Data Analytics

Assignment 5

Causal Discovery

Group members:

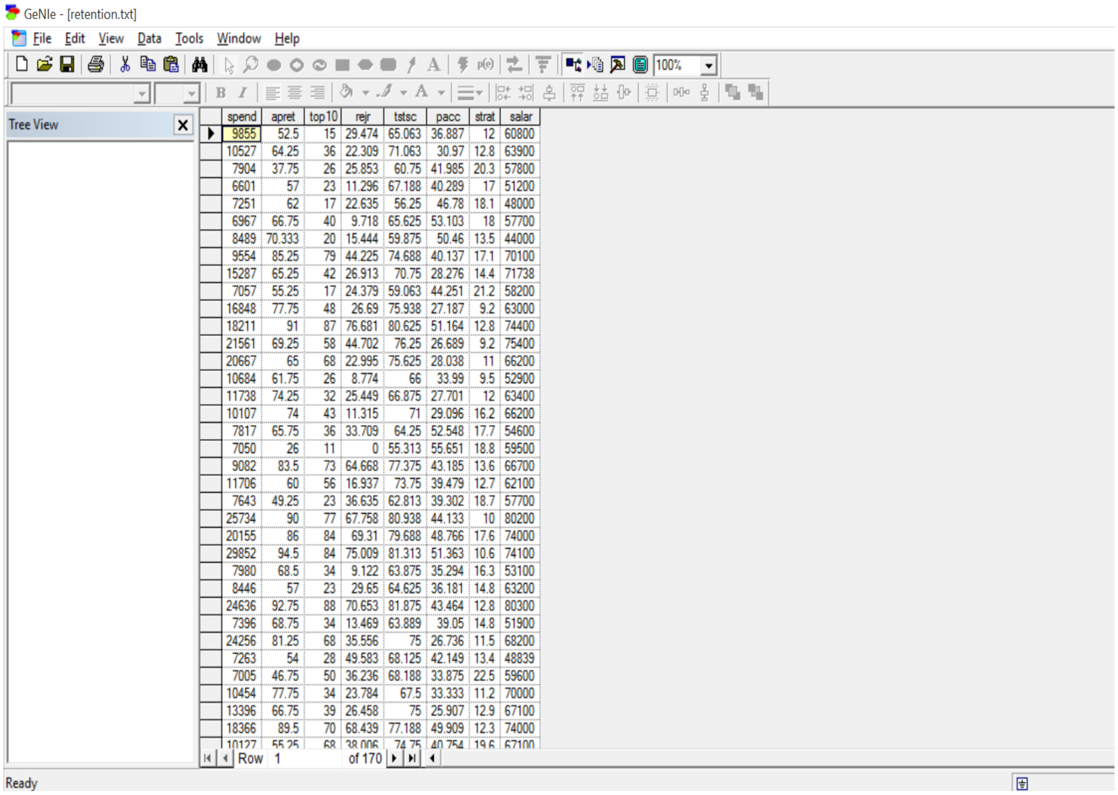
- Nanxun Xie
- Yifan Zhao

Using System: GeNIe

Preparation:

The reference reading article (kdd94.pdf) is concerning student retention in US colleges. The original study was performed on 1992 US News and World Report data, while the data that you will be studying is for the year 1993.

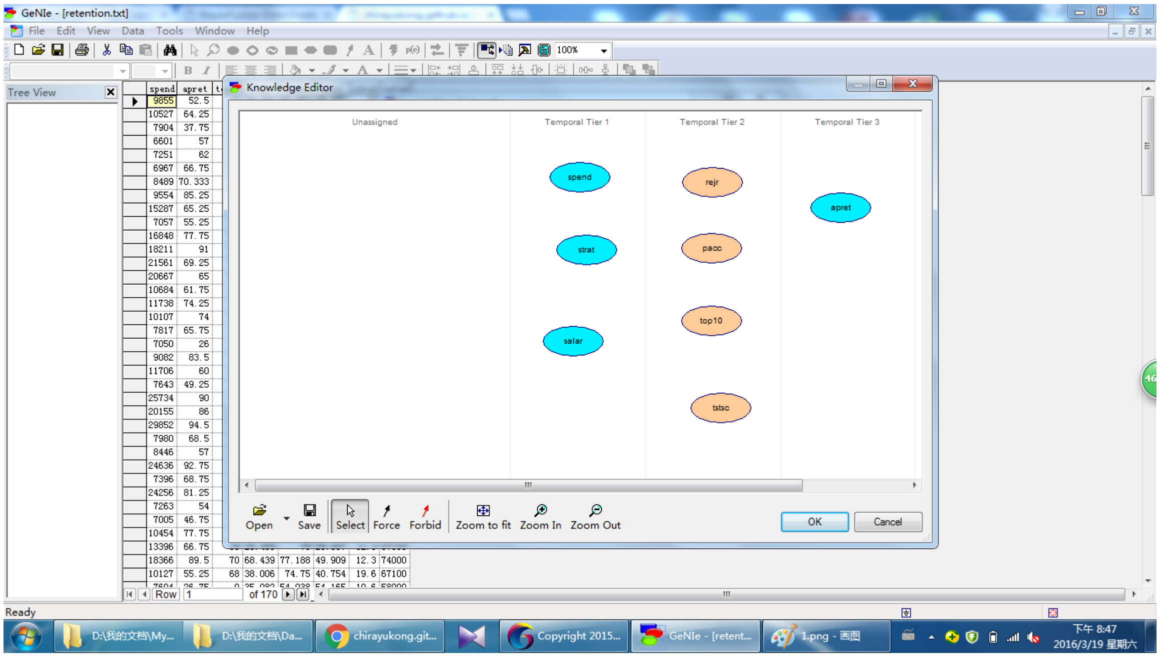
First of all, we need to import the data file (retention.txt) into GeNIe:



