Assignment 5 Causal Discovery

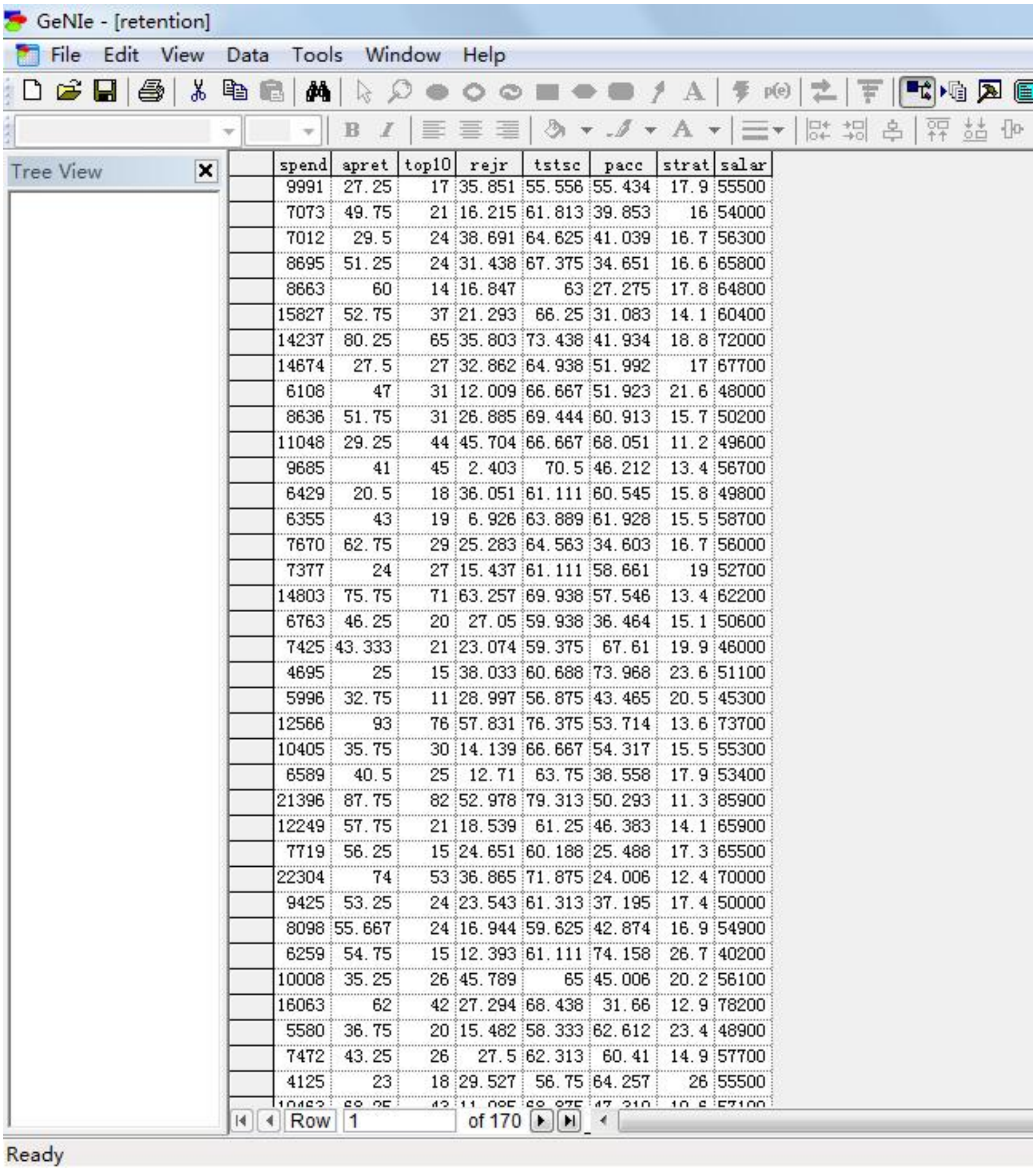
HAITAO ZHOU (HAZ59)

YINGZHI YANG(YIY50)

XIN JIN(XIJ21)

