Assignment 5

Groupmembers: Ying Ma, Weiyi Tuo, Kaiyang Lyu

1. Data Exploration

After import homework datasets, there are 8 columns and 170 rows in the data file, which means we have 8 variables and 170 instances. The variables are: spend, apret, top10, rejr, tstsc, pass, strat, salar. These variables have useful meaning for our further research, such as ordering variables in temporal tiers, and force and forbid causal connections. After checking original data description at <http://www.pitt.edu/~druzdzel/ftp/kdd94.pdf>, we can simply find the following meaning of these variables.

*apret = average percentage of freshmen retention*

*apgra = average percentage of graduation (it does not list on the datasets)*

*rejr = rejection rate*

*tstsc = average test scores of the incoming students*

*top10 = class standing of the incoming freshmen*

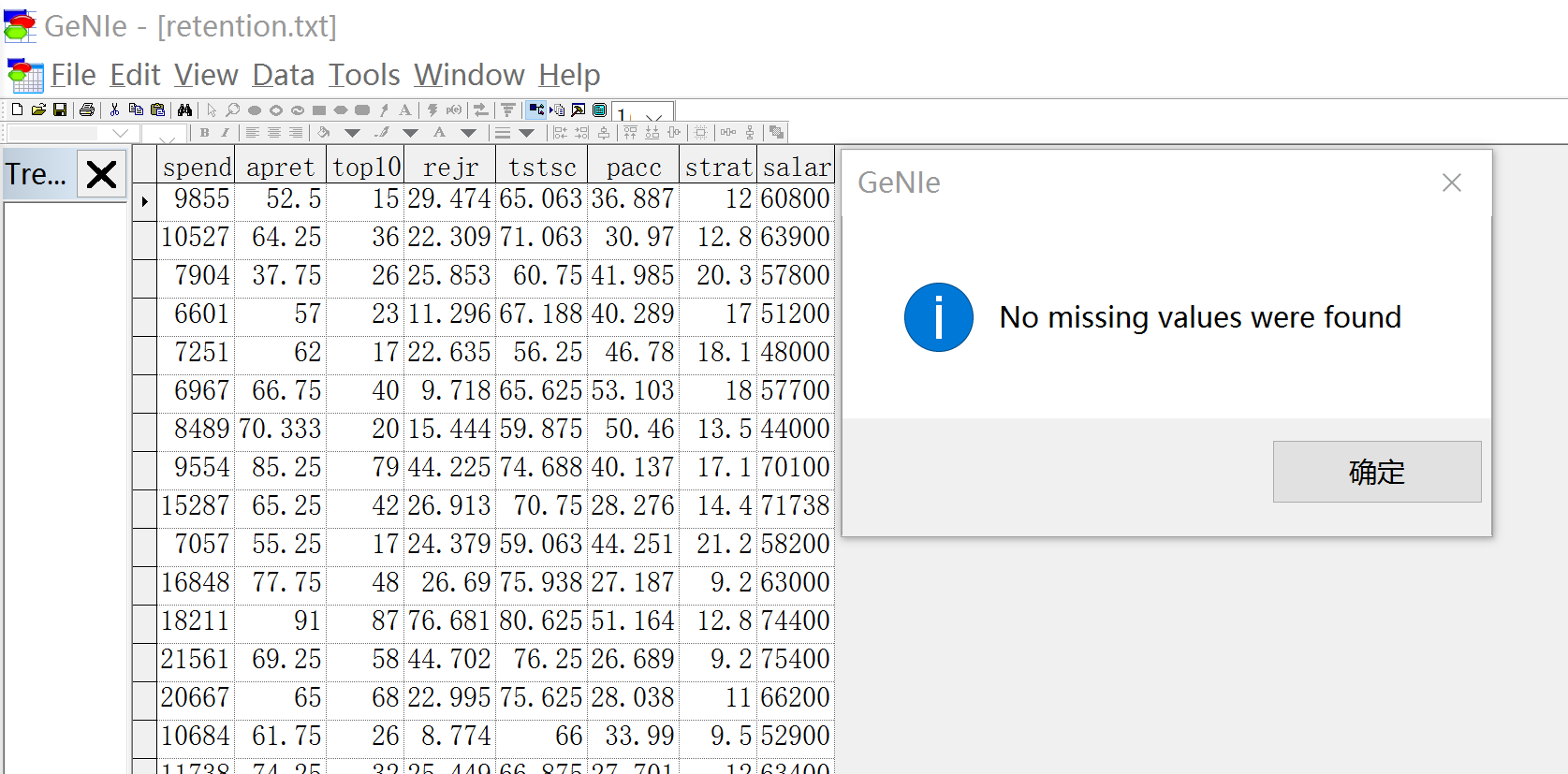
*pacc = percentage of admitted students*