spend	apret	top10	rejr	tstsc	pacc	strat	salar
9855	52.5	15	29.474	65.063	36.887	12	60800
10527	64.25	36	22.309	71.063	30.97	12.8	63900
7904	37.75	26	25.853	60.75	41.985	20.3	57800
6601	57	23	11.296	67.188	40.289	17	51200
7251	62	17	22.635	56.25	46.78	18.1	48000
6967	66.75	40	9.718	65.625	53.103	18	57700
8489	70.333	20	15.444	59.875	50.46	13.5	44000
9554	85.25	79	44.225	74.688	40.137	17.1	70100
15287	65.25	42	26.913	70.75	28.276	14.4	71738
7057	55.25	17	24.379	59.063	44.251	21.2	58200
16848	77.75	48	26.69	75.938	27.187	9.2	63000
18211	91	87	76.681	80.625	51.164	12.8	74400
21561	69.25	58	44.702	76.25	26.689	9.2	75400
20667	65	68	22.995	75.625	28.038	11	66200
10684	61.75	26	8.774	66	33.99	9.5	52900
11738	74.25	32	25.449	66.875	27.701	12	63400
10107	74	43	11.315	71	29.096	16.2	66200
7817	65.75	36	33.709	64.25	52.548	17.7	54600
7050	26	11	0	55.313	55.651	18.8	59500
9082	83.5	73	64.668	77.375	43.185	13.6	66700
11706	60	56	16.937	73.75	39.479	12.7	62100
7643	49.25	23	36.635	62.813	39.302	18.7	57700
25734	90	77	67.758	80.938	44.133	10	80200
20155	86	84	69.31	79.688	48.766	17.6	74000
29852	94.5	84	75.009	81.313	51.363	10.6	74100
7980	68.5	34	9.122	63.875	35.294	16.3	53100
8446	57	23	29.65	64.625	36.181	14.8	63200
24636	92.75	88	70.653	81.875	43.464	12.8	80300
7396	68.75	34	13.469	63.889	39.05	14.8	51900
24256	81.25	68	35.556	75	26.736	11.5	68200
7263	54	28	49.583	68.125	42.149	13.4	48839
7005	46.75	50	36.236	68.188	33.875	22.5	59600
10454	77.75	34	23.784	67.5	33.333	11.2	70000
13396	66.75	39	26.458	75	25.907	12.9	67100
18366	89.5	70	68.439	77.188	49.909	12.3	74000
10127	55.25	68	38.006	74.75	40.754	19.6	67100

