Optimization of Parallel Coordinates for Visual Analytics

By

Liang Fu Lu

Supervisor：Mao Lin Huang

A Thesis submitted in Fulfillment for the Degree of Doctor of Philosophy

in

Faculty of Engineering and IT

University of Technology, Sydney Australia

March 2016

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

UNIVERSITY OF TECHNOLOGY SYDNEY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

SIGNATURE OF STUDENT

_____

ACKNOWLEDGEMENTS

How time flies. It has been nearly five years since for my research study I first toke my feet on the grand and attractive university-University of Technology, Sydney, where I had been dreaming for years to go. Fortunately, I acquired a variety of knowledge in the short but precious semesters like a tiny boat loading to full capacity of great achievements, expectations and dreams was sailing backward in the sun shine through the ordeal from the sea storm. At the tranquil and solemn campus, everywhere is surrounded by the strong and profound academic atmosphere, resounded with vigorous study aspirations and left the track of our familiar figure shadows and good memories, which arouses a strong sense of gratitude in my heart.

Firstly, I would like to gratefully acknowledge Mao Lin Huang, an A/Prof in the field of information visualization, who has who has earned a great reputation at home and abroad for his numerous contributions to scientific researches and practical achievements, for the enthusiastic supervision in the process of the research. This thesis grew out of a series of dialogues with him. Each step of completing the thesis including selection, proposal, research and writing is dissolved in Prof. Huang's hard work and guidance. What was deeply impressed me is not the toil and sweat he scattered in numerous readings and modifications, but the wise academic vision, noble life faith and rigorous scholarship attitude he had, which will benefit me a lot in my rest whole life. He brought me closer to the reality I had initially perceived, and eventually enabled me to grasp its rich complexity.

Furthermore, I owe sincere thankfulness to the members of Visualization Team and fellow researchers and the professors of iNext Research centre. Their scholarly attainments and instructions in humor, serious and patient way gave me great help and

inspirations, and that besides gaining knowledge also benefitted me many like friendship, concern, trust and so on.

This research also benefited tremendously from many researchers and staffs in the University of Technology, Sydney. In addition, thank you to the participants in the usability study for the cooperation and valuable feedbacks which are taken as a key factor in providing a necessary guarantee on the datum I collected in accuracy and precision and the evidences in objectivity and scientificity for the thesis.

Last but not least, I owe my deepest gratitude to Huan Xu, my wife, who bravely took the burden of supporting the family alone without any complaints and totally poured her love onto my daughter and me and my whole family. All of her toil and sweat, tolerance and understanding, selfless support and devotion set up the strong will and belief for me in my study. I really appreciated sincerely my extended family for their continuous encouragement and support to make this PhD thesis possible. Thanks my daughter Wenxuan Lu for constantly delivering me angel smiles which accompanied me all the way. Her sweet smiles once encouraged me stand up again after fall a million times and then bravely and firmly faced the difficulties I had encountered. You are all my love most worthy of spending my lifetime loving.

Contents

Contents v

Figure List vii

Table List viii

Equation List ix

Algorithm List x

Abstract xi

Chapter 1. INTRODUCTION 1

1.1 INFORMATION VISUALIZATION 1

1.2 HIGH DIMENSIONAL DATA VISUALIZATION 5

1.3 PARALLEL COORDINATE PLOTS 9

1.4 RESEARCH CHALLENGES 13

1.5 RESEARCH OBJECTIVES 16

1.6 CONTRIBUTIONS 19

1.7 THESIS ORGANIZATION 23

Chapter2. VERTICES OPTIMIZATION IN PARALLEL COORDINATE PLOTS 27

2.1 CLUTTER DESCRIPTION IN PCP 27

2.2 NEW ALGORITHM FOR CLUTTER REDUCTION 30

2.3 CASE STUDIES 35

2.4 SUMMARY 45

Chapter 3. NEW AXES RE-ORDERING METHOD IN PARALLEL COORDINATE PLOTS 47

3.1 SIMLARITY MEASURE AND DIMENSION RE-ORDERING METHODS 47

3.1.1 SIMLARITY MEASURE 48

3.1.2 DIMENSION RE-ORDERING METHODS 49

3.2 NEW APPROACH FOR DIMENSION RE-ORDERING 53

3.2.1 Linear/Nonlinear Correlation 53

3.2.2 Similarity-based Reordering 56

3.3 CASE STUDIES 58

3.3.1 Cars dataset 59
3.3.2 Liver disorders dataset 61
3.4 SUMMARY 65
Chapter 4. USING ARCED AXES IN PARALLEL COORDINATE GEOMETRY 67
4.1 ANALYSIS OF PCP 67
4.2 OVERVIEW OF APPROACHES ON PCP 69
4.3 ARC-BASED PARALLEL COORDINATES GEOMETRY 74
4.3.1 Optimizing Length of Arced Axis 74
4.3.2 Arc-Coordinate Geometry 76
4.3.3 Contribution-Based Layout 80
4.4 CASE STUDIES 82
4.4.1 Random and Car Datasets 82
4.4.2 Case Study in Network Security Domain 84
4.5 SUMMARY 87
Chapter 5. CONCLUSION AND FUTURE WORK 89
PUBLICATION LIST 93
APPENDIX 95
REFERENCES 96

Figure List

Figure 1.Scientific visualization for Stanford bunny(Wijk 2002). 2

Figure 2. Parallel coordinate plots for car dataset. 2

Figure 3. Classification of information visualization.(Keim 2002)..... 4

Figure 4. Space-scale diagrams illustrating two experimental comparisons of zooming interfaces with display size varied between D and 2D. (Jakobsen and Hornbaek 2013)..... 5

Figure 5. Chernoff face visualization on high dimensional data(Kabulov and Tashpulatova 2010)..... 6

Figure 6.Scatterplot matrix for a 7-dimensional car dataset.(Elmqvist, Dragicevic et al. 2008) 7

Figure 7. 3D scatterplot matrix showing the 8D "olive oil" data set. (Sanftmann and Weiskopf 2012)..... 8

Figure 8. Five clusters in 2D visualization of 100,000 artificially generated data items(Keim and Krigel 1994).9

Figure 9. Examples of visual clutter in parallel coordinates visualization. See regions bounded by two ellipses that contain a large number of edge crossings. 29

Figure 10. The example display of uncertain values visualized in parallel coordinates. The lines in red and green behave realistic and uncertain data respectively. The dummy vertices are shown by green circle. . 32

Figure 11: Case 1 - Random data in parallel coordinates: (a) The initialization of incomplete data items; (b)Visualization of suboptimum positions of uncertain values; (c) The optimal positions of vertices..... 37

Figure 12. Case 2 - An incomplete dataset AMEX A visualized in parallel coordinate visualization: (a)The initial drawing of the given data with ten uncertain values; (b) The new drawing of the same given data after the implementation of our optimization method. The data source is available at: http://davis.wpi.edu/xmdv/datasets/amexa.html. 38

Figure 13: Case 3 - Forbes 94, a dataset with 5 variables visualized in parallel coordinate visualization: (a) Original plot; (b) after clutter reduction. Data from http://www-stat.wharton.upenn.edu/waterman/fsw/datasets/txt/Forbes94.txt 44

Figure 14. Cars dataset visualization in parallel coordinates. 63

Figure 15. Axes reordering visualization of Liver Disorders dataset. 64

Figure 16. Duality property between points and lines in Cartesian and parallel coordinate plots(Wegman 1990). 71

Figure 17. The rationale of arc coordinates plane. 76

Figure 18. Random data represented in two different coordinates systems 83

Figure 19. Car dataset visualized in PCP and ACP respectively 86

Figure 20. Detecting DDoS attacks using ACP:Red and green lines describe the Smurf and Neptune attacks respectively 86

Table List

TABLE 1 DETAILS OF DATA SETS USED IN THE THESIS. 25
TABLE 2 CLUTTER REDUCTION USING OPTIMAL ORDERING ALGORITHM FOR ALL CASES. 41
TABLE 3 CLUTTER REDUCTION IN DIMENSION DECOMPOSITION FOR CASE 3 FORBES94. 42
TABLE 4 THE COMPARISON OF THE SIMILARITY VALUES USING PCC AND NCC TO CARS DATASET. 65

Equation List

EQ. 1 33
EQ. 2 34
EQ. 3 54
EQ. 4 55
EQ. 5 55
EQ. 6 56
EQ. 7 58
EQ. 8 58
EQ. 9 74
EQ. 10 75
EQ. 11 75
EQ. 12 76
EQ. 13 79
EQ. 14 79
EQ. 15 79
EQ. 16 82

Algorithm List

Algorithm 1. Determination of positions of incomplete data 31
Algorithm 2. Similarity-based Reordering Algorithm 57

Abstract

The visualization and interaction of multidimensional data always requires optimized solutions for integrating the display, exploration and analytical reasoning of data into a kind of visual pipeline for human-centered data analysis and interpretation. However, parallel coordinate plot, as one of the most popular multidimensional data visualization techniques, suffers from a visual clutter problem. Although this problem has been addressed in many related studies, computational cost and information loss still hamper the application of these techniques, which leads to large high dimensional data sets. Therefore, the main goal of this thesis is to optimize the visual representation of parallel coordinates based on their geometrical properties.

At the first stage, we set out to find optimization methods for permuting data values displayed in parallel coordinate plot to reduce the visual clutter. We divide the dataset into two classifications according to the values and the geometric theory of the parallel coordinate plot: numerical data and non-numerical data, and missing data may exist between them occasionally. We apply Sugiyama's layered directed graph drawing algorithm into parallel coordinate plot to minimize the number of edge crossing among polygonal lines. The methods are proved to be valuable as it can optimize the order of missing or non-numerical value to tackle clutter reduction.

In addition, it is true that optimizing the order is a NP-complete problem, though changing the order of the axis is a straightforward way to address the visual clutter problem. Therefore, we try to propose in the research a new axes re-ordering method in parallel coordinate plot: a similarity-based method, which is based on the combination

of Nonlinear Correlation Coefficient (NCC) and Singular Value Decomposition (SVD) algorithms. By using this approach, the first remarkable axis can be selected based on mathematical theory and all axes can be re-ordered in line with the degree of similarities among them. We also propose a measurement of contribution rate of each dimension to reveal the property hidden in the dataset.

In the third stage, we put forward a new projection method which is able to visualize more data items in the same display space than the existing parallel coordinate methods. Moreover, it is demonstrated clearly in the research that the new method enjoys some elegant duality properties with parallel coordinate plot and Cartesian orthogonal coordinate representation. Meanwhile, the mean crossing angles and the amount of edge crossing between the neighboring axes are utilized in this research to demonstrate the rationale and effectiveness of our approaches.

Chapter 1. INTRODUCTION

1.1 INFORMATION VISUALIZATION

Nowadays people always acquire, cognize and process a variety of datum from their informational surroundings. With the enlargement of the data volume and the growing of complexity of data itself, it becomes more and more difficult to process such a vast data and reveal the pattern behind it in short time. Visualization, as the links with human mind and intelligent computer, transforms data, information and knowledge into a visual form which helps people discover the data pattern easily in a rapid and visual way. Visualization research and development has fundamentally changed the way we present and understand large complex data sets in such an increasingly information-rich age(Gershon, Card et al. 1998).

Scientific visualization, which aims to help people understand scientific phenomena by focusing on data (An example is shown in Fig. 1), and information visualization compose the stat-of–art visualization research and development. While information visualization mainly focuses on abstract information and nonphysical data in order to reveal the diverse patterns (See Fig. 2).

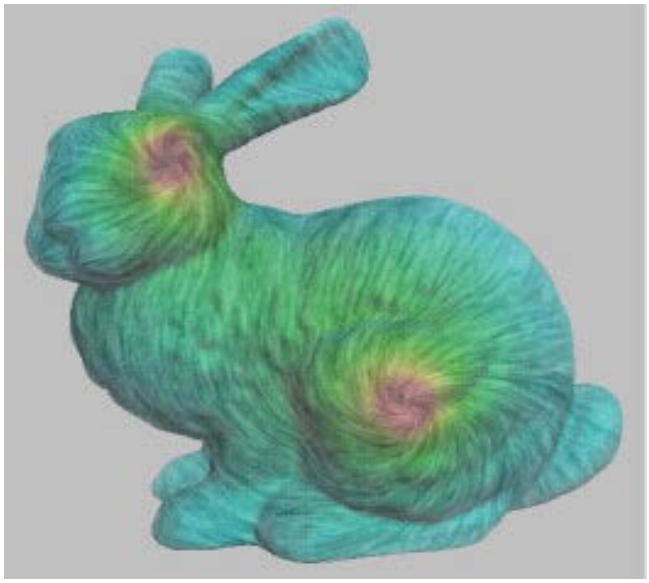


Figure 1.Scientific visualization for Stanford bunny(Wijk 2002).

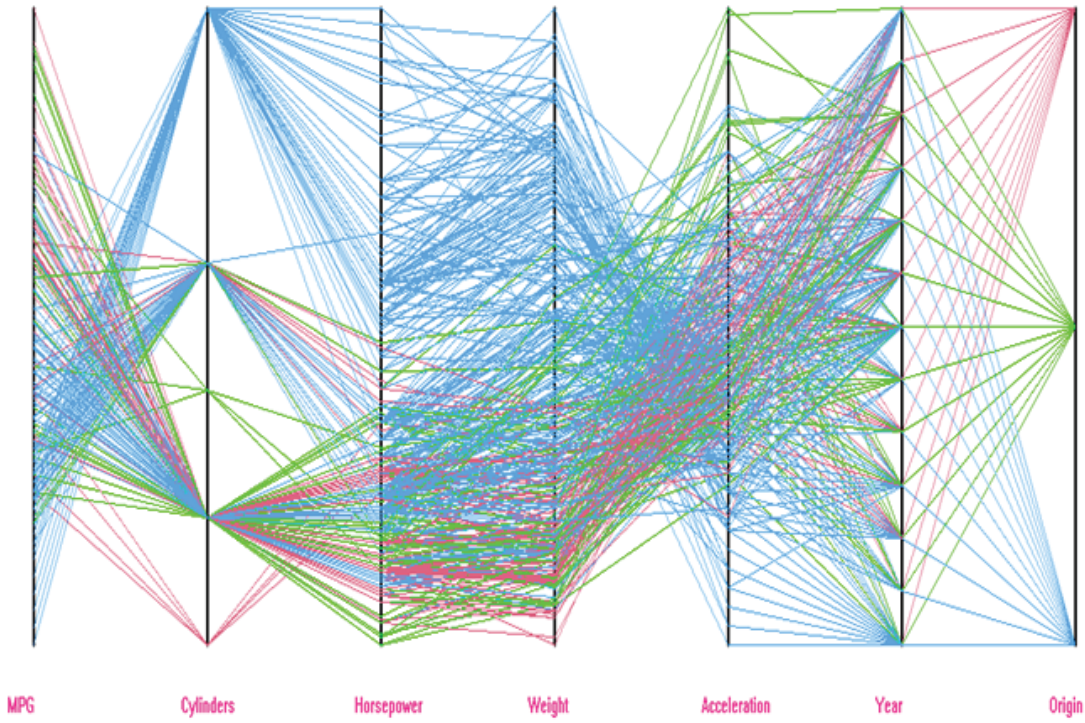


Figure 2. Parallel coordinate plots for car dataset.

Due to the increasing difficulty of exploring and analyzing the vast volume of data, information visualization can do great help in finding the valuable information hidden in the data. As Stuart K.Card described in literature “Using Vision to Think”(Card,

Mackinlay et al. 1999), information visualization is the use of computer–supported, interactive, visual representations of abstract data to amplify cognition, which means to explore the meaning of complicated large volume of data when they are represented in graphs rather than display them in letters or numbers according to the way the human brain processes information. Therefore, information visualization is a clear and effective way to perceive information graphically, especially when data patterns are not obvious to people.

In the past decades, a large number of well-known information visualization techniques have been developed in dealing with the multidimensional data. Daniel A. Keim presented some novel visualization methods from different aspects, such as data types, interaction and distortion techniques(Keim 2002). Fig. 3 shows the classification of information visualization in detail according to the three different criterions. The data in information visualization usually consists of large volume of records which own different attributes or dimensions especially in scale. Maybe one part of attributes is numerical, and the other part is non-numerical. Sometimes part of them is abstract or non-structural. For example, in the "Car Evaluation Data Set"(Repository) -a multivariate data set, each record consists of six attributes where the "origin" of the car is the only non-numerical attribute. Therefore, the data type to be visualized in information visualization research field can be classified into one-dimensional, two-dimensional, multidimensional data and other types like text, graphs and algorithms et al. And currently many kinds of visualization techniques have been proposed, which include the classic bar charts (standard 2D/3D displays), parallel coordinate plots (geometry-based visualization) and treemap (stacked display) et al. Besides the novel visualization techniques, interactive interfaces are usually utilized in information

visualization, such as “Zoom and Pan”, “Focus+Context” and “Overview+Detail” et al. techniques. In practical application, any information visualization technique may be combined with other techniques displayed in different orthogonal dimensions. For example, Mikkel R. Jakobsen et al. investigated the relation between information space and display size by implementing focus+context, overview+detail and zoom and pan et al. interactive visualization techniques for multi-scale navigation in maps. Fig. 4 shows the two experimental comparisons of zooming interface aiming to illustrating the interrelation between information space and display size (Jakobsen and Hornbaek 2013).

