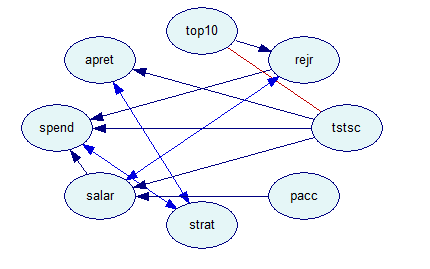
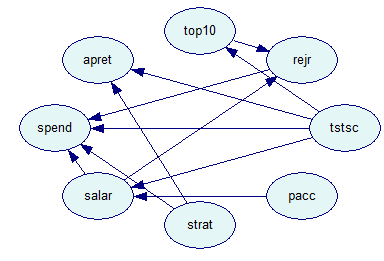
DA Assignment 5 Report

Jinrong Liu, Fengxi Liu

Use all of the things as GeNIe default (significance level = 0.05), we will get a pattern of PC algorithm like below:



1. The causal graph GeNIe suggests

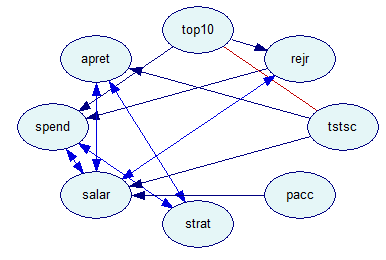


1. What causes student retention?

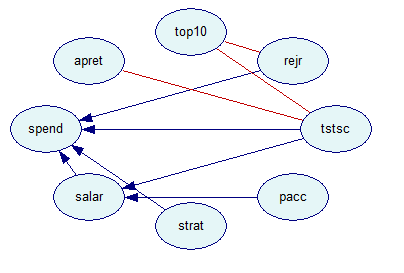
“apret” represents average retention rate. Based on the result of GeNIe’s causal graph, strat (student\_teacher ratio) and tstsc (average test scores of incoming freshmen)。

1. Check sensitivity of different significance level

Significance level 0.1



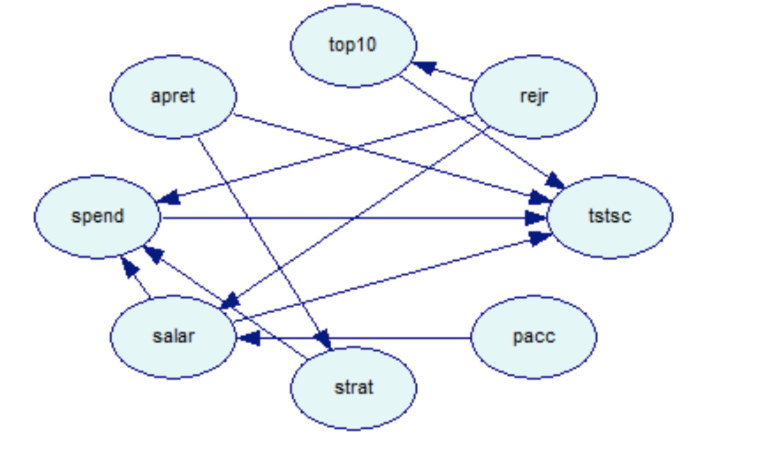
Significance level: 0.01



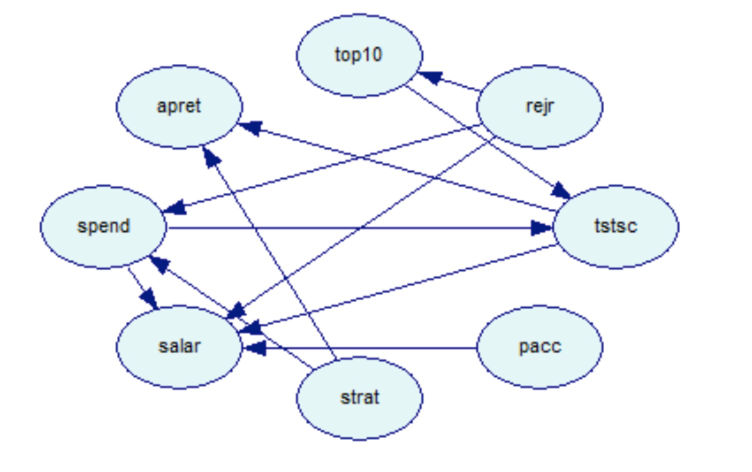
With the increase of significance level, the pattern will show more arcs and the current arcs will show more directions. In statistical hypothesis testing, statistical significance is attained when a p-value is less than the significance level. The p-value is the probability of obtaining at least as extreme results given that the null hypothesis is true whereas the significance level 𝛼 is the probability of rejecting the null hypothesis given that it is true. In this case, the higher significance level is, the less restrict to judge an assumption, so less arcs presented.

1. Background knowledge settings

Setting top10 and tstsc as temporal tier 1, and significance level = 0.05, we can get the pattern:



Setting top10, tstsc, apret, and salar as temporal tier 1, and significance level = 0.05, we can get the pattern:



From the graphs above, we can verify that none of the variables in the data set are directly causally related to freshmen retention except for test scores and class standing. This result, following directly from the fact that freshmen retention rate and graduation rate are, given average test scores and class standing, conditionally independent of all remaining variables, seems to be robust across varying significance levels, availability of prior knowledge, and data set size[[1]](#footnote-1).

1. Although the causal graph suggested by GeNIe give us some clue of

the current data. But we think there are some variables that can not be linked. For example: tstsc(average test scores of incoming freshmen) →salar(average faculty salary)

1. Marek J. Druzdzel and Clark Glymour. Causal inferences from databases:

   Why universities lose students. In Clark Glymour and Gregory F. Cooper

   (eds), Computation, Causation, and Discovery, Chapter 19, pages 521-539,

   AAAI Press, Menlo Park, CA, 1999 [↑](#footnote-ref-1)