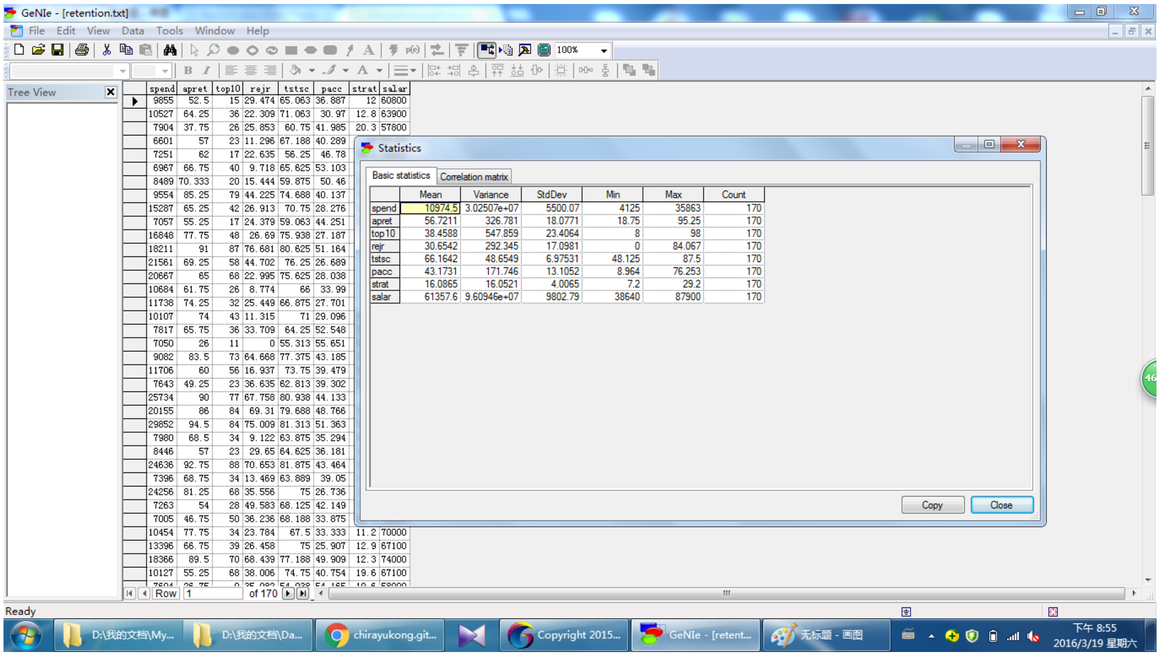
Observation

After importing the data file into the GeNIe, we need to learn new network. Clicking the "data" button on the top tools bar, then "learn new network". In the setting window, choosing the "PC" algorithms as our learning algorithm, next clicking the "background knowledge" button, here we separate the eight unsigned factors into 3 temporal tiers to restrict the model search for GeNIe, which is shown below.



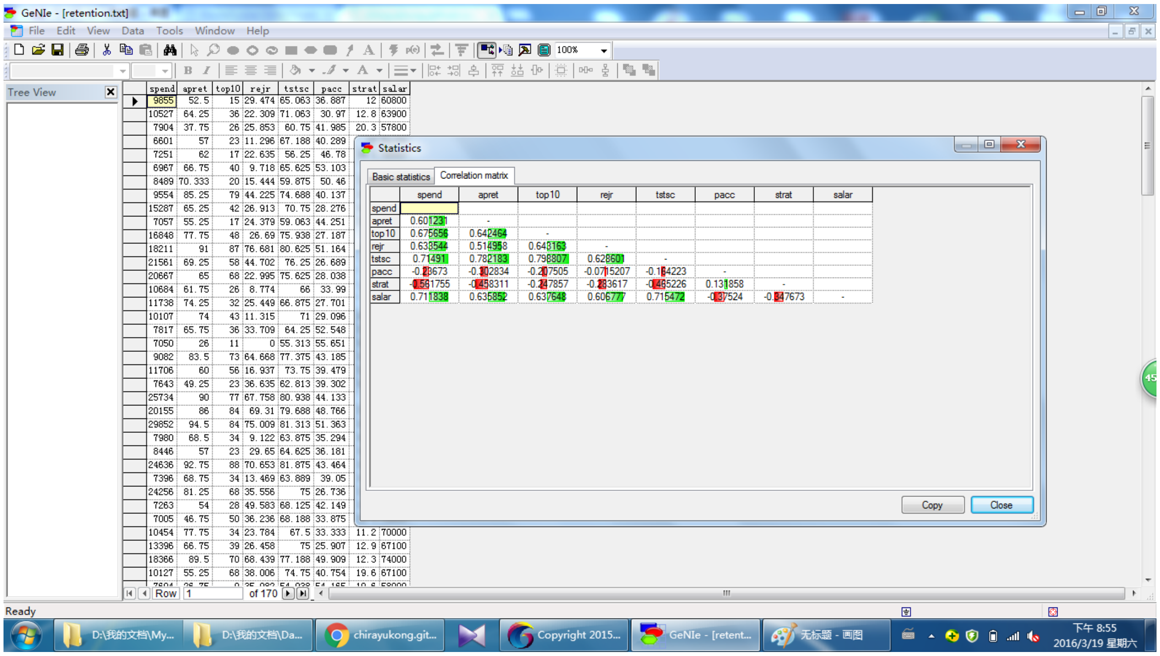
When GeNIe is running on normally distributed data with the assumption, it converts the raw data into a correlation matrix.

Following is the basic information of all raw 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

And the value of the elements of the matrix is all that matters in discovery. The corresponding matrix of all data points is shown below.



