INFSCI 2725: Data Analytics

Assignment 5: Causal Discovery Assignment

Team member: Tao Li
Yumeng Lu
Zhaoxuan Ren

To test the conclusion, we use the background-knowledge provided by Druzdzel at first, where variables are divided into 4 group as following:

spend, strat, salar
rejr, pacc
tstsc, top10
apret

Variables in lower tier (like apret) are considered to be the dependent variable, whereas there other variables are independent variables. In GieNe, it's feasible to add forced arc or forbid arc between any variable, but that's not what we intend to do in the first step.

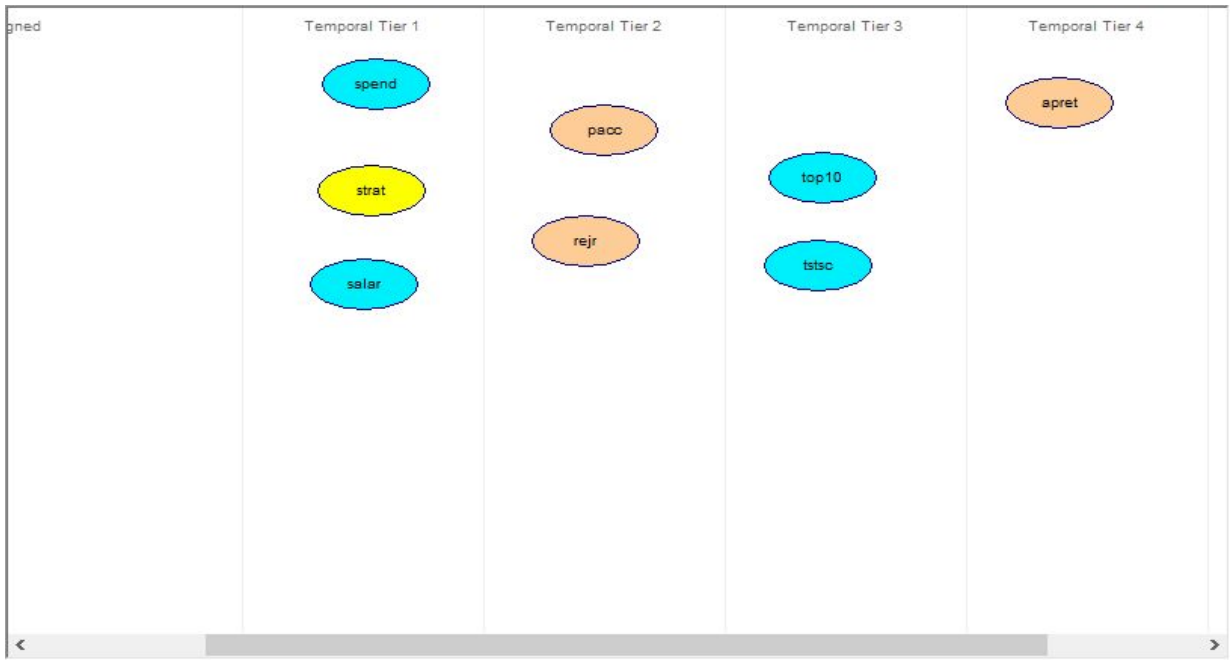

Figure 1: Variables grouped into 4 tiers

We use the structure without adding any forced or forbidden arc. Then we generate two causal graphs with different significance levels.

