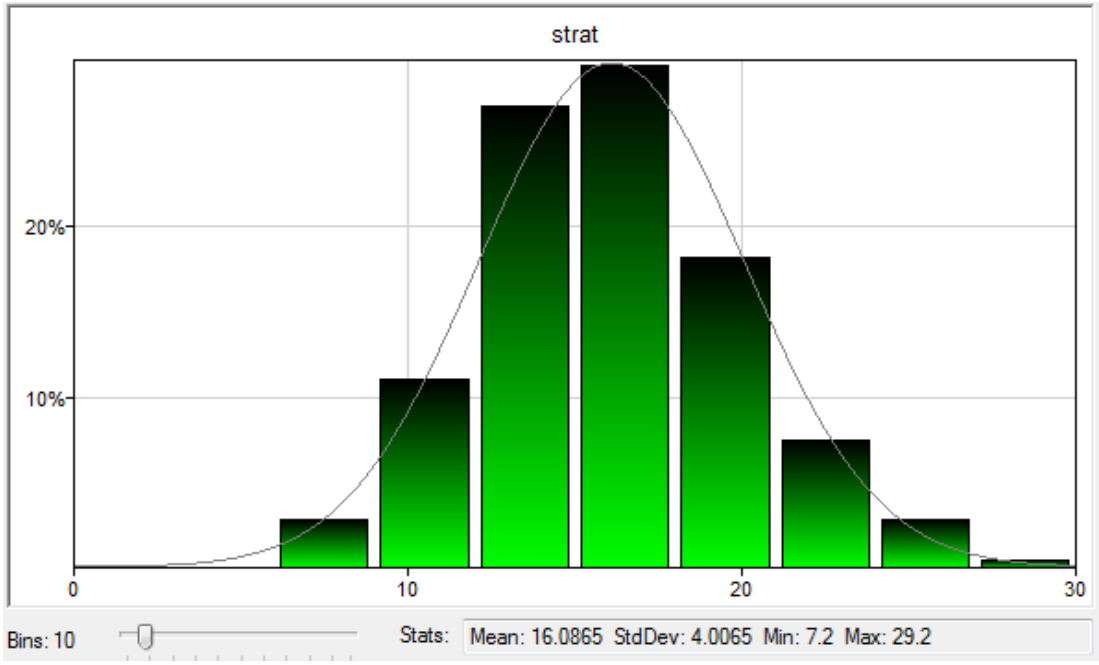


Figure 2.7 Histogram of "strat" – student-teacher ratio

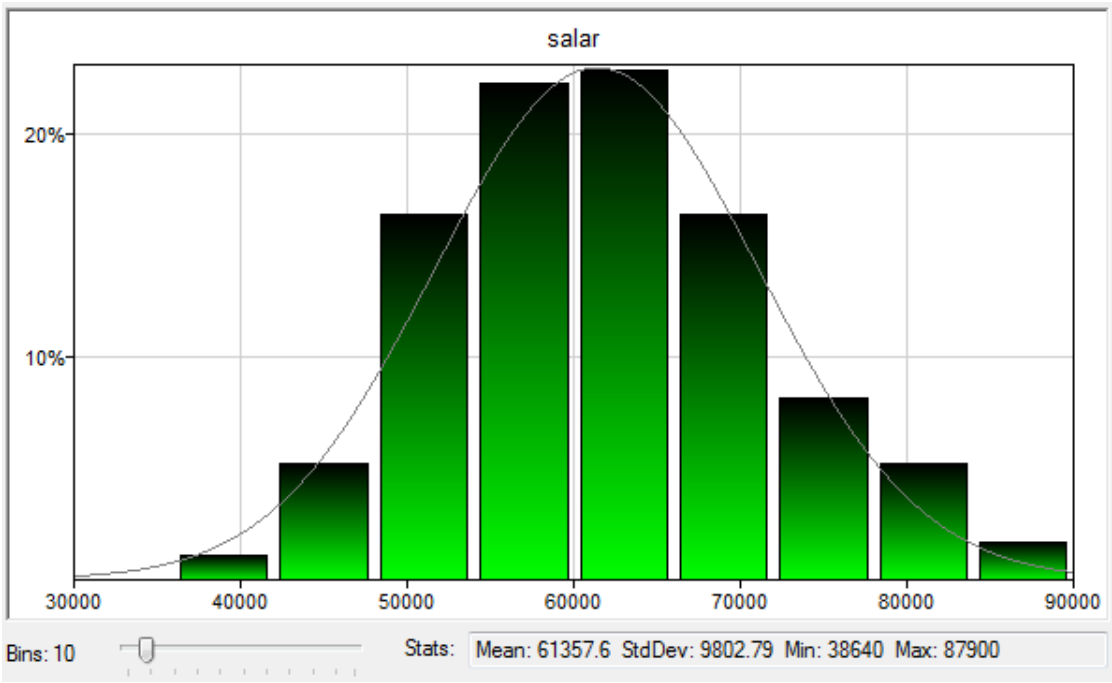


Figure 2.8 Histogram of "salar" – average faculty salary

From the histograms above, we can see that Figure 2.1, Figure 2.3, Figure 2.4, which are "spend", "top 10" and "rejr", are not symmetric unimodal distributions. It is different as the assumption made in the study. The differences will eventually cause the different casual results in subsequent analysis.

