# A Case Study of Data Analysis for Educational Management

Wasit Limprasert

Department of Computer Science, Faculty of Science and Technology, Thammasat University, Pathumthani, Thailand
wasit_l@sci.tu.ac.th

Somkiat Kosolsombat

Department of Computer Science, Faculty of Science and Technology, Thammasat University, Pathumthani, Thailand
somkiat.k@sci.tu.ac.th

***Abstract*— The recent technology development introduces many aspects of difficulty in education management. There are many new technologies and new businesses emerging in a single year. As a result of business sectors require a rapid changing of curriculum from academic institutions. However, the curriculum revision has to be done in a very short time comparing to a large amount of data to be processed, resulting in the teaching materials are often lack behind the business requirements. Education management has to be sensitive and responsive to the constantly changing of technology environment. In this paper, data analytic tools are applied for discovering a new dependence between courses, in order to improve the current sequence of prerequisites. The dependence detection is made of a feature importance technique based on an extra-trees classifier. We also compare various types of classifiers for predicting a grade of a next enrollment using the results of prior studies. The grade prediction is designed to suggest the best study path to match individual talent.**

***Keywords-component; Education management, Curriculum planning, Data analytic***

## I. INTRODUCTION

In 2000, a study [1] showed a low status of social and financial support has significant on school dropping out in Thailand. The study also suggested that government should make education more accessible for all eligible students. Consequently, Thailand government has progressively supported education in many ways to gain a number of students in higher education. Resulting in the tertiary education enrolment rate reaching 45% by 2008 [2]. However, the management in universities and the entire educational system both are under average compared all countries [2]. Another research in Netherlands [3] studied on a quality management of a curriculum by collecting and analyzing the data from various kinds of sources; students, faculty members and experts from outside the university. The result suggested that collecting data by a questionnaire in the qualitative study is not only expensive but also creating a biased result, depending on the list of topics in the questionnaire. There was no intense data analysis because of the limited technology in that period.

A success student in higher education mostly evolves some degree of planning. When a student is making a plan of something, it usually involves a metal exercise and a simulation of consequences in the future. In 2006, a case study[4] showed comparison of two approaches, in order to teach students to manage a project. The first approach is using a static traditional method compared to the second method, which is an e-learning simulation base approach. The result showed that the simulation base is significantly better than the static approach. Moreover, the simulation base is easier to gain interest from the students and promisingly makes them aware of difficulty and complexity of the system. Whereas, the traditional approach cannot effectively illustrate the dynamic and complexity of the system. These suggested that a proper tool can help a student to make an effective plan, not only for a project management but also for a study planning.

Recently there have been a dramatic change in the information and communication technologies affecting many employment patterns. This also introduces a new set of skills required from business sectors. In 2009, an article [5] by Wilf Altman suggested that business schools in UK must be responsive and interactive with the business activities outside universities. The article showed that a number of senior managers need re-training to equip with latest technologies to meet the company requirements and to prevent cutback. However, re-training in business schools does not meet the requirement because of most faculty members have a lack of business experience. Most of the case studies and publications were not practical and not responsive to business environment. Thus, he suggested that we need innovation methods to responsively manage the education system. The Similar phenomenon also affects other fields of study.

In the Computer Science field of study, there have been a lot of changes in a decade. We have seen some technologies thriving to the peaks of a hype cycle while another rapidly dropping out from the news. We need to response to the sudden changes and have a plan for long term development. The curriculum designers must have a robust tool to analyze the teaching processes in the institution and quickly comes out with a responsive plan.

In this paper, we propose a possible method to process a history of enrollments for detecting a dependency of knowledge between courses in a curriculum. This will help the curriculum designer to adjust the sequence of prerequisites effectively. Furthermore, various types of classifier are examined to prepare for a grade predicting application. The purpose of grade prediction in this work is not only to maximize the overall outcome but also to suggest a best suitable study path to match the characteristics of individual student.