*spend = total educational and general expenses per student*

*strat = student teacher ratio*

*salary = average faculty salary*

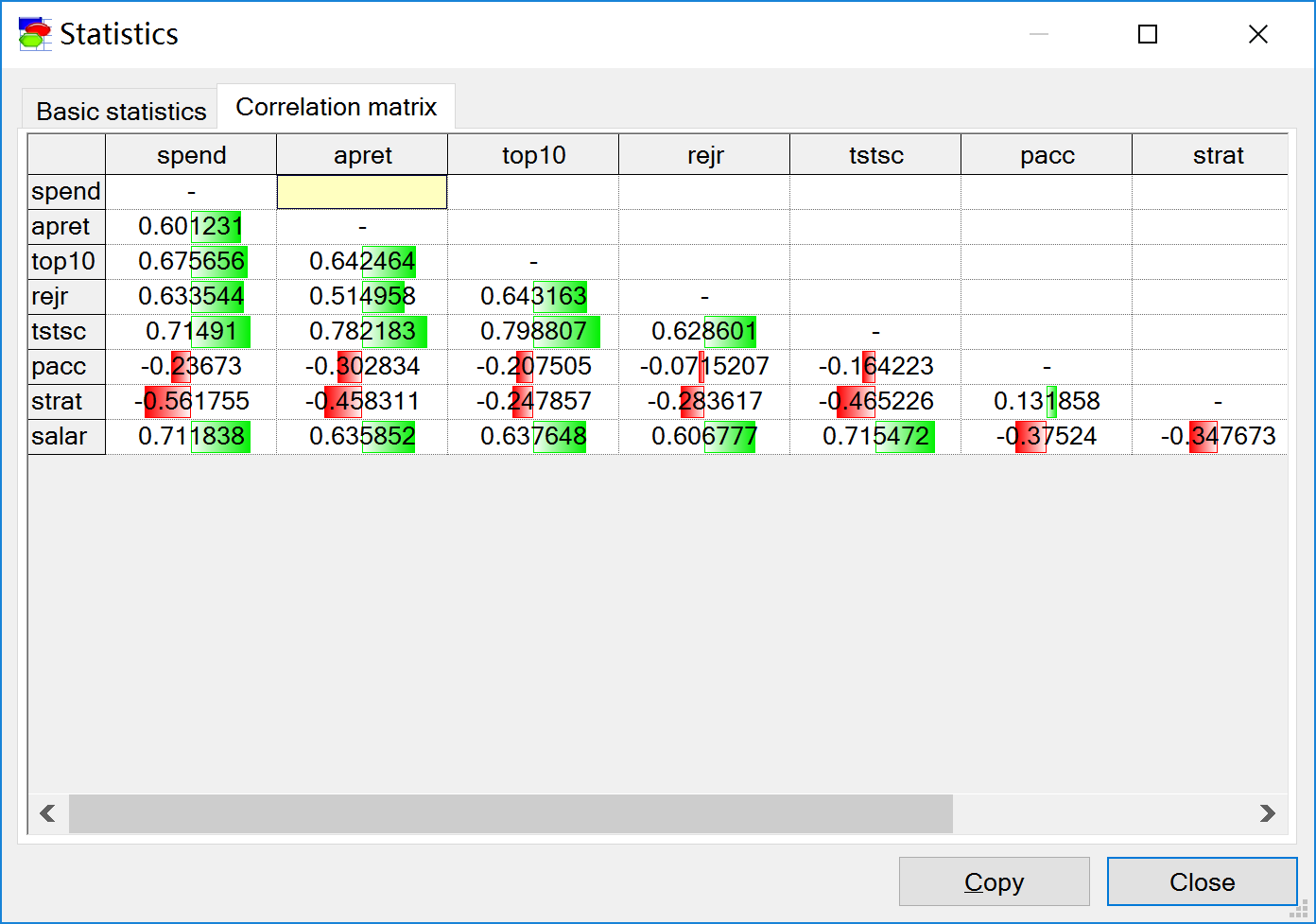
At the glimpse of data from GeNIe, we found there is no missing value in the datasets.



From statistic view of datasets, we also can find some useful information from itself.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Variance | StdDev | Min | Max | Count |
| spend | 10974.51176471 | 30250721.38743 | 5500.065580285 | 4125 | 35863 | 170 |
| apret | 56.72107647059 | 326.7814273847 | 18.07709676316 | 18.75 | 95.25 | 170 |
| top10 | 38.45882352941 | 547.8592412113 | 23.40639316963 | 8 | 98 | 170 |
| rejr | 30.65421764706 | 292.3451692955 | 17.09810426028 | 0 | 84.067 | 170 |
| tstsc | 66.16416470588 | 48.65489736916 | 6.975306256299 | 48.125 | 87.5 | 170 |
| pacc | 43.17310588235 | 171.7461420124 | 13.10519522985 | 8.964 | 76.253 | 170 |
| strat | 16.08647058824 | 16.05206439262 | 4.006502763336 | 7.2 | 29.2 | 170 |
| salar | 61357.64705882 | 96094622.31257 | 9802.786456542 | 38640 | 87900 | 170 |

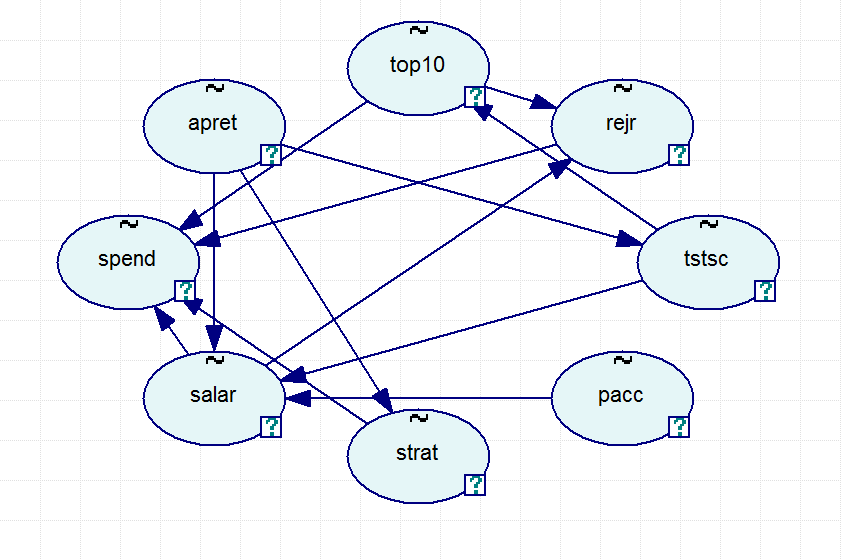
Then, we calculated correlation among these variables. The reason we made up correlation matrix is that we can force to create relationships between two variables which are actually weak relationships. And we can forbid some relationships where they are highly correlated.



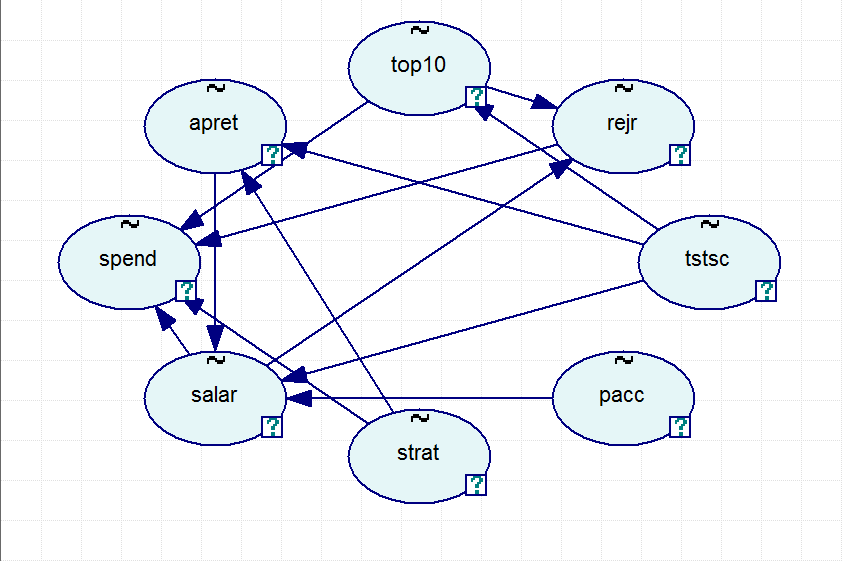
1. PC Algorithm Implementation from Different Aspect

In this part, we are using GeNIe to learn new networks with the same PC algorithm but different parameters. Then, we could get different patterns by learning parameters.

