Spring 2014

nanoHUB Database Analysis: Using Anomaly Detection Method and Principal Component Analysis

Mengyang Qi
Purdue University

Graduate School ETD Form 9
(Revised 01/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Mengyang Qi

Entitled
nanoHUB Database Analysis: Using Anomaly Detection Method and Principal Component Analysis

For the degree of Master of Science in Industrial Engineering

Is approved by the final examining committee:

Omid Nohadani

Steven Landry

Gerhard Klimeck

To the best of my knowledge and as understood by the student in the *Thesis/Dissertation Agreement. Publication Delay, and Certification/Disclaimer (Graduate School Form 32)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Omid Nohadani

Approved by Major Professor(s): _____

_____

Approved by: Abhijit Deshmukh 05/01/2014

Head of the Department Graduate Program Date

nanoHUB USAGE ANALYSIS:USING ANOMALY DETECTION AND PRINCIPAL COMPONENT ANALYSIS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Mengyang Qi

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Industrial Engineering

May 2014

Purdue University

West Lafayette, Indiana

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Omid Nohadani for his generous help throughout my master study at Purdue. Although he left for Northwestern University, he always gave me timely advise and encouragement. He also taught me the strict attitude and responsibility, which will benefit my whole life.

I also want to thank Jocelyn Dunn for her help and discussion about my master thesis. She gave me precious advise and support in difficult times. I really want to thank her for giving me advise on how to revise my thesis. This thesis cannot be completed without her generous help.

I wish to thank Professor Gerhard Klimeck in Electrical and Computer Engineering Department. His valuable advise and unique insights help me throughout my master study.

Finally, I would like to thank Professor Omid Nohadani, Professor Gerhard Klimeck and Professor Steve Landry to be my examination committee.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Qi, Mengyang M.S.I.E., Purdue University, May 2014. nanoHUB Usage Analysis:Using Anomaly Detection and Principal Component Analysis. Major Professor: Omid Nohadani.

This thesis analyzes usage data from nanoHUB.org, which is a web-based infrastructure for e-collaboration among nanotechnology simulation community. Previous analysis of nanoHUB database showed he nanoHUB usage data follows an unknown, heavy-tailed distributions. This thesis extends the analysis and develops an automatic anomaly detection method based on piece-wise linear approximation. The anomaly here refers to collective user behaviors different from others. The result shows that the method can accurately detect the anomalies in the unknown, heavily detailed distribution. This thesis also applies anomaly detection method and principal component analysis to other databases in nanoHUB and successfully reveals differences between different categories.

1. INTRODUCTION

1.1 Overview of the problem

The digital era has arisen from increasingly efficient data storage and networking capabilities coupled with the widespread usage of online resources and mobile devices. IBM has shown that 90 percent of the data in the world are created in the past two years with 2.5 quintillion bytes of data producing everyday [1]. The term "big data" refers to a large and complex data set that is difficult to be acquired, stored, searched, shared and analyzed by traditional analysis tools [2]. The examples are login information of a web-based software, purchase transactions of a retail company, and traffic records in a metropolitan region.

To take advantage of these rich data sources, predictive methods were established by researchers to gain useful information. Taking transactional data in retail industry as an example, Holt [3] discussed using exponentially-weighted moving averages method to forecast seasonal sales. Fildes and Beard [4] compared several forecasting method for production and inventory-control and summarized "ideal" system for production and inventory-control forecasting. Ni [5] built a two-stage dynamic sales forecasting model for the fashion retail, which is combined with long-term and short-term predictions, using autoregression method and decision tree to fulfill their goals.

The predictive approach, however, has inherent assumptions. One of the assumptions is that this approach depends on the size of data. More data means more ability to predict and understand, and the more accurate estimate of reality. Also, most predictive methods are based on statistical models and regression, such as linear regression. However, it requires the data to follow or approximately follow a specific distribution, which increases the limitation of this approach. In real-world contexts, it is hard to guarantee the size and distribution of data, such as the transactional

data in off seasons. Off seasons is the time of the year during which the demand is lowest. If original dataset does not satisfy the assumptions, the performance of predictive approach will decreases, and the information extracted from the data will be misleading.

This study concentrates on data-driven approach instead of predictive approach. Our approach allows researchers to learn directly from data and does not have assumptions in advance. No assumptions means this method can be used with most datasets independent of distributions of data and size of data. To study the data more completely and comprehensively, this study also concentrates on revealing the patterns or anomalies in the data.

1.2 Overview of "e-collaboration" and nanoHUB

The term "e-collaboration" refers to the collaboration among online user who share resources, knowledge and information through a web-based software [6]. It allows researchers to find others with similar research interests and promotes discipline collaboration. There are many advantages of e-collaboration. For example, it can speed up the efficiency and productivity among researchers [7]. It not only provides a resource-sharing platform, but also a communication platform. The term "e-collaboration" is also used in business world. Ma [8] argued that the benefits of e-collaboration are also being harnessed as an asset for the business world, providing a low cost and easy access strategy for resource-sharing and cooperation .

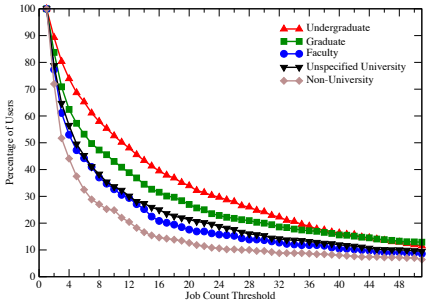
One such platform is nanoHUB, a web-based infrastructure for e-collaboration among the nanotechnology simulation community. This hub is maintained by the Network for Computational Nanotechnology (NCN) and supported by a state-of-the-art content management system [9]. It is the largest provider of nanotechnology simulation tools and educational materials, seving more than 56,000 users in 2007 [10] and 167,196 users in 2010 [11]. To date, the number of total users has rapidly increased to 309,146 which includes 28,292 registered users [12]. Through nanoHUB, users can

simulate models, download resources and interact with other users sharing the same interests. This study uses data in nanoHUB database, and analyzed user behavior in nanoHUB.

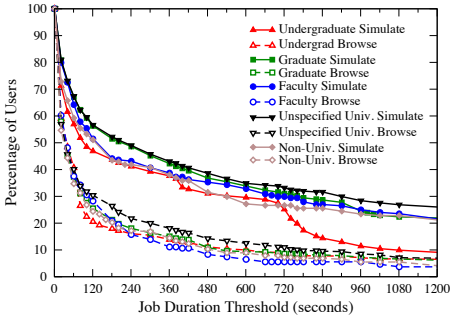
1.2.1 Previous Analysis of nanoHUB Database

Dunn et al. [13] analyzed nanoHUB usage data using threshold analysis and categorical boundary detection. Threshold Analysis is finding the number of users meet threshold by iterating over the range of threshold, and plot rate of change from largest to smallest [13]. Categorical boundary detection is selecting a critical threshold to separate data into groups. They analyzed usage data of nanoHUB.org for 2009-2010 academic year. Using these methods to analyze simulation job count, two data-driven categories were established: simulating and browsing users by simulation job count. Simulating users are defined as those users having five or more simulation jobs, while browsing users have less than five. The user profiles also provide information about professional and educational status: undergraduate, graduate, faculty, unspecified university and non-university. Figure 1.1(b) [13] shows the threshold analysis by simulation job count and simulation job duration count based on the categories.

The figure shows two steeply declines for undergraduate simulating usage data, which can be classified as anomalies. Their analysis concluded this information by examining Figure 1.1(b). This study starts based on these results and develops an anomaly detection method that can detect these anomalies only through the dataset, and extend the analysis to other databases. Chapter 2 is focused on automatic anomaly detection, which is a method to detect anomalies in a non-increasing curve to get the specific user group. Chapter 3 introduced analysis of data in nanoHUB database, using anomaly detection method developed in Chapter 2 and statistical methods.



(a) Threshold analysis by simulation job count



(b) Data-driven categories, defined by simulation job count, are compared over the metric simulation job duration and two steeply declines in undergraduate simulating usage data

Fig. 1.1. Previous Analysis of nanoHUB Database [13]

2. AUTOMATIC ANOMALY DETECTION METHOD

Anomaly detection is to detect anomalous behaviors or observations from a dataset. It has wide applications such as fraud detection for credit cards and insurance [14]. The anomalies in data convey critical information. For example, anomalies in loan application process could indicate loan fraud [15]. Anomalies in markets could indicates buying or selling opportunities [15]. Anomalies in MRI image could indicate malignant tumors [16]. In these examples, in order to find anomalies, the detection method should constantly monitor the systems. Hence the automatic anomaly detection method is extremely important. Section 1.2.1 shows Dunn et al. found that the nanoHUB usage data follows an unknown, heavy-tailed distributions [13], such as Figure 2.1. The starting point of this dataset is 1 second and the increment is 20 seconds. This automatic anomaly detection method introduces in this section is built on this study and can automatically detect anomalies in this distributions.

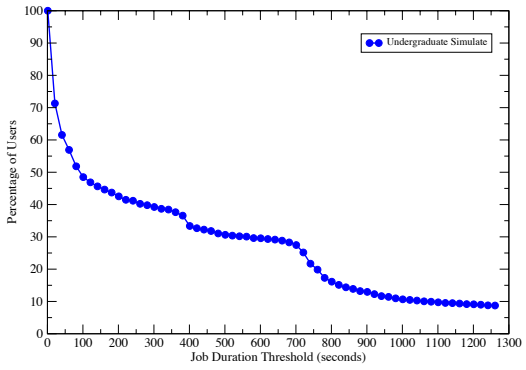


Fig. 2.1. Threshold Analysis of Percentage of Users of Undergraduate Simulating Users 2009-2010 vs. Job Duration. Two distinct declines occurred at job duration time 350-450s and 700-800s. These two groups of users were defined as two anomalies.