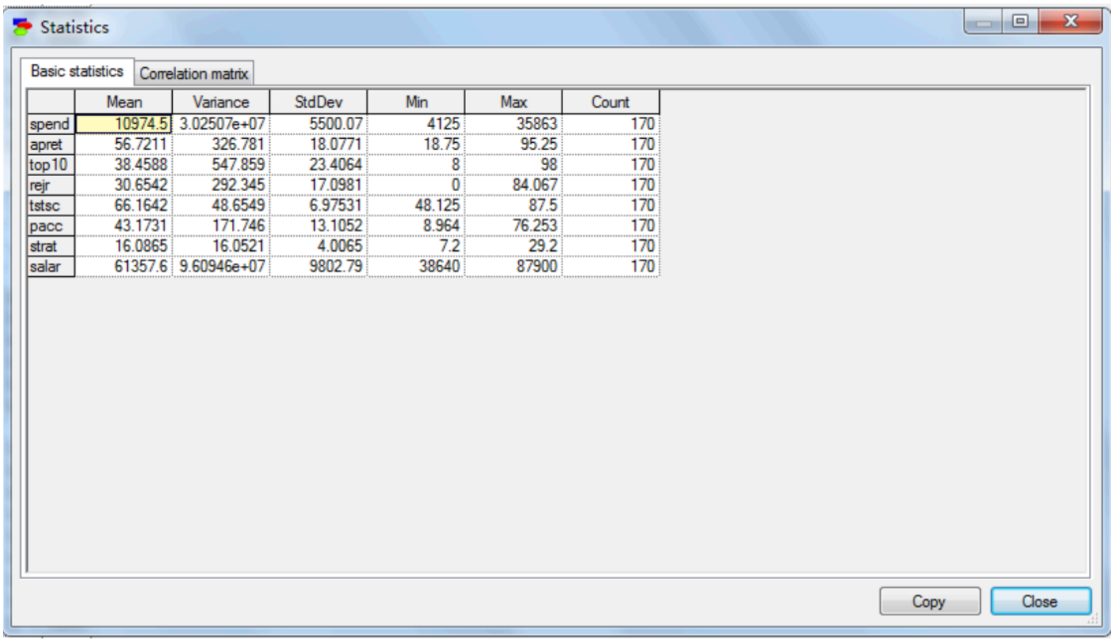
In this assignment, we use GeNIe to analyze the data.