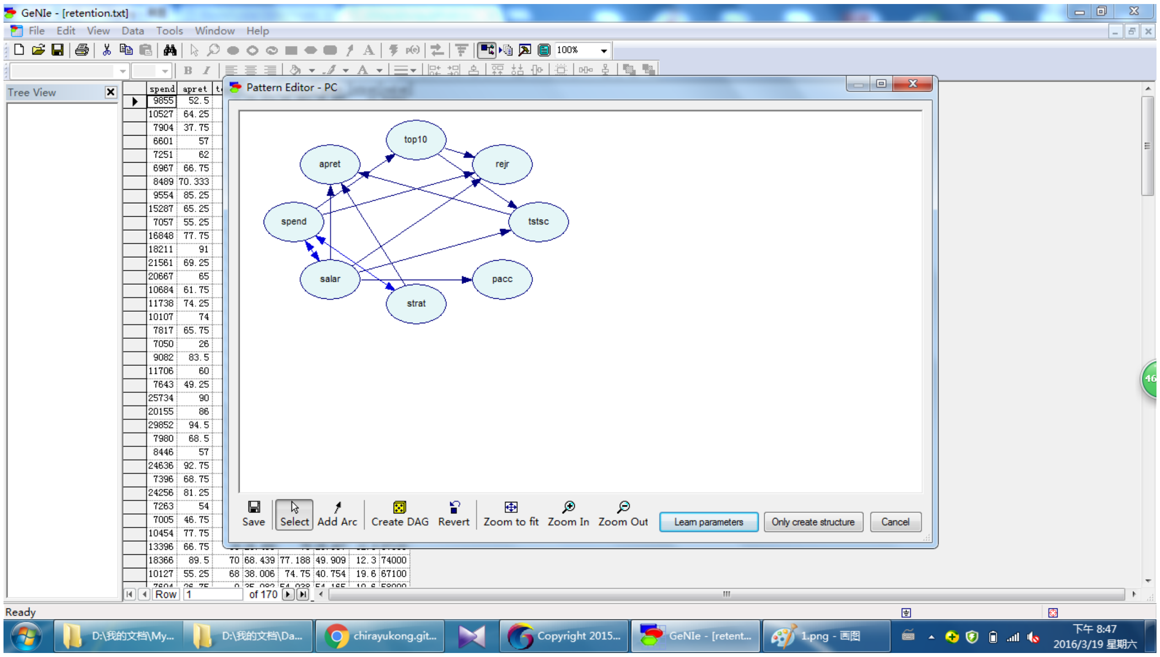
spend	apret	top10	rejr	tstsc	pacc	strat	salar
9855	52.5	15	29.474	65.063	36.887	12	60800
10527	64.25	36	22.309	71.063	30.97	12.8	63900
7904	37.75	26	25.853	60.75	41.985	20.3	57800
6601	57	23	11.296	67.188	40.289		
7251	62	17	22.635	56.25	46.78		
6967	66.75	40	9.718	65.625	53.103		
8489	70.333	20	15.444	59.875	50.46		
9554	85.25	79	44.225	74.688	40.137		
15287	65.25	42	26.913	70.75	28.276		
7057	55.25	17	24.379	59.063	44.251		
16848	77.75	48	26.69	75.938	27.187		
18211	91	87	76.681	80.625	51.164		
21561	69.25	58	44.702	76.25	26.689		
20667	65	68	22.995	75.625	28.038		
10684	61.75	26	8.774	66	33.99		
11738	74.25	32	25.449	66.875	27.701		
10107	74	43	11.315	71	29.096		
7817	65.75	36	33.709	64.25	52.548		
7050	26	11	0	55.313	55.651		
9082	83.5	73	64.668	77.375	43.185		
11706	60	56	16.937	73.75	39.479		
7643	49.25	23	36.635	62.813	39.302		
25734	90	77	67.758	80.938	44.133		
20155	86	84	69.31	79.688	48.766		
29852	94.5	84	75.009	81.313	51.363		
7980	68.5	34	9.122	63.875	35.294		
8446	57	23	29.65	64.625	36.181		
24636	92.75	88	70.653	81.875	43.464		
7396	68.75	34	13.469	63.889	39.05		
24256	81.25	68	35.556	75	26.736		
7263	54	28	49.583	68.125	42.149		
7005	46.75	50	36.236	68.188	33.875		
10454	77.75	34	23.784	67.5	33.333	11.2	70000
13396	66.75	39	26.458	75	25.907	12.9	67100
18366	89.5	70	68.439	77.188	49.909	12.3	74000
10127	55.25	68	38.006	74.75	40.754	19.6	67100
7604	26.75	0	35.082	54.938	54.165	10.6	58000