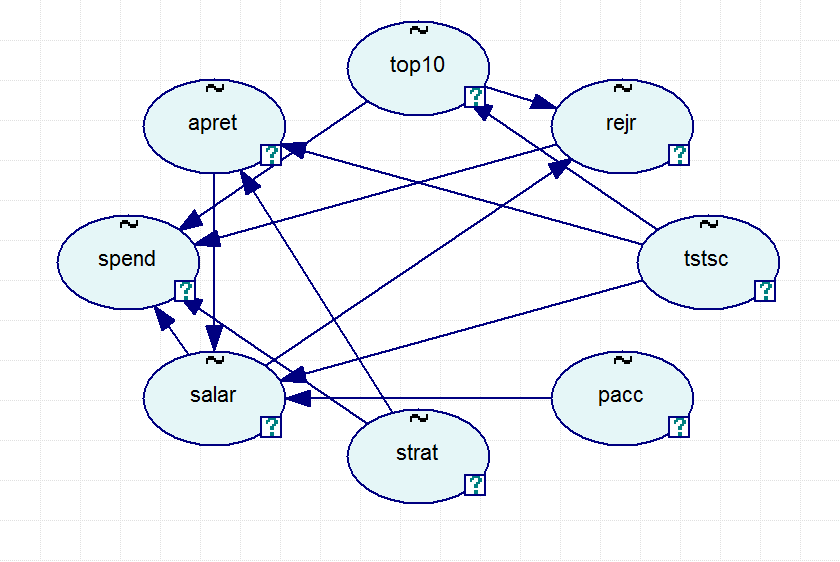
1. Different significant levels
2. Implement PC algorithm with significant level = 0.2



1. Implement PC algorithm with significant level = 0.15

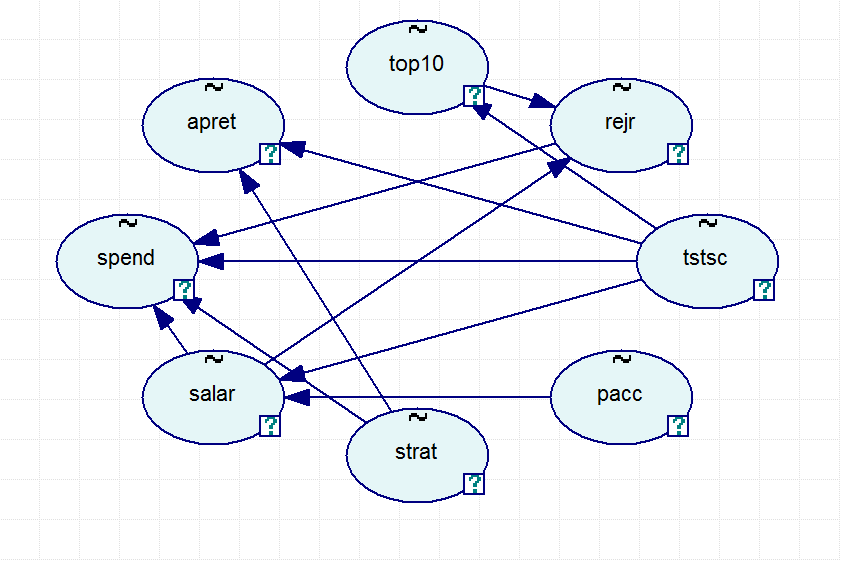


1. Implement PC algorithm with significant level = 0.10

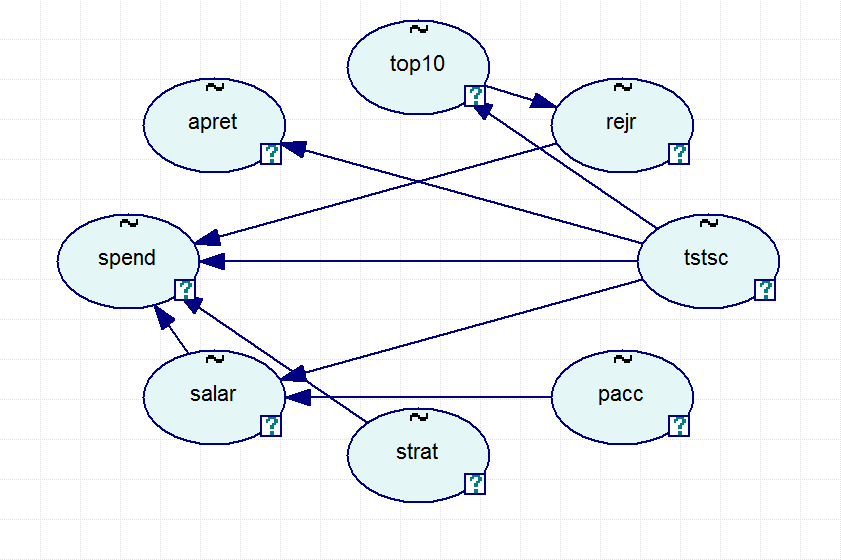


We can notice that there is no much difference between pattern with 0.15 significant level and with 0.10 significant level.