978-1-5090-2033-1/16/$31.00 ©2016 IEEE

978-1-5090-2033-1/16/$31.00 ©2016 IEEE

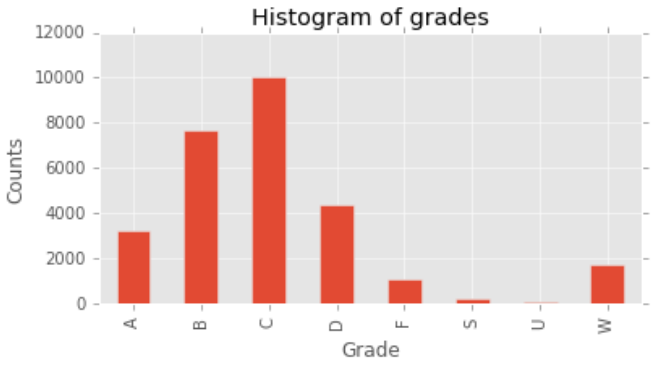


Figure 1: histogram of grade from the CS09

## II. RELATED WORK

### A. Grade prediction

A next term student grade prediction in [6] processed an enrollment history consist of 310,498 valid records, where a record contains an anonymous student ID along with a course detail. The grade of an enrollment is represented by an alphabet ranging from A to F, which were converted to a nominal range between 0 and 4. The study compared many prediction methods such as Factorization machine (FM) [7], Mean of Mean, SVD and Global means. The best root mean square error is from FM predictor in a range between 0.78 to 0.80. Other machine learning techniques can be used for predicting student's performance for example Bayesian network[8], Random forest[9], confidence-learning prediction algorithm[10].

## III. OUR DATA ANALYSIS METHOD

In this work, we attempt to predict a grade of each enrollment and also investigate the relationship between the prior grades on the grade of a particular subject in temporal order. In first experiment in section III.B, two sets of machine learning implementations; Scikit-learn [11] and Weka [12], are compared using enrolment dataset from Thammasat University. Various classifiers are compared in order to find the best classifier for the grade predicting application. In the second experiment in section III.C, Extra-trees classifier from Scikit-learn is used for detecting dependences from unsorted data. The results from the experiments are described in section IV and summarized in Section V.

### A. CS09 Dataset

In this paper, the anonymous dataset is acquired from a registrar of Thammasat University under a policy of privacy protection, the student ID is regenerated to prevent identification. We have acquired anonymous enrollment records of students from Computer Science department of Thammasat University. The student enrollment history has been constantly collected for 4 years between 2009 and 2013 consisting of over 28,272 records. A summarized histogram of grades is shown in Figure 1, where the best grade is 'A' descending to the worst grade 'F'. The binomial grades

*Before appending*

| STUDENTID | TERM | COURSEID | GRADE |
|---|---|---|---|
| 0001 | 2009/1 | X | A |
| 0001 | 2010/1 | Y | B |
| 0001 | 2010/2 | Z | C |
| 0002 | 2009/1 | X | D |
| 0002 | 2010/2 | Z | F |

*appending prior results*

*After appending*

| STUDENTID | TERM | COURSEID | GRADE | X | Y | Z |
|---|---|---|---|---|---|---|
| 0001 | 2009/1 | X | A | - | - | - |
| 0001 | 2010/1 | Y | B | A | - | - |
| 0001 | 2010/2 | Z | C | A | B | - |
| 0002 | 2009/1 | X | D | - | - | - |
| 0002 | 2010/2 | Z | F | D | - | - |

Figure 2: The temporal transformation

satisfied and unsatisfied are represented by letter 'S' and 'U'. The grade 'W' expressing withdrawing the subject.

The original dataset is a table consisting of many fields. Some unrelated fields are dropped out because they are similar throughout the dataset, which does not improve the discrimination power in the training process. The unrelated fields are for example; CAMPUS_ID and CURRICULUM_ID. The remained fields that used in further experiments are the following; STUDENT_ID, COURSE_ID, TERM, GRADE. A grade character is mapped to integer because most of classifiers require numeric input. The mapping is shown as the following;
{'A':8, 'B':7, 'C':6, 'D':5, 'F':4, 'W':3, 'S':2, 'U:'1, '-':0}, where '-' means that the student never enrolls in the particular course.