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



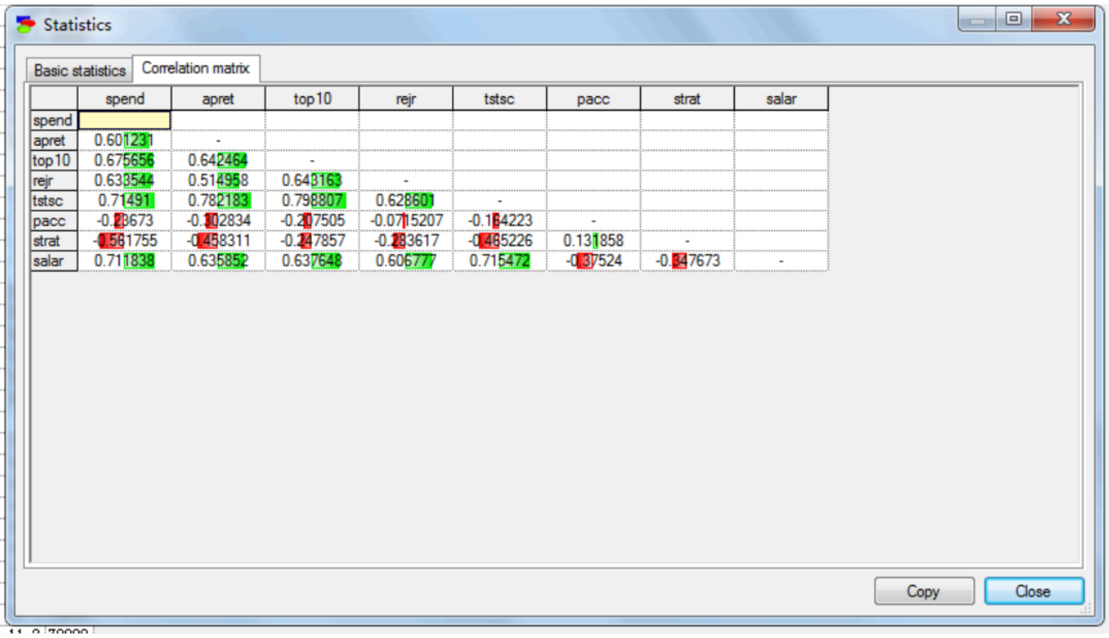
spend	apret	top10	rejr	tstsc	pacc	strat	salar
9991	27.25	17	35.851	55.556	55.434	17.9	55500
7073	49.75	21	16.215	61.813	39.853	16	54000
7012	29.5	24	38.691	64.625	41.039	16.7	56300
8695	51.25	24	31.438	67.375	34.651	16.6	65800
8663	60	14	16.847	63	27.275	17.8	64800
15827	52.75	37	21.293	66.25	31.083	14.1	60400
14237	80.25	65	35.803	73.438	41.934	18.8	72000
14674	27.5	27	32.862	64.938	51.992	17	67700
6108	47	31	12.009	66.667	51.923	21.6	48000
8636	51.75	31	26.885	69.444	60.913	15.7	50200
11048	29.25	44	45.704	66.667	68.051	11.2	49600
9685	41	45	2.403	70.5	46.212	13.4	56700
6429	20.5	18	36.051	61.111	60.545	15.8	49800
6355	43	19	6.926	63.889	61.928	15.5	58700
7670	62.75	29	25.283	64.563	34.603	16.7	56000
7377	24	27	15.437	61.111	58.661	19	52700
14803	75.75	71	63.257	69.938	57.546	13.4	62200
6763	46.25	20	27.05	59.938	36.464	15.1	50600
7425	43.333	21	23.074	59.375	67.61	19.9	46000
4695	25	15	38.033	60.688	73.968	23.6	51100
5996	32.75	11	28.997	56.875	43.465	20.5	45300
12566	93	76	57.831	76.375	53.714	13.6	73700
10405	35.75	30	14.139	66.667	54.317	15.5	55300
6589	40.5	25	12.71	63.75	38.558	17.9	53400
21396	87.75	82	52.978	79.313	50.293	11.3	85900
12249	57.75	21	18.539	61.25	46.383	14.1	65900
7719	56.25	15	24.651	60.188	25.488	17.3	65500
22304	74	53	36.865	71.875	24.006	12.4	70000
9425	53.25	24	23.543	61.313	37.195	17.4	50000
8098	55.667	24	16.944	59.625	42.874	16.9	54900
6259	54.75	15	12.393	61.111	74.158	26.7	40200
10008	35.25	26	45.789	65	45.006	20.2	56100
16063	62	42	27.294	68.438	31.66	12.9	78200
5580	36.75	20	15.482	58.333	62.612	23.4	48900
7472	43.25	26	27.5	62.313	60.41	14.9	57700
4125	23	18	29.527	56.75	64.257	26	55500

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



	Mean	Variance	StdDev	Min	Max	Count
spend	10974.5	3.02507e+07	5500.07	4125	35863	170
apret	56.7211	326.781	18.0771	18.75	95.25	170
top10	38.4588	547.859	23.4064	8	98	170
rejr	30.6542	292.345	17.0981	0	84.067	170
tstsc	66.1642	48.6549	6.97531	48.125	87.5	170
pacc	43.1731	171.746	13.1052	8.964	76.253	170
strat	16.0865	16.0521	4.0065	7.2	29.2	170
salar	61357.6	9.60946e+07	9802.79	38640	87900	170