1. Implement PC algorithm with significant level = 0.05



1. Implement PC algorithm with significant level = 0.01



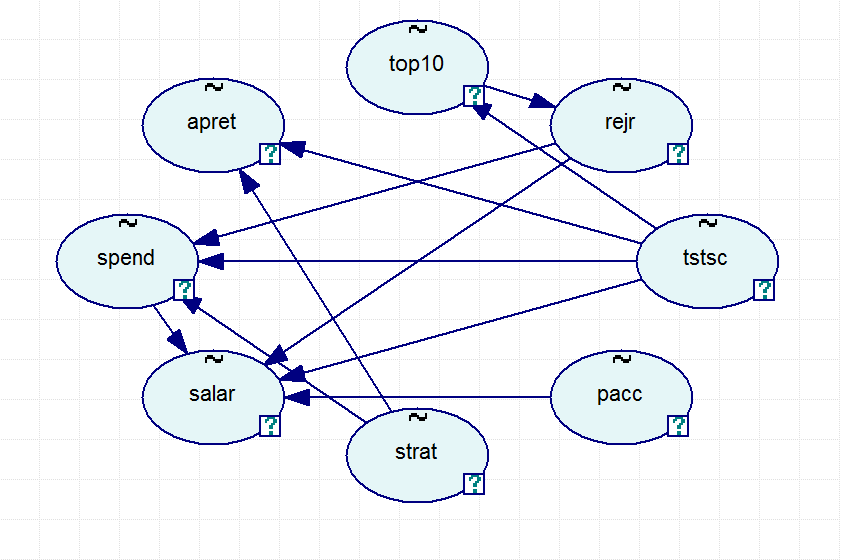
1. Different orders in temporal tiers

Here we interfere to make up different pattern with our background knowledge. (We consider significant level is set with default value 0.05)

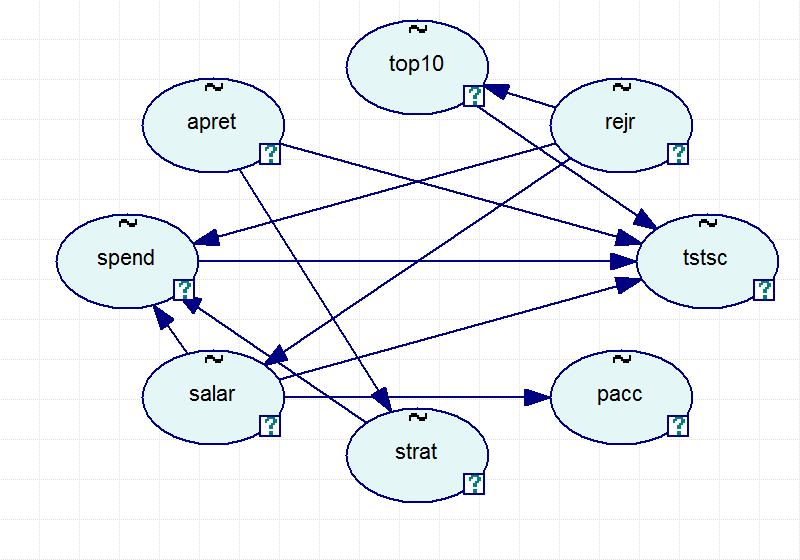
1. Suppose we have spend as course and salary as effect

(spend→salary)

**students spend more money on education, so faculties could earn more money from their work.**

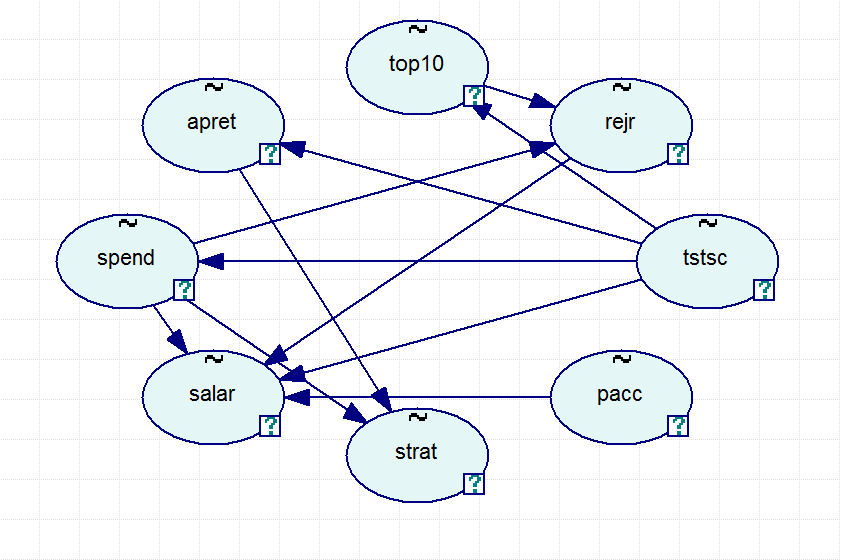


1. pacc→top10→tstsc



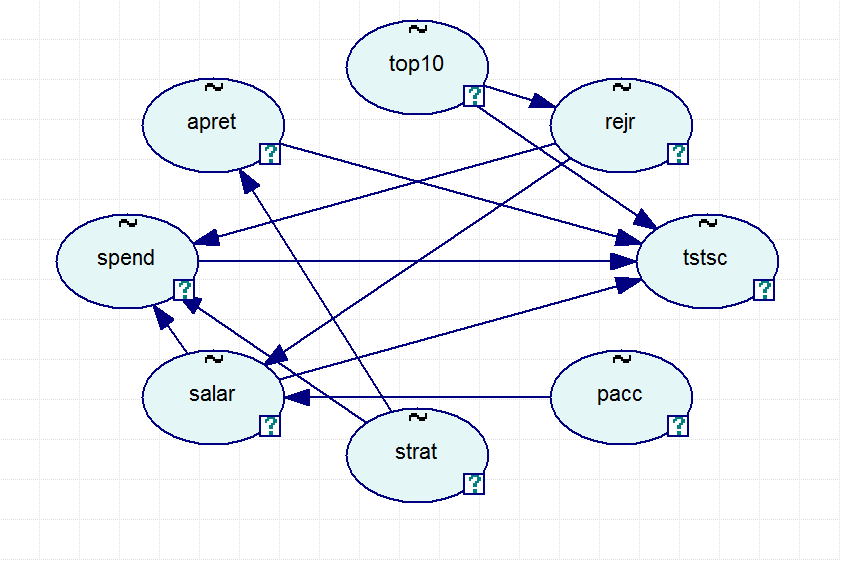
**As we can see, there is no arrow from pacc to top10 added but the arrow from tstsc to top has been changed. This result shows that we can only change the order based on existed causality. (Differ from forcing to create new connection)**