The grades of previous enrolments are transformed to create a feature vector for a training in order to construct a classification model. Figure 2 shows the temporal transformed data calculated by our transformation method. The input table of enrolments is appended with the new columns and initialized by '-'. Then it is transformed by assigned the value of the previous grade results. For each student, the records are sorted by temporal index (TERM) and then a grade from prior study is filled to the subsequent record. The process repeats until all records are visited. This temporal transformation constructs a pattern of prior grade, also known as a feature vector, which will be used in the later training processes.

### B. Classifier comparison

In order to predict a grade of a particular enrolled subject, we use the classifier to construct a model in training process. The proposed temporal transformed data is used for comparing many classifiers from Scikit-learn [11] and Weka [12]. In this experiment, we examine many classifiers as in the flowing list; DT ( Decision Tree [11] ), RF ( Random forest [11] ), ET ( Extra-trees [11] ), SVM ( support vector machine[11] ), RFw ( Random forest [12] ), SVMw ( support vector machine[12] ). The ensemble classifiers such as Random forest and Extra-tress are set to generate 10 estimators. The classifiers are evaluated by k-fold cross validation (k=5). For 5-fold cross validation, 80% of records are processed in the training and another 20% is used for evaluation.

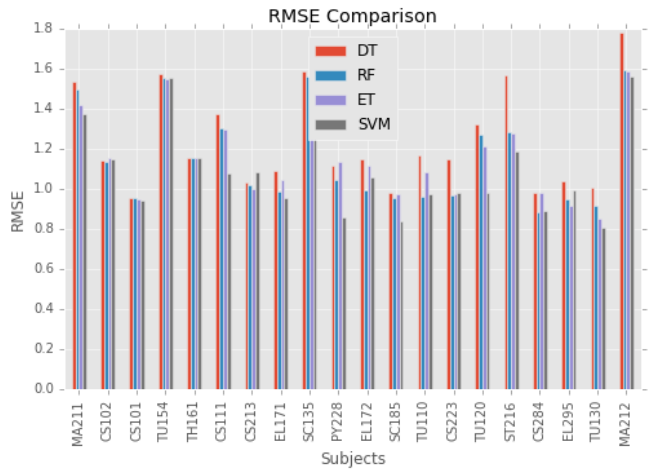


Figure 6: RMSE the first 20 subjects

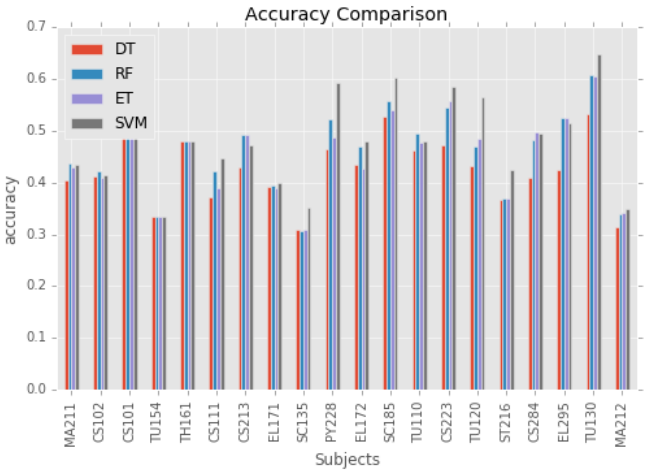


Figure 5: Accuracy of prediction for the first 20 subjects

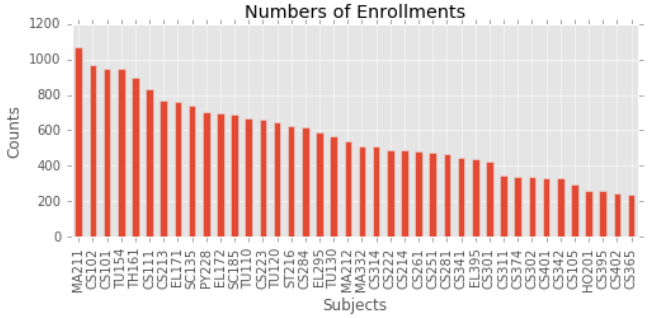


Figure 4: number of enrollments of the top 40 subjects

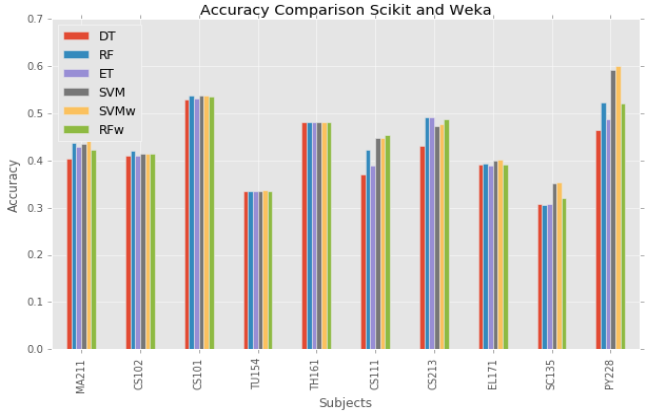


Figure 3: comparison between Scikit-learn and Wega

The predicted grades are evaluated by Root Mean Square Error (RMSE) and accuracy. For a record $i$, a numeric predicted grade $y_{pi}$ is compared to an actual grade $y_{ai}$ (ground-truth).

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_{pi} - y_{ai})^2}{N}}$$

In order to measure an accuracy of a classification model, the predicted grade is summarized using confusion matrix before calculating the accuracy. Where a true positive $T_p$ and a true negative $T_n$ are on the diagonal. The accuracy is the summation along the diagonal divided by the total number of test samples N.

$$Accuracy = \frac{T_p + T_n}{N}$$

The transformed data as shown in Figure 2 are filtered to select only a single subject at a time for training. Therefore, with this training approach, the prediction model for an individual subject can be constructed without difficulty. However, the number of appended columns can be very large when increasing a number of available courses. After transformation, the number of records in the remaining dataset is 28,272 rows and 199 columns with 195 subjects. This transformation increases the number of data element to be a squared of the original data. Applying this method on a very large dataset may take a very long time for data transformation.