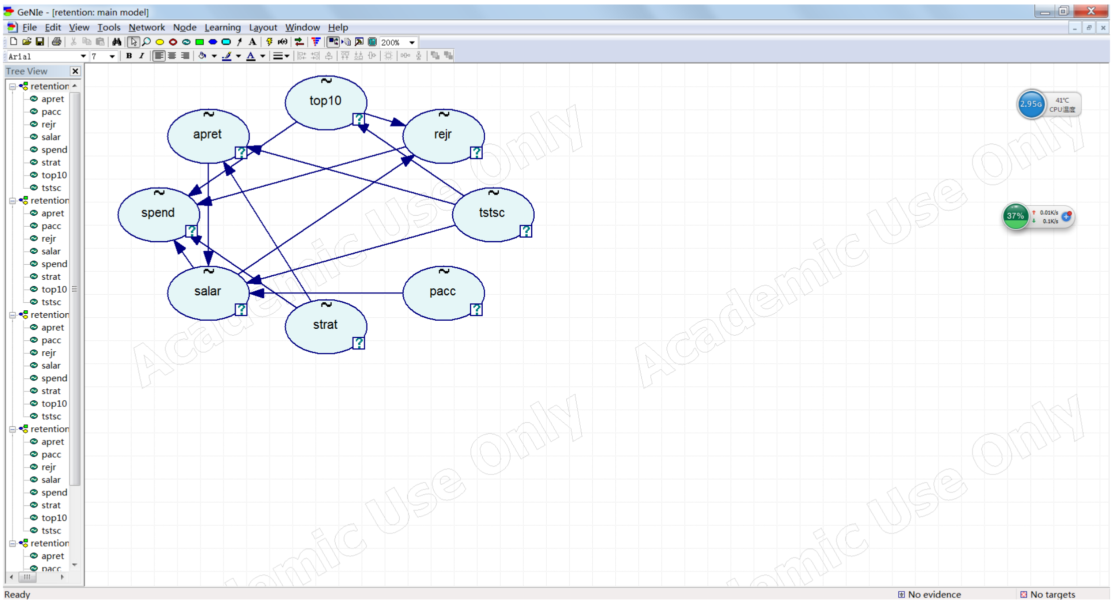
3.With the correlation matrix, we can see the correlation ratio between every two features. We can tell from this chart that apret may be positively correlated to top10, rejr, tstsc, salar and negatively correlated to pacc and strat.



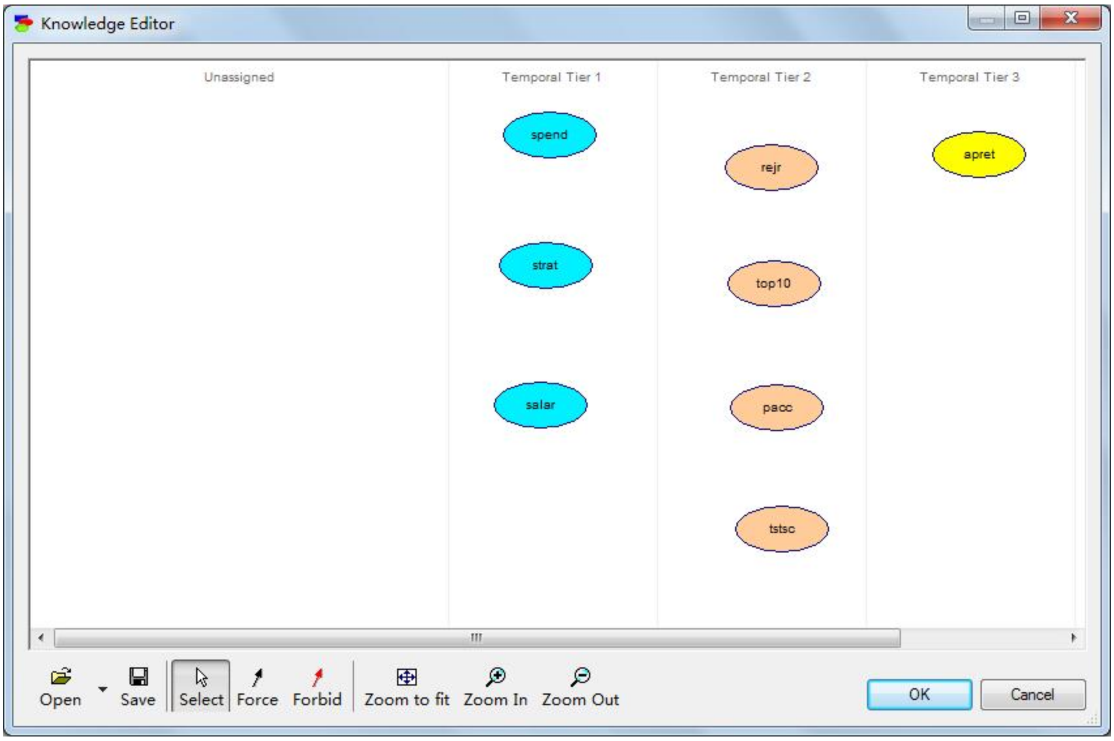
	spend	apret	top10	rejr	tstsc	pacc	strat	salar
spend								
apret	0.601231	-						
top10	0.675656	0.642464	-					
rejr	0.633544	0.514958	0.643163	-				
tstsc	0.71491	0.782183	0.798807	0.628601	-			
pacc	-0.23673	-0.302834	-0.207505	-0.0715207	-0.164223	-		
strat	-0.561755	-0.458311	-0.247857	-0.283617	-0.465226	0.131858	-	
salar	0.711838	0.635852	0.637648	0.606777	0.715472	-0.37524	-0.347673	-