There are many regression based anomaly detection methods. These methods used outliers to define anomalies, and outliers refer to non-conforming observations. Rousseeuw and Leroy [17] argued robust regression was very useful to detect anomalies. They observed that the big breakdown point can capture large fraction of outliers as they tend to stay far away from robust fits. The breakdown point is an estimator which estimates the proportion of incorrect observations an estimator can bear before giving inaccurate result. This method is also well known as least trimmed squares estimation [17]. This method is designed to not heavily rely on normality assumptions about original data. However, if data are not normally distributed, then the performance declines. Galeano [18] developed projection pursuit methods to detect anomaly in a multivariate time series. The author argued outliers can be more powerful in some projection directions, and he proposed an iterative method to detect anomalies based on autoregression model. This method is useful if the dataset could be fitted as a straight line as it detected outliers based on this line. This requires the data set should follow approximate linear relationship. The distribution here is a heavily-tailed curve and can not be fit as one line. Moen [19] argued that non-parametric median methods and regression methods can be used to detect anomalies. The author argued outliers can be identified by comparing y-deviation from regression line. Again, this method required linear relationship of data set and cannot fit the distribution here. The Finite Difference method can be used to approximate first derivative of curve [20]. The outlier occurs when there is a sudden change of first derivative. This method can be very accurate, but it will detect every change of derivatives, some of which can not be regarded as anomalies.

To guarantee the performance of this method, three criteria are set to test this method:

- Criterion 1: The anomaly intervals it detects should be as accurate as possible,
- Criterion 2: The output should be the same regardless of the order of data, and

- Criterion 3: The output stays the same when adding or omitting few non-interrupting data.

The anomalies may convey important information, the more accurate the intervals are, the more accurate information can be acquired from the database. So it relies on the accuracy of the detected anomaly intervals. The order of data means the starting point of the method. For example, in Figure 2.1, if the first data point (1,100) is data1 and the last data point (1281, 8.6) is dataN, the direction of approaching can be from data1 to dataN or from dataN to data1. But no matter the order of approaching , it should detect the anomalies and give the same result. Futhermore, the result should not be too sensitive to few non-interrupting changes. The non-interrupting changes refer to changes that will not significantly alter the original curve.

2.1 Method

The main concept of this method is piece-wise linear approximation, which means using several line segments to approximate the curve. The algorithm starts from one end of the curve, estimates a line from first three data points, and continue adding data points until an outlier is found. The occurrence of outliers means the changing of line properties, and the current fitted line is not suitable for the following data points. Hence a new line will be constructed from this data to continue approximating the curve.

The least square method is used to estimated line properties. Karl Gauss proposed this method in 1801 [21]. Here x stands for the value of x-axis and y stands for value of y-axis. Each data record has these two values (x_i, y_i) . The least square method finds the best linear fit by minimizing sum of squares error between the data points and the fitted line. [22]. Here the least square method is used to pieces-wise approximate curve and to determine the line properties (slope and intersection). The Eq.(2.1) and Eq.(2.2) shows how to calculate the slope β_1 and the intersection β_0 using least square method.

$$\beta_1 = \frac{n\sum y_i x_i - \sum y_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2} \tag{2.1}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{2.2}$$

The method uses Sum Square Error (SSE) to decide the outlier and compares the value with the error rate, which is inputed by the user. SSE is used here as it is a direct representative of error of the fitted line. SSE is given by

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{2.3}$$

After obtaining all the fitted piece-wise lines, the algorithm compares SSE of consecutive lines, the anomaly can be found when SSE has a sudden change.

The steps can be summarized as following:

For each data point, also called current pivoting point:

- Step 1: Estimate line property (Slope and Intersection) and SSE,
- Step 2: Search to left: consider adding one more data point by checking the line property and SSE. If SSE exceeds threshold value, do not add this point to the fitted line,
- Step 3: Search to right: consider adding one more data point by checking the line property and SSE. If SSE exceeds threshold value, do not add this point to fitted line, and
- Step 4: Repeat Step 2 and Step 3 until the maximum length fitted line is reached. Record left boundary, right boundary and SSE.

After obtaining all intervals for each data point,

- Arrange each interval from highest SSE to lowest SSE and delete intervals that have overlapped.
- Compare SSE of consecutive lines. If greater than SSE threshold, consider as anomaly.

- Connect to nanoHUB database, obtain information of the anomaly users.

This method allows user to set a specific SSE threshold. It also allows user to set a starting point of SSE and the increment of SSE. The algorithm will automatically output all the corresponding intervals until no anomalies are found.

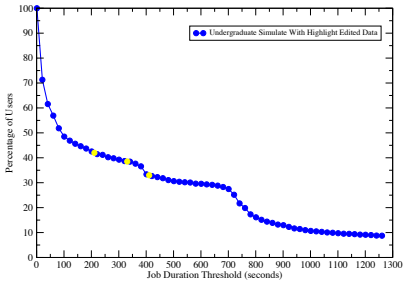
2.2 Result

After applying the method to Figure 2.1, the result is shown below. The first number stands for the current pivoting point, the interval stands for the approximate line and the last number stands for SSE value. Based on Figure 2.1, me obtain:

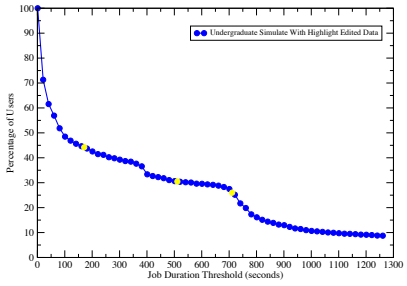
21: [21, 61]: 4.30733 381: [381, 421]: 1.07683 721: [721, 761]: 0.433035

The interval [21, 61] occurs because the rapid decay in the beginning of the data set. This study does not regard it as anomaly.

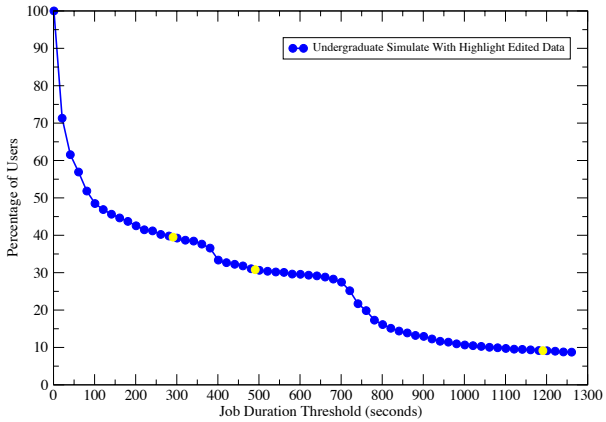
To check whether this method satisfied the criterion 2 mentioned, three "testing data" are randomly added to the original data set to form three scenarios.



(a) Scenario 1: Adding Points 211,331,411



(b) Scenario 2: Adding Points 171,511,711



(c) Scenario 3: Adding Points 291,491,1191

Fig. 2.2. Three Testing Scenarios

Table 2.1
Result Summary of Figure 2.2(a) to Figure 2.2(c)

Figure	Result
Figure 2.2(a)	21: [21, 61]: 4.30733 361: [361, 401]: 0.784724 721: [721, 761]: 0.433035
Figure 2.2(b)	21: [21, 61]: 4.30733 381: [381, 421]: 1.07683 721: [721, 761]: 0.433035
Figure 2.2(c)	21: [21, 61]: 4.30733 381: [381, 421]: 1.07683 721: [721, 761]: 0.433035

Table 2.1 shows that scenario 2 and scenario 3 have the same results as the original dataset. The second interval in scenario 1, however, is slightly different. In original dataset, the algorithm " regards" the data point 361 as part of line segment before [381, 401]. But the testing point 331 smoothes that line segment, which makes the algorithm "regards" the data point 361 as an outlier, and a new line start from data point 361. After comparing all line segments, the algorithm outputs [361, 401] as an anomaly. Since the intervals are almost the same, it is safe to conclude the anomaly intervals that this method detects are quite accurate and satisfied the criteria.

The tests show that this method is useful to detect anomalies in the heavy-tailed, decaying distributions. Since this method does not have any assumptions of original data sets, it should be able to applied to any other ordered dataset. But it needs further validations. In order to explore more about the anomalies, it is necessary to apply the anomaly information to other databases of nanoHUB. The next chapter introduces applying the anomaly detection method and anomaly information to another database of nanoHUB.

3. *ANDMORE* DATABASE ANALYSIS

The *andmore* database has records of downloadable content from nanoHUB. This database starts in 2010 and has not been fully studied. By studying the document that users downloaded, it may reveal more connections between users or files. Furthermore, it may reveal more specific user patterns. All these will allow people to come up with suggestions as to how nanoHUB can be improved, built or even restructured for the future in order to generate more impact. Analysis of nanoHUB *andmore* database builds on previous analysis done by Dunn et al. [13], which is introduced in section 1.2.1. This work builds on previous results and to further dissect these categories, anomaly detection is employed to establish when a subset of users within a category exhibit a collective behavior that departs from the overall trend of the category. To further explore anomalies in these analyses, such as the undergraduate usage decline over job duration at 720 seconds (as seen in Figure 1.1(b)), the interactions between users and resources were analyzed for the ranges of a metric where this anomaly is present. For a given range of job durations, if the usage declines and a subset of tools can account for the usage decline, then it is determined that this constitutes a group anomaly. The same distinction is made between browsing and simulating users, based on the number of simulation jobs recorded in the job log tables.

3.1 Data Filters

The *andmore* database has records of downloadable content from nanoHUB. Although nanoHUB requires user account to run simulation tools, it is not necessary to have an account to download resources. Also in order to maintain normal function of nanoHUB, employees of nanoHUB can download from nanoHUB, and all their login information are recorded in an exclude list of nanoHUB database. Hence this thesis

only focuses on download records of registered users between July 1, 2010 and June 21, 2011 and not in the exclude list.