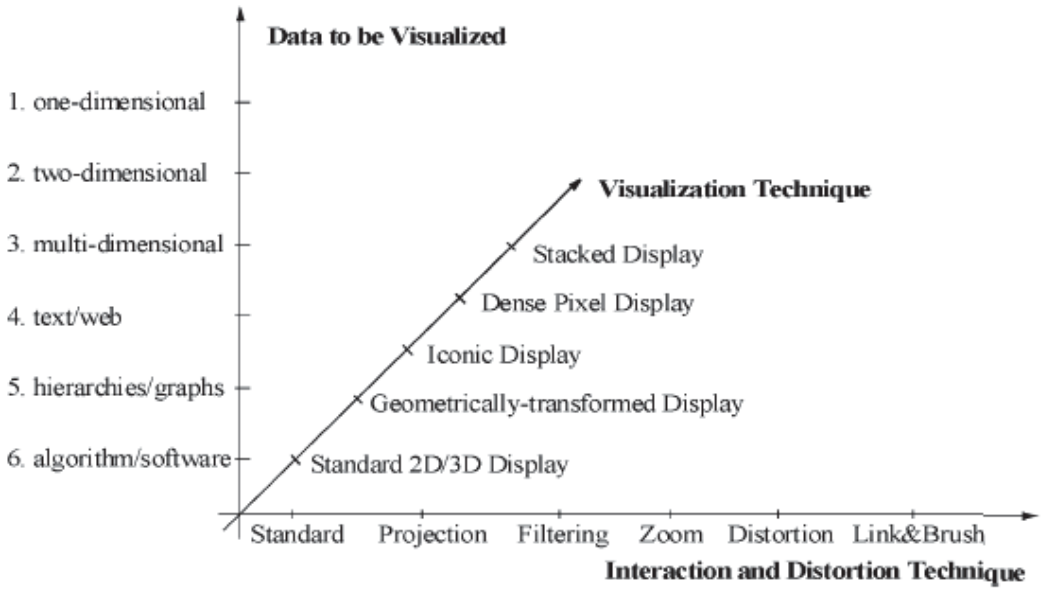


Figure 3. Classification of information visualization.(Keim 2002)

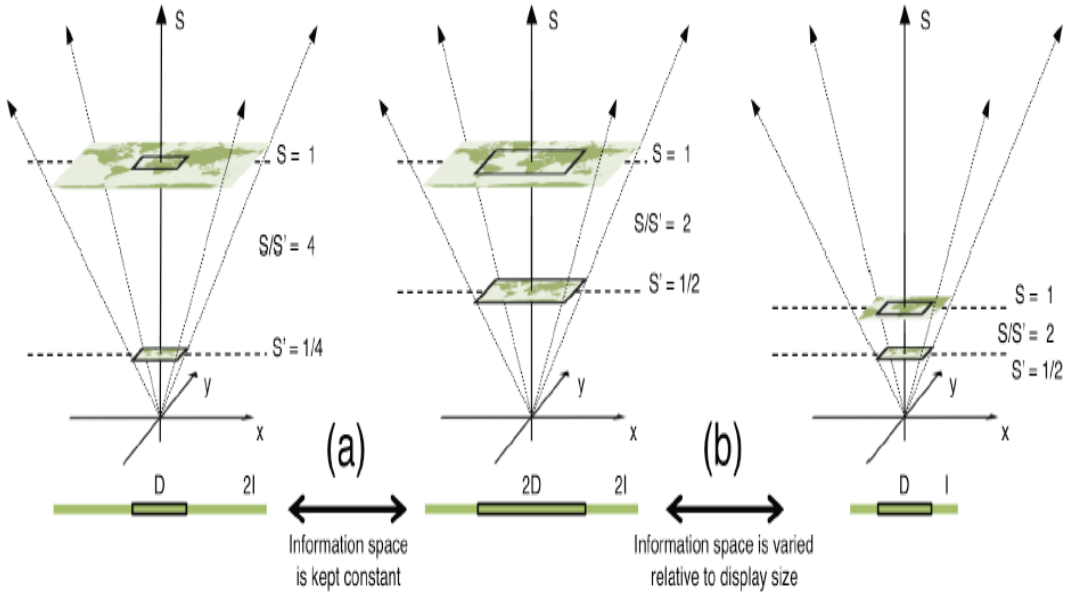


Figure 4. Space-scale diagrams illustrating two experimental comparisons of zooming interfaces with display size varied between D and 2D. (Jakobsen and Hornbaek 2013)

1.2 HIGH DIMENSIONAL DATA VISUALIZATION

In our daily life, we have to deal with the problems occurred in the data processing, especially when the dimensions of data are too high. For example, in the process of analyzing the gene expression microarray data set, we have to handle with ten or hundreds of experimental conditions which are considered to be different dimensions. Though many models and algorithms have been built and developed to mine the data patterns, the curse of dimensionality and the meaningfulness of the similarity measure in high-dimensional space are still the key challenges need to be solved(Wang and Yang 2005).

Herman Chernoff proposed that multivariate data less than 18 dimensions can be presented by the features in a cartoon face such as length of nose, curvature of mouth and size of eyes(Chernoff 1973). Fig. 5 shows the high dimensional data visualization by Chernoff face(Kabulov and Tashpulatova 2010). Some industrial applications based

on Chernoff face are proposed in recent years such as analysis of oil company activity(Bruckner 1979), potable water quality testing(A. Astel 2006) et al. However, the existing Chernoff faces are only oriented to comparison between two different objects or different aspects of single object. B.T. Kabulov et al. proposed an enhanced visualization version to make the face contain more information and easier to show the interval estimation of the values of parameters (Kabulov and Tashpulatova 2010).

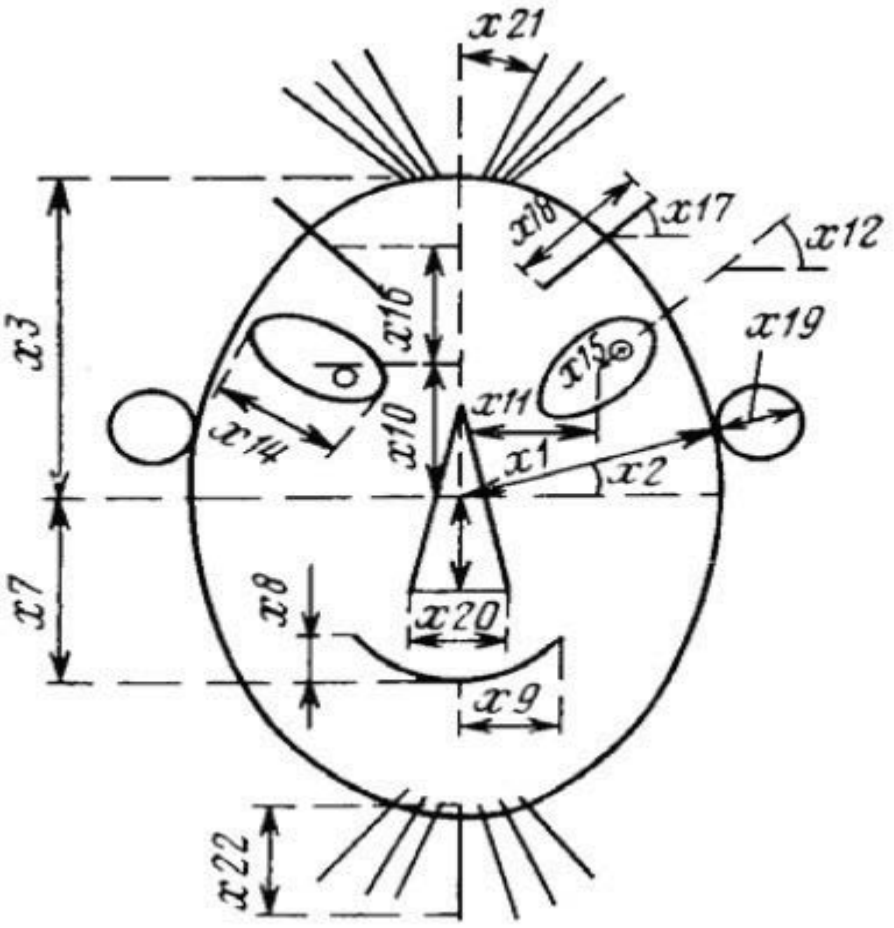


Figure 5. Chernoff face visualization on high dimensional data(Kabulov and Tashpulatova 2010).

Scatterplot, as a statistically graphical visualization method, displays data items as points in the Cartesian space while the dimension of the data are represented as

graphical axes (See Fig. 6). Though 3D scatterplot can visualize some high dimensional data by assigning the points with color, size and shape, the data with much more dimensions are always visualized by traditional 3D scatterplot combing with some interactive navigations. (Elmqvist, Dragicevic et al. 2008).

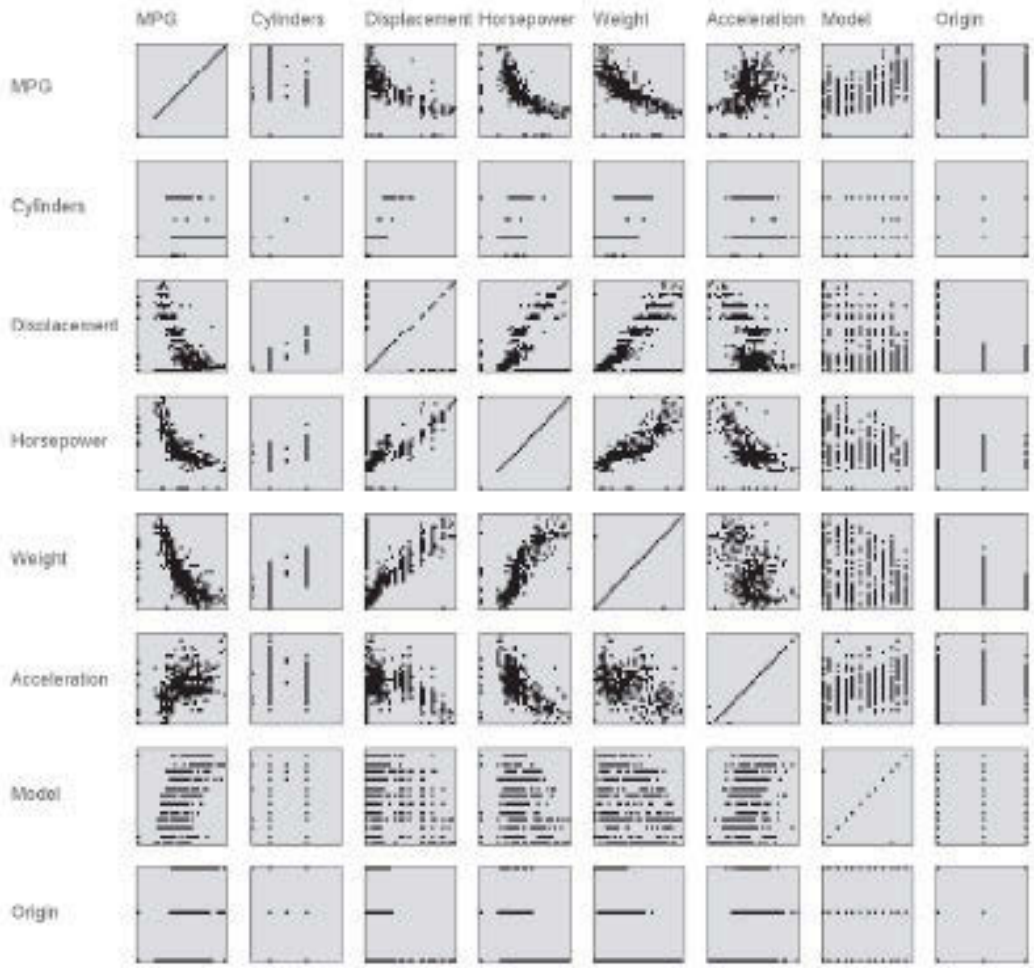


Figure 6.Scatterplot matrix for a 7-dimensional car dataset.(Elmqvist, Dragicevic et al. 2008)

Motivated by the navigation method proposed in the literature (Elmqvist, Dragicevic et al. 2008), Harald Sanftmann et al. extended the approaches to 3D axes by swapping one or two axes during transitions (Sanftmann and Weiskopf 2012). For example, Fig. 7 displays us the 8D oil dataset in 3D scatterplot views, where the third dimension of the data is mapped to the y-axis and all 2D projections of the 3D scatterplot matrices that

preserve the y-axis mapping are projected to the back face perpendiculartо the y-axis of the cube(Sanftmann and Weiskopf 2012).

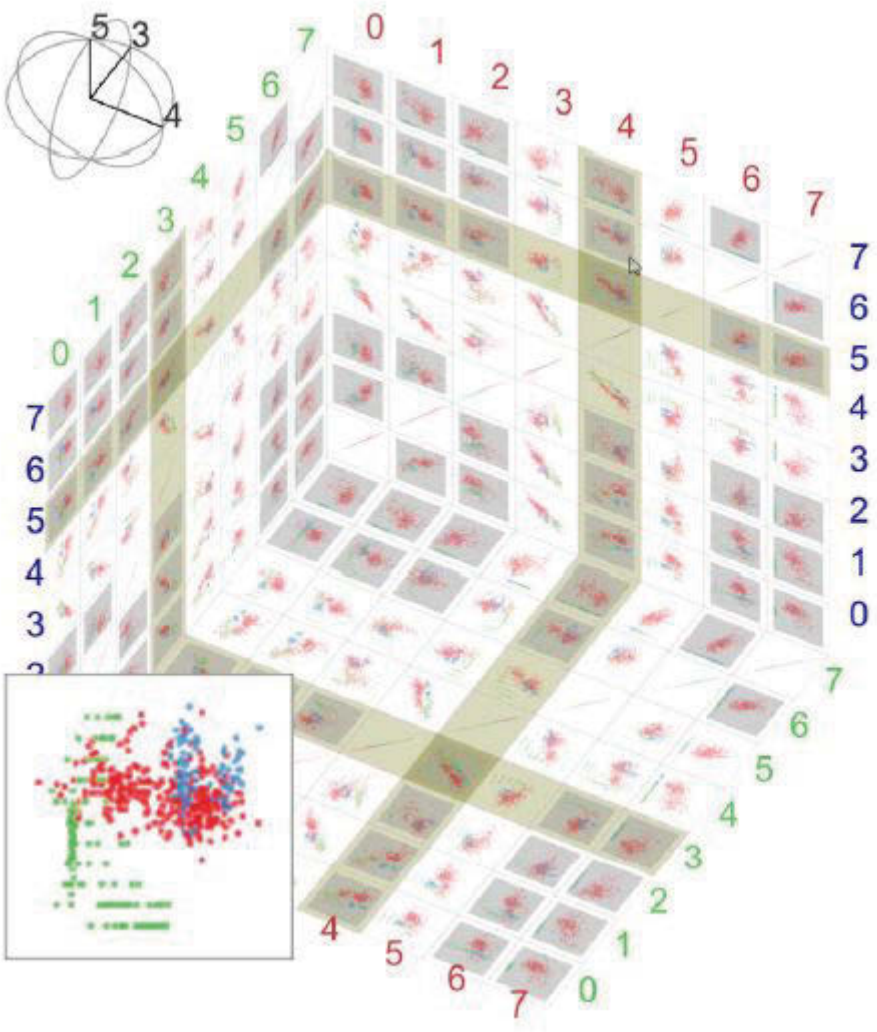


Figure 7. 3D scatterplot matrix showing the 8D "olive oil" data set. (Sanftmann and Weiskopf 2012)

Daniel A. Keim etc. (Keim and Krigel 1994) utilized visualization techniques to explore database by combining traditional database querying and information retrieval techniques and proposed a system termed VisDB, which provides valuable feedback in querying the database and allows the user to find results even hidden in the database by using of visualizing large amount of data on current displays. Fig. 8 visualizes the 100,000 data items with five clusters. The comparable distances of different clusters are denoted by various regions with different colors in the visualization.

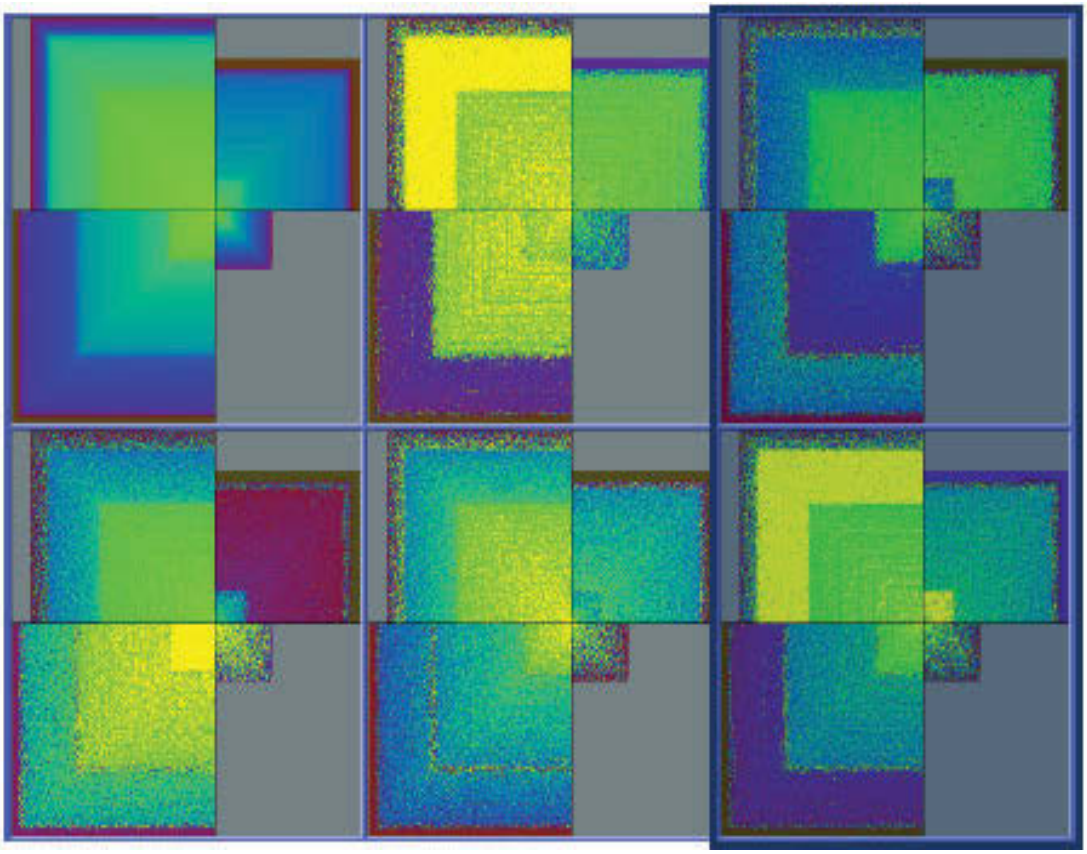


Figure 8. Five clusters in 2D visualization of 100,000 artificially generated data items(Keim and Krigel 1994).

1.3 PARALLEL COORDINATE PLOTS

High-dimensional data and multivariate data are becoming commonplace as the number of applications increases, such as statistical and demographic computation, digital libraries and so on. Though it can provide flexible and cost-saving IT solutions for the end users, it is much easier in causing a great deal of problems such as network and system security issues due to its sharing and centralizing computing resources.