### C. *Dependency detection*

An unsuccessful learning path depends on many factors such as student attention, teaching method and background of knowledge. To guarantee that a student has a sufficient background of knowledge, prerequisites are required before registering. However, most of curriculum revisions base on human intuition relying on an inadequate set of data to support the decision. It this section we propose a method to delicately detect hidden dependency pattern in the enrolment history.

In order to extract dependency of courses that influencing on a particular a new enrolled course, a decision tree base classifier is trained and dissected to expose all features that occurring in the prediction model. The Gini importance [13] is measured for acquired impurity reduction of data during the process of feature space partitioning to form a decision tree. There is an evidence showing that Gini importance in decision tree base classifier such as Random Forest [13] has better performance [14] in feature selection compared to principal component regression. The process of partitioning reduces impurity of data from the parent node to the children nodes in the decision tree, in another word it increases information gain. The Gini score produced by particular feature is then exported as feature importance score (FIS).

```
Feature important ranking:
CS284 is under the influence of ['SC135', 'MA211', 'CS111']

subject: SC135, importance:0.060, acc_sum:0.060
subject: MA211, importance:0.060, acc_sum:0.120
subject: CS111, importance:0.059, acc_sum:0.179
subject: TU154, importance:0.055, acc_sum:0.234
subject: EL171, importance:0.054, acc_sum:0.288
subject: EL172, importance:0.052, acc_sum:0.340
```

Figure 7: example output of feature importance ranking.

In this specific experiment, a list of prior grades of all courses are converted to numeric data and assigned to be a feature vector. A value in an element in the feature vector represents a grade of a single enrolment. The collection of feature vectors are used for training in order to generate the ExtraTree classifier[11]. Then the feature importance score (FIS) is extracted from the classifier. The most import feature is the most frequently occurring and contributing to largest impurity reduction. Then the features are sorted according to FIS. The feature on top of the rank in the first 20% is presumably to be the most significant course that heavily influent on the considered subject. However, there are many subjects that never be influenced by any other subjects in the past. Therefore, we select only the subject (feature) that has at least 0.05 of FIS in order to filter out any independent subjects. Figure 7, shows the FIS of CS284, where the first three important subjects are highlighted; SC135, MA211 and CS111. The first three subjects are necessary for constructing the classifier. Please note that the FIS does not indicate positive correlation as similar to other correlation analysis. The FIS shows importance of the feature that contributes to accuracy of a classification.