3.2 Resource Types

The downloadable content from nanoHUB.org include the file types shown in Table 3.2. For the analysis presented here, file types of interest are broadly categorized as Document or Multimedia file types. Document files are printable materials. Multimedia files are downloadable videos, online videos, and interactive Java applications. For now, other files are excluded from analysis, these include downloadable images, codes, softwares, and packages. For this dataset, there was a total of 2280 Document files and 3048 Multimedia files that were accessed for a total of 155,600 downloads during this one-year time period.

Table 3.1
Downloadable Content

Document	.pdf, .doc(.docx), .txt, .ppt(.pptx), .xls(.xlsx)
Image	.gif, .jpg, .JPG, .png
Video	.mp4, .wmv, .wav, .m4v, .mp3, .mov, .asx, .asf, .avi
Online play	.swf, play (watching online)
Java Apps	.jar (watching online), .jnlp (download and run on own computer)
Code	.xml, .tcl, .m, .f, .c
Software	.exe
Package	.zip, .rar, .tgz, .gz

3.3 Result

3.3.1 Threshold Analysis over Download Count

Threshold Analysis over Download Count For All User Category

Downloadable content data were separated by file type: Print or Multimedia (see definitions in Section 3.2). If the same user downloads the same material consecutively, then it is considered as only one download to eliminate double counts due to probable download errors. For each user, it was determined how many times they downloaded from each file type, and then threshold analysis was performed over this metric, download count, for each user category (threshold analysis is introduced in section 1.2.1). Figure 3.1 shows threshold analysis over Download Count for Print and Multimedia file types for each simulating user category with scaled y-axis, percentage of simulating users.

No matter for which pre-defined category, users prefer print files, even there are less print files in nanoHUB website. Figure 3.1 has some misleading anomalies. For example, after applying the anomaly detection method introduced in Chapter2, a anomaly is found between 7 to 10 in non-university users, print files curve, but there are only 14 users in this range. Because of the small size, no useful information can be found.

Figure 3.2 shows the same information of each browsing user category with percentage of users. In Figure 3.2, all curves overlaps each other, which means although the downloads numbers are different, every category has the same downloads characteristic. Graduate users have the highest downloads but with higher sample size. Browsing faculty, non-university users and unspecified university users downloads more than simulating ones, but undergraduate users almost stays the same regardless of browsing and simulating status. The downloads difference of print files and multimedia files is not as clear as simulating users.

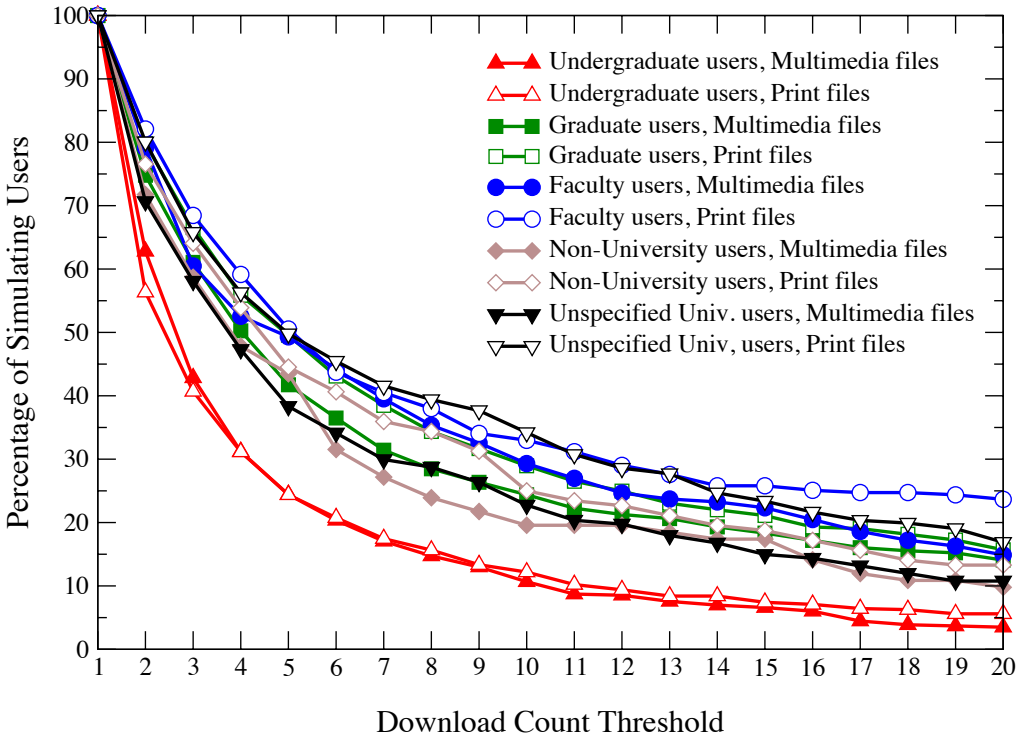


Fig. 3.1. Threshold analysis of Percentage of Simulating Users vs. Download Count. The number of simulating users downloading print files from *andmore* database was 607 for undergraduate students, 777 for graduate students, 279 for faculty, 231 for unspecific university and 128 for non-university users. The number of simulating users downloading multimedia files from *andmore* database was 516 for undergraduate students, 611 for graduate students, 215 for faculty, 167 for unspecific university and 92 for non-university users.

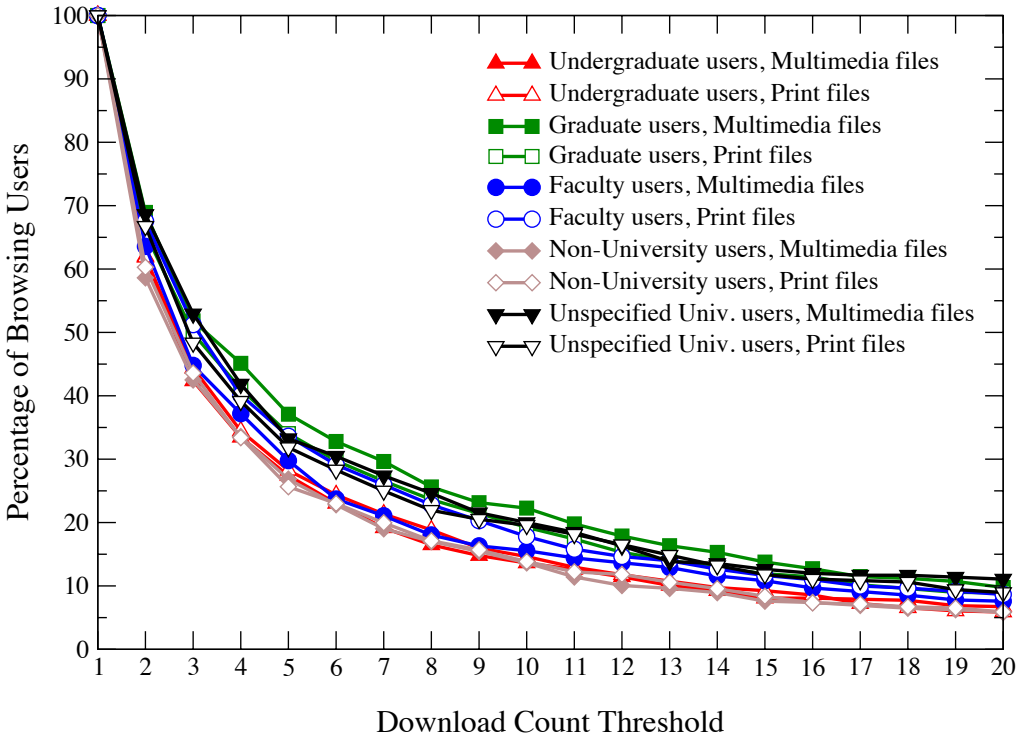


Fig. 3.2. Threshold analysis of Percentage of Browsing Users vs. Download Count. The number of simulating users downloading print files from *andmore* database was 595 for undergraduate students, 1190 for graduate students, 601 for faculty, 424 for unspecific university and 491 for non-university users. The number of simulating users downloading multimedia files from *andmore* database was 595 for undergraduate students, 1046 for graduate students, 527 for faculty, 325 for unspecific university and 447 for non-university users.

3.3.2 Threshold Analysis of Undergraduate Classroom and Non-Classroom Users

New Data-Driven Categories: Classroom and Non-Classroom Users

New Data-Driven Categories of classroom and non-classroom users are found in a study of the similarity between each user and found anomaly based user-user similarity [23]. Mike et al. [24] used a similar concept as Levenshtein edit distance and calculated overall edit distance for each user by taking a set of transformations of their longitudinal records. The longitudinal records of two users appeared identical after the set of transformations were taken, and the penalties occurred for each transformation sum to an overall edit distance [24]. Most of classroom users are undergraduate users and all anomaly information is stored in nanoHUB database. This section is exploring the difference of downloading behavior of classroom users and non-classroom users.

Section 3.3.1 shows user centered analysis of downloadable content data. To study the download behavior of each user category on each file types, resources centered analysis was conducted. All data were separated into three mutually exclusive sets:

- Print Only: Number of print files when only print files were downloaded,
- Multimedia Only: Number of multimedia files when only multimedia files were downloaded, and
- Both: Max number between number of print files and number of multimedia files when both file types were downloaded.

Same as in previous sections, threshold analysis of download count for different download file types are studied of the new data-driven categories. Figure 3.3 shows the threshold analysis of classroom and non classroom users. In Figure 3.3, most users downloaded both print and multimedia files, and non–classroom users downloaded more than classroom users. It is because of there are more non-classroom users

than classroom users. Also, the classroom users can get additional materials and instructions from professors while non-classroom users has to learn by themselves.