As pointed out in the literature (Claessen and van Wijk 2011), many methods have been proposed to provide insight into multivariate data using interactive visualization

techniques. Parallel coordinate plots (PCP), as a simple but strong geometric high-dimensional data visualization method, represents N-dimensional data in a 2-dimensional space with mathematical rigorousness. PCP is proposed firstly by Inselberg (Inselberg 1985) and Wegman suggested it as a tool for high dimensional data analysis (Wegman 1990). Coordinates of n-dimensional data can be represented in parallel axes in a 2-dimensional plane and connected by linear segments. (See Fig. 2).

In order to reduce the edge clutter and avoid the over-plotting in PCP, Dasgupta et al. (Dasgupta and Kosara 2010) propose a model based on screen-space metrics to pick the axeslayout by optimizing arranges of axes. Huh et al. present a proportionate spacing between two adjacent axes rather than the equally spaced in conventional PCP parallel axes. Moreover, the curves possessing some statistical property linking data points on adjacent axes are described in literature (Huh and Park 2008) as well. Zhou et al. (Hong Zhou 2008)convert the straight-line edges into curves to re-duce the visual clutter in clustered visualization. They also utilize the splatting framework (Zhou, Cui et al. 2009) to detect clusters and reduce visual clutter. To achieve the aim of avoiding over-plotting and preserving density information, Dang et al. proposed a visualization and interaction method for stacking overlapping cases (Tuan Nhon, Wilkinson et al. 2010).To filter out the information to be presented to the user and reduce visual clutter, Artero et al. develop a frequency and density plots from PCP(Artero, de Oliveira et al. 2004), which can uncover clusters in crowded PCP. Yuan et al. (Xiaoru, Peihong et al. 2009)combine the parallel coordinate method with the scatter-plots method to reduce the visual clutter. It plots scattering points in parallel coordinates directly with a seamless transition between them. The shapes of poly lines are remodeled to cooperate with the scattering points, resulting in the diminution of their inherent visual effects.

Many efforts have been put to deal with the visual clutter in completed data visualization (Laguna, Mart et al. 1997; Ankerst, Berchtold et al. 1998; Wei-xiang and Jing-wei 2001; J. Yang 2003; Peng, Ward et al. 2004; Ellis and Dix 2006; Pascale, Bruno et al. 2006; Ellis and Dix 2007; Hong Zhou 2008; Peihong, He et al. 2010; Huang and Huang 2011). In multi-dimensional data visualization, especially in the parallel-coordinates visualization, there are three directions for the current research of visual clutter reduction: 1) dimension reduction(J. Yang 2003; Peng, Ward et al. 2004) ; 2) data clustering(Ankerst, Berchtold et al. 1998; Hong Zhou 2008; Peihong, He et al. 2010); and 3) minimization of edge crossings(Laguna, Mart et al. 1997; Wei-xiang and Jing-wei 2001; Pascale, Bruno et al. 2006; Huang and Huang 2011). The first approach is to find and filter some less important data attributes for simplifying the decision making or problem solving processes, and in this case,, the visual clutter and visual complexity will be reduced accordingly in the corresponding visual representation of the data; The second is data clustering, which is one of the data mining approaches to group data items based on a variety of rules, such as data similarity, and through data clustering, we can display some abstract clusters instead of the data details, and then the visual complexity and clutter can be reduced accordingly; The minimization of edge crossings is a problem of graph drawing. Through the optimization of geometric positions of nodes and edges to reduce the edge crossings among edges is the main method for visual clutter reduction.

To enhance the high-dimensional data visualization, some studies on dimension reordering have been done to find good axes layouts in visualization techniques both in one- or two-dimensional arrangement (J.Bertin 1983; Johansson and Johansson 2009; Tatu, Albuquerque et al. 2009; Dasgupta and Kosara 2010; Bertini, Tatu et al. 2011)

(Hahsler, Hornik et al. 2008; Hurley and Oldford 2010) (Ankerst, Berchtold et al. 1998; Friendly and Kwan 2003; J. Yang 2003; Peng, Ward et al. 2004; Artero, Oliveira et al. 2006). Mihael Ankerst et al. (Ankerst, Berchtold et al. 1998) defined similarity measures which determined the partial or global similarity of dimensions and argued that the reordering based on similarity could reduce visual clutter and do some help in visual clustering. Wei Peng et al.(Peng, Ward et al. 2004) introduced the definition of the visual clutter in parallel coordinates as the proportion of outliers against the total number of data points and they tried to use the exhaustive algorithm to find the optimal axes order for minimizing the member of edge crossings (or visual clutter). As mentioned in (Artero, Oliveira et al. 2006), the computational cost $o(n \cdot n!)$ hampers applications of this technique to large high dimensional data sets. Almir Olivette Artero et al. (Artero, Oliveira et al. 2006) introduced the dimension configuration arrangement based on similarity to alleviate clutter in visualizations of high-dimensional data. They proposed a method called SBAA (Similarity-Based Attribute Arrangement), which is a straightforward variation of the Nearest Neighbor Heuristic method, to deal with both dimension ordering and dimensionality reduction. Other studies have been done on the dimension reordering based on the similarity(Friendly and Kwan 2003; J. Yang 2003; Tatu, Albuquerque et al. 2011) (Guo 2003) (Albuquerque, Eisemann et al. 2010). Michael Friendly et al.(Friendly and Kwan 2003) designed a framework for ordering information including arrangement of variables. However, the arrangement of variables is decided mainly according to the users' desired visual effects. J. Yang et al. (J. Yang 2003) established a hierarchical tree structure over the attributes, where the similar attributes were positioned near each other. Diansheng Guo (Guo 2003) developed a hierarchical clustering method, which was based on comparison and sorting of dimensions by use of the maximum conditional entropy. Georgia Albuquerque et al.

(Albuquerque, Eisemann et al. 2010) introduced the quality measures to define the placement of the dimensions for Radviz and also to appraise the information content of pixel and Table Lens visualizations.

In addition, few approaches have been done for extensions of axes in PCP. Claessen et al. (Claessen and van Wijk 2011) develop flexible linked axes to enable users to define and position coordinate axes freely. Axes-based techniques with radial arrangements of the axes are developed by Tominski(Tominski, Abello et al. 2004), termed as TimeWheel and the MultiComb, which can combined with some conventional interaction techniques. With the combination between interaction techniques and PCP, Hauser et al. (Hauser, Ledermann et al. 2002) design a angular brushing technique to select data sub-sets which exhibit a data correlation along two axes.

Even though the data can be represented in a novel and meaningful visualization system without losing any features, PCP always suffer from crowded dimensions, hardly figuring out the relationship between attributes in non-adjacent positions, over-plotting and clutter et al, which are mainly caused by the increasing size of datasets and the large number of dimensions. Therefore, this thesis is to study the optimization methods for PCP in order to improve the performance of it.

1.4 RESEARCH CHALLENGES

Though visualization techniques greatly help viewers to find the patterns and structures which are hidden in the large scale datasets, most visualization techniques fail to present the incomplete datasets. The incompleteness of data would be difficult for further

analysis. Therefore, in this research, we propose novel methods to reduce the visual clutter of incomplete multi-dimensional data (or non-numerical multi-dimensional data) in parallel coordinate plot.

Parallel coordinate plot, as a geometry-based visualization technique, visualizes the dimensions of datasets into different axes. During the past years, many researches on the similarity of dimension have been proposed with this visual technique. However, these methods have been proved to be a NP-complete problem. Though the traditional heuristic algorithms can help in finding an optimal order of axes for one- or two-dimensional visualizations, most studies have not been done to a deeper investigation on how to determine the first dimension (the most significant dimension) in multi-dimensional data visualizations, and the first dimension always attracts much more user's attention than the others. Therefore, we may consider the one with the highest contribution rate as the first dimension to simplify the traditional similarity-based re-ordering method and to find out the optimal order of parallel axes in a short time period; we will also propose the method to find out the contribution of each dimension in the dataset. And then, we present a similarity-based re-ordering method in parallel coordinate plot, which is sensitive to any relationships, including the linear dependence.

As parallel coordinate plots was firstly introduced and suggested as a tool for high dimensional data analysis in the approximately 30 years, little improvements have been done to the algorithm itself. In our research, we also propose a new projection method which is able to visualize more data items in the same display space than the existing parallel coordinate methods. Moreover, we make a mathematical demonstration of the method to show our method that can enjoy some elegant duality properties with parallel

coordinate plot and Cartesian orthogonal coordinate representation.

To reduce the visual clutter and achieve the optimized representations of data in parallel coordinate plot, we need conclude the following questions:

- **RQ1.** How to optimize the visual representation and reduce the visual clutter when we visualize the non-numerical data in parallel coordinate plot?

- **RQ2.** How to determine the positions of the incomplete data attribute values firstly when we visualize them in parallel coordinate plot?

- **RQ3.** How to permute the incomplete data attribute values to find the optimal positions to reduce the visual clutter when we visualize the incomplete numerical data in parallel coordinate plot?

- **RQ4.** To measure the similarity of the data items (dimensions), which algorithm is suitable and rational to reveal the correlation of the dimensions?

- **RQ5.** In parallel coordinate plot, the first remarkable axis always attracts much more visual attention. How to determine the first axis in visualization when we propose similarity-based re-ordering method?

- **RQ6.** Many extended parallel coordinate methods are proposed by the researchers. How to transplant and revise the methods we proposed in the above to the new

techniques based on the parallel coordinate plot?

- **RQ7.** Except the above methods to optimize the positions of values or dimensions, what interactive techniques can be used in parallel coordinate to reduce the visual clutter?

- **RQ8.** Combined with reducing edge crossing and enlarging the mean crossing angles in parallel coordinate plot, can we propose a new method to reveal the main characteristic of visual clutter?

1.5 RESEARCH OBJECTIVES

This research will be designed to focus on four primary research objectives based on the above research problems to optimize parallel coordinates.

- **RO1.** ***To optimize the positions of attribute values in parallel coordinate plot, it is necessary to propose a novel method to position the dummy vertices and permute the order of vertices (aim to answer RQ1, RQ2 and RQ3).***

For dummy vertices with uncertain values, we initially position each of them at the crossing point of its axis and a polygonal line connecting the vertex of its left hand side neighboring axis through the maximization of crossing angles to increase the readability of graphs.

Supposed that the data values in the second axis of parallel coordinates are non-numerical data, we can permute the order of all the crossing points (vertices) in this

axis to minimize the number of edge crossings between two neighboring axes, and then determine the best positions of data values to reduce the visual clutter and increase the readability effectively. In fact, if the data values in the second axis are numerical data, not all the data points can be reordered because of their unequal relations and practical meanings, whereas the new algorithms can be used to permute the order of the missing datum.

- **RO2.** ***To measure the similarity, it is necessary to propose a method based on contribution to determine the first remarkable dimension in parallel coordinate plot; and further present a new method which can detect linear and nonlinear relationships between two dimensions sensitively (aim to answer RQ4 and RQ5).***

The first dimension always attracts much more user's attention than the others. Therefore, we may consider the one with the highest contribution rate as the first dimension to simplify the traditional similarity-based re-ordering methods and to find out the optimal order of parallel axes in a short time period. Therefore, we will propose a method based on the contribution, which not only can give the theoretical support for the selection of the first dimension but also can visualize a clear and detailed structure of the dataset with the contribution of each dimension. Consequently, the computational complexity of clutter reduction methods can be greatly reduced and much more time correspondently could be saved through the new method than any other traditional reordering ones.

The correlation of two variables (dimensions/attributes) is a statistical technique that can indicate the magnitude relationship between the two variables. It also shows the way how the two variables interact with each other. It can be easily seen in the research

that linear correlation can detect the dependence of two variables, while in the real world the correlations can also be nonlinear. So we will propose a method to measure the linear or nonlinear relationship between the two dimensions in multidimensional datasets and it is sensitivity to any relationship.

- **RO3.** ***To extend the proposed methods to the improved parallel coordinate plot techniques, it is necessary to develop a system based on the proposed methods and some interaction techniques which can reduce the visual clutter easily and quickly (aim to answer RQ6 and RQ7).***

To propose a new projection method which is able to visualize more data items in the same display space than the existing parallel coordinate methods, we do some researches on polar coordinates to demonstrate that our method can enjoy some elegant duality properties with parallel coordinate plot and Cartesian orthogonal coordinate representation.

Continuous parallel coordinates are designed and studied for visualizing the datasets on a continuous domain in recent years. So far the structure of the visualization in all parallel coordinate methods is more or less fixed, and the user can only change some properties of the given representations. Some researchers also propose methods to freely define and position coordinate axes, suitably to specify visualizations and flexibly to link the axes of parallel coordinate plot. To transplant our methods into these new techniques, we will combine some interaction techniques with our proposed methods to reduce the visual clutter on continuous domain, and propose a self-adapting visual clutter reduction method to the new improved parallel coordinate plot techniques.

- **RO4.** ***To demonstrate the effectiveness of our methods in visual clutter reduction, it is necessary to propose the method based on statistics to evaluate the visual representations except the traditional reducing edge crossing and enlarging the mean crossing angles (aim to answer RQ8).***

To show the advantages of the readability and understandability of our method, we propose a formula to calculate the mean angles occurring among the polygonal lines between two neighboring attributes. Except this method to evaluate the visual clutter, we will design a user study and analyze it using statistical theory. Finally we will develop a new evaluation system to demonstrate the effectiveness of the methods in reducing the visual clutter.

1.6 CONTRIBUTIONS

The uncertainty in data visualization is a new research field, which represents incomplete data for analysis in real scenarios. In many cases, datasets, especially multi-dimensional datasets, often contain either errors or uncertain values. To address this challenge, we may treat these uncertainties as scalar values like probability. For achieving visual representation in parallel coordinates, we draw a small “circle” to temporarily define a dummy vertex for an uncertain value of a data item at the crossing point between polylines and the axis of certain dimension. Furthermore, these temporary positions of uncertainty could be permuted to achieve visual effectiveness. Further, optimizing the order to uncertain values can provide a great opportunity to tackle another important challenge in information visualization: clutter reduction. As a result, optimizing the order of uncertain values will have a great opportunity to tackle another important challenge in information visualization: clutter reduction. Visual

clutter always obscures the visualizing structure even in small datasets. In this thesis, we apply Sugiyama's layered and directed graph drawing algorithm into parallel coordinates visualization to minimize the number of edge crossing among polylines, which has been proved to significantly develop the readability of visual structure. The experiments made in case studies have shown clearly the effectiveness of our new methods for clutter reduction in parallel coordinates visualization. And they also have implied that besides visual clutter, the number of uncertain values and the type of multi-dimensional data are important attributes to affect visualization performance in this field. The visualization and interaction of multidimensional data always require optimized solutions to integrate the data presentation, exploration and also analytical reasoning into one visual pipeline for human-centered data analysis and interpretation. Parallel coordinate, as one of the most popular multidimensional data visualization techniques, is suffered from the visual clutter problem. Though changing the ordering of axis is a straightforward way to work it out, optimizing the order of axis is a NP-complete problem. In this thesis, we propose a new axes re-ordering method in parallel coordinates visualization, a similarity-based method, which is created on the basis of Nonlinear Correlation Coefficient (NCC) algorithm and Singular Value Decomposition (SVD) algorithm. By using this approach, the first remarkable axis can be selected on mathematical theory and then all axes will be re-ordered in line with the degree of similarities among them. Meanwhile, we would also propose a measurement of contribution rate of each dimension to reveal the properties hidden in the dataset. At last, case studies demonstrate the rationale and effectiveness of our approaches: NCC reordering method can enlarge the mean crossing angles and reduce the amount of polylines between the neighboring axes. It can reduce the computational complexity greatly in comparison with other re-ordering methods.

With the rapid growth of data communications in size and complexity, it is available to get data on shared data cloud computing platform, and meanwhile the threat of malicious activities and computer crimes have increased as well. Thus, it is urgently required to develop efficient data visualization techniques for visual network data analysis and visual intrusion detection over data intensive cloud computing. In this thesis, we first propose a new parallel coordinates visualization method that is characterized by arced-axis for high-dimensional data representation. This new geometrical scheme can be efficiently used to identify the main features of network attacks by displaying recognizable visual patterns. In addition, with the aim of visualizing a clear and detailed structure of the dataset according to the contribution of each attribute, we propose a meaningful layout for the new method based on the singular value decomposition (SVD) algorithm, which possesses the statistical property and can overcome the curse of dimensionality. Finally, we design a prototype system for network scan detection, on the basis of our visualization approach. The experiments have shown that our approach is effective in visualizing multivariate datasets and detecting attacks from a variety of networking patterns, such as significantly distinguishing the features of DDoS attacks.

An arc-based parallel coordinates visualization method, termed arc coordinate plots (ACP), is developed to extend the axes in parallel coordinate plots. Because the length of arc is longer than the line segments, the density of points displayed in each axis of our method could be enlarged. Moreover, ACP can preserve much more geometric structures of the data, such as the circular data. At the second stage, we leverage singular value decomposition algorithm to provide a new way of looking into the dimensions within datasets. We propose the contribution-based visualization method

and a formula for contribution rate of each dimension. At last, the experimental evaluations demonstrate the effectiveness and rationale of our approaches especially the applications in security domain.

To sum up, this study applies optimization techniques to reduce the visual clutter caused by the positions of attribute values and the order of dimensions in parallel coordinate plot. To combine tasks with research objectives, an intelligent visual analytics system will be designed for clutter reduction, which can enhance the visual readability and understandability. Based on the above objectives, the research expected outcomes and contributions will be as follows:

- A new algorithm to reach optimization of reducing edge crossings by determining the optimal positions of attribute values. -

- A new method to guide the parallel coordinate plot to visualize the datasets in line with the contribution of each dimension.

- A new similarity-based reordering method, through calculating the similarity between the two dimensions which is sensitive to any relationships between the two dimensions to optimize dimension order according to the similarity in parallel coordinate plot. A new similarity-based reordering method: That is to optimize dimension order through calculating the similarity of two dimensions in parallel coordinate plot for being sensitive to any relationships of two dimensions.

- A new method to combine the proposed methods with the improved parallel coordinate plot techniques.

1.7 THESIS ORGANIZATION

In this thesis, we present some clutter reduction methods, propose our optimization approaches, discuss the implementation of optimization algorithms and demonstrate their capability in processing the practical problems by using of visual analysis. And then we evaluate the approaches based on optimization criteria and graph drawing aesthetics. Finally, we conduct some case studies in network security and assess the performance of optimized parallel coordinate plots in comparison with traditional parallel coordinate plots.

The thesis is organized as follows: Chapter 1 describes the introduction of parallel coordinate plots and the visual clutter occurred in process the data using PCP; and related work and the challenges of these latest technologies are presented in this Chapter as well; the enclosure approaches on high dimensional data visualization are concluded in this chapter.