Row 1 of 170

	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	-

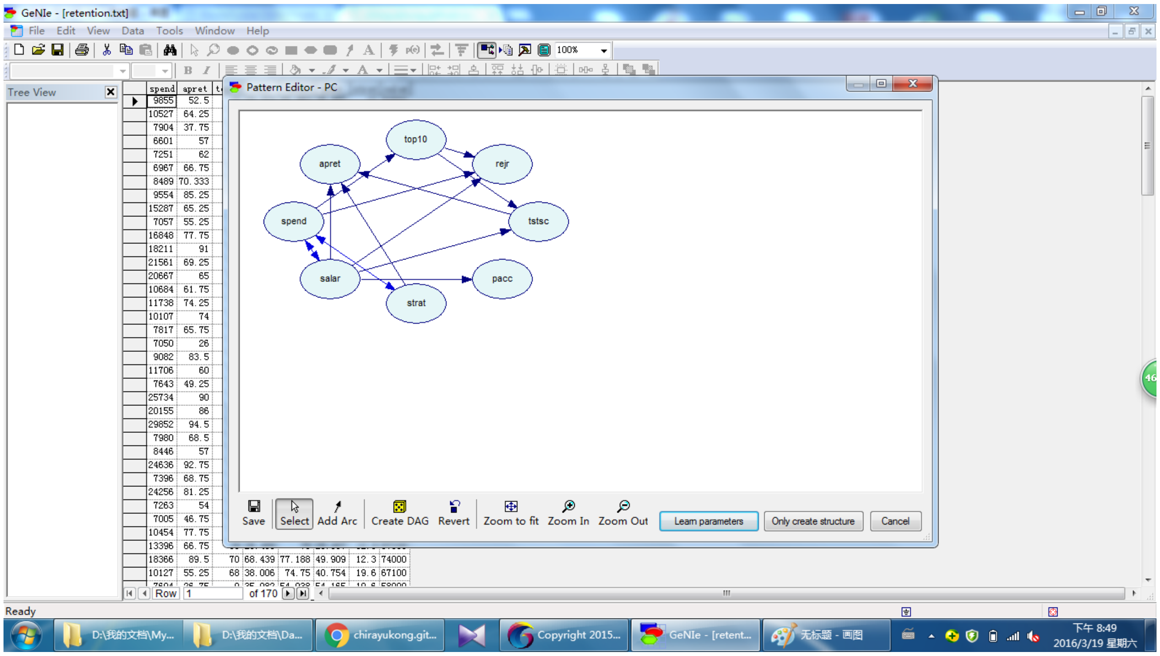
When finishing setting the background knowledge, we need to set different significance level. Depending on the significance level used in the independence tests, GeNIe's decisions regarding independence may be different and different class of causal structures may result. Therefore, it is a good practice to run the program at several significance level. Here we ran with the following significance level: p=0.2, 0.15, 0.1, 0.05, 0.01and 0.001. The graphs proposed by GeNIe for these significance level are presented following. The edges of the graph have the following meaning: A single arrow means a direct causal influence. A double arrow between two variables means presence of a latent common cause of these two variables.

P=0.2:



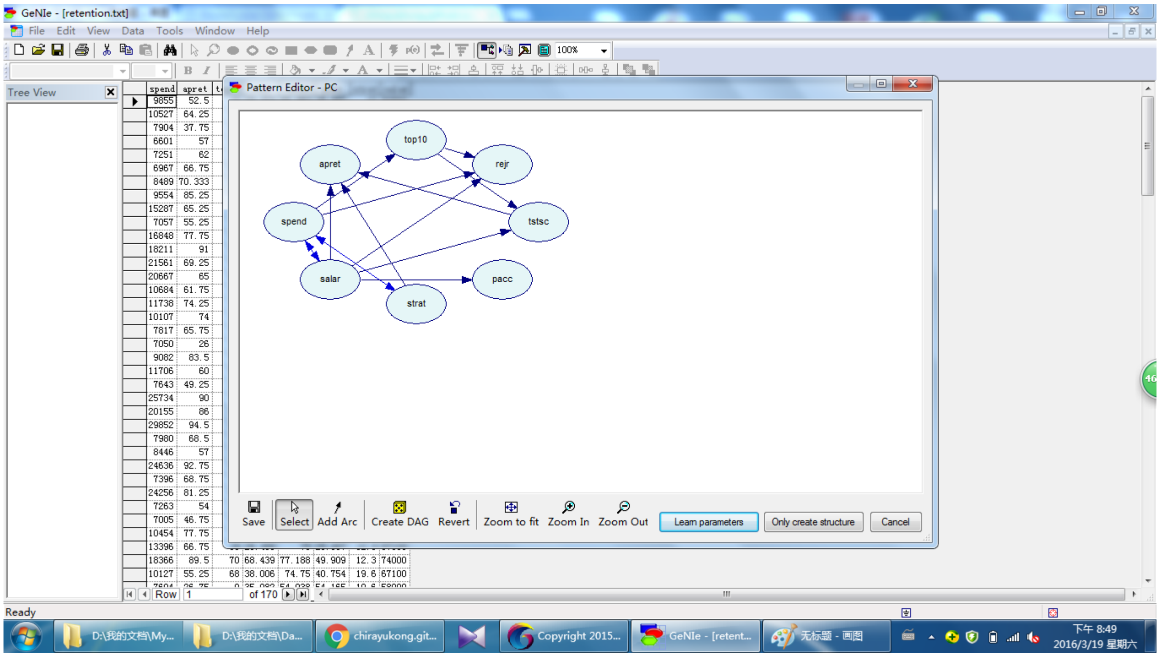
spend	apret						
9855	52.5						
10527	64.25						
7904	37.75						
6601	57						
7251	62						
6967	66.75						
8489	70.333						
9554	85.25						
15287	65.25						
7057	55.25						
16848	77.75						
18211	91						
21561	69.25						
20667	65						
10684	61.75						
11738	74.25						
10107	74						
7817	65.75						
7050	26						
9082	83.5						
11706	60						
7643	49.25						
25734	90						
20155	86						
29852	94.5						
7980	68.5						
8446	57						
24636	92.75						
7396	68.75						
24256	81.25						
7263	54						
7005	46.75						
10454	77.75						
13396	66.75						
18366	89.5	70	68.439	77.188	49.909	12.3	74000
10127	55.25	68	38.006	74.75	40.754	19.6	67100