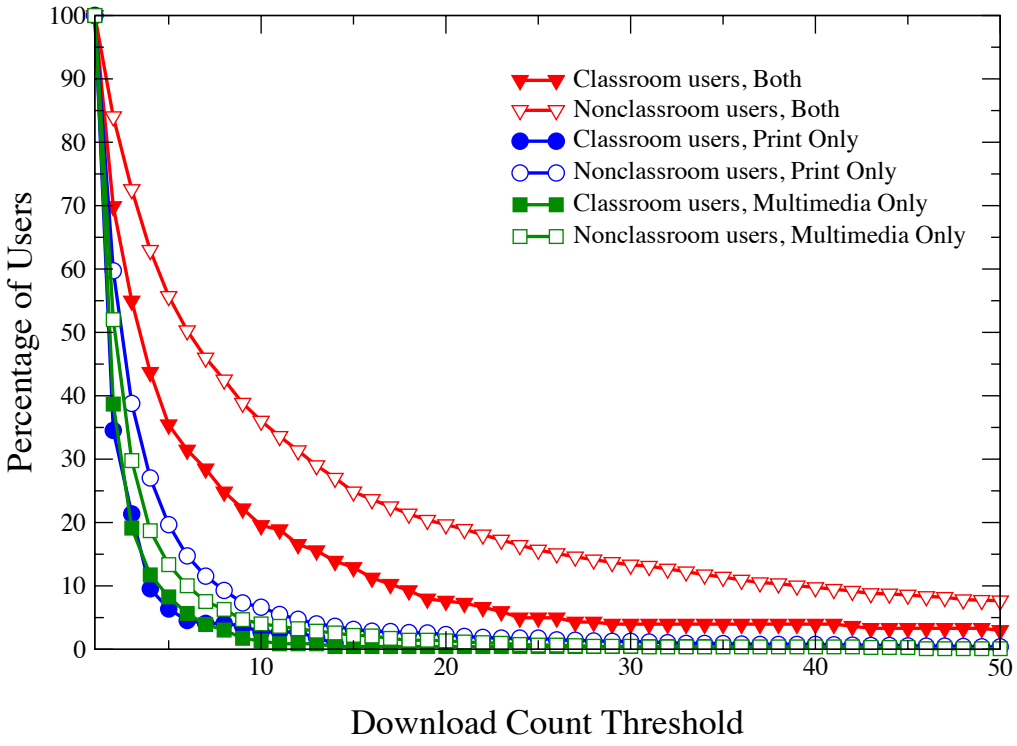


Fig. 3.3. Threshold analysis of Percentage of Users vs. Download Count. The number of users downloading only print files from *andmore* database was 230 for classroom users and 1820 for nonclassroom users. The number of users downloading only multimedia files from *andmore* database was 220 for classroom users and 1048 for non classroom users. The number of users downloading both files from *andmore* database was 302 for classroom users and 2973 for non classroom users.

3.3.3 Threshold Analysis of Group Users

As in the previous section, the anomalies of undergraduate 2009-2010 are detected. The study now concentrates on the group user behavior of the users in the anomaly.

Group Users in 2009-2010: Major Tools and Major Downloads from *And-more* Database

Applying the anomaly detection method in previous section to the Job Duration Threshold curve, two groups stand out: GroupA and GroupB Users. The figure is shown in Figure 3.4. To fully understand why these two groups stands out and are there any differences between the users in group and other users, the following studies are focused on group users.

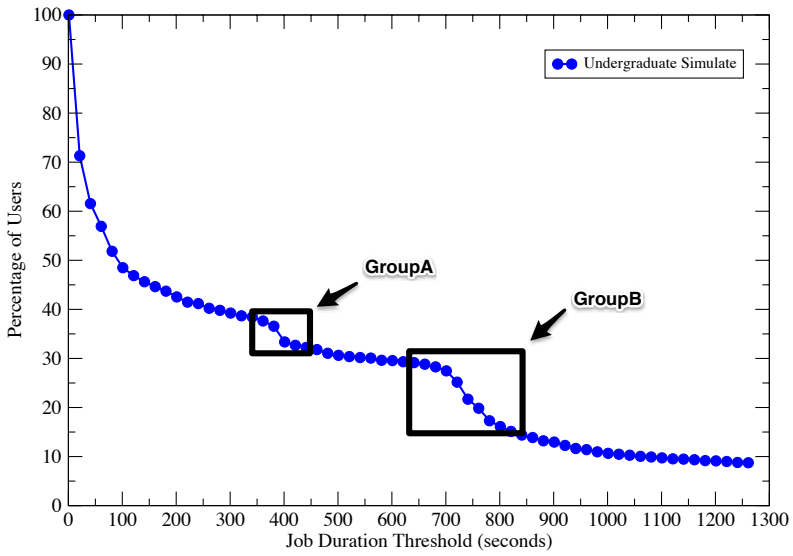


Fig. 3.4. Data from nanoHUB Undergraduate Simulating Users Usage for 2009-2010 academic year.

The major tools GroupA user used are crystal viewer and mosfet while the majority of GroupB users used cndo and mosfet [13]. It turns out these tools are classroom tools. But as the *andmore* database started in 2010, only two downloads were found for GroupA and GroupB users. It is not insightful to study only two downloads, so this study continues the method to all simulating users in 2010-2011.

Group Users in 2010-2011: Major Tools and Major Downloads from *Andmore* Database

This GroupA and GroupB Users in 2010-2011 is shown in Figure 3.5. After applying the anomaly detection method, two groups stand out. The GroupA users are around 3 to 5 second while the GroupB users are around 10 to 15 seconds.

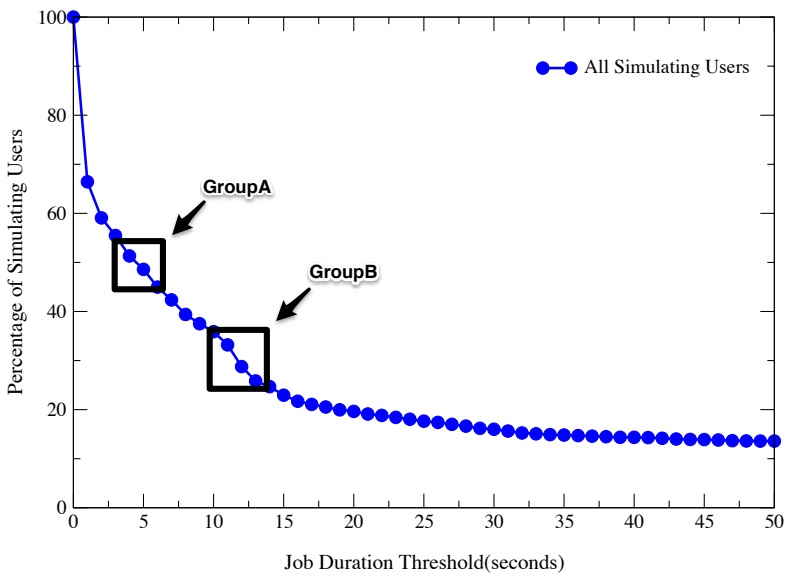


Fig. 3.5. Data from nanoHUB Undergraduate Simulating Users Usage for 2010-2011 academic year.

The major tools GroupA user used are cndo, nsoptics and qclab, while the majority of GroupB users used bandstrlab, cndo and cntbands-ext. Similarly, these two groups users are also classroom users. There are 495 GroupA users and 229 GroupB users. The detailed summary is shown in Table 3.2.

Table 3.2
Result Summary of Group Users Behavior in *Andmore* Database

Group Type	Number of Users	Number of Downloading Users	Number of Downloads	Number of Distinct Download
GroupA	483	130	918	398
GroupB	229	107	1926	1185

Although there are more users in GroupA, only 1/5 of the users actually downloaded files from *andmore* database and about 3/5 of the downloads are duplicates. For GroupB user there are only 299 users, but 1/2 of the users downloaded files from *andmore* database and most of the downloads are distinct files.

Table 3.3 shows the top three downloads of Group users and Table 3.4 shows the top three distinct downloads of Group users.

In the two tables, "springer lundstrom.jpg" is a image for a book named Nanoscale Transistors by Mark Lundstrom and Jing Guo. The file named "refresh.gif" no longer exists. It seems that GroupA users mainly download Cndo Supporting Document while GroupB user mainly download BioSensor Lanb User Manual. Although "springer lundstrom.jpg" had downloaded 57 times for GroupA and 41 times for Group B, there were only 8 users and 4 users actually downloaded, which means most of the downloads were replicates. The top three distinct downloads also indicate the major tools used by each group. The top two distinct downloads for GroupA users are files for tool cndo, and the third distinct download is file for tool qclab.

Table 3.3
Top Three Downloads of Group Users

Group Type	File Name	Number of Downloads	Number of Users
GroupA	Springer lundstrom.jpg	57	8
	Cndo Supporting Docs	55	42
	Quantum Dot Lab Learning Module	20	8
GroupB	Springer lundstrom.jpg	46	4
	BioSensorLab User Manual	37	9
	refresh.gif	34	8

Table 3.4
Top Three Distinct Downloads of Group Users

Group Type	File Name	Number of Users
GroupA	Cndo Supporting Docs	42
	Theoretical Analysis of Gold nanoparticles.pdf	11
	Quantum Dot Lab Learning Module	8
GroupB	BioSensorLab User Manual	19
	CNTbands Supporting Docs	9
	Introduction to CNTbands	9

The top one download for GroupB users is file for tool bandstrlab, and the other two downloads are files for tool cntbands.

Threshold Analysis of Group User

Threshold Analysis also performed on Group Users. Figure 3.6 shows the results of Downloads Count vs. Percentage of Group Users. It shows that GroupB users downloaded more documents from *andmore* database than GroupA users. Although these two Group Users are all classroom users, the users in GroupB had longer job durations and had more download than GroupA users.