In Chapter 2, we introduce a layered directed graph drawing algorithm into parallel coordinates for visualization of uncertainty. Clutter and corresponding reduction methods are firstly described in Section 2.1. And then we propose our vertices optimization method in parallel coordinate plots and develop a multi-objective optimization algorithm as well in Section 2.2. Section 2.3 explains the algorithms of the approach by using of illustrations, examples and experimental results. At last, the conclusion of this chapter is described in Section 2.4.

We propose a new method to improve the readability and understandability of parallel coordinates visualization theoretically, i.e. a new axes re-ordering method in PCP in

Chapter 3. Firstly, in Section 3.1, we explain the similarity measure of all axes, which are different attributes visualized in PCP. In the same section, we present the dimension re-ordering methods based on similarity as well. We present a method, named similarity-based reordering method, for calculating the similarity between the two dimensions based on the nonlinear correlation coefficient and singular value decomposition algorithms instead of the traditional Pearson's correlation coefficient, and then visualize the optimal dimension order according to the similarity in parallel coordinates in Section 32. In Section 3.3, we conduct the experimental evaluations to demonstrate the effectiveness and rationale of our approaches: NCC reordering method enlarges the mean crossing angles of the whole data set and reduces the amount of polylines among some neighboring dimensions. In the final Section 3.4, we conclude the whole chapter and present the future work of this approach.

As to another innovative part of this thesis, Chapter 4 presents a novel approach to extend the parallel axes in parallel coordinates plane theoretically, which is termed arc coordinate plots (ACP). Because the length of arc is longer than the line segments, the density of points displayed in each axis of our method could be enlarged. Moreover, ACP can preserve much more geometric structures of the data, such as the circular data. At the second stage, we leverage singular value decomposition algorithm to provide a new way of looking into the dimensions within datasets. We propose the contribution-based visualization method and a formula for contribution rate of each dimension. At last, the experimental evaluations demonstrate the effectiveness and rationale of our approaches especially the applications in security domain.

From Chapter 2 to Chapter 4, we have done many case studies and achieved some good

results to demonstrate the effectiveness of our algorithms. As regards the data sets that we have used in our experiments, random data sets and some popular data sets available online for data mining research [(Irvine Machine Learning Repository) (http://archive.ics.uci.edu/ml/datasets/Car+Evaluation)]. In order to know about the data comprehensively, I list the following table to illustrate the details of the data sets we used in the thesis, which includes the number of dimensions and the size of the data sets.

Table 1 Details of data sets used in the thesis.

Data sets	Size	Selected Dimensions	Used in Chapter
Random 1	100	2	Chapter 2
AMEXA	55	2	Chapter 2
Forbes94	100	5	Chapter 2
Cars	406	7	Chapter 3,4
Liver Disorders	345	7	Chapter 3
Random 2	50	2	Chapter 4
KDD Cup 1999	1113	42	Chapter 4

As the last part of my thesis, the conclusions and contributions what I have achieved during my studies in our university are presented in Chapter 5. Furthermore, I proposed the research problems which occurred during my research in the last part of this chapter. Following this chapter, I listed the papers what I have finished and published about this research topic.

Chapter2. VERTICES OPTIMIZATION IN PARALLEL COORDINATE PLOTS

2.1 CLUTTER DESCRIPTION IN PCP

Today high dimensional data analysis is becoming more and more important as the number of analyzing applications increases, such as the analysis of bio-informatics data, networking data, social network data and so on. But an important challenge is that the real-world datasets are often incomplete, and various reasons cause this issue, such as the missing of data values when collect data source, and this incompleteness of data is undesirable for the analysts. Though visualization techniques greatly help the viewers to find the hidden patterns and structures in the large scale datasets, most visualization techniques fail to present the incomplete datasets.

Uncertainty visualization is a new research field which is dealing with the representation of uncertainties in datasets, and it has attracted many researchers during the last few years (Martin Theus 1997; Pang, Wittenbrink et al. 1997; Swayne and Buja 1998; Johnson and Sanderson 2003; Cyntrica Eaton 2005; Popov 2006; Feng, Kwock et al. 2010; Skeels, Lee et al. 2010). In AVI'08 conference, senior Microsoft research scientist George Robertson and his team classified the uncertainty into five different levels(Skeels, Lee et al. 2010). 1) Measurement Precision; 2) Completeness; 3) Inference; 4) Disagreement and 5) Credibility levels. Especially inn the second level (completeness level), the main concern is how to recover and represent the missing data (or missing information), which truly exist, but their specific values are unknown by

people. And the uncertainties in completeness level are quite often occurred in high dimension datasets.

Visual clutter is one of the most significant problems in visualization; especially in large scale data visualization, it seriously damages the quality of visualization in readability and understandability, and most visualization techniques need strategies to deal with this problem, which includes overcrowded displays, and fitting large volume of data in small display space. Many efforts (Laguna, Mart et al. 1997; Ankerst, Berchtold et al. 1998; Wei-xiang and Jing-wei 2001; J. Yang 2003; Peng, Ward et al. 2004; Ellis and Dix 2006; Pascale, Bruno et al. 2006; Ellis and Dix 2007; Hong Zhou 2008; Peihong, He et al. 2010; Huang and Huang 2011) have been put to deal with the visual clutter in completed data visualization. In multi-dimensional data visualization, especially in the parallel-coordinates visualization, there are three directions for the current research of visual clutter reduction: 1) dimension reduction(J. Yang 2003; Peng, Ward et al. 2004) ; 2) data clustering(Ankerst, Berchtold et al. 1998; Hong Zhou 2008; Peihong, He et al. 2010); and 3) minimization of edge crossings(Laguna, Mart et al. 1997; Wei-xiang and Jing-wei 2001; Pascale, Bruno et al. 2006; Huang and Huang 2011). The first approach is to find and filter some less important data attributes for simplifying the decision making or problem solving processes, and in this case,, the visual clutter and visual complexity will be reduced accordingly in the corresponding visual representation of the data; The second is data clustering, which is one of the data mining approaches to group data items based on a variety of rules, such as data similarity, and through data clustering, we can display some abstract clusters instead of the data details, and then the visual complexity and clutter can be reduced accordingly; The minimization of edge crossings is a problem of graph drawing. Through the optimization of geometric

positions of nodes and edges to reduce the edge crossings among edges is the main method for visual clutter reduction. Our research is focusing on the reduction of edge crossings in parallel-coordinates visualization. Visual clutter often occurs in parallel-coordinates visualization of multidimensional data (see Fig. 9) along with the growing of dimensionality and the number of data items as well. When the dimensions or number of data items grows higher, it is inevitable to display some clutters, no matter what visualization method is used(J. Yang 2003).

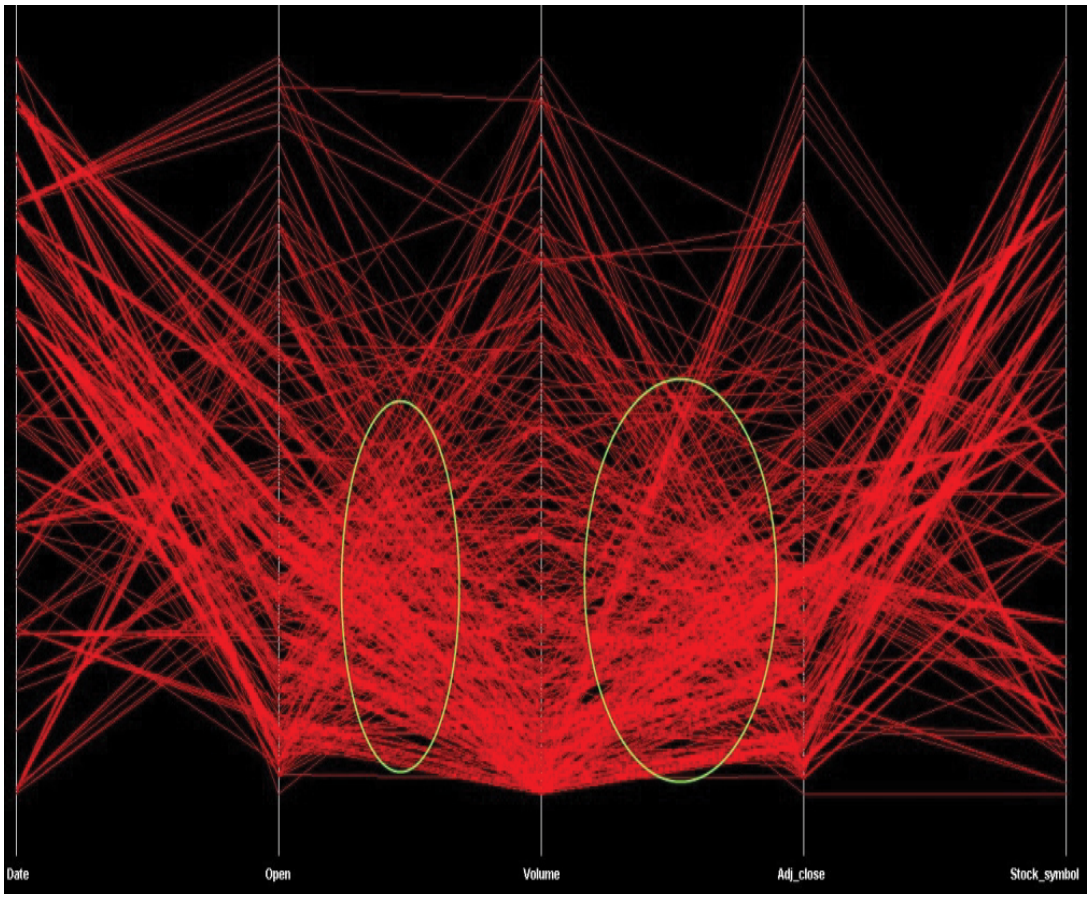


Figure 9. Examples of visual clutter in parallel coordinates visualization. See regions bounded by two ellipses that contain a large number of edge crossings.

In this chapter, we also use Sugiyama algorithm, which is known as layered graph drawing method and is one of the most effective methods that can be used to reduce edge crossings in visualization. It forms all edges into polylines and places all vertices

on a number of horizontal layered for optimization. It produces clear and intelligible layouts of hierarchical digraphs theoretically and heuristically. Vertices of each layer are reordered to reduce crossing numbers while holding the vertex orderings on the other layers. We select this method in our parallel-coordinates visualization because we could naturally transplant the polyline property from Sugiyama layout method into the parallel-coordinates visualization. In this thesis, we propose a novel method for clutter reduction of incomplete multi-dimensional data through the reduction of edge (polyline) crossings.

2.2 NEW ALGORITHM FOR CLUTTER REDUCTION

In parallel coordinates visualization, we use m vertical axes to represent a $m-$dimensional space. The ordering of these axes in the visualization is a random set. However, as the axes ordering changes, the drawing of polylines will also be changed. Different orders of axes always reveal different visual structures in multi-dimensional datasets. Wei Peng etc. (Peng, Ward et al. 2004) defined the visual clutter in parallel coordinates as the proportion of outliers against the total number of data points, and they tried to use the exhaustive algorithm to find the optimal axes order for minimizing the member of edge crossings (or visual clutter). Different from Wei Peng's re-ordering approach, we propose a new method to permute the ordering of polylines (the visual representation of data items) to minimize the number of edge crossing. As we focus on the incomplete dataset, which some values are missed in particular axes, it is possible to alter the positions of polylines, and because the crossing points between those polylines with missing (or uncertain) values and the corresponding axes are changeable,we could optimize the positions of those crossing points to achieve the reduction of edge crossing

(or visual clutter).

Many sources could produce uncertain (or missing) values, such as uncollected data, data source confidentiality, redefined data categories, mutually exclusive multivariate combinations and uncertainty deemed excessive(Cyntrica Eaton 2005), and normally they can be classified into non-numerical values and numerical values. In this section, we will discuss our new approaches to handle these data. Here all the algorithms are considered to be used for two-dimensional datasets. And these methods can be easily extended to the multi-dimensional datasets.

Supposed that the data values in the second axis of parallel coordinates are non-numerical data, we can permute the order of all the crossing points (vertices) in this axis to minimize the number of edge crossings between two neighboring axes, and then determine the best positions of data values to reduce the visual clutter and increase the readability effectively. In fact, if the data values in the second axis are numerical data, not all the data points can be reordered because of their unequal relations and practical meanings, whereas the new algorithms can be used to permute the order of missing data. The following is the theoretical algorithm for the optimal positioning of vertices for reducing the visual clutter in parallel coordinates.

Algorithm 1. Determination of positions of incomplete data

In our methods, we use a small "circle" to create a dummy vertex for presenting an uncertain value $v_{i,j}^{e}$ of the i^{th} data item in the crossing point between polyline $L(i^{th})$ and

the j^{th} axis. The position of the small "circle" could be permuted. This feature provides a great opportunity to optimize the positions of these dummy vertices for the reduction of edge crossings (or visual clutter).

Suppose that a $n-$dimensional parallel coordinates' system is given, where the points on the first axis are fixed and some of the vertices are dummy vertices that represent the missing or uncertain values with changeable positions in the second axis. We assume that all the values in the second axis are non-numerical in our visualization. Given a value set of the i^{th} multi-dimensional data item is

$$V_i = (v_{i,1}, v_{i,2}, \cdots, v_{i,n}), i = 1,2,3,\cdots,m,$$

where m is the total number of data items. The j^{th} axis (or dimension) in parallel coordinates is defined to be ξ_j, $j = 1,2,3,\cdots,n$ while $v_{i,j}$, $j = 1,2,3,\cdots,n$ is the value of the i^{th} data item in the crossing point between polyline $L(i^{th})$ and the j^{th} axis. Suppose that $v_{i,j}^e \in \{v_{i,j}, i = 1,2,3,\cdots,m;\ j = 2,3,\cdots,n\}$ behaves that the value of $v_{i,j}$ is uncertain, see the example visualization in Fig. 10.

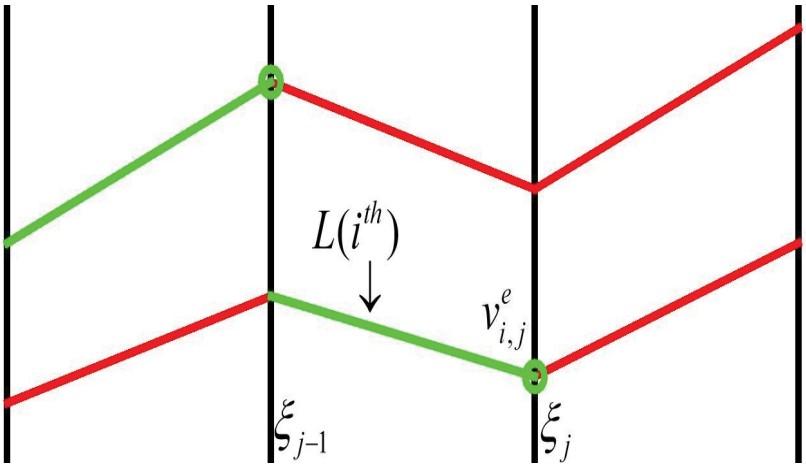


Figure 10. The example display of uncertain values visualized in parallel coordinates. The lines in red and green behave realistic and uncertain data respectively. The dummy vertices are shown by green circle.

Step1: Initial positions of dummy vertices

We first equidistantly map vertices with certain values of all data items to axis in the visualization. For dummy vertices with uncertain values, we initially position each of them at the crossing point of its axis and use a horizontal straight line to connect the vertex of neighboring axis on the left hand side.

If we draw a straight line between these two points, we could find that the angle between this straight line and two neighboring axes is 90°. This is exactly expected by Huang, etc.(Huang and Huang 2011)in their evaluation result, which is using maximization of crossing angles to increase the readability of graphs. In our experiments, we also examine the rationality and correctness of graphs.

Suppose that s_j denotes the total number of uncertain values in the j^{th} dimension ξ_j. The initialized positions $v^e_{i,j} = [v^{\circ}_{i,j}]$ of uncertain values can be calculated by the following nonlinear equation system:

$$v^{normal}_{i,j-1} = v^{\circ}_{i,j} \Big/ \sqrt{\sum_i (v^{\circ}_{i,j})^2 + \sum_i (v^{normal}_{i,j})^2}, \quad i = 1, 2, \cdots, s_j \qquad \textbf{eq. 1}$$

Here we assume that the datasets in parallel coordinates are visualized after the values are normalized into the interval $[0, 1]$. $v^{normal}_{i,j-1}$ and $v^{normal}_{i,j}$ denote the normalized complete values in $(j-1)^{th}$ and j^{th} axes separately.

Step2: Suboptimum positions of dummy vertices

In general, the number of uncertain values is always less than certain values in data sets.

Therefore, we could optimize the positions of dummy vertices (uncertain value $v^e_{i,j}$) by altering their initial positions. We calculate the number of edge crossings through the following formula and assign a suboptimum position for the dummy vertex with minimum number of edge crossings with other polylines:

$$\arg\min_{0\le e\le s_j} \sigma(v^e_{i,j}) \qquad \textbf{eq. 2}$$

where

$\sigma(\cdot)$: the function of edge crossing number.

In the above function, the independent variable is the number of edge crossings with the straight line $L(v_{i,j-1}, v^e_{i,j})$ and other straight lines between $(j-1)^{th}$ and j^{th} neighboring axes.

Step3: Reducing edge crossings to reach optimization

In this step, we permute the order of vertices to minimize the number of edge crossings between each pair of the neighboring axes using Penalty Minimization method (PM method), which is proposed by Sugiyama in (Sugiyama, Tagawa et al. 1981). The method is summarized below:

Let $P(\xi_{j,}\ \xi_{j+1})$ denotes a set of possible pairs of values (or positions) in two neighboring axes ξ_j and ξ_{j+1}, where some of the values in ξ_{j+1} are uncertain and the corresponding vertices are dummy. However, after the implementation of step 2 all dummy positions in ξ_{j+1} are sub-optimized. We consider the further permutation of the order of straight lines linking each pairs of vertices between axes ξ_j and ξ_{j+1}.