1. apret→rejr→salary→strat

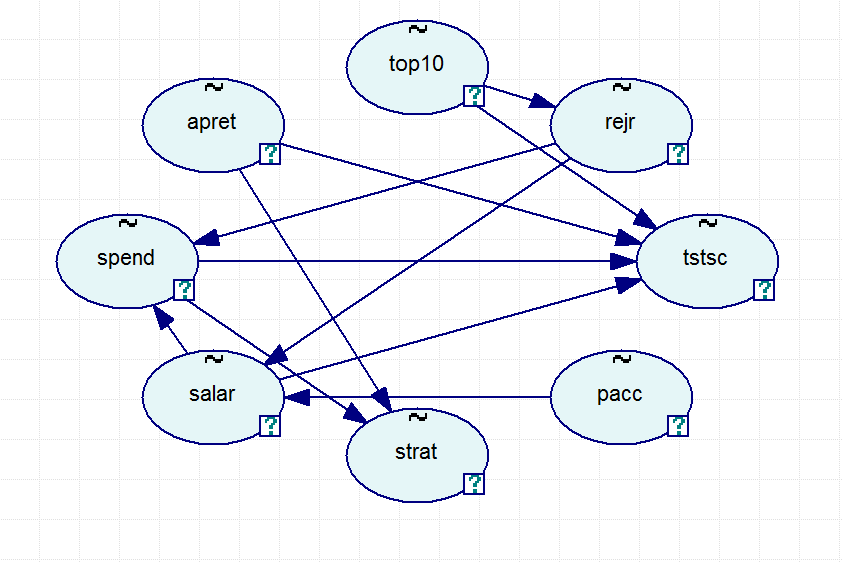


**combine the average percentage of freshman retention with rejection rate are equal to 100% which represent the whole group of fresh man. And apret and rejr have direct influence on school’s revenue, also impact on professor’s salary. Also we assume that salary is an important factor influence teacher stay or not**

1. apret→spend→tstsc



1. strat→tstsc

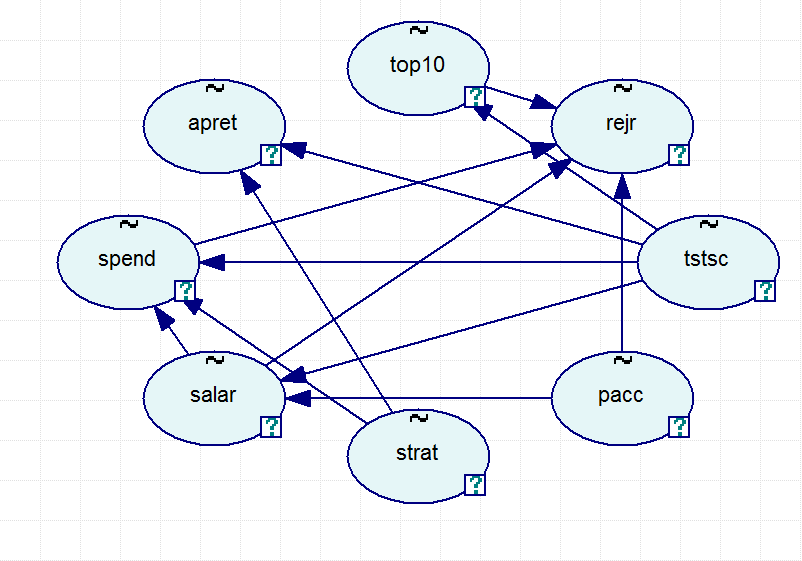


**lower retention represent that school has relative strict selective mechanism, and the stay students maintain higher academic quality. This is the main factor in determining admission scores**

1. Different forced connections

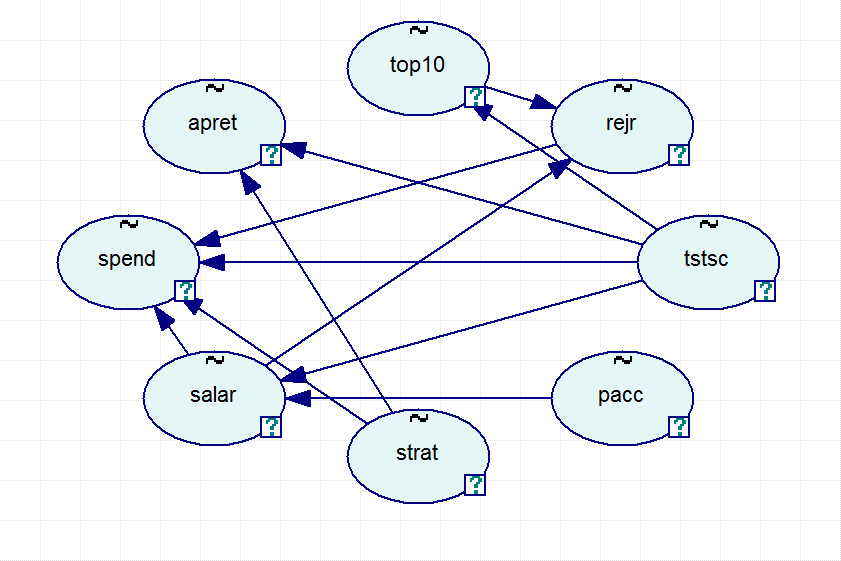
We changed those relationships based on weak or negative relation in correlation matrix. Before changing connection, we remained other parameters unchanged. (Default significant value and default order)

1. pacc→rejr



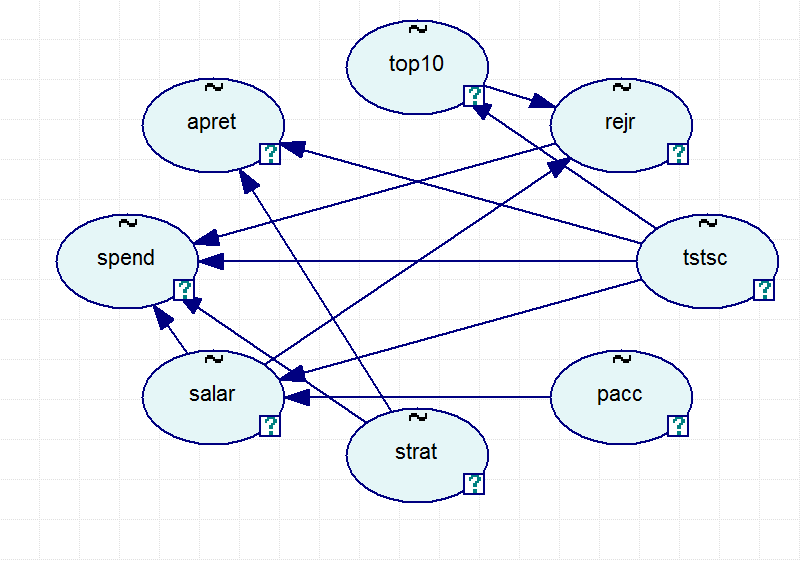
**regard the all application as a whole group, there are only two options for applicant, get admission or be rejected.**