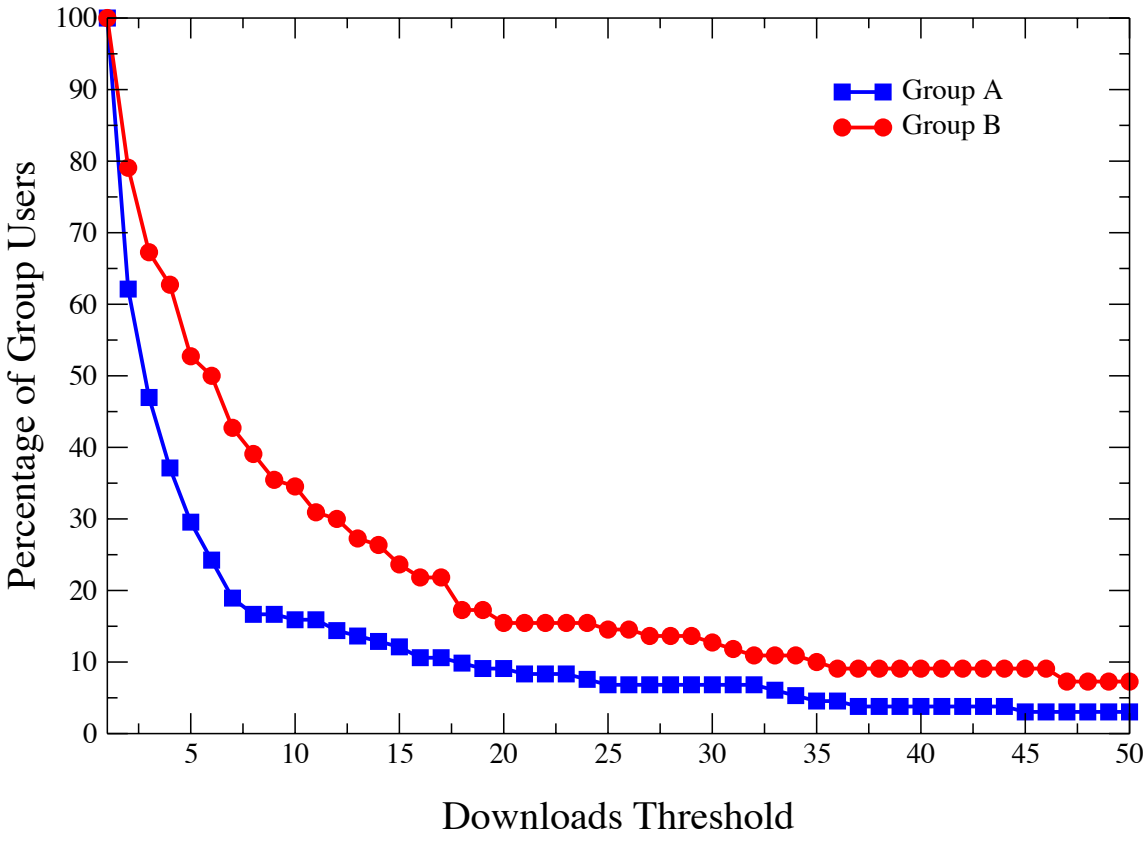


Fig. 3.6. Threshold Analysis of Percentage of Group Users vs. Download Count.

3.3.4 Analysis of Variance(ANOVA) of Andmore Database

ANOVA is a basic statistical analysis method of dataset. Here this method is used to give a brief understanding of andmore database. Each user has several factors, which are

- Total number of downloads,
- Number of download in each month,
- Organization type, and
- Data-driven category

Total number of downloads stands for how many files did the user download in *andmore* database during 2010-2011 academic year. The time period that being studied is from July 2010 to June 2011, so number of download in each month is counting how many files did the user download in each month. Organization type is whether the user is undergraduate, graduate, faculty, unspecified university or non-university user. Data-driven category is whether the user is simulating or browsing users.

The ANOVA Analysis result is shown in the table below. Table 3.5 shows that the data-driven category differs in total number of downloads, number of download in July 2010, November 2010, February 2011 and April 2011. Organization type differs in total number of downloads, number of download in September 2010, November 2010, December 2010, February 2011 and March 2011. It also shows that Organization type is a better category separation in *andmore* database than data-driven category as it reveals more different between users. This result may not be accurate as the data here does not follow approximately distribution. It is just a first insight to analysis *andmore* database.

Table 3.5
Result Summary of ANOVA Analysis Andmore Database

Factor	Source	P Value	Conclusion
Total number of downloads	Data-driven category	0.002	Different
	Organization type	0.000	Different
Number of downloads in July 2010	Data-driven category	0.017	Different
	Organization type	0.256	No Different
Number of downloads in August 2010	Data-driven category	0.127	No Different
	Organization type	0.100	No Different
Number of downloads in September 2010	Data-driven category	0.133	No Different
	Organization type	0.042	Different
Number of downloads in October 2010	Data-driven category	0.375	No Different
	Organization type	0.415	No Different
Number of downloads in November 2010	Data-driven category	0.030	Different
	Organization type	0.001	Different
Number of downloads in December 2010	Data-driven category	0.301	No Different
	Organization type	0.001	Different
Number of downloads in January 2011	Data-driven category	0.486	No Different
	Organization type	0.326	No Different
Number of downloads in February 2011	Data-driven category	0.023	Different
	Organization type	0.041	Different
Number of downloads in March 2011	Data-driven category	0.072	No Different
	Organization type	0.047	Different
Number of downloads in April 2011	Data-driven category	0.005	Different
	Organization type	0.740	No Different
Number of downloads in May 2011	Data-driven category	0.143	No Different
	Organization type	0.167	No Different
Number of downloads in June 2011	Data-driven category	0.56	No Different
	Organization type	0.155	No Different

3.3.5 Principle Component Analysis of *Andmore* Database

Principle Component Analysis (PCA) is a powerful method to analyze data described by several variables, trying to extract important information from data and reduce the dimension of data set by finding relationship between variables [25]. In *andmore* database, each user may download several different files and this method can be used to analyze the relationship between files and the relationship between download time. Minitab is used in this analysis and the method can be separated in several steps [26],

- Step 1: Subtract the mean,
- Step 2: Calculate correlation matrix,
- Step 3: Calculate eigenvectors and eigenvalues of the correlation matrix,
- Step 4: Choose how many principal components needed,
- Step 5: Derive the new data set.

Suppose the data set is

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{bmatrix}$$

The first step is to subtract the mean from each of the data dimension and produce a data set whose mean is zero. In the data, each column represent one dimension or one variables. So after first step the matrix becomes

$$X' = \begin{bmatrix} x'_{1,1} = x_{1,1} - \overline{x_{,1}} & x'_{1,2} = x_{1,2} - \overline{x_{,2}} & \dots & x'_{1,n} = x_{1,n} - \overline{x_{,n}} \\ x'_{2,1} = x_{2,1} - \overline{x_{,1}} & x'_{2,2} = x_{2,2} - \overline{x_{,2}} & \dots & x'_{2,n} = x_{2,n} - \overline{x_{,n}} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x'_{m,1} = x_{m,1} - \overline{x_{,1}} & x'_{m,2} = x_{m,2} - \overline{x_{,2}} & \dots & x'_{m,n} = x_{m,n} - \overline{x_{,n}} \end{bmatrix}$$

The second step is calculate the correlation matrix. The equation to calculate correlation matrix is shown below.

$$r_{i,j} = \frac{\sum_{a=1}^{m}(x'_{i,a})(x'_{j,a})}{\sqrt{\sum_{a=1}^{m}(x'_{i,a})^2 \sum_{a=1}^{m}(x'_{j,a})^2}}$$

As the data set is $m \times n$ matrix, the correlation matrix should be a $n \times n$ matrix. Minitab will not show the results of first two steps as they are just simple calculations.

The third step is to calculate the eigenvalues and eigenvectors of the correlation matrix. The equation is given by

$$Corr(X)v - \lambda v = 0,$$

where v is eigenvectors and λ is the corresponding eigenvalues. In PCA, the eigenvectors are called principal components (PCs). As the data set is a $m \times n$ matrix, PC should be a $n \times n$ matrix and PC1 represents the first column of this matrix. Also it should produce n eigenvalues. In Minitab, this step is called eigenanalysis and it will produce a loading plot, which plots the n variables based on the first two PCs using PC1 as x axis and PC2 as y axis. This plot can show the relationship between variables. If some variables are close in this plot, it shows these variables are related.

The forth step is to choose how many principal components needed. It is based on how much variance does each principal component explain and this can be done by examining the eigenvalues [27]. Suppose the eigenvalues of one data set are

$$\lambda_1, \lambda_2, .., \lambda_n,$$

the proportion of variation explained by each principal components are

$$P_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_n},$$

where P_i is the proportion of variation explained by *ith* principal components and i is from 1 to n. The cumulative proportion of variation explained by the first m principal components can be represented as

$$P_{cumulative} = \sum_{i=1}^{m} P_i,$$

If adding from 1 to n, the cumulative proportion is 1. Also in PCA, it named the eigenvector that explained the most amount of variation as PC1 and continuines naming the eigenvectors as the amount of variation explained decreasing. The basic idea to determine how many PCs to use is minimizing the number of PCs while maximizing the amount of cumulative proportion of variation they explained.

The fifth step is to derive the new data set based on the PCs, which is called principal component scores in PCA. The scores are just another representative way of original data, which means the score matrix should be a $m \times n$ matrix. Each column of scores matrix is calculated as following,

$$ith\ column\ of\ scores\ matrix = PCi \times [x_{i,1}, x_{i,2}, ...x_{i,n}] + [\overline{x_{,1}}, \overline{x_{,2}}, ..., \overline{x_{,n}}]$$

To study the relationship between the download month and the relationship between downloadable files, PCA method is performed among monthly download of all users and most of the downloadable files. Figure 3.7 is the loading plot of PCA method for monthly download. The plot shows that first principle component is strongly correlated with number of downloads in April, May and June, which means the PC1 increases when number of downloads increases in these three months. It suggests that if a user downloaded files in April, it is likely that he/she also downloaded files in May and June. The second principal component is correlated with number of download in August and September, which means the PC2 increases when number of download decreases in August and September. Although the value of second component is negative, it suggests that downloads in August and September are related.