P=0.15:



spend	apret						
9855	52.5						
10527	64.25						
7904	37.75						
6601	57						
7251	62						
6967	66.75						
8489	70.333						
9554	85.25						
15287	65.25						
7057	55.25						
16848	77.75						
18211	91						
21561	69.25						
20667	65						
10684	61.75						
11738	74.25						
10107	74						
7817	65.75						
7050	26						
9082	83.5						
11706	60						
7643	49.25						
25734	90						
20155	86						
29852	94.5						
7980	68.5						
8446	57						
24636	92.75						
7396	68.75						
24256	81.25						
7263	54						
7005	46.75						
10454	77.75						
13396	66.75						
18366	89.5	70	68.439	77.188	49.909	12.3	74000
10127	55.25	68	38.006	74.75	40.754	19.6	67100

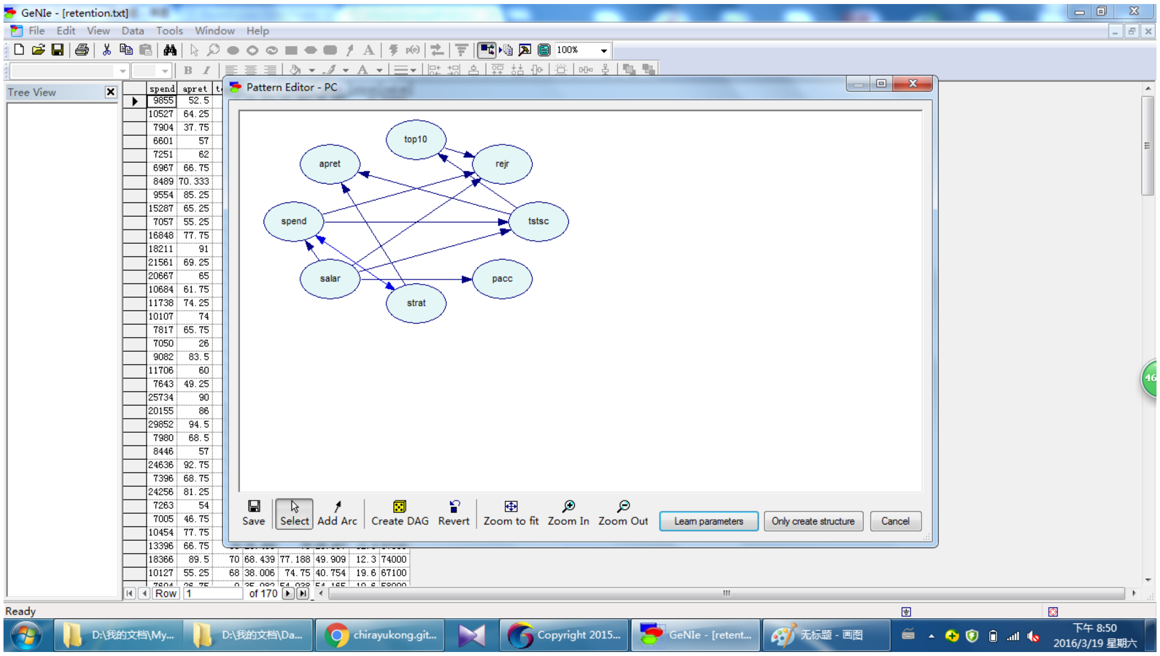
P=0.1:



spend	apret					
9855	52.5					
10527	64.25					
7904	37.75					
6601	57					
7251	62					
6967	66.75					
8489	70.333					
9554	85.25					
15287	65.25					
7057	55.25					
16848	77.75					
18211	91					
21561	69.25					
20667	65					
10684	61.75					
11738	74.25					
10107	74					
7817	65.75					
7050	26					
9082	83.5					
11706	60					
7643	49.25					
25734	90					
20155	86					
29852	94.5					
7980	68.5					
8446	57					
24636	92.75					
7396	68.75					
24256	81.25					
7263	54					
7005	46.75					
10454	77.75					
13396	66.75					
18366	89.5	70	68.439	77.188	49.909	12.3	74000
10127	55.25	68	38.006	74.75	40.754	19.6	67100

