**Assignment 5 Causal Discovery**

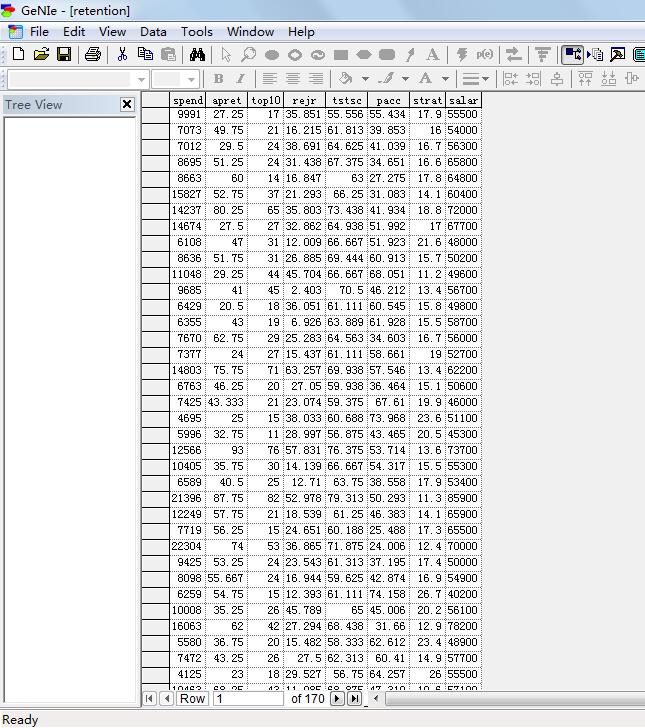
HAITAO ZHOU (HAZ59)

YINGZHI YANG(YIY50)

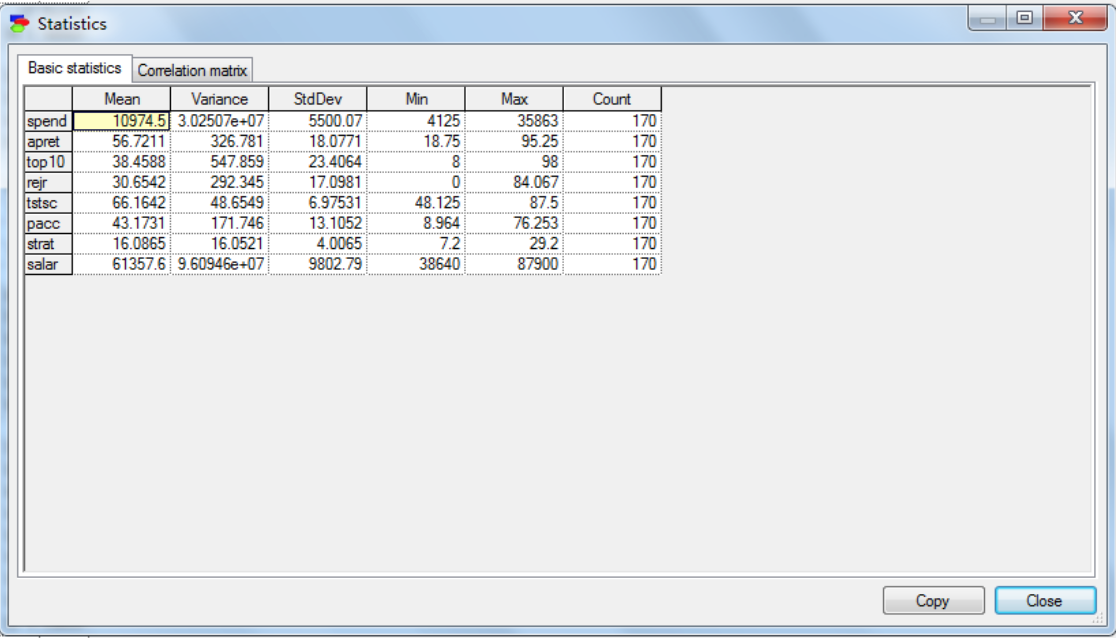
XIN JIN(XIJ21)

In this assignment, we use GeNIe to analyze the data.