1. strat→spend



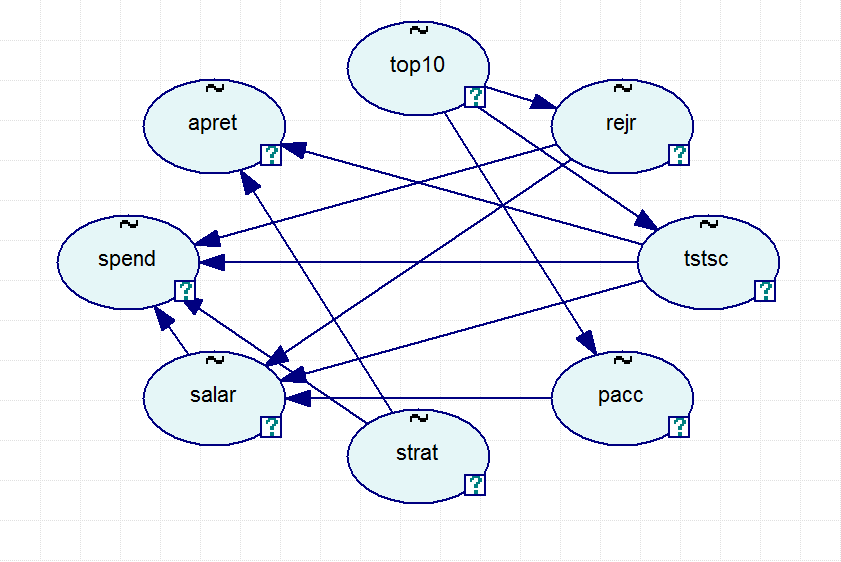
**in certain degree, lower student-teacher ratio means every single student could receive better education and more attention from teacher. In the meantime, this kind of benefit means students have to spend more on their education.**

1. pacc→salary



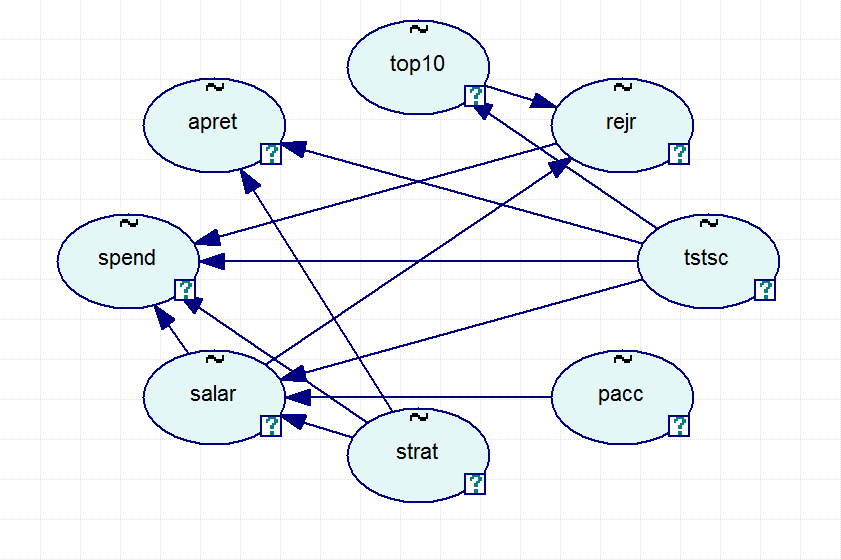
**it’s quite simple, in the short time and the provided conditions remain unchanged, the more students get admission the higher revenue of school will achieve. This means teachers will receive more salary.**

1. salary→pacc→top10



**to adjust teacher’s salary, school may want to offer more admissions to students, this means admitted students may have relatively lower academic ability, so they have poor performance in entrance exam.**

1. salary→strat→apret

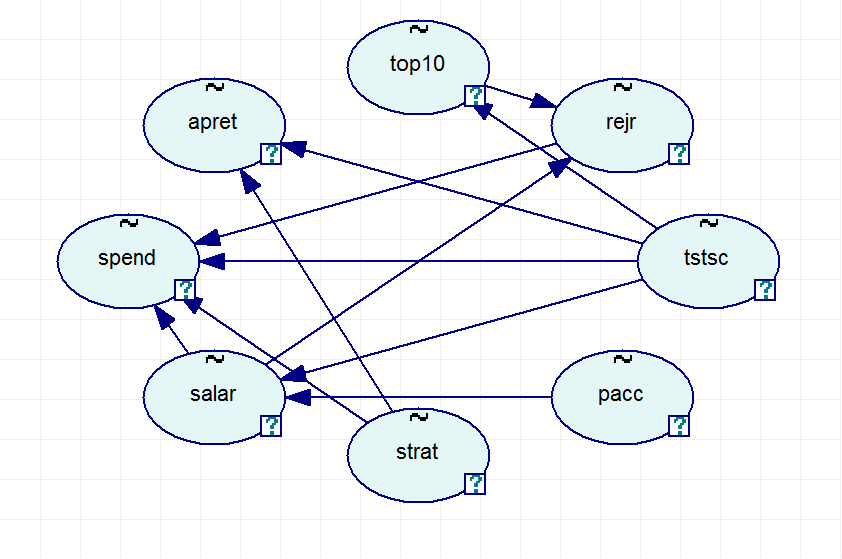


**in the certain degree, salary is the most important factor in determining stay or leave for almost teachers. This will influence the student-teacher ratio in the short time, and higher S-T ratio represent poor teaching quality, this will lead fresh student leave.**