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)
```



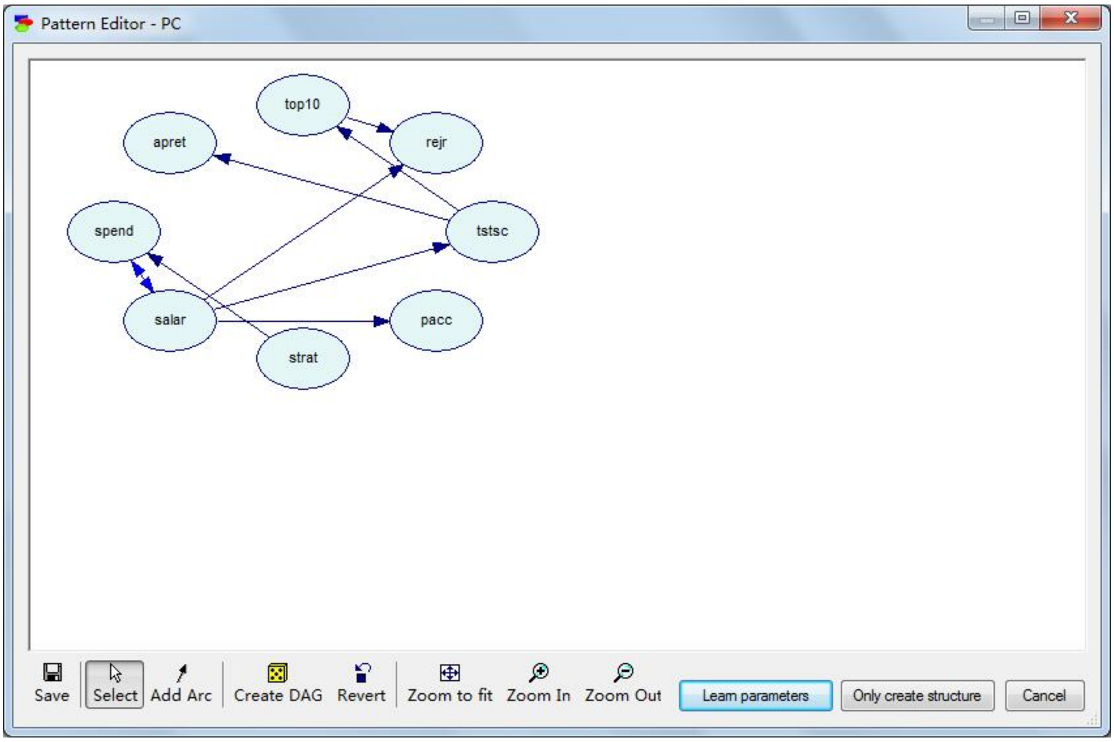
5. In order to force and forbid causal connections, we separate these 8 features into 3 temporal ties, which is “spend, start, salar” , “rejr, pacc, top10, tstsc” and “apret”.