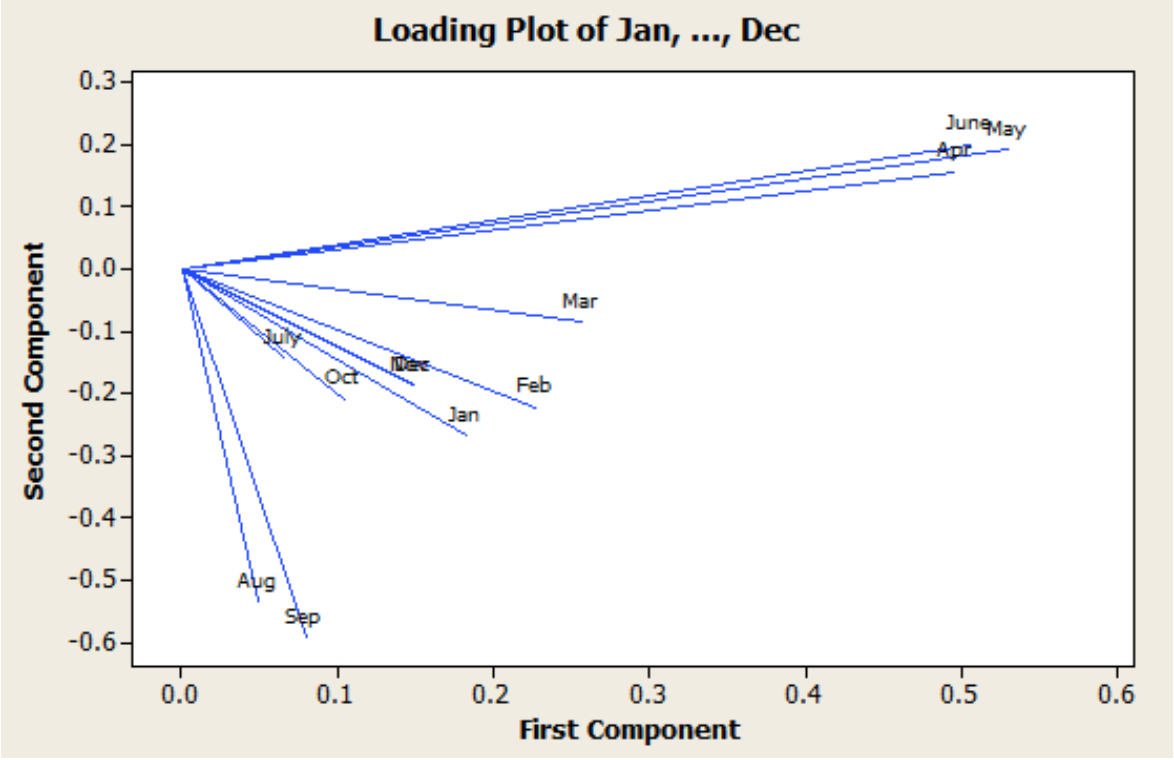


Fig. 3.7. Loading plot of principle component analysis of monthly downloads.

The *andmore* database has mass downloadable files, according to Secion 3.3.1, the number of downloads for simulating undergraduate student is 1123 and some of the files are only download less than 5 times. Therefore, there is no need to study all the files. So to minimize the size of matrix, this study only focuses on the files have more than 40 downloads. Also, to study the difference between data-driven category and professional status, this method is performed on each categories separately.

Figure 3.8 shows the loading plot of the analysis for simulating users. To make the plot more readable, numbers are used instead of file names. Each number represents one file and there are 105 distinct file downloads having more than 40 times for simulating users.

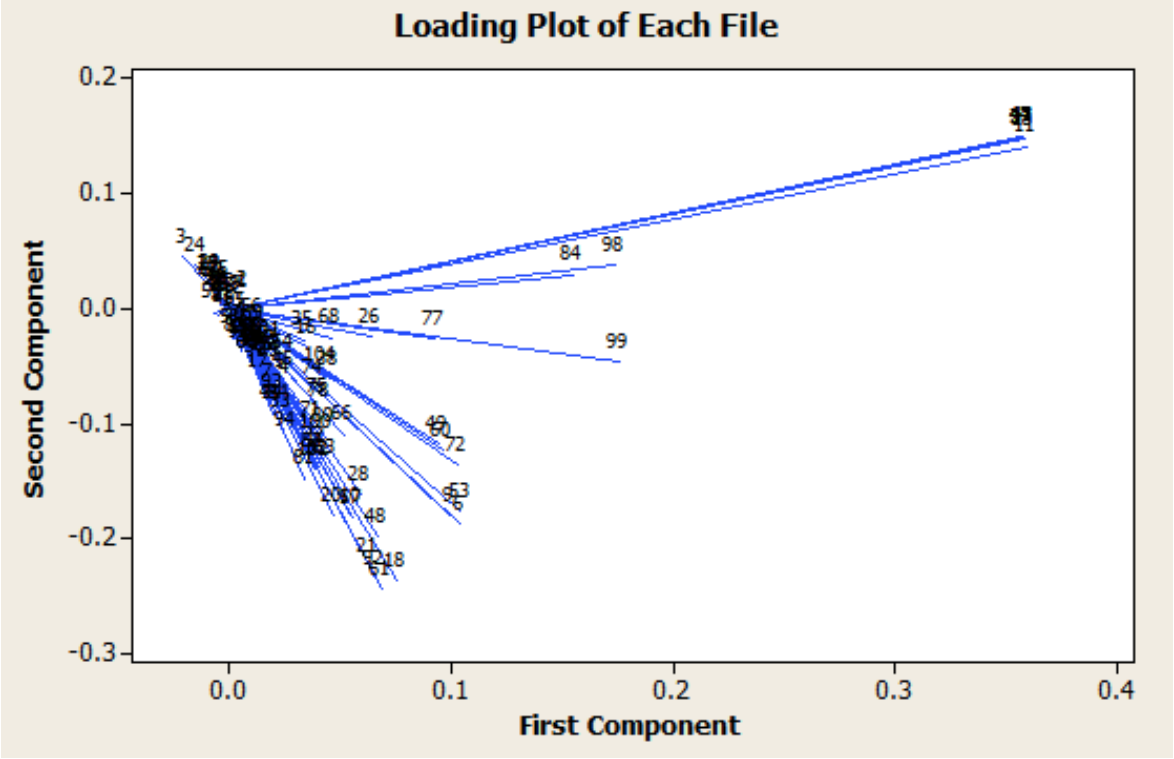


Fig. 3.8. Loading plot of principle component analysis of number of downloads of each distinct file among all simulating users with more than 40 total download, which has 105 distinct files in total.

This plot shows that several files are related. For example, numbers on top right are file 1, 2, 11, 23, 27, 43, 44, 84 and 98. These files are related as they are all lecture notes of ECE495. File 6, 9 and 53 are related as file 6 is a MATLAB Scripts for "Quantum Transport: Atom to Transistor" and file 53 is a thesis article discussed quantum transport problem (file 9 can not be traced back).

Figure 3.9 shows the result among browsing users. There are 174 distinct files downloads satisfy the criterion. Numbers on bottom right are file 13, 27, 33, 44, 48, 58, 65, 73, 76, 79, 80, 81, 87, 91, 92, 101. They are related as they are all lecture notes of ECE606. File 69, 74, 85, 95, 46, 59, 97, 53, 50, 16 on the top right are also ECE606 lecture notes. The reason why they separate is the former ones are all pdf files and the latter ones are all mp4 files.

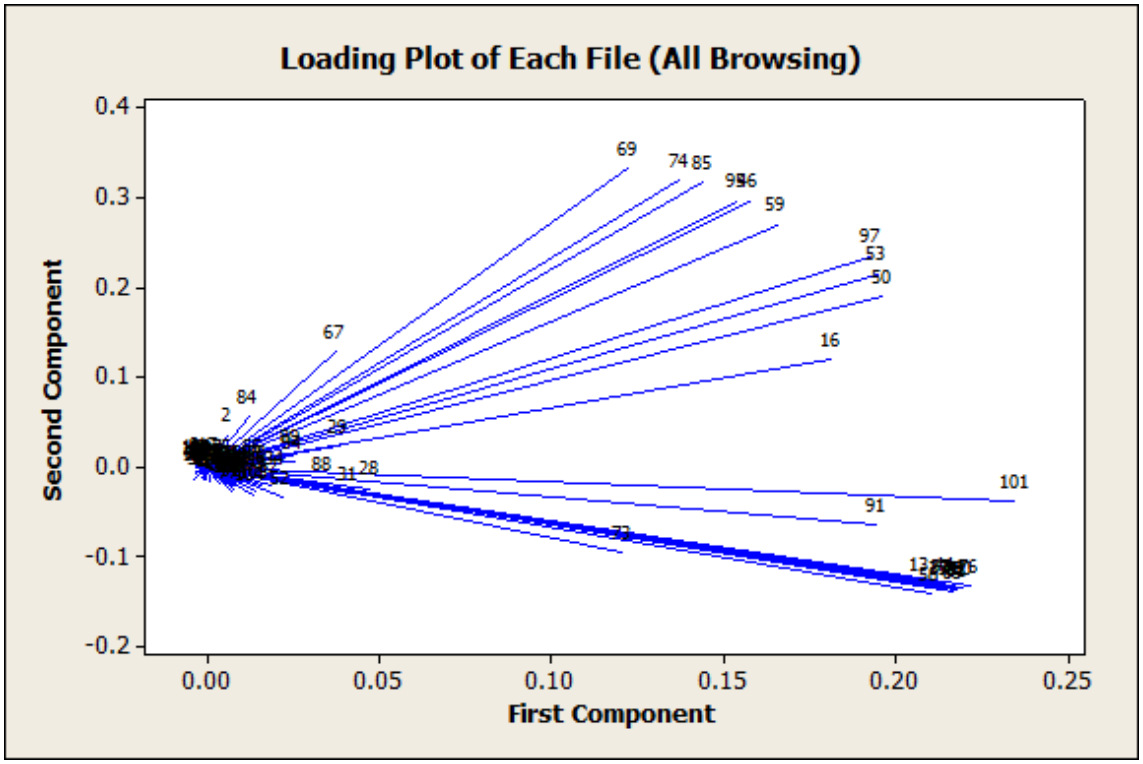


Fig. 3.9. Loading plot of principle component analysis of number of downloads of each distinct file among all browsing users with more than 40 total download, which has 174 distinct files in total.