1. Different forbidden connections

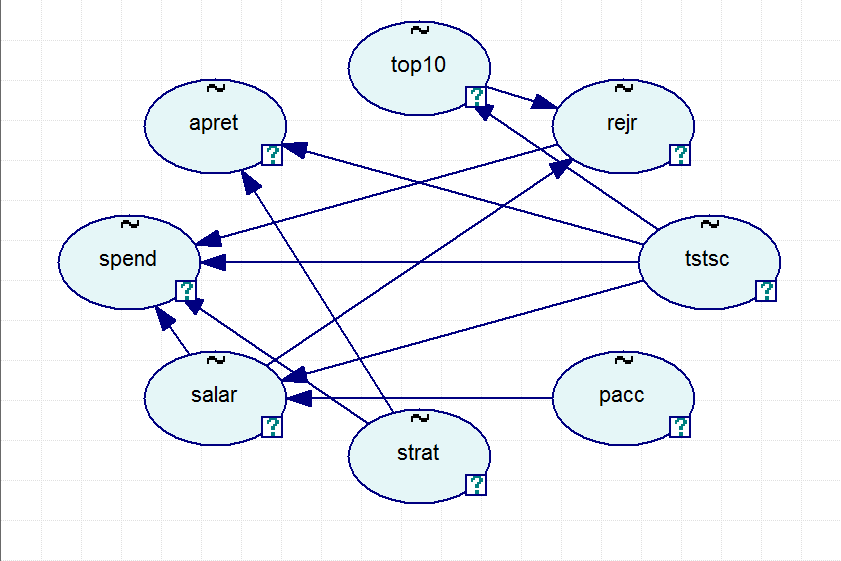
We forbade those relationships based on strong or positive relation in correlation matrix. Before changing connection, we remained other parameters unchanged. (Default significant value and default order)

1. rejr→salary



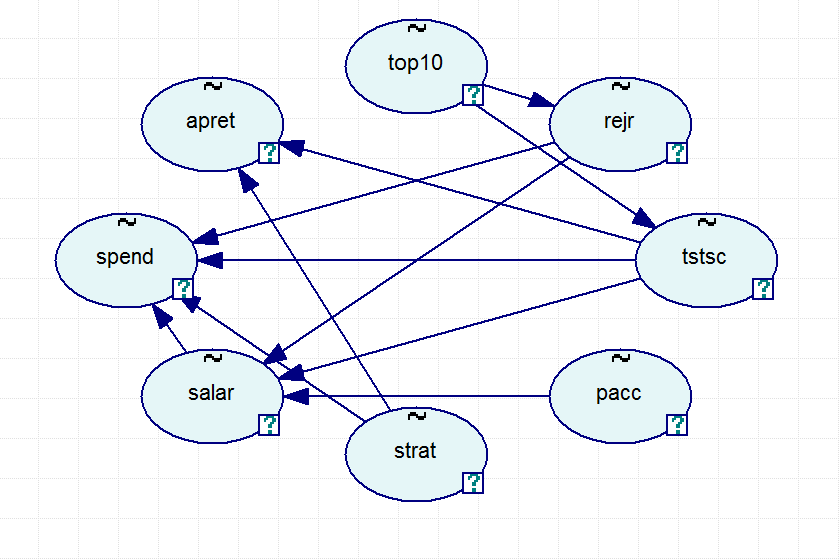
**in the short time, under certain conditions unchanged, higher rejection rate will cause school earn low avenue which directly decrease teachers’ salary.**

1. apret→rejr



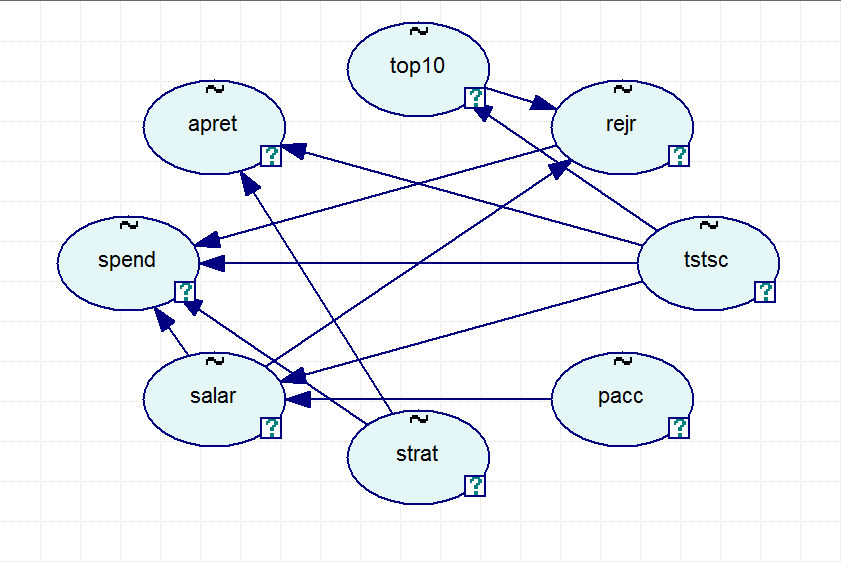
**the rejection rate is not only consists of the reject number in application part but also maintain in the entrance exam part or the beginning period of new term. It’s possible for school withdraw the admission for certain reasons.**

1. tstsc→top10



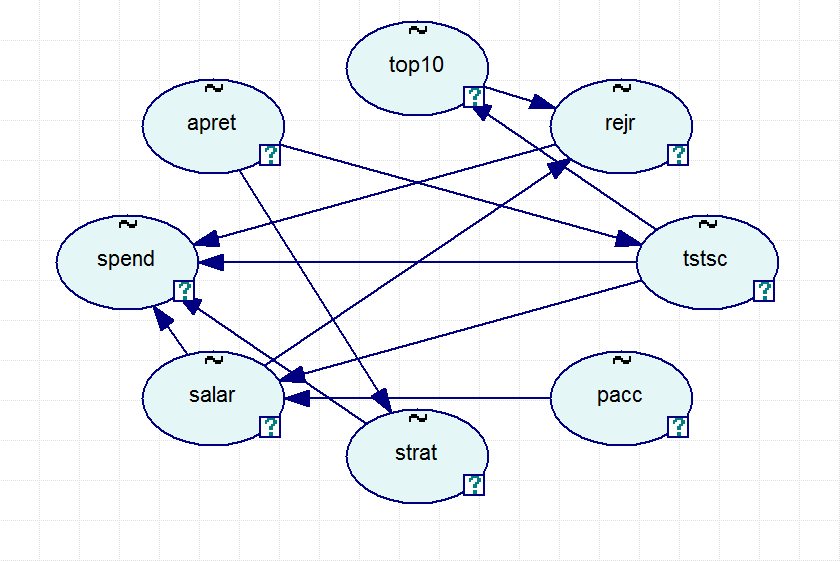
**if school want to evaluate fresh student, the average test scores is the most important reference. This could directly influence class standing of incoming students.**