Row 1 of 170

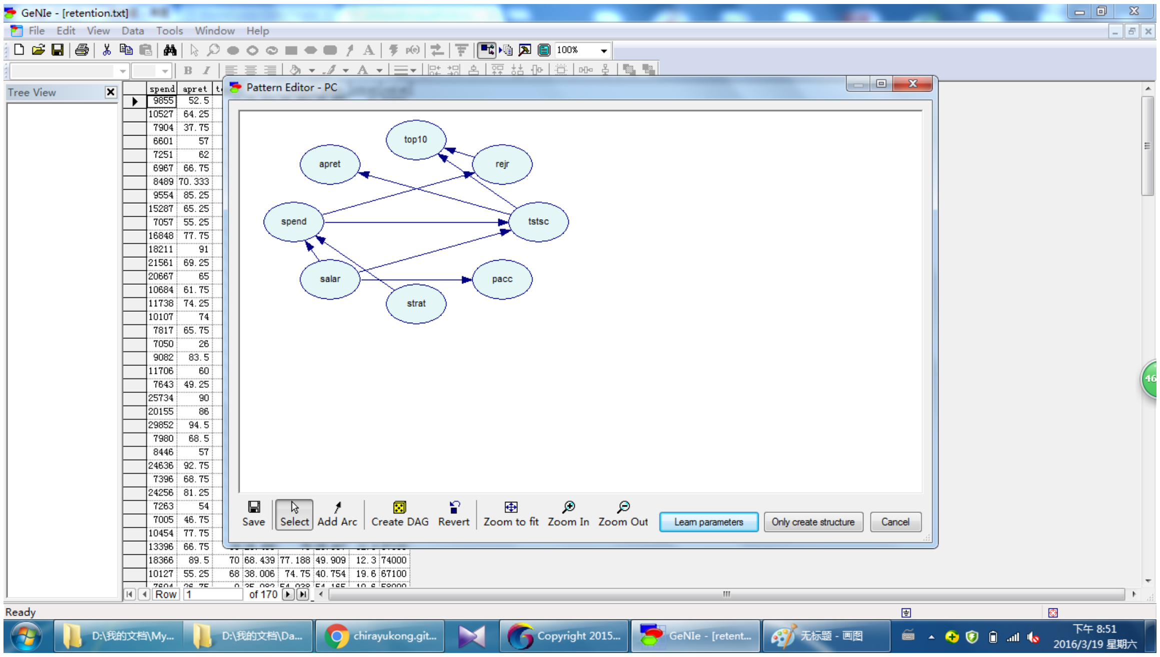
P=0.05:



spend	apret					
9855	52.5					
10527	64.25					
7904	37.75					
6601	57					
7251	62					
6967	66.75					
8489	70.333					
9554	85.25					
15287	65.25					
7057	55.25					
16848	77.75					
18211	91					
21561	69.25					
20667	65					
10684	61.75					
11738	74.25					
10107	74					
7817	65.75					
7050	26					
9082	83.5					
11706	60					
7643	49.25					
25734	90					
20155	86					
29852	94.5					
7980	68.5					
8446	57					
24636	92.75					
7396	68.75					
24256	81.25					
7263	54					
7005	46.75					
10454	77.75					
13396	66.75					
18366	89.5	70	68.439	77.188	49.909	12.3	74000
10127	55.25	68	38.006	74.75	40.754	19.6	67100

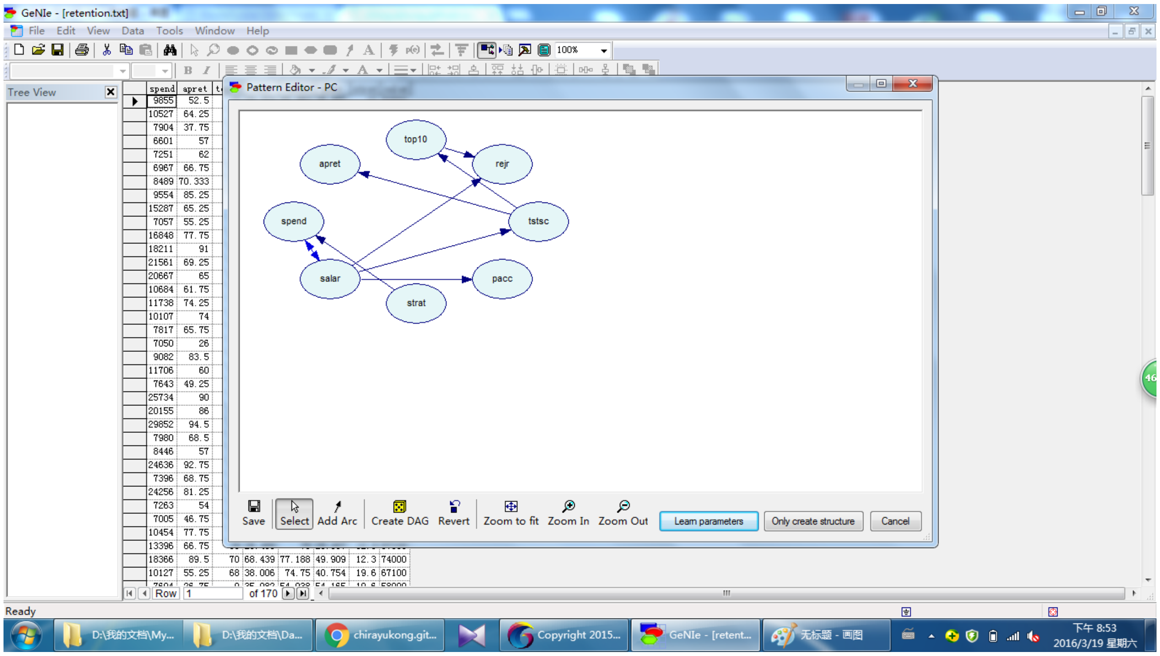
Row 1 of 170

P=0.01:



spend	apret						
9855	52.5						
10527	64.25						
7904	37.75						
6601	57						
7251	62						
6967	66.75						
8489	70.333						
9554	85.25						
15287	65.25						
7057	55.25						
16848	77.75						
18211	91						
21561	69.25						
20667	65						
10684	61.75						
11738	74.25						
10107	74						
7817	65.75						
7050	26						
9082	83.5						
11706	60						
7643	49.25						
25734	90						
20155	86						
29852	94.5						
7980	68.5						
8446	57						
24636	92.75						
7396	68.75						
24256	81.25						
7263	54						
7005	46.75						
10454	77.75						
13396	66.75						
18366	89.5	70	68.439	77.188	49.909	12.3	74000
10127	55.25	68	38.006	74.75	40.754	19.6	67100

P=0.001:



spend	apret						
9855	52.5						
10527	64.25						
7904	37.75						
6601	57						
7251	62						
6967	66.75						
8489	70.333						
9554	85.25						
15287	65.25						
7057	55.25						
16848	77.75						
18211	91						
21561	69.25						
20667	65						
10684	61.75						
11738	74.25						
10107	74						
7817	65.75						
7050	26						
9082	83.5						
11706	60						
7643	49.25						
25734	90						
20155	86						
29852	94.5						
7980	68.5						
8446	57						
24636	92.75						
7396	68.75						
24256	81.25						
7263	54						
7005	46.75						
10454	77.75						
13396	66.75						
18366	89.5	70	68.439	77.188	49.909	12.3	74000
10127	55.25	68	38.006	74.75	40.754	19.6	67100

As the correlation matrix of the raw data shown firstly, our result is similar with the reading article's observation.

According to the reading and the GeNie graphs all above, we found that the core of the structure, i.e., how freshmen retention rate is related to the remaining variables, is insensitive to changes in significance. This suggests that the proposed by GeNIe is robust. The graphs for different significance levels are shown above, we can readily to find that there are also some slight differences of the connection between each two indicators even if the structure is robust. And the connection between factors in the causal structure is varying as we change the significance level, however, we found the first three changes of level (p=0.2, p=0.15, p=0.1) do not influence any connections between factors, the graph is totally the same. After we change the level to p=0.05, the connection starts to change, the edge is decreasing which means some kinds of connection are missing. Besides, the graph is also keep changing when changing the levels to p=0.01 and 0.001,

Based on the first three graphs, patterns with the significance level p=0.2, 0.15 and 0.1 are the same, we found a new demonstration beyond the original article, which indicates that the average retention rate is directly influenced by three factors: salar, strat, tstsc in the significance level 0.2, 0.15 and 0.1. But after observation, we found that the direct connection between salar and apret disappears in the significance level equal to 0.05 and the direct connection between strat and apret is also disappears after the significance level p<0.04, so we suggest these two factors are not robust. And most graphs contained the direct causal connection between the tstsc and the apret, which means that the connection between tstsc and apret never changed. There are also other direct connections between other factors but we have to ignore those connections because they are not related to our aim which is about the causal structure of apret (average retention rate).

In the first four graphs, the "latent common cause" connection between spend and strat, disappears at p<0.05, while the connection between spend and salar disappears at p=0.05 and p=0.01 while re-appears at p=0.001.

In running the GeNIe, it proposed different orderings of variables, all direct links, and the direct link between test scores and retention and graduation in particular, were the same in both cases. It seems that none of the variables in the data set are directly causally related to retention except for test scores. Given by average test scores are conditionally independent of all remaining variables, seems to be robust across varying significance levels, which is similar to the result of the article. The average test scores seem to have a high predictive power for student retention. Average test scores can be viewed as the main indicator of the quality of incoming students. It seems that retention rate in an individual college can be improved by increasing the quality of the incoming students. Changing factors such as faculty salary, student/teacher ratio, or spending per student should, according to GeNie result, have no direct effect on freshmen retention. After finishing the procedure, we think that the student retention is mainly caused by the average test scores.

In a word, all raw data points and the test of 1993 data almost support all the result of the Druzdzel & Glymour's conclusions, and we also also found out some similar results by GeNie. By using the PC algorithm, we have already set the time sequence of given data to meet the consideration of the time precedence.