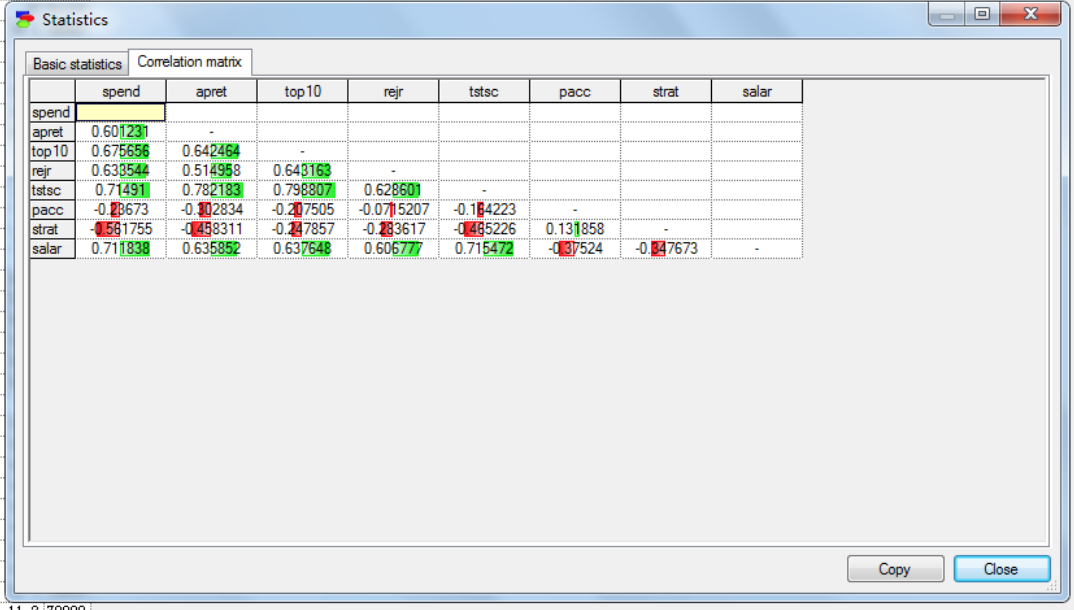
1. We import the data retention.txt into GeNIe:



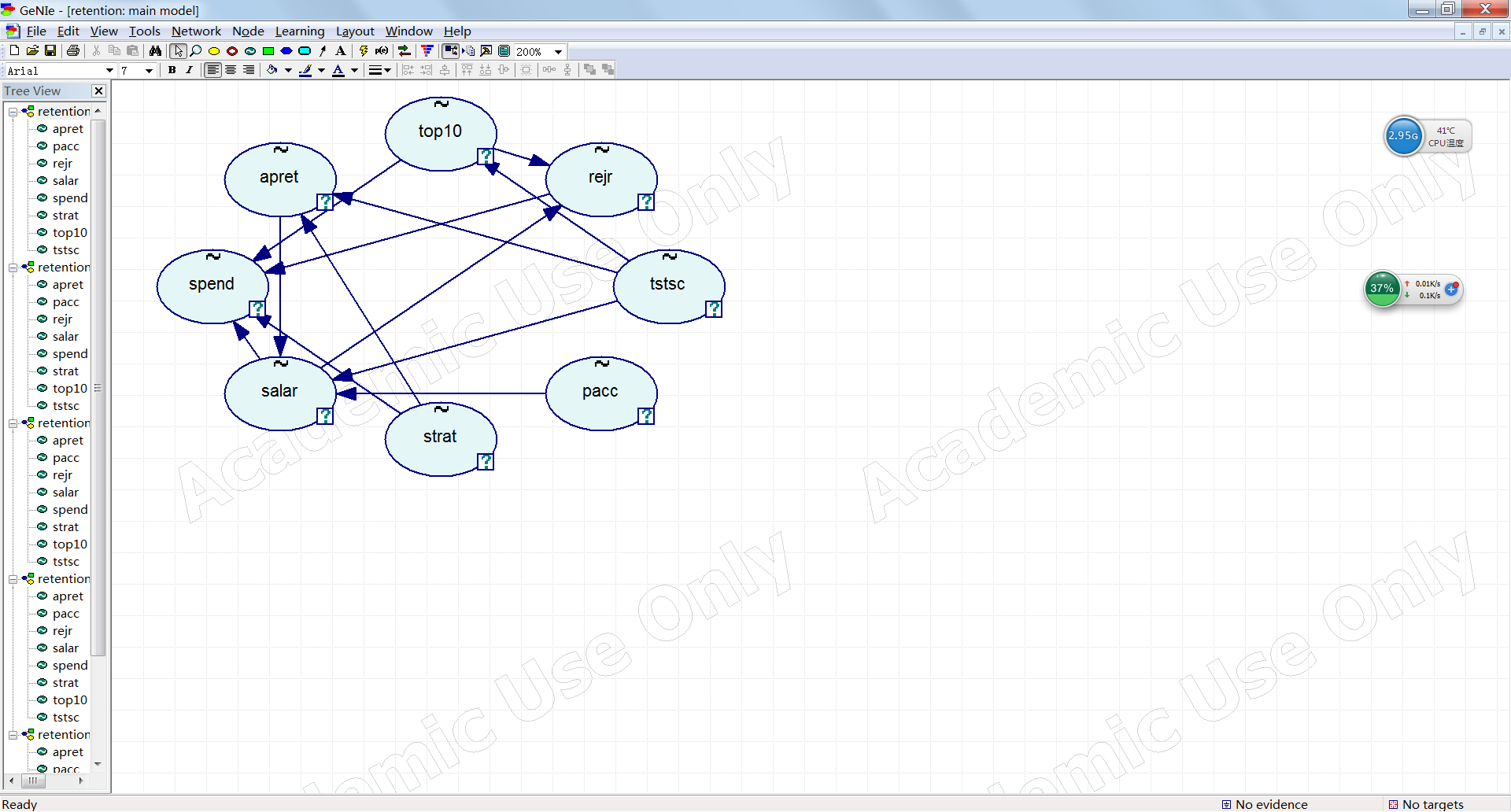
2. We can learn from some statistics of these natural data:

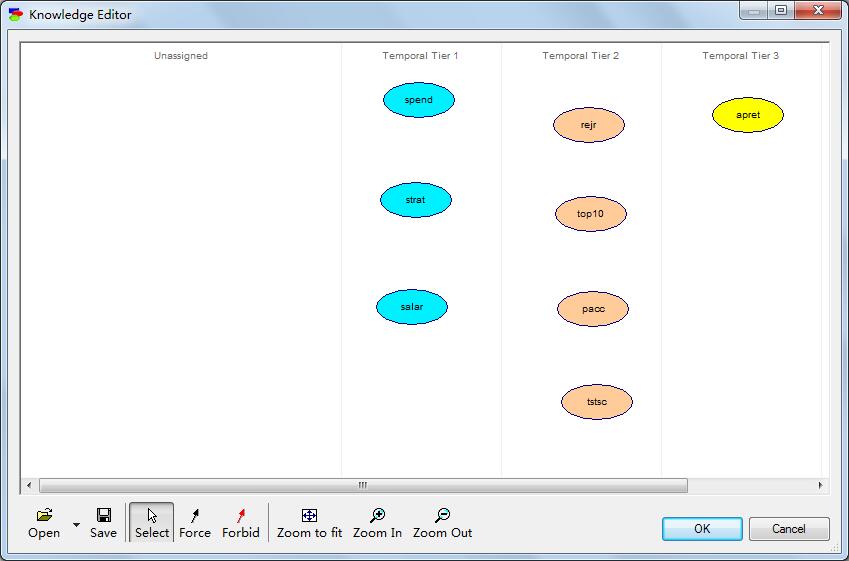


3.With the correlation matrix, we can see the correlation ratio between every two features. We can tell from this chart that apret may be positively correlated to top10, rejr, tstsc, salar and negatively correlated to pacc and strat.



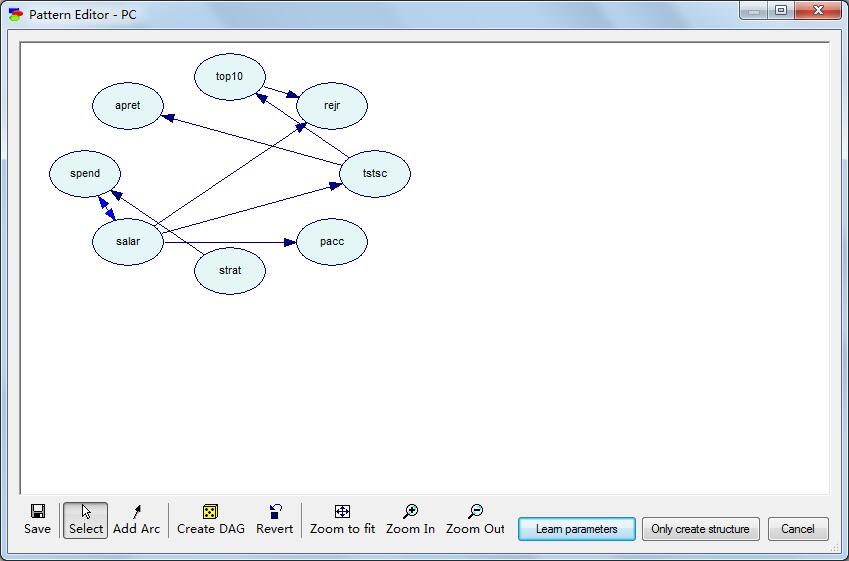
4. To have an overview of the data, we develop the pattern without setting any background or changing the significant level. In this case, apret=1.88181\*tstsc-0.543687\*strat+Normal(-59.0413,11.0966)

5. In order to force and forbid causal connections, we separate these 8 features into 3 temporal ties, which is “spend, start, salar” , “rejr, pacc, top10, tstsc” and “apret”.