Figure 3.10 shows the result among group users. If cooperated with Table 3.4 in Section 3.3.3, for GroupA user, the top one download is "Cndo Supporting Docs" and the corresponding number in the figure is 24. The top two download is "Theoretical Analysis of Gold nanoparticles.pdf" and the corresponding number is 79.

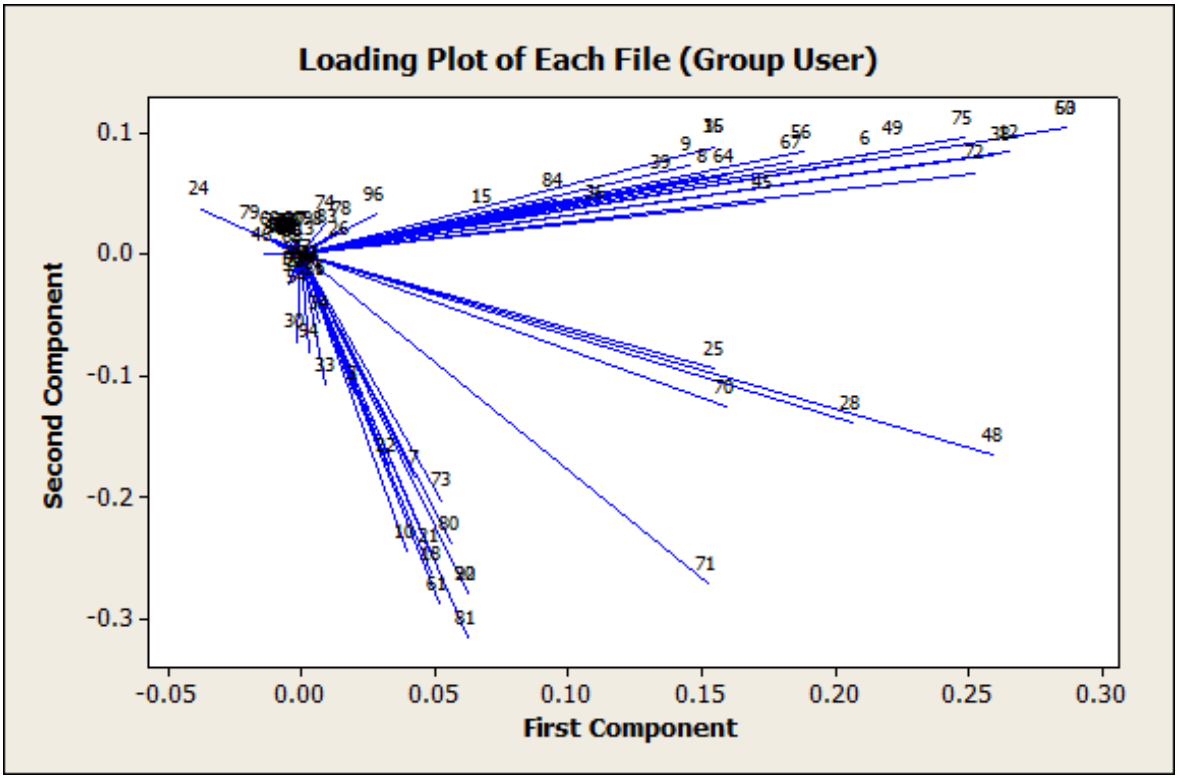


Fig. 3.10. Loading plot of principle component analysis of number of downloads of each distinct file among all group users.

The figure shows these two files are related, which makes sense as the author of "Theoretical Analysis of Gold nanoparticles.pdf" is the author of "Cndo Supporting Docs" and these two articles both address metal nanoparticles.

For GroupB user, the top two download is "CNTbands Supporting Docs" and the corresponding number in the figure is 45. The top three download is "Introduction to CNTbands(Video)" and the corresponding number is 64. The figure shows these two files are related as they all talk about CNTbands. The figure also shows that several files related with them, such as file 49, 72 and 75. File 49 is thesis article talking about carbon nanotube electronics written by Jing Guo, who is one of the author of CNTbands. File 72 is an article also talking about carbon nanotube transistors written by Jing Guo, Supriyo Datta, and Mark Lundstrom. Among the three people, Jing Guo and Mark Lundstrom are author of CNTbands. File 75 is a ppt named "Introduction to Carbon Nanotube Electronics" talking about CNTbands tools.

Figure 3.11 shows the PCA result for all classroom users in Fall 2010(see section 3.3.2). There are 91 different classes with 1555 classroom users in Fall 2010. As cluster user has less number of download, this study focuses on the files have more than 10 downloads. There are 50 distinct files satisfy this criterion. The figure shows several files are related. For example, numbers on the left are file 12, 14, 28, 29, 38, 45 and 47. File 14 is a teaching material for tool PN Junction Lab and file 47 is a online video for PN Junction Lab Demonstration. The PN Junction Lab is powered by PADRE, which is a 2D/3D simulator for electronic devices. So file 12 is a tutorial for PADRE and file 29 is an introduction to PADRE simulator. File 28 and file 38 are a lecture about semiconductor device simulation prepared by Professor Mark Lundstrom in Purdue University. In this lecture, he used PADRE to simulate examples. (file 45 can not be traced back).

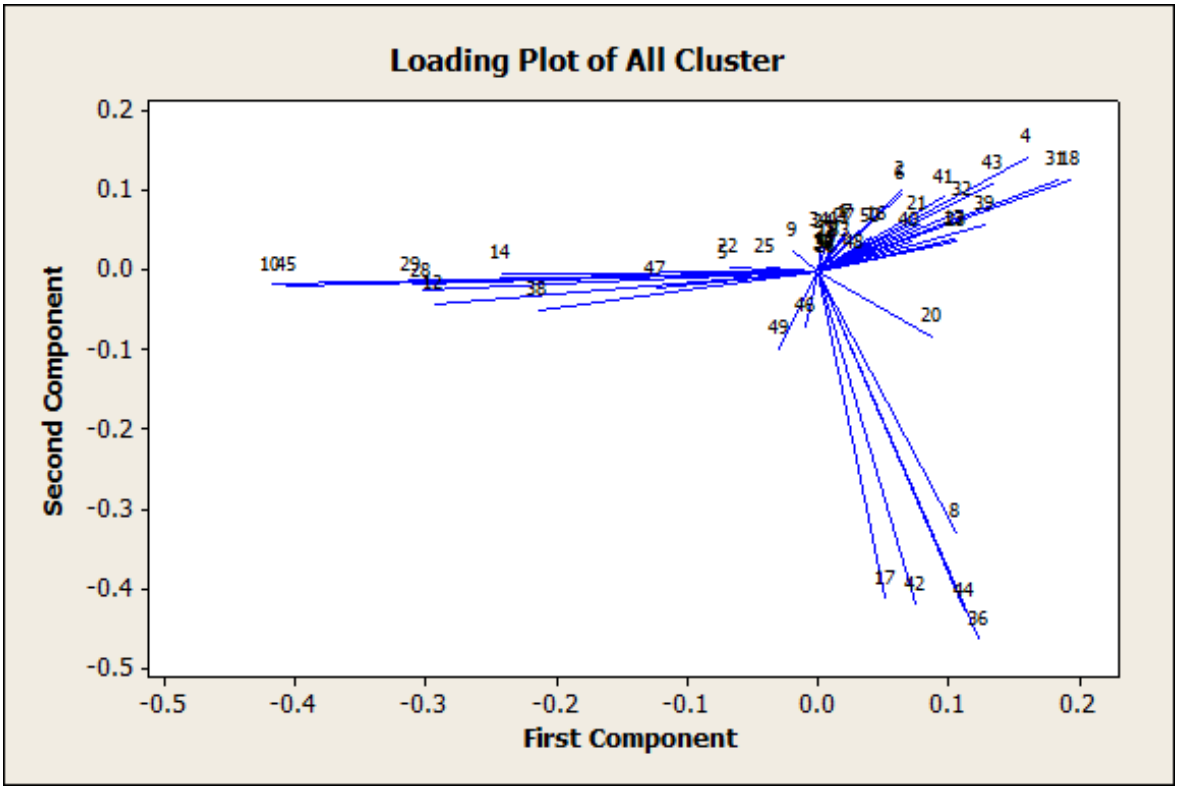


Fig. 3.11. Loading plot of principle component analysis of number of downloads of each distinct file among all classroom users in Fall 2010.

To further study the differences between different classes in different universities, this study narrows down to classes in three areas, namely Evanston, Illinois(Northwestern University), Urbana, Illinois(University of Illinois at Urbana-Champaign) and West Lafayette, Indiana(Purdue University), as these three universities are parts of NCN group. In Fall 2010, there are 28 distinct files having more than 10 downloads and 216 classroom users, 31 from 5 classes in Evanston, 56 from 12 classes in West Lafayette and 129 from 7 classes in Urbana, downloaded files from this database. Figure 3.12 is the loading plot of PCA method for three biggest classes. The plot shows that first principle component is negative correlated with file 5, 10, 12, 15, 22 and 26, which are all about graph bandsructure. The second principal component is negative correlated with file 3, 16 and 21, which are all introduction about Quantum Dot Lab (qdot).