A second test of the study is the correlation matrix. In the paper, the authors generated a following matrix:

	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 2.9 Correlation Matrix in the Study

In the study, the datasets is of year 1992, and in our datasets, the data of year 1993 will has some differences. The matrix is shown below.

	spend	apret	top10	rejr	tstsc	pacc	strat	salar
spend	-							
apret	0.601231	-						
top10	0.675656	0.642464	-					
rejr	0.638544	0.514958	0.648163	-				
tstsc	0.71491	0.782183	0.798807	0.628601	-			
pacc	-0.28673	-0.302834	-0.207505	-0.0715207	-0.164223	-		
strat	-0.581755	-0.458311	-0.247857	-0.283617	-0.485226	0.131858	-	
salar	0.711838	0.635852	0.637648	0.606777	0.715472	-0.37524	-0.347673	-

Figure 2.10 Correlation Matrix of year 1993

We can see the highest three correlations with "apret" are tstsc, top10, and salar. Back to the matrix in study, the highest three correlations were still the tstsc, top10, and salar. However, the values of correlations have changed. The correlations are decreased from year 1992 to year 1993.

III. PC Algorithm

By using the temporal ordering of variables mentioned in the study to generate a model in PC algorithm, the casual graphs with p = 0.05 and p = 0.001 are shown below.

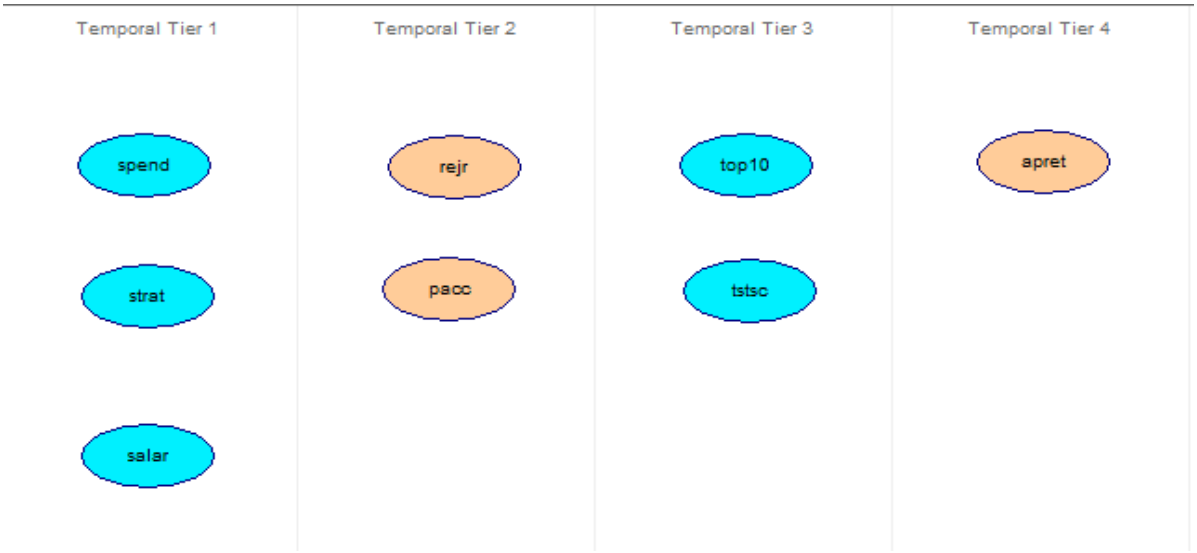


Figure 3.1 Temporal Ordering of Variables

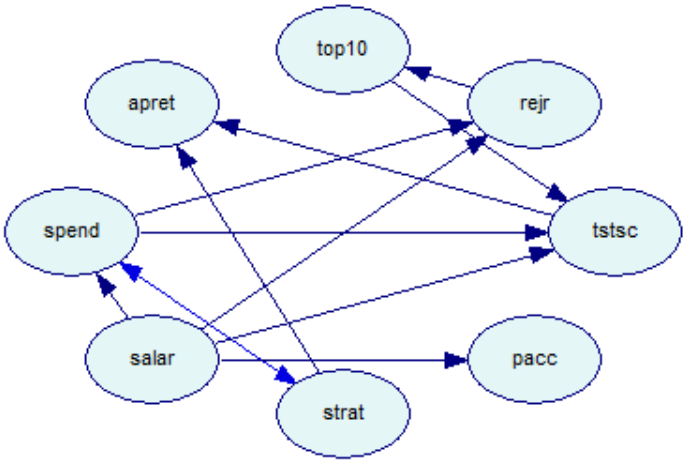


Figure 3.2 Casual Graph by PC Algorithm (p = 0.05)