Step 3.1 Get the map m and the matrix realization M of the data items between the two coordinates, and calculate the number of edge crossings;

Step 3.2 Calculate the penalty digraph D for ξ_{j+1} ;

Step 3.3 Obtain all the strongly connected components in penalty digraph D ;

Step 3.4 Eliminate all the cycles in the strongly connected component, and get the minimum feedback arc set;

Step 3.5 Get the optimal orders of the vertices in ξ_{j+1} by reversing the directions of straight lines;

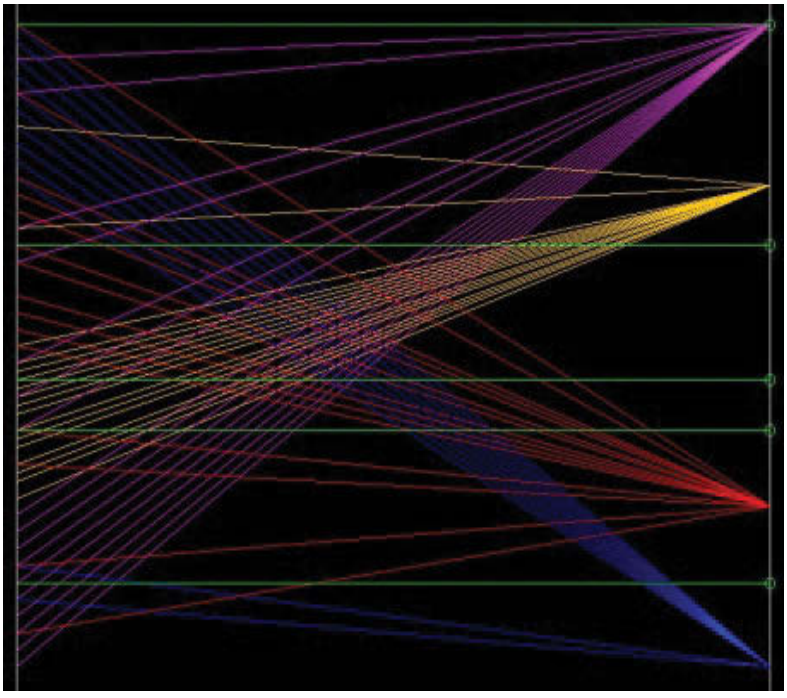
Step 3.6 Form the new map m' and matrix realization M' for the solution, and calculate the reduced crossing number.

2.3 CASE STUDIES

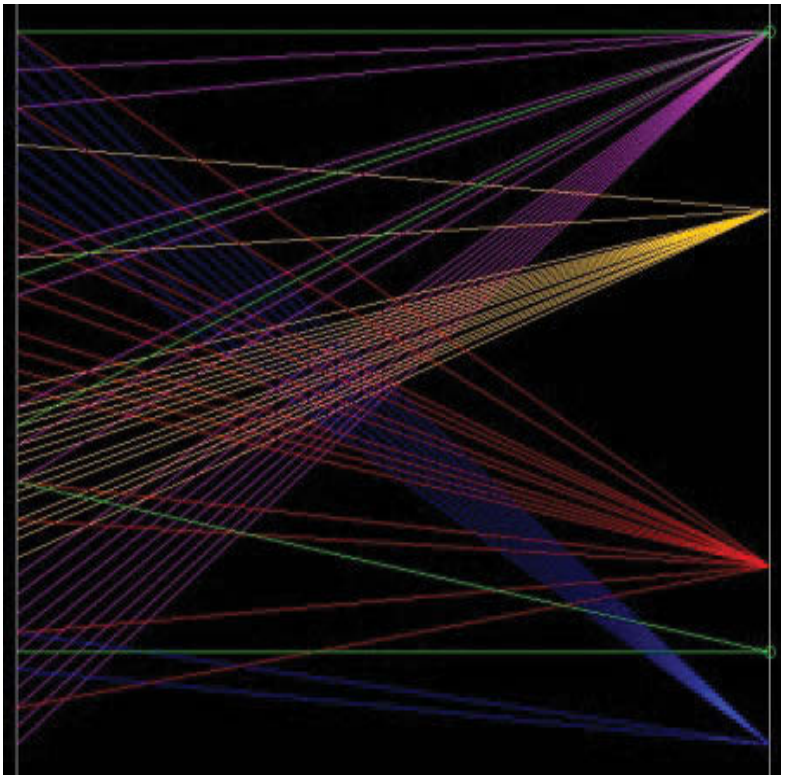
In this section, we will explain the case studies in incomplete random and real world datasets. The experiments were implemented in java and run on a standard Microsoft Windows 7 desktop PC with an Intel(R) Core(TM) 2 Duo CPU @ 2.16 GHz, 2.0GB memory. The comparable visualization results are illustrated in Fig. 11, 12 and 13, and the measurements of clutter reduction are summarized in Table 1 and 2. These examples will demonstrate the effectiveness of our clutter reduction approaches in multi-dimensional visualization of uncertainty.

Taking data with two-dimension as the simplest multidimensional example, the experiments are conducted into two cases. First we use a random dataset (See Table 1) with five missing values in the second dimension ξ_2 , where all the values are non-numerical, and there are mainly three steps to complete data in this case. In step 1,

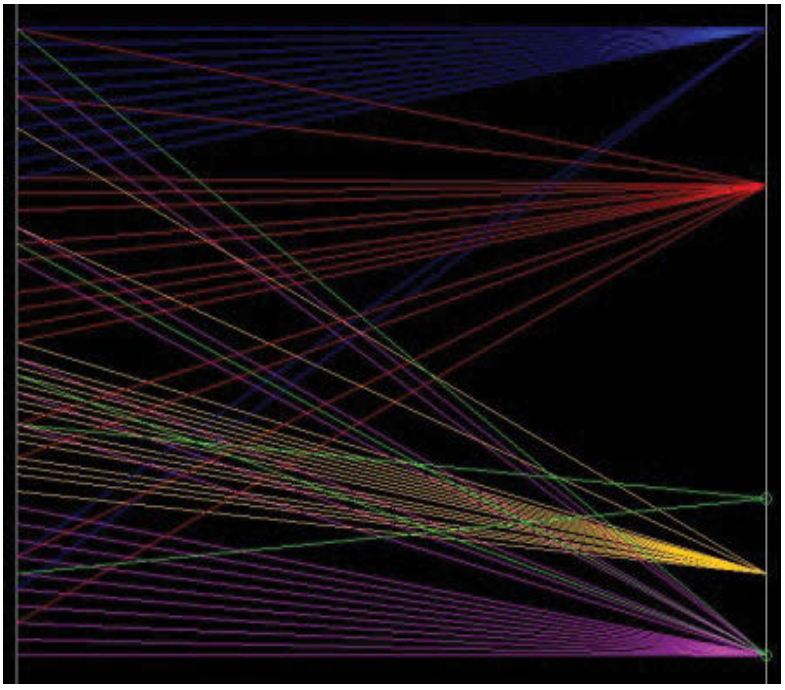
the initial representation indicates four groups of clustered polylines, refer to Fig. 11(a)which shows the initial positions of complete data in polylines colored with pink, red, yellow and blue and incomplete data colored with green, which were mapped from corresponding values in each dimension. And then Fig. 11(b) illustrates the suboptimum result of incomplete data after step 2 of our new method, and the number of edge crossings reduces to 1256 from the original 1312, the final step is by the use of permuting the vertices order in the second axis (see in Fig. 11 (c)), the number of edge crossings drops by 59.5% to 532. In the final visualization result, the number of dummy vertices has merged from 5 to 2. In the second example, we applied new method into stock market analysis. The dataset used is historical stock prices for AMEX stocks beginning with the letter A (AMEXA dataset), with two dimensions named "Stock_symbol" and "Volume". Volume attribute contains numerical values with ten unknown values. The comparison of initial and final visualization results is presented in Fig. 12(a) and (b). The number of edge crossing reduces from 519 to 464 and the number of dummy vertices decreases from 10 to 3.



(a)

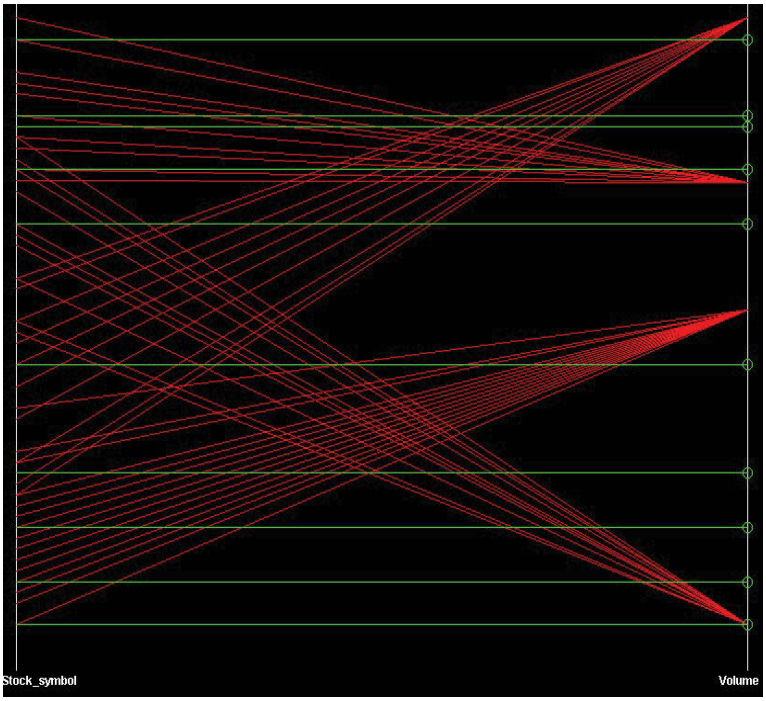


(b)

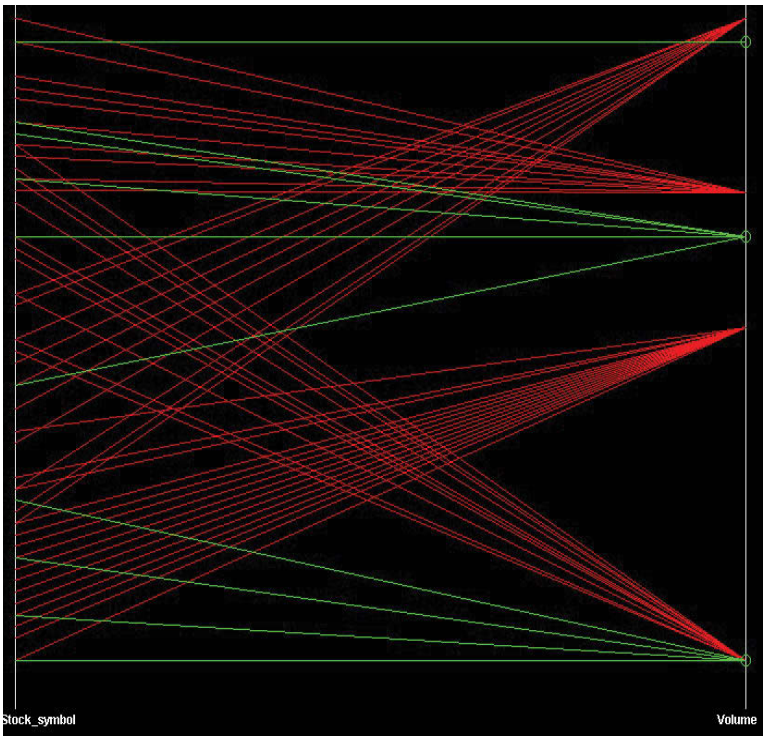


(c)

Figure 11：Case 1 - Random data in parallel coordinates: (a) The initialization of incomplete data items; (b)Visualization of suboptimum positions of uncertain values; (c) The optimal positions of vertices.



(a)



(b)

Figure 12. Case 2 - An incomplete dataset AMEX A visualized in parallel coordinate visualization: (a)The initial drawing of the given data with ten uncertain values; (b) The new drawing of the same given data after the implementation of our optimization method. The data source is available at: http://davis.wpi.edu/xmdv/datasets/amexa.html.

To apply the method into more complicated scenario, the third experiment has been extended in real data with more dimensions. Fig. 13 shows a visual performance for incomplete Forbes list in 1994(Forbes94 dataset), which contains five dimensions. In this case, we assume that the order of dimensions is fixed from company CEO to wide industry, industry, city of birth and age of graduate. The last dimensional attribute is numerical, while the rest are non-numerical. There are 5, 10, and 20 uncertain values in the last three dimensions ξ_3, ξ_4, ξ_5 respectively. As a result, total 35 dummy vertices on axes have reduced to 7. The clutter reduction has been achieved by decreasing the number of edge crossings from 7979 to 4176. The individual reduction measurement of each pair of dimensions $(\xi_j, \xi_{j+1}), j = 1,2,3,4$ has been detailed in decomposition Table 2.

Review the final visualization results, we found that：

- The approach merged dummy vertices into fewer positions on every axis, linked by clustered polylines in parallel coordinates, which significantly reduce the noise of incomplete data items with missing values.
- The number of edge crossing become fewer and the visual quality has been improved, as edge crossing numbers is an important attribute for visualization aesthetics. Therefore, the clutter has been reduced for both complete and incomplete data items in visualization.

In addition, Table 1 and 2 provide following insights:

- In the first case, the total reduction percentage is almost 60% much higher than 10.6% in the second case. In decomposition of the third case, the reductions of (ξ_1, ξ_2) (ξ_2, ξ_3) (ξ_3, ξ_4) are also greater than (ξ_4, ξ_5) . The reason is that non-numerical values have more opportunities to be re-ordered, so the order of

incomplete numerical data is relatively fixed. Therefore, for numerical attributes, step 3 of the method may not be necessary in most cases.

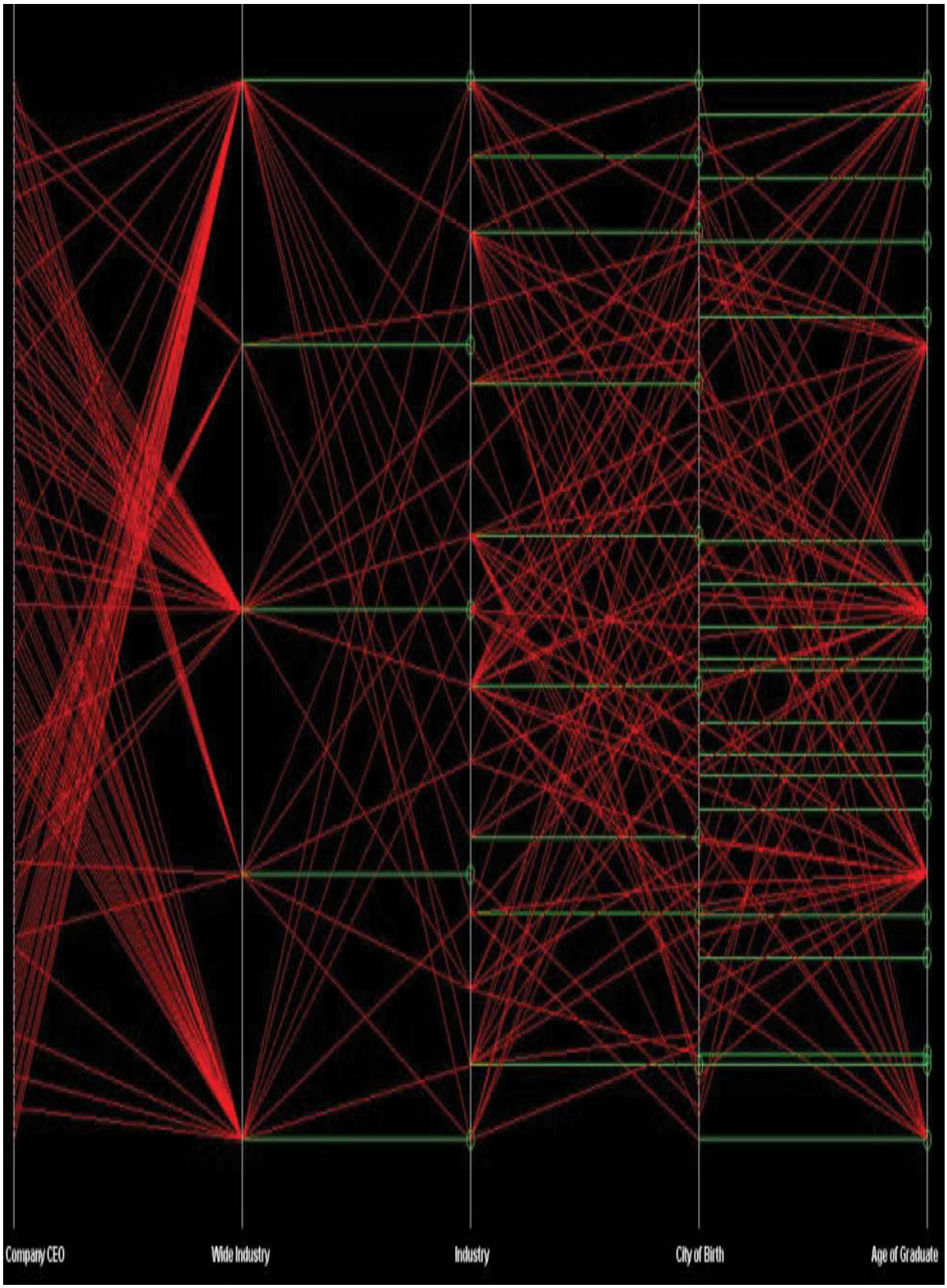
- Accordingly, the comparison results of clutter reduction reveal that multi-dimensional data with non-numerical values has more potential to reduce clutter than the data with numerical values.
- The number of missing values accordingly increases from 5, to 10 and to 20, meanwhile, the clutter reduction of (ξ_1, ξ_2) (ξ_2, ξ_3) (ξ_3, ξ_4) (ξ_4, ξ_5) is decreased respectively. This illustrates that the number of uncertain values is an important attribute to significantly affect visualization quality as well.

Table 2 Clutter reduction using optimal ordering algorithm for all cases.