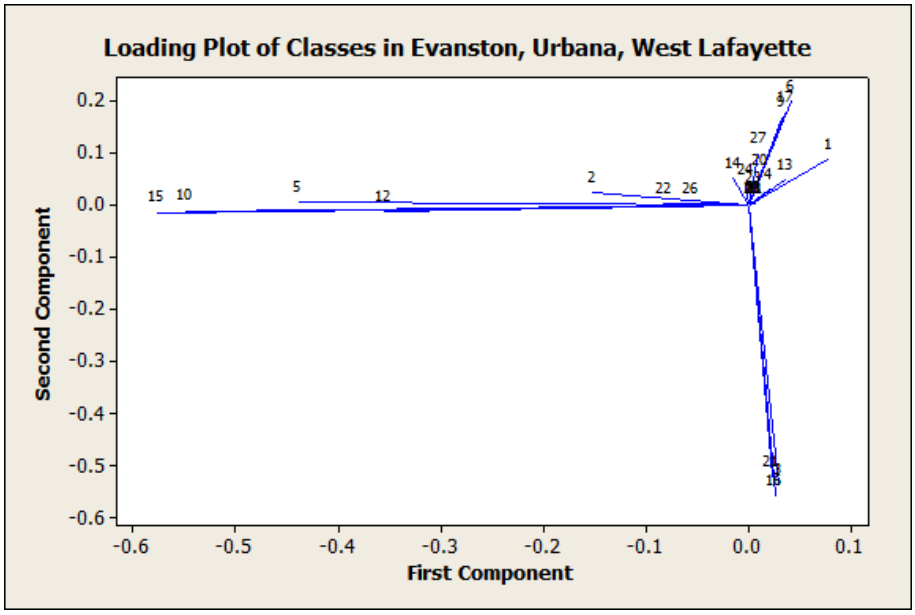


Fig. 3.12. Loading plot of principle component analysis of number of downloads of each distinct file of users from three universities in Fall 2010.

Figure 3.13 shows the score plot for The score plot is transferring the original data set based on the PCs, and using PC1 as x axis, PC2 as y axis. The figure shows

that several users from Urbana related file 3, 16 and 21. These users were from the same class in University of Illinois at Urbana-Champaign and used qdot. The figure also shows that several classroom users from Purdue University related to file 6, 9 and 17. File 6 and 9 are lecture notes of MSE235 taught by Professor Alejandro Strachan. File 17 is a lab handout of the same course, which introduced using nano-Materials Simulation Toolkit to perform molecular dynamics simulations. The users related to these files are from same class in Purdue University and used tool nano-Materials Simulation Toolkit, which is developed by Alejandro Strachan, Amritanshu Palaria, Ya Zhou and Janam Jhaveri.

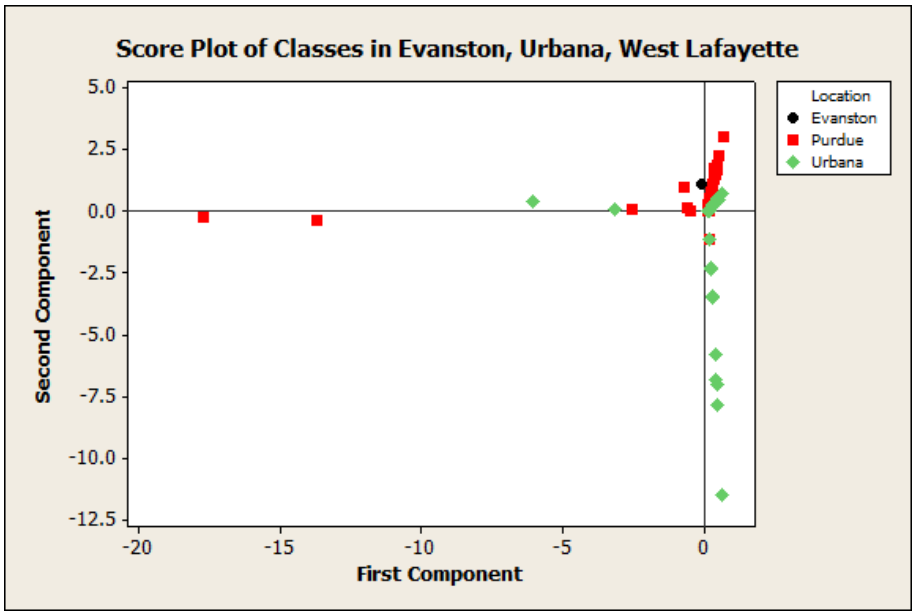


Fig. 3.13. Score plot of principle component analysis of number of downloads of each distinct file of users among three biggest classes in Fall 2010.

The principal component analysis of classroom users reveals the relationship between files, shows major downloads for each class, and proves the accuracy of classroom users.

The analysis results show that PCA method is quite useful to find relationship between download files and download times. The loading plots here only present PC1 and PC2. If looking at all the PCs, the analysis can reveal the relationship between each file. Using this information, nanoHUB can provide extra related documents for each tool which can save users' time and help user to learn each tool more comprehensively. Also, this method can be used for different matrices extracted from nanoHUB and study the relationship between different variables.

4. CONCLUSION

This study develops a automatic anomaly detection method that can apply to a heavy-tailed, unknown distribution, which is generated by Dunn et al. [13]. Chapter 2 showed the details of this method and proved that this method detects accurate anomaly intervals and method is not sensitive to few non-interrupting changes. Since this method does not have any assumptions of original data sets, it should be able to applied to any other ordered dataset. But it needs further validations. Chapter 3 is introduces *andmore* database analysis using the anomaly detection method, threshold analysis developed by Dunn et al. [13] and statistical methods. The study used ANOVA to give a first intuition of *andmore* database and use principal component analysis method to do further exploration. The PCA results shows the relationship between download files from simulating users, browsing users, group users and classroom users. The principal component analysis of classroom users reveals the relationship between files, shows major downloads for each class, and proves the accuracy of classroom users.

5. FUTURE WORK

The two possible extensions to this work are listed below.

1. Chapter 1 introduces method for anomaly detection. This method allows user to setup an error rate or iterating from a starting point until no output is available. However, users still need self judgment on which error rate is desirable. In the future, this work will be extended to full automation. The algorithm will decide which threshold is the most proper one and users only need to provide input data.

2. Chapter 2 introduces analysis of andmore database. However, this database includes unexplored data and unknown user behavior. Future works will focus on other parameters and find more interesting user categories or patterns. For example, PCA method can be applied to download files from classes in different years and different universities. Also, by studying the document that users downloaded, one may find problems users are encountering when using nanoHUB. All these will allow us to come up with suggestions as to how nanoHUB can be improved, built or even restructured for the future in order to generate more impact.

LIST OF REFERENCES

LIST OF REFERENCES

[1] IBM, "Ibm what is big data to the enterprise." http://www-01.ibm.com/software/data/bigdata/.

[2] Oracle, "Oracle information architecture: An architect's guide to big data," tech. rep., Oracle, 2012.

[3] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International Journal of Forecasting*, vol. 20, no. 1, 2004.

[4] R. Fildes and C. Beard, "Forecasting systems for production and inventory control," *International Journal of Operations and Production Management*, vol. 12, no. 15, 1992.

[5] Y. Ni and F. Fan, "A two-stage dynamic sales forecasting model for the fashion retail," *Expert Systems with Applications*, vol. 38, no. 3, 2011.

[6] N. Kock, "What is e-collaboration?," *International Journal of e-Collaboration*, 2005.

[7] T. Berners-Lee, "The many forms of e-collaboration: Blogs, wikis, portals, groupware, disoussion boards, and instant messaging," 1999.

[8] C. Ma, "E-collaboration: A universal key to solve fierce competition in tourism industry?," *International Business Research*, vol. 1, no. 2, 2008.

[9] G. Klimeck, M. McLennan, M. Lundstrom, and G. Adams, "nanohub.org - online simulation and more materials for semiconductors and nanoelectronics in education and research," *IEEE*, 2008.

[10] G. Klimeck, M. McLennan, S. Brophy, G. Adams, and M. Lundstrom, "nanohub.org: Advancing education and research in nanotechnology," *IEEE*, 2008.

[11] G. Klimeck, G. Adams, K. Madhavan, N. T. Denny, M. Zentner, S. Shivarajapura, L. Zentner, and D. Beaudoin, "Social networks of researchers and educators on nanohub.org," *IEEE*, 2011.

[12] nanoHUB.org. https://nanohub.org/usage#fn1.

[13] J. Dunn, G. Klimeck, and O. Nohadani, "From predefined to data-driven categories: The case study of nanohub users." submitted (2013).

[14] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, 2009.

[15] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, 2004.

[16] C. Spence, L. Parra, and P. Saida, "Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model," *IEEE*, 2001.

[17] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection.* Wiley, 1988.

[18] P. Galeano, D. Pea, and R. S. Tsay, "Outlier detection in multivariate time series by projection pursuit," *Journal of the American Statistical Association*, vol. 101, no. 474, 2006.

[19] M. Moen, K. Griffin, and A. Kalantar, "Simple regression and outlier detection using the median method," *Analytica Chimica Acta*, 1993.

[20] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations, Second Edition.* SIAM, 2004.

[21] http://scienceworld.wolfram.com/biography/Gauss.html.

[22] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers, Second Edition.* Wiley, 2011.

[23] N. Denny, D. McKay, S. Shivarajapura, S. Snyder, and M. Zentner, "Elevating nanohub to the next level." https://hubzero.org/resources/1082/download/.

[24] M. G. Zentner, N. Denny, K. Madhavan, S. Shivarajapura, G. Adams, and G. Klimeck, "If you build it, they will not necessarily come: Using automatic detection and characterization to plan value-added services that retain newcomers in online scientific communities.".

[25] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, 2010.

[26] L. I. Smith, "A tutorial on principal components analysis," tech. rep., University of Otago, COSC453, 2002.

[27] http://onlinecourses.science.psu.edu/stat505/node/53.