## IV. RESULT

### A. Classifier comparision

Figure 4 and Figure 6 show a comparison between various classifiers evaluated by 20 most popular courses (based on a number of enrollments). The average result acquired from the experiment are concluded in Table 1. DT has the lowest performance in the test. While, SVM and Random forest perform as the first and the second in the comparison. Figure 5 shows classifier comparison between various implementations from Scikit-learn[11] and Weka[12] evaluated by the first 10 subjects sorted by the number of enrollments. The result illustrates no significant different between those implementations.

### B. Dependency detection

Figure 7 shows an output from FIS processing. The result ranking is sorted from the most influent course on the top to the least. In this particular case the course CS284 *Introduction to Software Engineering* is under the influence of the following subjects;

- SC135 *General Physics*
- MA211 *Calculus I*
- CS111 *Object-Oriented Programming*

Table 1: mean RMSE and accuracy using 5-fold cross validation

| **Classifier** | DT | RF | ET | SVM |
|---|---|---|---|---|
| **mean RMSE** | 1.14 | 1.00 | 1.03 | 0.98 |
| **mean accuracy** | 0.48 | 0.53 | 0.52 | 0.54 |

Table 2: The most influent subject sorted by the number of impacts

| **Subject** | **Description** | **#Impacts** |
|---|---|---|
| TU154 | *Foundation of Mathematics* | 17 |
| MA211 | *Calculus I* | 13 |
| CS102 | *Computer Programming Fundamentals* | 12 |
| EL171 | *English Course II* | 9 |
| CS111 | *Object-Oriented Programming* | 8 |
| SC135 | *General Physics* | 6 |
| CS101 | *Discrete Structures* | 5 |
| TH161 | *Thai Usage* | 4 |
| TU110 | *Integrated Humanities* | 3 |
| CS284 | *Introduction to Software Engineering* | 3 |

According to the current curriculum, CS111 is a prerequisite of CS284. This makes our dependence detection is reasonably correct because the method can pick a dependence signal from enrollment history and the detection follows the suggestion in the curriculum. Therefore, a successful student in CS111 is likely to perform well in CS284.

Furthermore, we found some prerequisites have no influent on a subsequent subject, for example, CS251 *Database System I* has a prerequisite of CS213 *Data Structures*. However, our dependence detection found a strong signal of influence from CS223 *Computer Organization and Architecture*. This suggest there must be an unordered arrangement of knowledge appearing in the curriculum. The problem could come from syllabus, student background, or teaching method, which requires further investigation. When sorting FIS and selecting the top 10 influences to other subjects, we found the most influent subjects as shown in Table 2. According to the result, the system suggests a student in this curriculum should study three groups of knowledge in the early year;

- Math and Physics: TU154, MA211, SC135, CS101
- Programming languages: CS102, CS111
- Others: EL171, TH161, TU110, CS284

We found that subjects in the group Math and Physics provide a secure foundation for Computer Science field of study. Whereas only few subjects in Computer Programming have effect on the entire curriculum.

## V. DISCUSSION

In the first experiment, the comparison between state of the art classifiers shows comparable predictive ability to [6], which has RMSE between 0.78 and 0.8 compared to our result in range of 0.98 to 1.14. The higher RMSE from our result is the effect from a larger number of categories of grades and a smaller number of dataset. As in many previous research articles, e.g. [15], [16], show that increasing a number of dataset can improve accuracy.

In the second experiment, we found that the numbers of dependences are corresponding to the original curriculum. However, there are some subjects, which do not follow the prerequisite guidelines suggesting a further investigation. The

discovered influences from other subjects, which is not in a prerequisite list, may infer the strong dependences between those connected courses. The discovery of these influences probably have profound effect on the future curriculum revision.

## ACKNOWLEDGMENT

The authors would like to thank Methinee Methavanich and Piyaporn Jaiareerob for data cleaning and preliminary analysis.

## REFERENCES

[1] I. Nicaise, P. Tonguthai, and I. Fripont, "School dropout in Thailand: causes and remedies," 2000.

[2] *Global Competitiveness Report: Data from World Economic Forum*. .