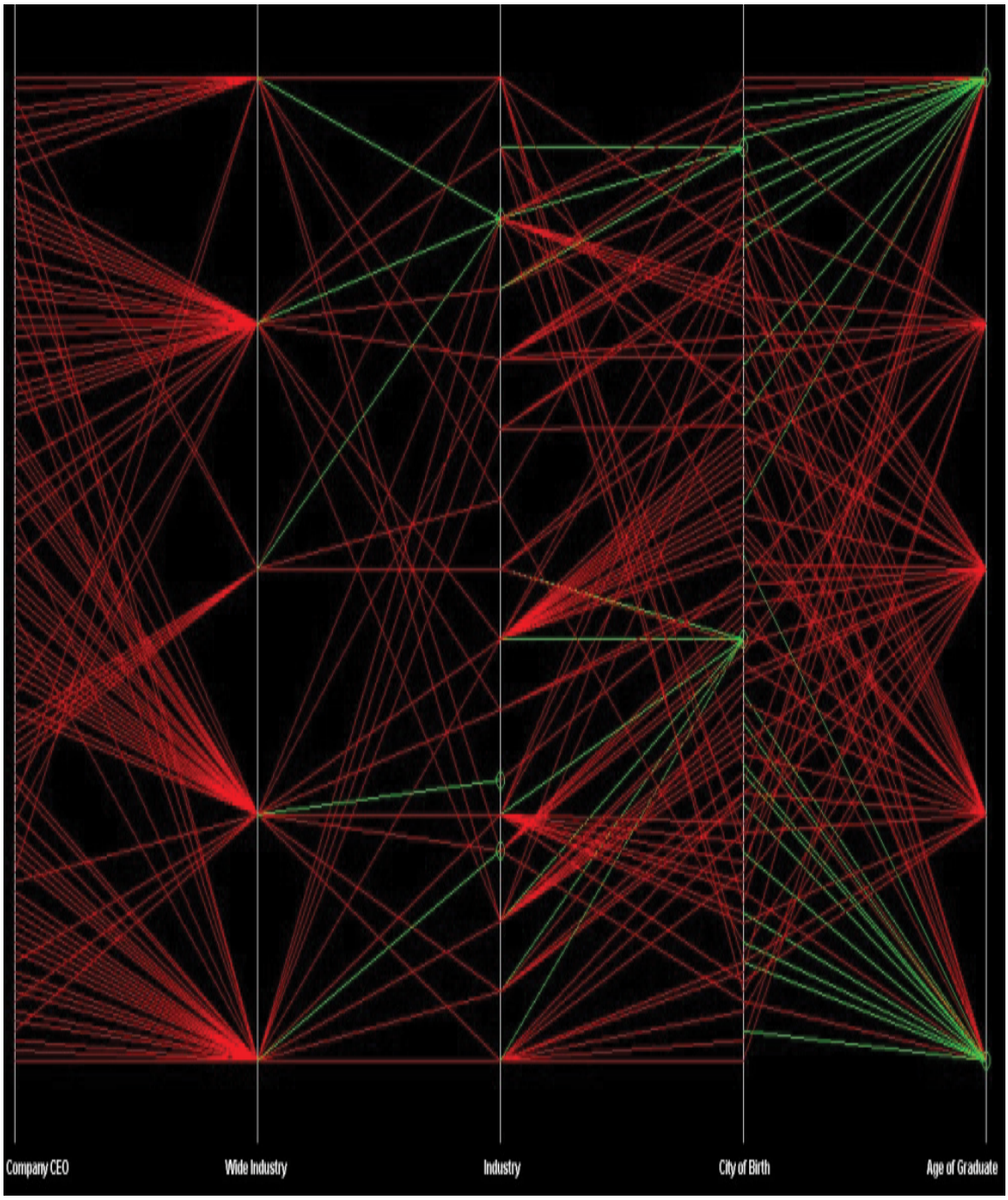
Data properties	Case				Random	AMEX	Forbes94		
	Data type				Non-numerical	Numerical	Combination		
	Dimensionality				2	2	5		
Uncertain vertex measurement	Uncertain attribute								
	Uncertain	attribute value			5	10	5	10	20
	Final dummy vertex				2	3	3	2	2
	Uncertain	vertex	reduction	%	60%	70%	80%		
Edge crossing measurement	Step1				1312	519	7979		
	Step2				1256	464	N/A		
	Reduction after step2				56	55	N/A		
	Reduction	%			4.30%	10.60%	N/A		
	Step3				532	464	N/A		
	Overall	reduction			780	55	3803		
	Overall	%			59.50%	10.60%	47.70%		

Table 3 Clutter reduction in dimension decomposition for Case 3 Forbes94.

Forbes94 Case study							
Dimension decomposition			Each dimension clutter reduction measurement				
Dimension	Data type	Uncertain value	Edge crossing number	After optimization	Crossing Reduction	Reduction in dimension %	Overall reduction
(ξ_1, ξ_2)	Non-numerical	0	2329	814	1515	65.0%	19.0%
(ξ_2, ξ_3)	Non-numerical	5	2087	636	1451	69.5%	18.2%
(ξ_3, ξ_4)	Non-numerical	10	2024	1257	767	37.9%	9.6%
(ξ_4, ξ_5)	Combination	20	1539	1469	70	4.5%	0.9%
Total	Combination	35	7979	4176	3803	N/A	47.7%


Company CEO
Wide Industry
Industry
City of Birth
Age of Graduate

(a)



(b)

Figure 13: Case 3 - Forbes 94, a dataset with 5 variables visualized in parallel coordinate visualization: (a) Original plot; (b) after clutter reduction. Data from http://www-stat.wharton.upenn.edu/ waterman/fsw/datasets/txt/Forbes94.txt

2.4 SUMMARY

In this section, we leverage a layered directed graph drawing algorithm into parallel coordinates for visualization of uncertainty. In the first stage, a nonlinear equation system is deployed to obtain the initial positions for all uncertain values. In the second stage, a multi-objective optimization algorithm is adapted to relocate positions for the dummy vertices. In the final stage, the penalty minimization method finalizes ordering of all vertices.

The case studies showed the clutter reduction among polylines and demonstrated the effectiveness of the method with better visual structure in parallel coordinates visualization. These experiments also illustrated that the number of edge crossings, uncertain values and the attributes of multi-dimensional data could play important roles in affecting visualization performance.

In order to advance technology towards information visualization of uncertainty, formal evaluation needs to be conducted in the future. Based on the evaluation results, we will not only work on visual quality but also further modify the algorithm in order to reduce complexity of computing time for larger datasets.

Chapter 3. NEW AXES RE-ORDERING METHOD IN PARALLEL COORDINATE PLOTS

3.1 SIMLARITY MEASURE AND DIMENSION RE-ORDERING METHODS

Multi-dimensional data analysis is becoming a commonplace as the number of applications increases, such as statistical and demographic computation, digital libraries and so on. However, traditional visualization techniques for these datasets usually require dimensionality reduction or selection to generate the meaningful visual representations. Dimensionality reduction, as Sara Johansson et.al pointed out, is always employed prior to visualization for dealing with the data with a large number of attributes(Johansson and Johansson 2009). Currently, many dimensionality reduction methods are able to preserve the information inside the data as much as they can by removing some less relevant data items or attributes from the original dataset. While dimension selection is mainly referred to dimension re-ordering which means that the corresponding axe of the dimension in a parallel coordinate visualization can be positioned in accordance to some effective rules such as similarity of dimensions to achieve good visual structures and patterns. This chapter focuses on the dimension re-ordering rather than dimension reduction to address the problems of visual clutter and computational complexity.

3.1.1 SIMLARITY MEASURE

In 1998, Mihael Ankerst et al. [2] presented a method which uses the similarity of dimensions to improve the quality of visualization of multidimensional data, using global and partial similarities for one or two-dimensional visualization methods. Pearson's Correlation Coefficient (PCC) is one of the most commonly used measurements for measuring similarity between two dimensions. PCC can be used for dimension reduction, clutter reduction and clustering et al. in visualization. At the same time, it has also been proved that the PCC based re-ordering problem is a NP-complete problem. Therefore, many researchers have applied heuristic algorithms to find out an optimal order of axes (or dimensions) in multi-dimensional visualization.

Though the traditional heuristic algorithms can help in finding an optimal order of axes for one- or two-dimensional visualizations, most studies have not done a deeper investigation on how to determine the first dimension (the most significant dimension) in multi-dimensional data visualizations. The first dimension always attracts much more user's attention than the others. Therefore, we may consider the one with the highest contribution rate as the first dimension to simplify the traditional similarity-based re-ordering methods and to find out the optimal order of parallel axes in a short time period.

In the section, we firstly propose a method based on the Singular Value Decomposition (SVD) to find out the contribution of each dimension in the dataset.

And then, we present a similarity-based re-ordering method in parallel coordinates which is based on the combination of a Nonlinear Correlation Coefficient (NCC) and the SVD algorithms, which is sensitive to any relationship, including the linear dependence (Zhiyuan, Qiang et al. 2011). In our experiments, we show the effectiveness of our new method by visualizing the patterns and enlarging the mean crossing angles for better visual representation. It should be noticed that the proposed method can be easily applied to the other visualization techniques.

3.1.2 DIMENSION RE-ORDERING METHODS

An effective way to improve the quality of multi-dimensional visualizations is to re-order the dimension axes in parallel coordinates based on similarity of data attributes. In this section, we summarize the previous research works done in the area of high-dimensional visualization.

Parallel coordinates(Inselberg 1985; Wegman 1990), scatter plot matrix(Becker and Cleveland 1987), table lens(Rao and Card 1994) and pixel-oriented display(Keim 2000) et al. are well-known and accepted visualization techniques for high-dimensional datasets. Similarity measure as one aspect of quality metrics in high-dimensional data visualization has been addressed in the past few years (Johansson and Johansson 2009; Tatu, Albuquerque et al. 2009; Dasgupta and Kosara 2010; Bertini, Tatu et al. 2011; Tatu, Albuquerque et al. 2011). It is worth noting that

Enrico Bertini et al.(Bertini, Tatu et al. 2011) systematically presented an overview of quality metrics in many visualization techniques through a literature review of nearly 20 papers and considered correlation between two or more dimensions to be the main characteristic of similarity measure. Sara Johansson(Johansson and Johansson 2009) introduced a weighted quality metrics to their task-dependent and user-controlled dimensionality reduction system, where small correlation values were ignored to reduce the dataset that preserved the important structures within the original dataset. Andrada Tatu et al. proposed similarity-based function for classified and unclassified data based on Hough Space transform on the resulting image of parallel coordinates (Tatu, Albuquerque et al. 2009; Tatu, Albuquerque et al. 2011). Aritra Dasgupta et al. (Dasgupta and Kosara 2010) introduced binned data model and branch-and-bound algorithm as the screen-space metrics for parallel coordinates to reduce the computations and find the optimal order of axes.

To enhance the high-dimensional data visualization, some studies on dimension reordering have been done to find good axes layouts in visualization techniques both in one- or two-dimensional arrangement(J.Bertin 1983; Johansson and Johansson 2009; Tatu, Albuquerque et al. 2009; Dasgupta and Kosara 2010; Bertini, Tatu et al. 2011) (Hahsler, Hornik et al. 2008; Hurley and Oldford 2010) (Ankerst, Berchtold et al. 1998; Friendly and Kwan 2003; J. Yang 2003; Peng, Ward et al. 2004; Artero, Oliveira et al. 2006). Mihael Ankerst et al. (Ankerst, Berchtold et al. 1998) defined similarity measures which determined the partial or global similarity of dimensions

and argued that the reordering based on similarity could reduce visual clutter and do some help in visual clustering. Wei Peng et al.(Peng, Ward et al. 2004) introduced the definition of the visual clutter in parallel coordinates as the proportion of outliers against the total number of data points and they tried to use the exhaustive algorithm to find the optimal axes order for minimizing the member of edge crossings (or visual clutter). As mentioned in (Artero, Oliveira et al. 2006), the computational cost $o(n \cdot n!)$ hampers applications of this technique to large high dimensional data sets. Almir Olivette Artero et al. (Artero, Oliveira et al. 2006) introduced the dimension configuration arrangement based on similarity to alleviate clutter in visualizations of high-dimensional data. They proposed a method called SBAA (Similarity-Based Attribute Arrangement), which is a straightforward variation of the Nearest Neighbor Heuristic method, to deal with both dimension ordering and dimensionality reduction. Other studies have been done on the dimension reordering based on the similarity(Friendly and Kwan 2003; J. Yang 2003; Tatu, Albuquerque et al. 2011) (Guo 2003) (Albuquerque, Eisemann et al. 2010). Michael Friendly et al.(Friendly and Kwan 2003) designed a framework for ordering information including arrangement of variables. However, the arrangement of variables is decided mainly according to the users' desired visual effects. J. Yang et al. (J. Yang 2003) established a hierarchical tree structure over the attributes, where the similar attributes were positioned near each other. Diansheng Guo (Guo 2003) developed a hierarchical clustering method, which was based on comparison and sorting of dimensions by use of the maximum conditional entropy. Georgia Albuquerque et al. (Albuquerque,

Eisemann et al. 2010) introduced the quality measures to define the placement of the dimensions for Radviz and also to appraise the information content of pixel and Table Lens visualizations.

To sum up, most of the current dimension reordering methods are based on Pearson's correlation coefficient. From the statistics point of view, PCC is a method for measuring the linear correlation between the two random variables. Therefore, it is partial that we reorder the dimensions according to their similarity only depending on the calculation of PCC. Though Pargnostics, proposed by Aritra Dasgupta et al. in(Johansson and Johansson 2009), is the most similar with our approach, the probability and joint probability during the computational process are both denoted as their special axis histograms, which lack the support by mathematical theories. Moreover, it can be seen from the definition of the mutual information that it does not range in a definite closed interval as the correlation coefficient does, which ranges in $[-1,1]$.

Hence, it is of great importance that a comprehensive and useful method should be proposed for correlation analysis among the dimensions for conveying better visual structures and patterns. In this thesis, we propose similarity-based reordering method for dimensions reordering in parallel coordinates to solve the above problems.

3.2 NEW APPROACH FOR DIMENSION RE-ORDERING

The ordering of dimensions has large impact on how easily we can perceive different structures in the data(Ankerst, Berchtold et al. 1998). Completely different displays and conclusions may be obtained if we interactively switch between different dimension reordering. How to reorder the dimensions in high-dimensional datasets meaningfully is one of the most significant problems of the researches on quality metrics in data visualization due to its influences on the quality of visualization in terms of readability and understandability. In this chapter, we visualize them in a more rational way rather than arrange them only according to the empiricism.

Throughout this chapter the following notation is used: a dataset D is composed of n dimensions (variables) with m data items for each one. In some cases we need to measure the statistical characters between the two dimensions X and Y, where

$X = (x_1, x_2, \cdots, x_n)^T$, $Y = (y_1, y_2, \cdots, y_n)^T$.

3.2.1 Linear/Nonlinear Correlation

The correlation of two variables (dimensions/attributes) is a statistical technique that can indicate the magnitude relationship between the two variables. It also shows how the two variables interact with each other. In this section, we present the reordering methods based on the two correlation measures: Pearson's correlation and nonlinear

correlation information measures.

Pearson's Correlation Coefficient (Rodgers and Nicewander 1988), as one of the most popular similarity measures in visualization of multidimensional data, is a linear correlation measurement for each pair of random variables:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \qquad \textbf{eq. 3}$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$, $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$, $\bar{x}$ and $\bar{y}$ behave the mean of variables X and Y respectively. The value of PCC ranges in the closed interval [-1, 1], which indicates the linear correlation degree of the two variables. When the PCC value is close to 1 or -1 it denotes a strong relationship and if it close to 0 it means a weak relationship between the two variables. A positive and negative correlation coefficient denotes that both variables are in the same way or in the opposite way

Although, linear correlation can detect the relationship between two dependence variables, in the real world the correlations can also be nonlinear. Mutual Information can be thought of as a generalized correlation analogous to the linear correlation coefficient, but sensitive to any relationship, not just linear correlation. Moreover, NCC is a method that can measure nonlinear relationship based on mutual information(Matsuda 2000; Zheng Rong and Zwolinski 2001) and redundancy(Drmota and Szpankowski 2004), which is sensitive to any relationship, not just the linear

dependence(Zhiyuan, Qiang et al. 2011). Zhiyuan Shen et al.(Wang, Shen et al. 2005; Zhiyuan, Qiang et al. 2011) did further researches on the effects of statistical distribution to it and made it range in a closed interval $[0,1]$.

Corresponding to the literature(Ankerst, Berchtold et al. 1998), we mainly apply NCC to compute the partial similarity measures of dimensions in multidimensional data visualization, while SVD is used for measuring the global one. We introduce the detailed NCC in the following paragraphs. Mutual information plays an important role in the computation of NCC, which is defined as

$$I(X;Y) = H(X) + H(Y) - H(X;Y) \qquad \textbf{eq. 4}$$

where $H(X)$ is the information entropy of variable X:

$$H(X) = -\sum_{i=1}^{n} p_i \ln p_i$$

$H(X;Y)$ is the joint entropy of the variables X and Y:

$$H(X;Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \ln p_{ij}$$

p_i denotes the probability distribution that random variable X takes the value x_i, and p_{ij} denotes the joint probability distribution $p(X = x_i, Y = y_i)$ of the discrete random variables X and Y.

After revising joint entropy of the two variables X and Y,

$$H^r(X;Y) = -\sum_{i=1}^{b}\sum_{j=1}^{b} \frac{n_{ij}}{n} \log_b \frac{n_{ij}}{n} \qquad \textbf{eq. 5}$$

in which $b \times b$ rank grids are used to place the sample pairs $\{(x_1, y_i)\}_{1 \le i \le n}$. n_{ij} is the number of samples distributed in the $ijth$ rank grid, Wang et al. (Wang, Shen et al. 2005) proposed the calculation method for nonlinear correlation coefficient as follows:

$$NCC(X;Y) = H^r(X) + H^r(Y) - H^r(X;Y)$$
$$= 2 + \sum_{i=1}^{b}\sum_{j=1}^{b} \frac{n_{ij}}{n} \log_b \frac{n_{ij}}{n} \qquad \textbf{eq. 6}$$

In the following section 3.2.2, we apply the above formula to measure the linear or nonlinear relationship between the two dimensions in multidimensional datasets because of its sensitivity to any relationship.

3.2.2 Similarity-based Reordering

Since the problem of dimension reordering is similar to the Traveling Salesman problem, many researchers applied heuristic algorithms, such as genetic algorithms, colony optimization and nearest neighbor heuristic method etc. (Ankerst, Berchtold et al. 1998; Artero, Oliveira et al. 2006), to overcome exhaustive time. In the method SBAA proposed by Almir Olivette Artero et al.(Artero, Oliveira et al. 2006), the largest value $s_{i,j}$ in their similarity matrix s (lower diagonal) is considered to be the initial dimension "ij" in the new order. And then, they try to search for the dimensions which will be positioned in the left and the right of it. It seems rational that we just reorder all the dimensions in line with this similarity. However, some dimensions always attract much more concentrations from the whole visual structure. For example, in parallel coordinates, the first and the last dimensions can draw much more attention than the other axes do. Therefore, different from the existed methods, we propose a new dimensions reordering algorithm based on the NCC and SVD algorithms. These methods help users reduce the computation complexity and

improve the visual readability greatly.