1. salary→strat



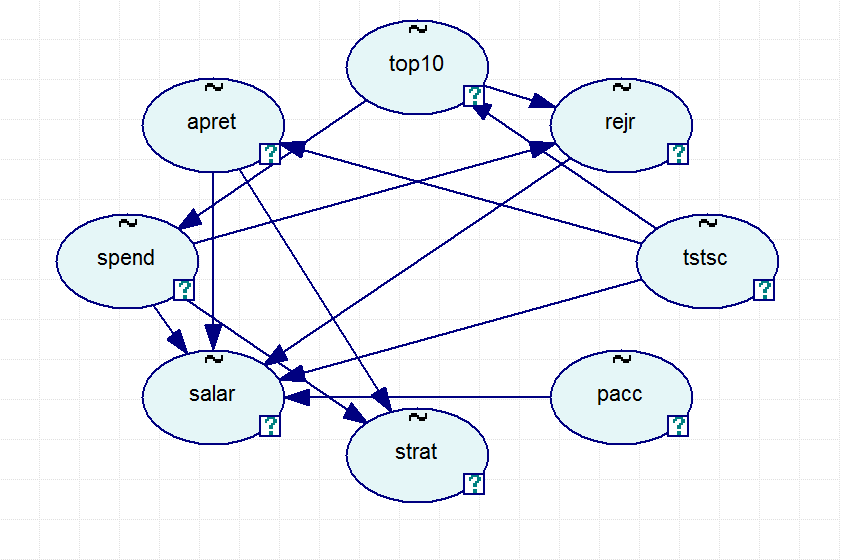
**as we mentioned before, in the short time, salary could be the only factor in determining stay or leave for teacher. So this could direct rewrite the ratio of student and teacher.**

1. tstsc→apret

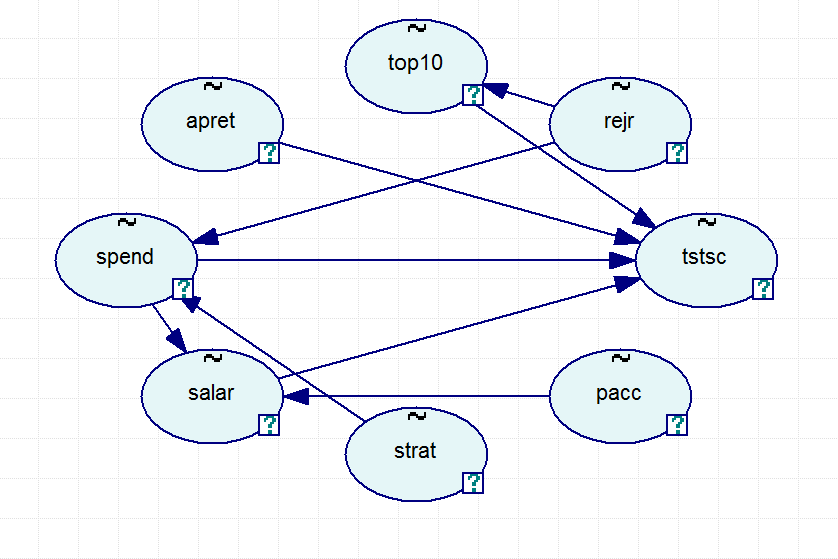


**could we think that if fresh students have poor performance in the entrance exam, school may school will withdraw the admission.**

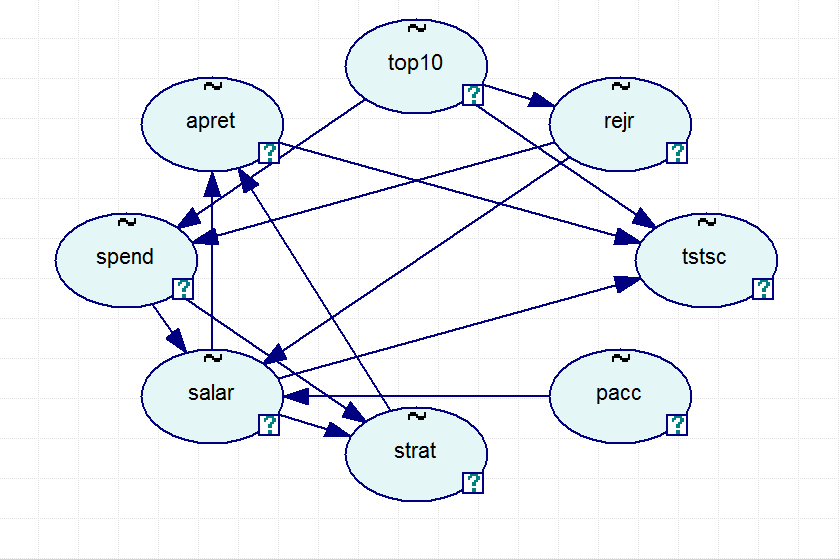
1. Different combinations of all aspects
2. Significant level: 0.15, order: apret→rejr→salary→strat, forced link: pacc→salary, forbidden link: salary→strat



1. Significant level: 0.01, order: pacc→top10→tstsc, forced link: pacc→salary, forbidden link: rejr→salary



1. Significant level: 0.2, order: strat→tstsc, forced link: salary→strat→apret, forbidden link: tstsc→apret



1. Conclusion

Overall, starting with data explorationwe, we found there is no missing data, which means we can perfectly implement our PC algorithm. Then we obtained some useful statistic information through datasets, especially correlation matrix. We can manipulate relationships among variables based on correlation matrix. We ran PC algorithm with different aspects at 23 times in total. We found that significant level has impact on pattern generated from PC algorithm. In our research, we tested different levels from 0.2 to 0.01. The lower the level is, the lower probability of incorrectly rejecting the null hypothesis would be acceptable according to the type I error. Then we configured different orders in temporal tier and different relations among variables. We found that the prior tier could not be the effect of later tier. In addition, forced relationships can be created even without existed relationships in the original pattern and forbidden relationships could be changed either by changing the direction of the arrow or by eliminating the original arrow. Last but not least, we tested different combination of all parameters. We found some interesting pattern during the research. All of our work is done with GeNIe which is useful statistical and graphical software.