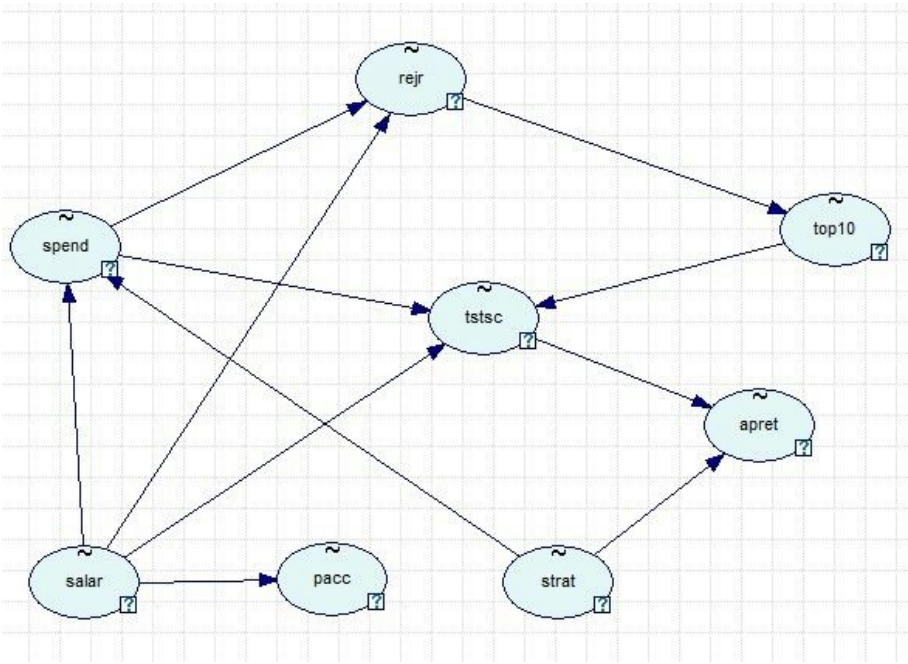

Figure 2: the causal graph when significance level = 0.05

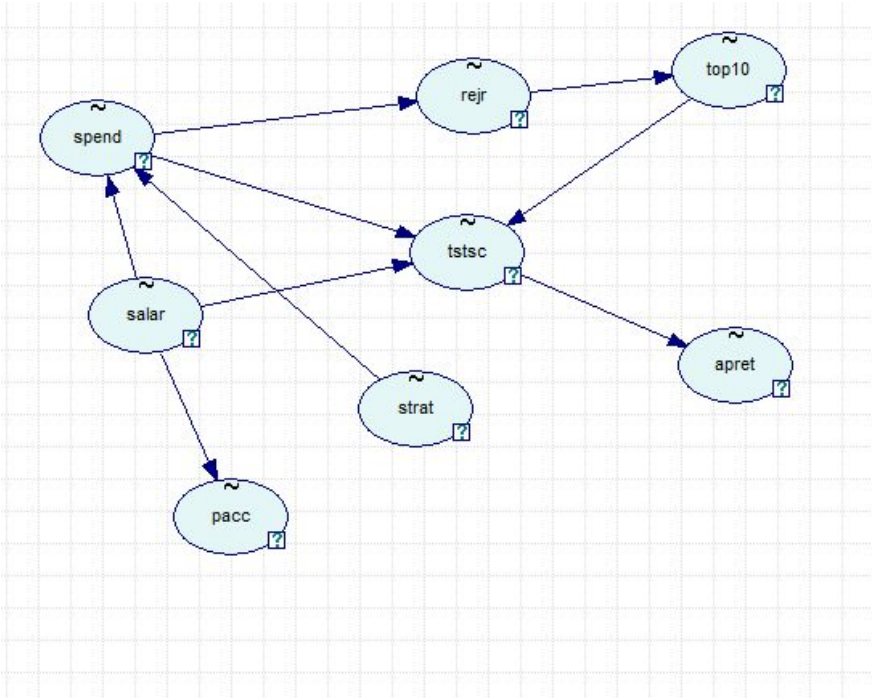

Figure 3: the causal graph when significance level = 0.01

Apparently tstsc (test score of incoming students) is the main factor among studied variables.

Then we are going to use our common sense knowledge to help GieNe improving the result. We add force arc from *rejr* to *top10* and *tstsc* because the stricter the admission is, the higher the academic ability is observed from incoming students.

However, adding common sense knowledge does not influence the result, therefore we still keep our original conclusion that *tstsc* is the main factor.

Additionally, we also use GieNe to test whether the input data are normally distributed.

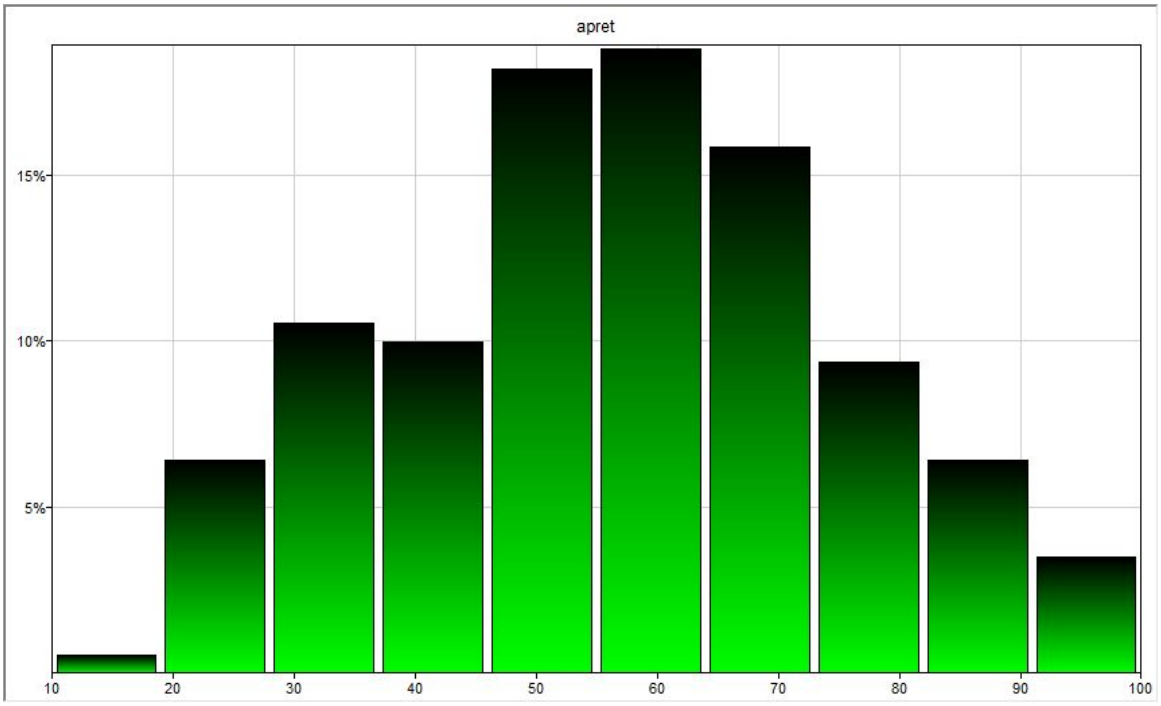

Figure 4: distribution of variable *apret*

From the histogram we can prove the assumption that input variable ***apret*** used in this study is distributed normally, so the conclusion is statistically valid.

Conclusion: We test the university retention data by PC algorithm with different background knowledge and significance levels. The result verifies that the findings in [Druzdzel & Glymour 1994].