We define the similarity matrix s, which is symmetric, as follows:

$$s = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix}$$

where $s_{ij} = s_{ji}\ (i \neq j)$, which are calculated by use of nonlinear correlation coefficient. $s_{ii}\ (i = 1, 2, \cdots, n)$ (We also can denote them as v_{1i}) behaves the contribution value of the ith dimension to the whole data values, which is calculated by SVD algorithm.

Algorithm 2. Similarity-based Reordering Algorithm

Step1. Form the matrix D of the data sets.

Step2. Calculate the singular value decomposition (Golub and Loan 1983) of matrix D, and get the contribution factors $s_{ii}, i = 1, 2, \cdots, n$.

Step3.Compute the other elements s_{ij} of similarity matrix s, using our nonlinear correlation coefficient method, besides s_{ii}, $i = 1, 2, \cdots, n$ which have calculated in step2.

Step4. Choose the largest value of $s_{ii}, i = 1, 2, \cdots, n$ as the extreme left attribute to start display the data sets. We denote this attribute as

S_{ll}, $l \in \{1, 2, \cdots, n\}$.

Step5. Get the largest value S_{ll} from $\{S_{li}, l < i\}$. Therefore, the $r_1 th$ attribute is appended to the lth attribute. We get the first two elements of neighbouring sequence $NS = \{l, r_1\}$.

Step6. Repeat step5 using the $r_1 th$ attribute as the left neighbouring attribute

from $\{S_{\eta i}, r_1 < i\}$ until inserting all attributes into the NS.

It is worth noting that this visualization method can not only provide us the similarities between each pair of dimensions, but also express some ideas of the self-property of each dimension. During the computation process of the nonlinear correlation coefficient, we chose the $b \times b$ rank grids according to the empirical formula, which is mentioned in (Zhuang Chu Qiang 1992):

$$b = 1.87 \times (n-1)^{2/5} \qquad \textbf{eq. 7}$$

Moreover, it is natural that we can compute the contribution rate of each dimension to the whole dataset using the following possible measure:

$$C_i = \frac{v_{1j}}{\sum_{j=1}^{n} v_{1j}} \times 100\% \qquad \textbf{eq. 8}$$

This approach not only provides us a new reordering method helping us take much more insights into the dataset but also gives rise to the following new method which can help in determination of the first dimension with the most contribution.

3.3 CASE STUDIES

To demonstrate the effectiveness of our rational dimension reordering methods, we analyzed many datasets in this section, Cars and Liver Disorders data set for our similarity-based reordering method. All of these data sets we tested come from the

literature (Irvine Machine Learning Repository).According to the literature [30], larger crossing angle the two polylines make, the less cognitive load and the better visualization efficiency is. Therefore, to show the advantages in the readability and understandability of our method, we calculated the mean angles occurred among the polylines between two neighboring attributes using the following formula:

$$mean\_angle = \frac{total\_angle}{total\_edge\ cros\sin g}.$$

3.3.1 Cars dataset

Based on the theory in section 3, the similarity matrix of Cars data set was calculated as the following S.

$$S = \begin{bmatrix} 0.0067 & 0.5950 & 0.3236 & 0.0561 & 0.9078 & 0.8104 & 0.0302 \\ 0.5950 & 0.0018 & 0.5806 & 0.5028 & 0.8944 & 0.0261 & 0.6288 \\ 0.3236 & 0.5806 & 0.0354 & 0.1313 & 0.5223 & 0.9544 & 0.0104 \\ 0.0561 & 0.5028 & 0.1313 & 0.9991 & 0.3389 & 0.6968 & 0.0302 \\ 0.9078 & 0.8944 & 0.5223 & 0.3389 & 0.0047 & 0.9598 & 0.0197 \\ 0.8104 & 0.0261 & 0.9544 & 0.6968 & 0.9598 & 0.0235 & 0.0117 \\ 0.0302 & 0.6288 & 0.0104 & 0.0302 & 0.0197 & 0.0117 & 0.0004 \end{bmatrix}$$

After positioning the first dimension “Weight”, which enjoys its significant contribution to the whole data set, we try to find out the one from the unordered dimensions with the largest similarity value to this dimension: $s_{46} = 0.6968$. Therefore, the $6-th$ dimension is considered to be the strongest correlation with the $4-th$ one. And then, we make the $6-th$ attribute to be appended to the $4-th$ one. Similar to this process, we can get the final rational dimension order, which is

$$4 \to 6 \to 5 \to 1 \to 2 \to 7 \to 3$$

Corresponding to the initial Cars dataset, the reordering dimensions calculated using our algorithm is

$$\text{Weight} \rightarrow \text{Year} \rightarrow \text{Acceleration} \rightarrow \text{MPG} \rightarrow \text{Cylinders} \rightarrow \text{Origin} \rightarrow \text{Horsepower}$$

The reordering results after our analysis are visualized in parallel coordinates in Fig. 14(a).

We visualize Car dataset using the traditional reordering method-Pearson's Correlation Coefficient in Fig. 14(b). The corresponding order of dimensions is as follows:

$$\text{Weight} \rightarrow \text{Cylinders} \rightarrow \text{Horsepower} \rightarrow \text{MPG} \rightarrow \text{Year} \rightarrow \text{Acceleration} \rightarrow \text{Origin}$$

Comparing with these two images in Fig. 14, we can find that visualization structures between the "Cylinders" and "Origin" dimensions with our method are clear and simple. In the visualization graph of NCC, the mean angle between the attributes "Acceleration" and "MPG" gets to 22.359°. Moreover, the mean angle between "Cylinders" and "Origin" attributes is 28.162°.Compared to the mean angle of the overall polylines produced in the PCC reordering method, 0.422°, the angle in NCC reordering one is 21.2 times larger than it. Therefore, we can find the visual effect of our reordering method is much better than the traditional one.

Table 3 presents the detailed comparisons between the similarity values of attributes, which are calculated using PCC and NCC. The numbers from 1 to 7 denote the dimensions: "MPG, Cylinders, Horsepower, Weight, Acceleration, Year and Origin" separately. Note that no matter which method we use, the similarities between the two

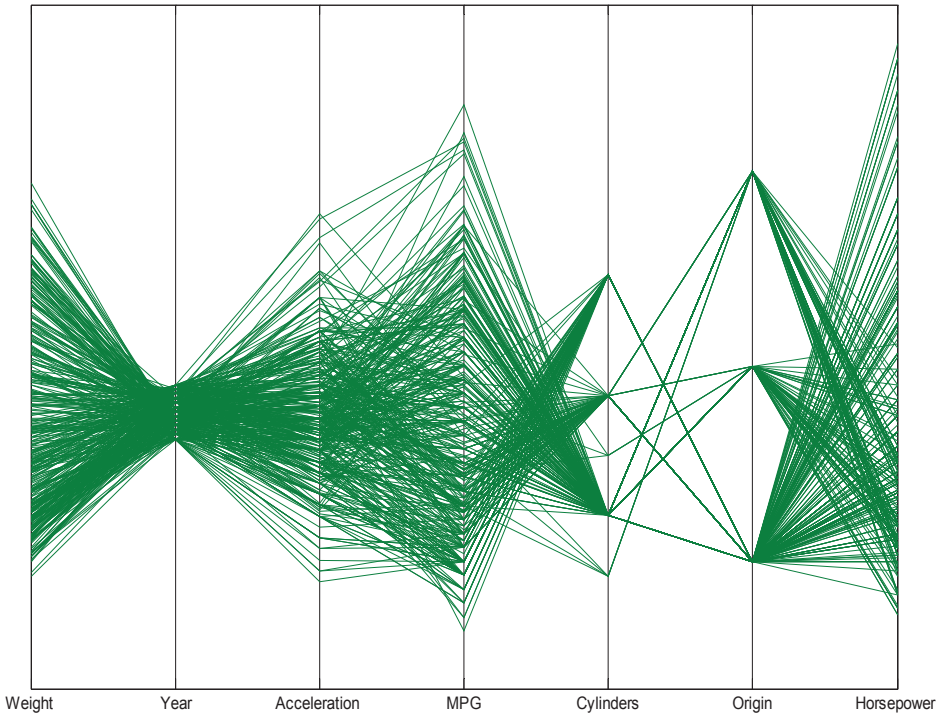
dimensions are the same, that is $s_{ij} = s_{ji}$ $(i \neq j)$. It is obvious that there are big differences between the similarity values with two methods. For example, to our knowledge, the similarity between the $3-th$ ("Horsepower") and $7-th$ ("Origin") dimensions of the dataset is not strong enough at all. However, the result of PCC is 0.4552, comparing to ours result 0.0104.

3.3.2 Liver disorders dataset

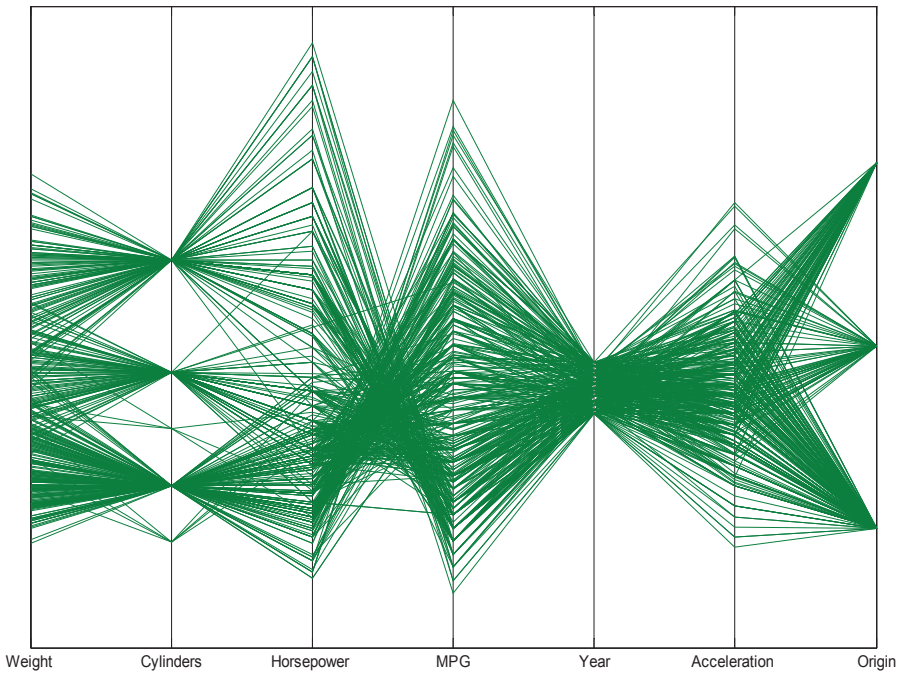
Liver Disorders dataset consists of 345 instances with 7 dimensions. Fig. 15 illustrates us the final visualization result of the whole dataset according to their similarities calculated by NCC and PCC methods respectively, where the dimension "MCV" enjoys its most significant contribution to the whole dataset and occupies the first place in the two reordering visualization.

It is easy to find that the polylines among the "SF" and "DN" attributes are much less than any others among the neighboring attributes no matter in Fig. 15 (a) or (b). The mean crossing angle of these two dimensions, 43.515°, as the largest one in the dimensions reordering visualizations as well, simplifies the visual representation greatly. The mean crossing angle of our NCC reordering method to this dataset is 12.322°, which is 3.722° larger than the result calculated using PCC method.

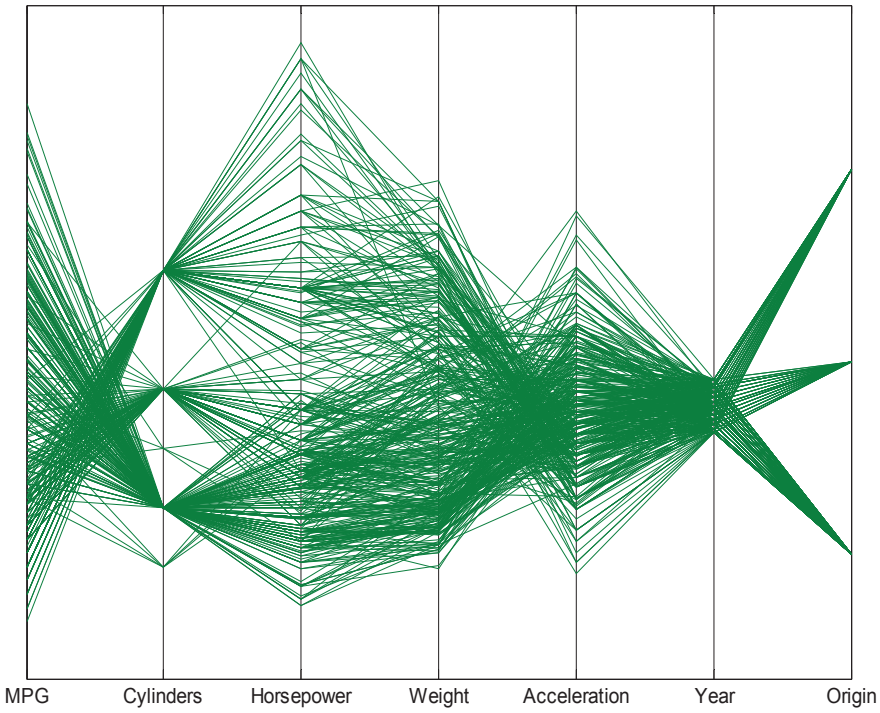
We also tested the other datasets such as Nursery, Iris et al. large scale ones to illustrate the advantages of our methods, which all showed us that our methods can enlarge the mean crossing angles for better visualization.



(a) Measurement with NCC

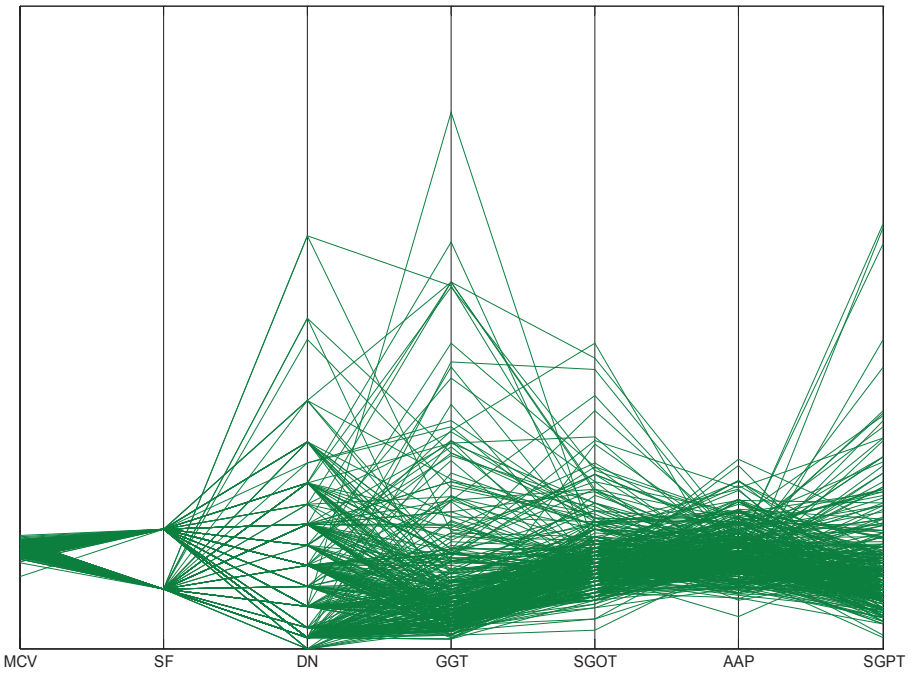


(b) Measurement with PCC.

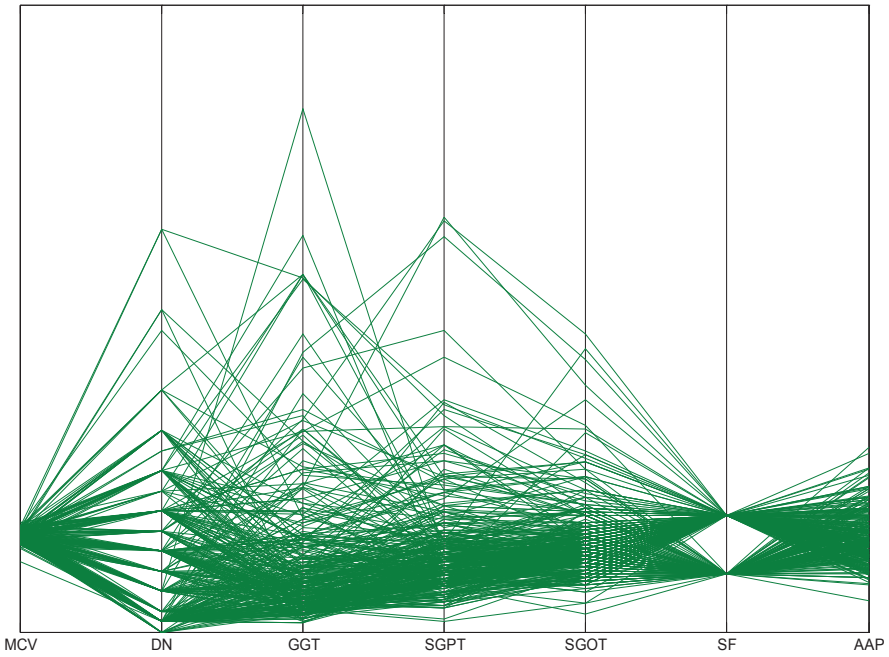


(c) Original dimension arrangement visualization

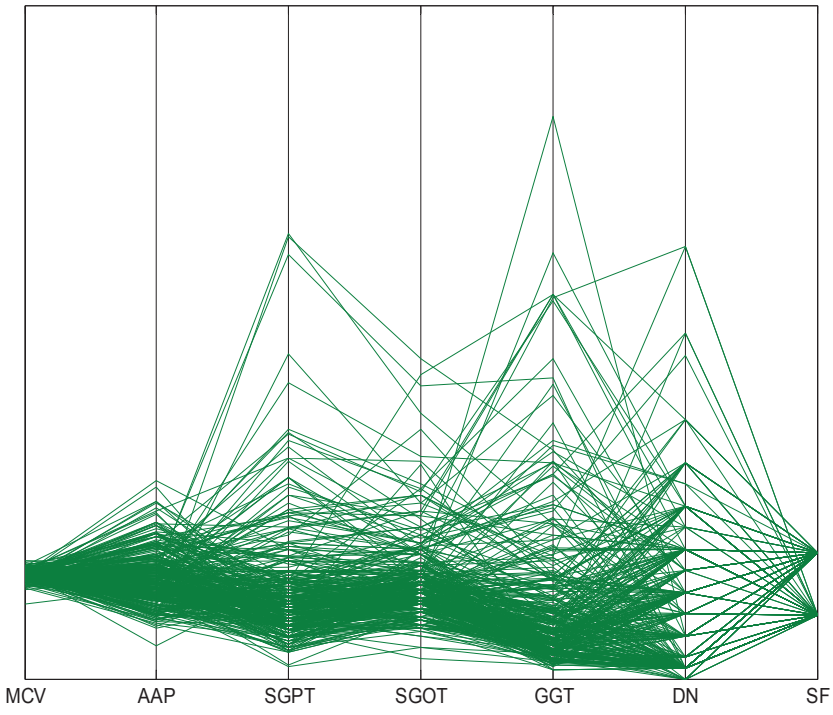
Figure 14. Cars dataset visualization in parallel coordinates.