[3] A. van Peppen and M. R. van der Ploeg, "Practising what we teach: quality management of systems-engineering education," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 30, no. 2, pp. 189–196, May 2000.

[4] D. Rodrguez, M. n Sicilia, J. J. Cuadrado-Gallego, and D. Pfahl, "e-Learning in Project Management Using Simulation Models: A Case Study Based on the Replication of an Experiment," *IEEE Trans. Educ.*, vol. 49, no. 4, pp. 451–463, Nov. 2006.

[5] W. Altman, "What's wrong with management education?," *Eng. Technol.*, vol. 4, no. 6, pp. 76–79, Apr. 2009.

[6] M. Sweeney, J. Lester, and H. Rangwala, "Next-term student grade prediction," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 970–975.

[7] S. Rendle, "Factorization Machines with libFM," *ACM Trans Intell Syst Technol*, vol. 3, no. 3, p. 57:1–57:22, May 2012.

[8] A. Sharabiani, F. Karim, A. Sharabiani, M. Atanasov, and H. Darabi, "An enhanced bayesian network model for prediction of students' academic performance in engineering programs," in *2014 (EDUCON)*, 2014, pp. 832–837.

[9] D. Petkovic, K. Okada, M. Sosnick, A. Iyer, S. Zhu, R. Todtenhoefer, and S. Huang, "Work in progress: A machine learning approach for assessment and prediction of teamwork effectiveness in software engineering education," in *(FIE), 2012*, 2012, pp. 1–3.

[10] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Predicting Grades," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 959–972, Feb. 2016.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *J Mach Learn Res*, vol. 12, pp. 2825–2830, Nov. 2011.

[12] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 10-Apr-2016].

[13] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[14] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics*, vol. 10, p. 213, 2009.

[15] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *2011 (CVPR)*, 2011, pp. 1297–1304.

[16] W. Limprasert, "Parallel random forest with IPython cluster," in *2015 International Computer Science and Engineering Conference (ICSEC)*, 2015, pp. 1–6.

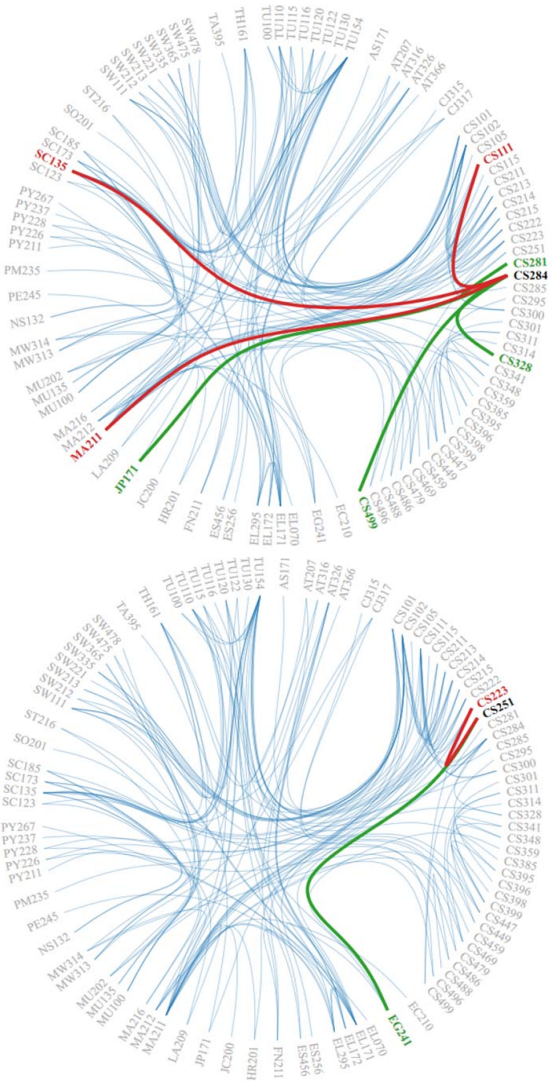


Figure 8: result from dependency detection (top) considering CS284 requires CS111, SC135 and MA211. CS284 are required for CS281 CS328 CS499 and JP171. (bottom) CS251 requires CS223 and it is a main influence no EG241