6.Furthermore, we train the data by setting different significance level into pattern editor, and watching the connection’s change:

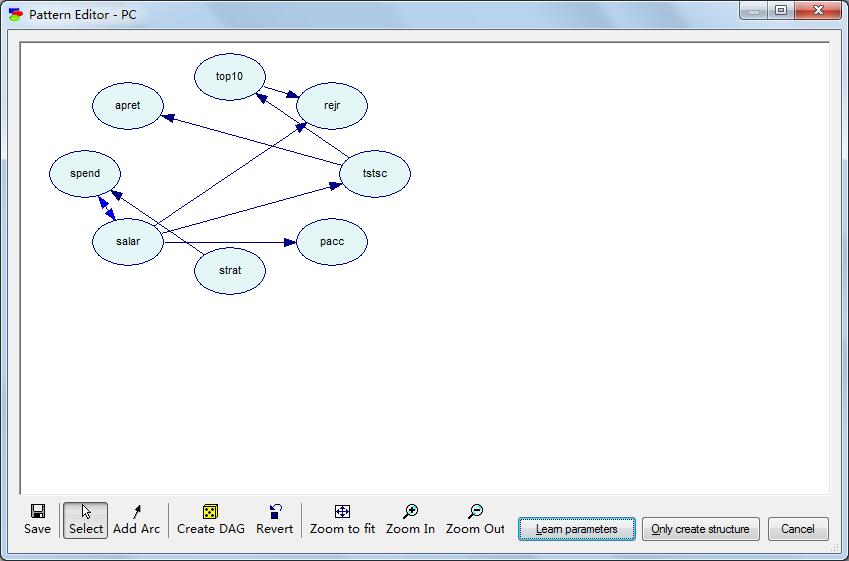
(1)significance level = 0.001:

apret=2.02709\*tstsc+Normal(-77.3999,11.2629)



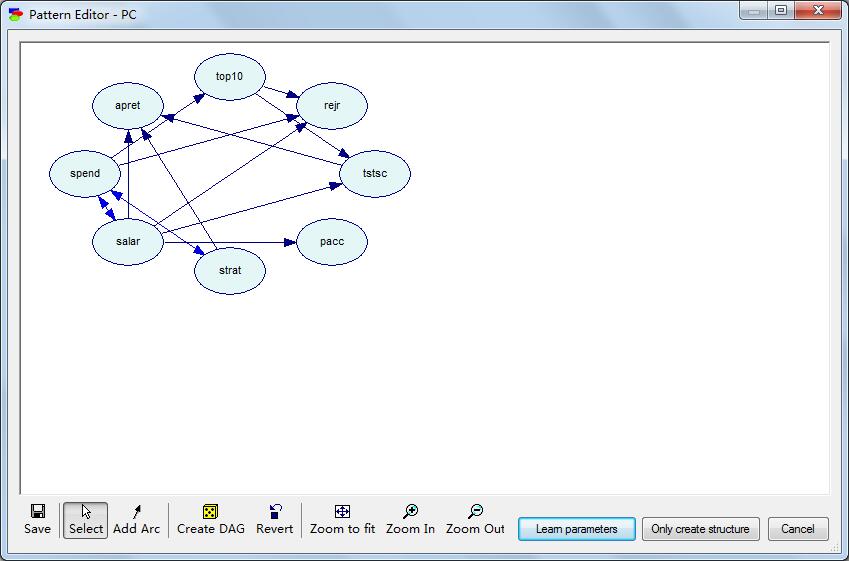
(2)significance level = 0.01:

apret=2.02709\*tstsc+Normal(-77.3999,11.2629)



(3)significance level = 0.1:

apret=1.60236\*tstsc-0.530669\*strat+0.000281388\*salar+Normal(-58.0262,10.9281)



Conclusion:

The diagrams showed that “tstsc” affects the “apret” mostly. In Druzdzel & Glymour’s conclusion, “top 10” and “tstsc” are the factors determine the value of “apret”. And in our research, when the significance level is low, the “top 10” factor will not affect “apret” directly, but it does affect the “tstsc”, and “tstsc” always affects “apret”. And when the significance level becomes as high as 0.1, “strat” becomes another important factor which will influence “apret” and “salar” has very little effect on “apret”. So, as far as we concern, the test score is the main reason for retention, and the quality of high school education will affect the freshman year test score for the college student. On the other hand, we can also see that the lower the significance level is, the more accurate the relationships are. To sum up, we believe that the data support the Druzdzel & Glymour’s conclusions in some way.