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

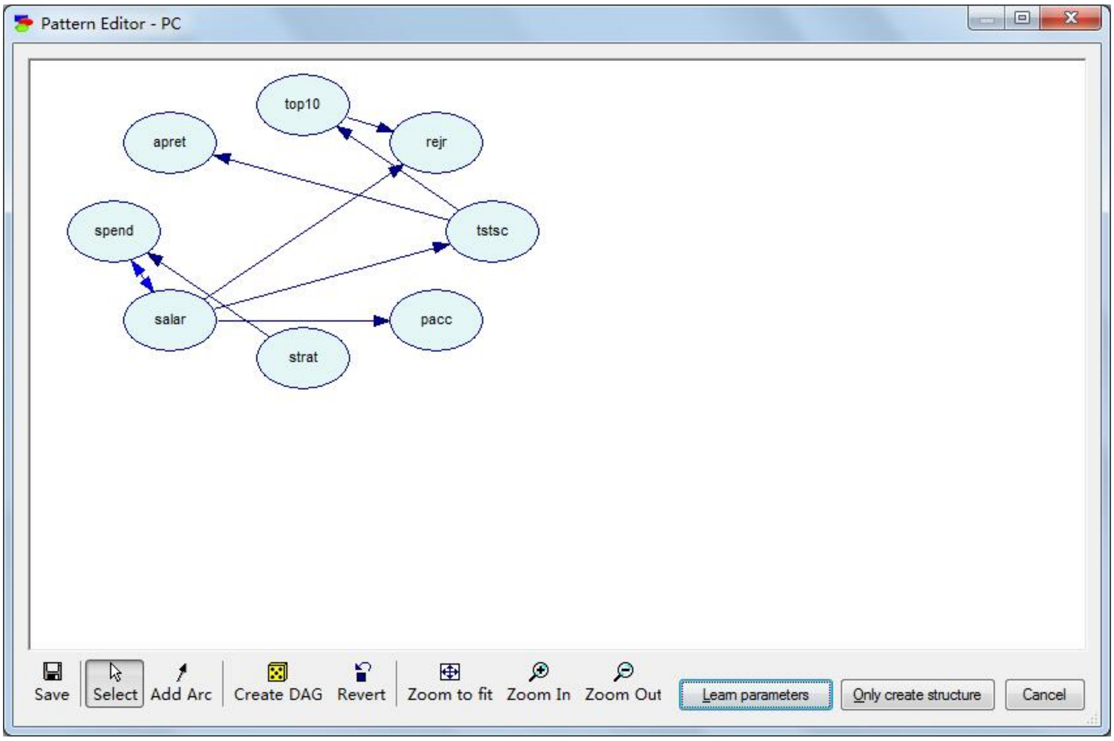
(1)significance level = 0.001:

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



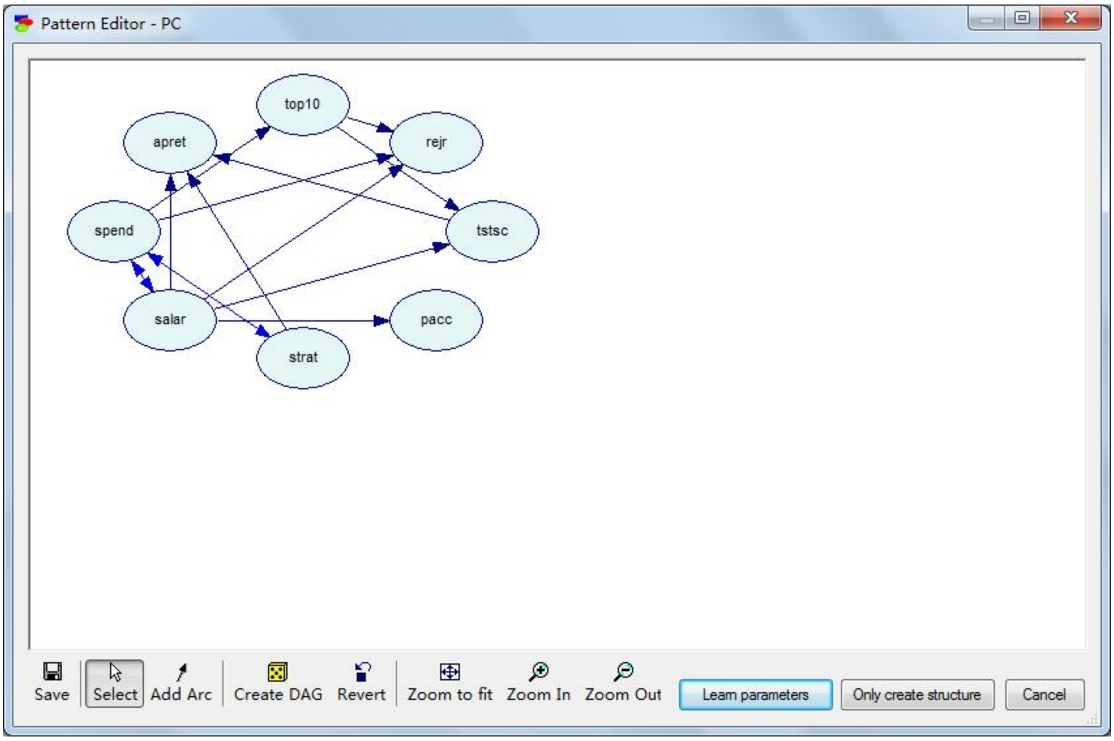
(2)significance level = 0.01:

```
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.