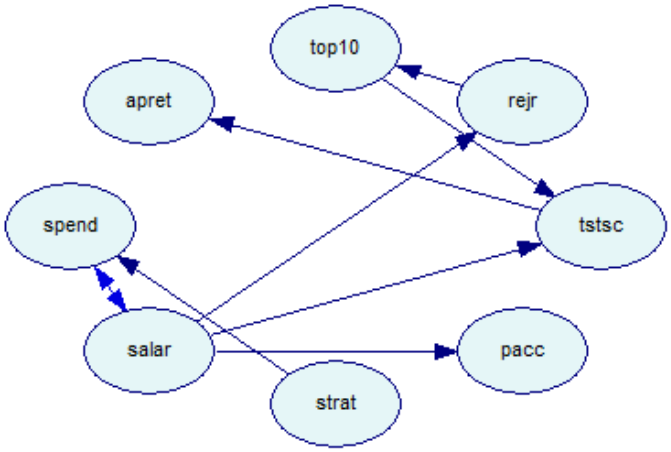
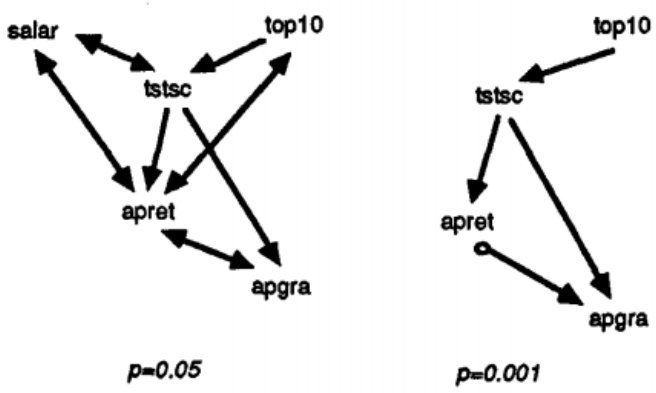


Figure 3.3 Casual Graph by PC Algorithm (p = 0.001)

From the two figures above we can see that there are some differences between the two casual graphs with the change of p. The relationship from “spend” to “rejr” and to “tstsc” disappeared, as well as the relationship from “strat” to “apret”. Look back to the graphs generated in the study.



With the comparison of the two pair of graphs for datasets year 1993 and year 1992, we can see when p = 0.001, the three variables' relationships illustrated including "top10", "tstsc" and "apret" are the same. "spend", "strat" and "salar" are also as same as the graph in the study, are connected with "apret" through "tstsc".

Then we will follow the authors in the study that do not set temporal ordering of variables. PC algorithm is still be used to generate two casual graphs with p = 0.05 and p = 0.001.

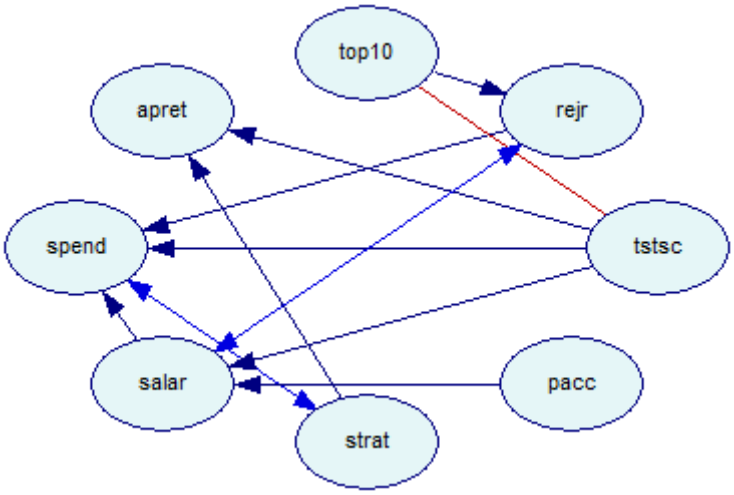


Figure 3.4 Casual Graph 2 by PC Algorithm (p = 0.05)

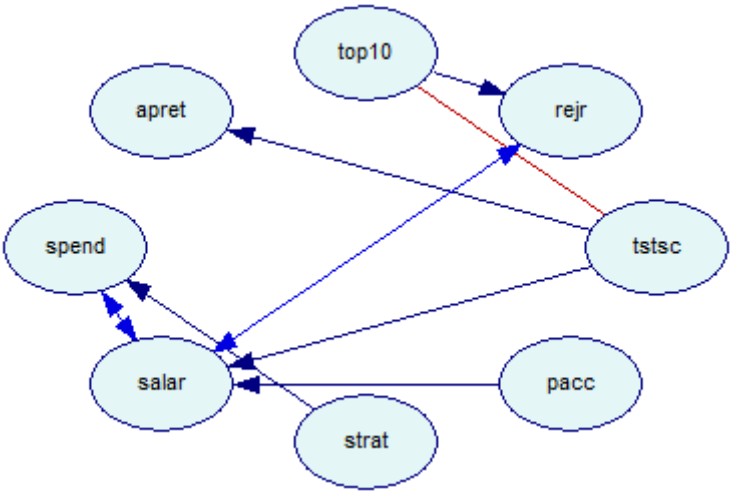


Figure 3.5 Casual Graph 2 by PC Algorithm (p = 0.001)

These two graphs support the prior knowledge supplied to PC algorithm that 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" were the same in both graphs.