(a) Measurement with NCC.



(b) Measurement with PCC.



(c) Original dimension arrangement visualization

Figure 15. Axes reordering visualization of Liver Disorders dataset.

Table 4 The comparison of the similarity values using PCC and NCC to Cars dataset.

PCC / NCC	2	3	4	5	6	7
1	0.7776 / 0.5950	0.7784 / 0.3236	0.8322 / 0.0561	0.4233 / 0.9078	0.5805 / 0.8104	0.5652 / 0.0302
2		0.8429 / 0.5806	0.8975 / 0.5028	0.5046 / 0.8944	0.3456 / 0.0261	0.5689 / 0.6288
3			0.8645 / 0.1313	0.6892 / 0.5223	0.4163 / 0.9544	0.4552 / 0.0104
4				0.4168 / 0.3389	0.3091 / 0.6968	0.5850 / 0.0302
5					0.2903 / 0.9598	0.2127 / 0.0197
6						0.1815 / 0.0117

3.4 SUMMARY

In this chapter, we proposed a new method to improve the readability and understandability of parallel coordinates visualization theoretically. At the first stage, we propose a new way of looking into the dimensions within datasets based on the singular value decomposition algorithm. At the second stage, we present a method, named similarity-based reordering method, for calculating the similarity between the two dimensions based on the nonlinear correlation coefficient and singular value decomposition algorithms rather than the traditional Pearson's correlation coefficient, and then visualize the optimal dimension order according to the similarity in parallel coordinates. We have conducted the experimental evaluations to demonstrate the effectiveness and rationale of our approaches: NCC reordering method enlarges the mean crossing angles of the whole data set and reduces the amount of polylines between some neighboring dimensions.

During the process of calculation for nonlinear correlation coefficient, the more exact choice of rank grids will do much more help in the speed of calculation. Therefore, we consider this problem to be our first future work. And then we will apply our methods with interactive techniques to more real-world datasets and help users analyze the datasets using visualization.

Chapter 4. USING ARCED AXES IN PARALLEL COORDINATE GEOMETRY

4.1 ANALYSIS OF PCP

The rapid growth of data communication through the Internet and World Wide Web has led to vast amounts of information available online. In addition, business and government organizations create large amounts of both structured and unstructured information which needs to be processed, analyzed, and linked. Cloud computing plays a popular and important role in providing on demand services for handling such large volumes of online datasets. Some previous research works have well done in data intensive cloud computing, especially in the field of data privacy in cloud computing (Xuyun, Chang et al. 2013; Zhang, Liu et al. 2013; Xuyun, Yang et al. 2014).

Consequently, high-dimensional data and multivariate data are becoming commonplace as the number of applications increases, such as statistical and demographic computation, digital libraries and so on. Though it can provide flexible and cost-saving IT solutions for the end users, it is much easier in causing a great deal of problems such as network and system security issues due to its sharing and centralizing computing resources.

However, there is no absolute way to secure the data and data transformations in large

scale networking systems. The existing techniques and tools of securing a network system still rely heavily on human experiences. Most of them require human involvement in analyzing and detecting anomalies and intrusions. To enhance the human perception and understanding of different types of network intrusions and attacks, network visualization has become a hot research field in recent years that attempts to speed up the intrusion detection process through the visual analytics. Unlike the traditional methods of analyzing textual log data, visualization approach has been proven that can increase the efficiency and effectiveness of network intrusion detection significantly by the reduction of human cognition process.

As pointed out in the literature (Claessen and van Wijk 2011), many methods have been proposed to provide insight into multivariate data using interactive visualization techniques. Parallel coordinate plots (PCP), as a simple but strong geometric high-dimensional data visualization method, represents N-dimensional data in a 2-dimensional space with mathematical rigorousness. PCP, together with scatterplot and the radar chart have been widely adopted for visualizing multivariate datasets (Claessen and van Wijk 2011). In this thesis, we propose an arc-based parallel coordinates visualization method, termed as arc coordinate plots (ACP). In our novel method, segments of curve, rather than the line segments, are considered as the coordinate axes. In the same coordinates system, such as Cartesian coordinates system, the length of arc is longer than the line segments if their x-coordinates are set with the same interval and thus it can visualize much more data items in the same screen space.

Furthermore, besides visualizing the original data sets to reveal the patterns, ACP can preserve much more geometric structures of the data, especially the data with circular properties. In addition, we leverage singular value decomposition algorithm to provide a new way of looking into the dimensions within datasets. We propose the contribution-based visualization method and a formula for contribution rate of each attribute. Finally, we visualize some real data sets such as KDD99 security data by use of ACP to detect the intrusions. In our system, abnormal network activities are extracted from a large volume of network flows and their patterns. By using our approach, we can easily detect the unusual patterns from network scans, port scans, the hidden scans, and DDoS attacks et al.

The rest of this chapter is organized as follows. Section 2 gives an overview of existing enhancements in PCP. Section 3 presents the arc-coordinate geometry theoretically in the novel coordinates system in detail. To overcome the curse of dimensionality in coordinates visualization methods, this section describes the attributes contribution method as well. The experimental evaluation of our new approaches is explored in Section 4. Finally, conclusions and future work are presented in section 5.

4.2 OVERVIEW OF APPROACHES ON PCP

Parallel coordinate plots, as one of the most popular methods, is proposed firstly by Inselberg(Inselberg 1985) and Wegman suggested it as a tool for high dimensional

data analysis(Wegman 1990). Coordinates of n-dimensional data can be represented in parallel axes in a 2-dimensional plane and connected by linear segments. It enjoys some elegant duality properties with Cartesian plots. As mentioned in the literature(Wegman 1990), a point $M(a, ma+b)$ in Cartesian coordinates (See Fig.16) can be mapped into a line l in parallel coordinates. While line $l': y = mx + b$ joining the two points in Cartesian coordinates can be mapped into point P in parallel coordinates. In fact, lines l and $l': y = mx + b$ in parallel coordinates which are determined by two points in Cartesian plot, intersect at a point $P(\frac{b}{1-m}, \frac{1}{1-m})$; where $m \neq 1$. Moreover, the literature (Wegman 1990)also discussed some other cases of m. Therefore, duality property between two different coordinates is shown in theory. Thus parallel coordinate representation is often used to visualize and analyze high dimensional data.

Even though the data can be represented in a novel and meaningful visualization system without losing any features, PCP always suffer from crowded dimensions, hardly figuring out the relationship between attributes in non-adjacent positions, over-plotting and clutter et al, which are mainly caused by the increasing size of datasets and the large number of dimensions. Several enhancements have been proposed to overcome these artifacts.

In order to reduce the edge clutter and avoid the over-plotting in PCP, Dasgupta et al. (Dasgupta and Kosara 2010)propose a model based on screen-space metrics to pick

the axeslayout by optimizing arranges of axes. Huh et al. present a proportionate spacing between two adjacent axes rather than the equally spaced in conventional PCP parallel axes. Moreover, the curves possessing some statistical property linking data points on adjacent axes are described in literature (Huh and Park 2008) as well. Zhou et al. (Hong Zhou 2008)convert the straight-line edges into curves to reduce the

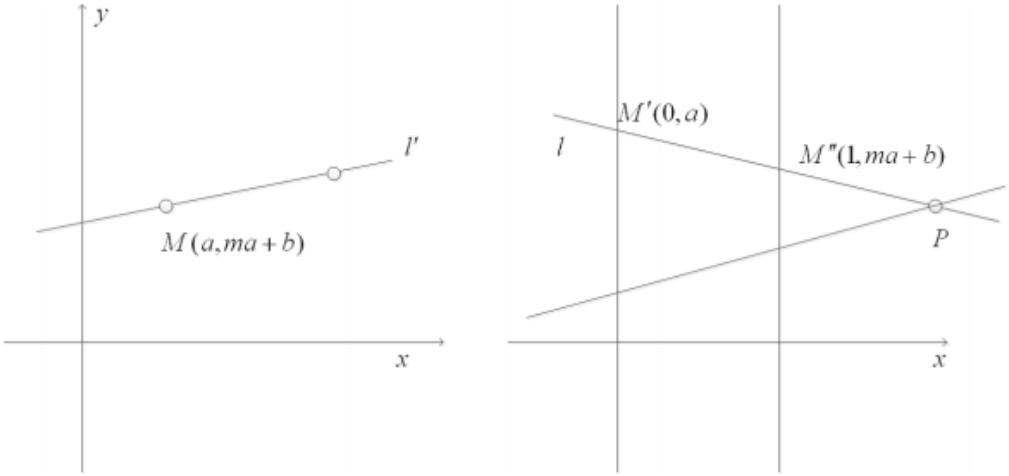


Figure 16. Duality property between points and lines in Cartesian and parallel coordinate plots(Wegman 1990).

visual clutter in clustered visualization. They also utilize the splatting framework (Zhou, Cui et al. 2009) to detect clusters and reduce visual clutter. To achieve the aim of avoiding over-plotting and preserving density information, Dang et al. proposed a visualization and interaction method for stacking overlapping cases (Tuan Nhon, Wilkinson et al. 2010).To filter out the information to be presented to the user and reduce visual clutter, Artero et al. develop a frequency and density plots from PCP(Artero, de Oliveira et al. 2004), which can uncover clusters in crowded PCP. Yuan et al. (Xiaoru, Peihong et al. 2009)combine the parallel coordinate method with the scatter-plots method to reduce the visual clutter. It plots scattering points in

parallel coordinates directly with a seamless transition between them. The shapes of poly lines are remodeled to cooperate with the scattering points, resulting in the diminution of their inherent visual effects.

To enhance the high-dimensional data visualization, some studies on dimension reordering have been done to find good axes layouts in PCP. Wei Peng et al. (Peng, Ward et al. 2004)introduce the definition of the visual clutter in parallel coordinates as the proportion of outliers against the total number of data points and they tried to use the exhaustive algorithm to find the optimal axes order for minimizing the number of edge crossings. As mentioned in(Artero, de Oliveira et al. 2006), the computational cost $o(n \cdot n!)$ hampers applications of this technique to large high dimensional data sets. Almir Olivette Artero et al. (Artero, de Oliveira et al. 2006)present the dimension configuration arrangement based on similarity to alleviate clutter in visualizations of high-dimensional data. They propose a method called Similarity-Based Attribute Arrangement (SBAA), which is a straightforward variation of the nearest neighbor heuristic method, to deal with both dimension ordering and dimensionality reduction. By analyzing the structures displayed in subspaces of the full feature space in PCP to obtain the dimension ordering, Ferdosi et al.(Ferdosi and Roerdink 2011) argue that they can identify the cluster and noise dimensions to improve the readability.

In addition, few approaches have been done for extensions of axes in PCP. Claessen et al.(Claessen and van Wijk 2011)develop flexible linked axes to enable users to define

and position coordinate axes freely. Axes-based techniques with radial arrangements of the axes are developed by Tominski(Tominski, Abello et al. 2004), termed as TimeWheel and the MultiComb, which can combined with some conventional interaction techniques. With the combination between interaction techniques and PCP, Hauser et al. (Hauser, Ledermann et al. 2002)design a angular brushing technique to select data sub-sets which exhibit a data correlation along two axes.

Even though these approaches can enhance the quality of visualization to some extent, the corresponding extensions for the axes in PCP still focus on the line segments and the main contributions lie in the ordering of axes and the curves jointing the vertices between the two adjacent axes. In this thesis, we propose a novel axes in parallel coordinates visualization, termed as arc coordinate plots (ACP). We take arc axes into account, rather than the line segments, as the coordinate axes in parallel coordinates plane. Besides visualizing the original data sets for reveal the patterns, ACP can preserve much more geometric structures of the data and can visualize much more data items in the same screen space than PCP. On the other hand, we propose a contribution-based layout of our ACP to overcome the curse of dimensionality by filtering the less important features among the original ones. Finally, we test our models in several datasets and find they are effective in revealing the patterns in the perspectives of density of points and rationale of the original geometric data properties.

4.3 ARC-BASED PARALLEL COORDINATES GEOMETRY

4.3.1 Optimizing Length of Arced Axis

To avoid the intersecting between the line segments and arc-axes, it is a nontrivial thing to calculate curvature of these parallel arcs.

Without loss of generality, we argue that the origin of Cartesian coordinates is $(0,0)$, where the center of the first axis in PCP lies in it. The distance between axes X_1 and X_2 is one. Given the length of each axis in PCP is $2T$, where T is a non-negative real number. It is another hypothesis that these axes are divided into two equal line segments vertically by one horizontal line. Meanwhile, we set the point $O_1(-x_0,0)$ as the center of the circle which generates the first arc axis, where x_0, as its radius, is a non-negative real number as well. The geometric structures of these three coordinates have been displayed in Fig.17.From the notification we can find that the coordinates of X_1 is $(0,-T)$. The position of upper end point of the second axis X_2 is $(1,T)$. Hence, we try to find the optimum radius based on the theory that arc axes do not act as a hindrance to data visual representation.

Based on the Cartesian coordinates, the equation of the first circle is

$$\left(x+x_0\right)^2+y^2=\left(x_0\right)^2 \qquad \textbf{eq. 9}$$

Similarly, the second one is

$$\left(x + x_0 - 1\right)^2 + y_2 = (x_0)^2 \qquad \textbf{eq. 10}$$

In equation (1), we assume that $y = -T$, then

$$\left(x + x_0\right)^2 + T^2 = \left(x_0\right)^2 \qquad \textbf{eq. 11}$$

Therefore, $x_M = -x_0 + \sqrt{\left(x_0\right)^2 - T^2}$. The position of points M and N can be defined in x and y coordinates system as:

$M\left(-x_0 + \sqrt{\left(x_0\right)^2 - T^2}, -T\right); N\left(1 - x_0 + \sqrt{\left(x_0\right)^2 - T^2}, T\right)$

The slope of line O_1M is $k_{O_1M} = \dfrac{-T}{\sqrt{\left(x_0\right)^2 - T^2}}$; Meanwhile, $k_{MN} = \dfrac{2T}{1} = 2T$ .If there is only one intersection point between any line segment and arc axis, what the line O_1M is perpendicular to line segment MN is the boundary condition. i.e,

$$k_{O_1M} \cdot k_{MN} = -1$$

Therefore, we have $x_0 = T\sqrt{4T^2 + 1}$.

Referring to the knowledge of geometry, we know that curvature of one circle is the reciprocal of the radius. Larger the radius is, the less the degree of bending is. This reduces the amount of intersections occurred by arc axis. Hence, we have the following property:

Property: To plot the arc-axes, the radius of the circle needs to satisfy: $x_0 \geq T\sqrt{4T^2 + 1}$.

For simplicity, we assume the distance of each pair of neighboring parallel coordinate axes equals to one, i.e. $T = \dfrac{1}{2}$. Based on the above analysis, the equations of the first

and second arc-axes can be simplified as $\left(x+\frac{\sqrt{2}}{2}\right)^2+y^2=\frac{1}{2}$ and $\left(x+\frac{\sqrt{2}}{2}-1\right)^2+y^2=\frac{1}{2}$ respectively. Likewise, the equation of $i-th$ arc axis can be termed

$$\left(x+\frac{\sqrt{2}}{2}-i\right)^2+y^2=\frac{1}{2}\ ,\ \text{where}\ \ i=0,1,2,\cdots n.\ n\in N\ . \qquad \textbf{eq. 12}$$

Without loss generality, we propose arc-coordinate plots based on this simplified equation.

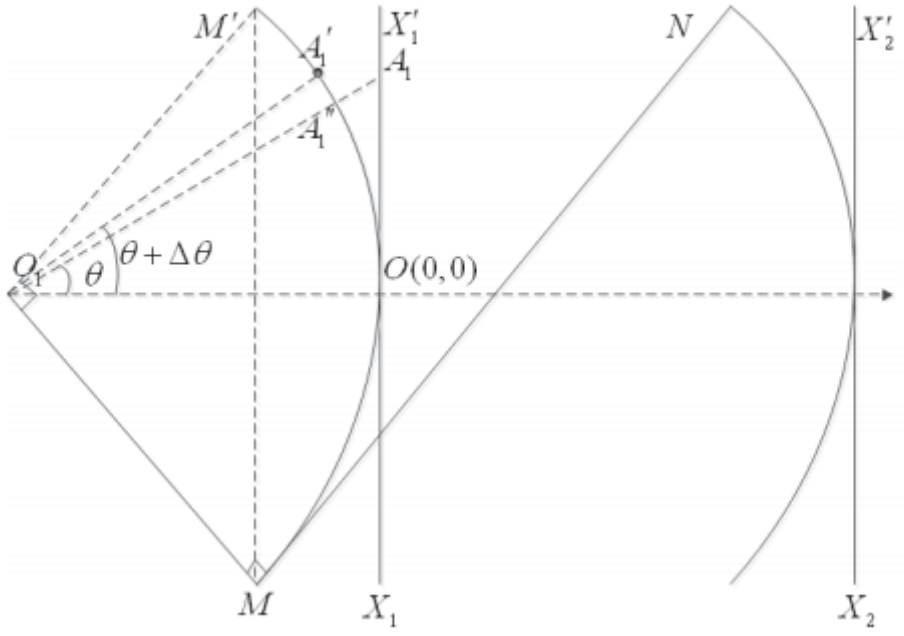


Figure 17. The rationale of arc coordinates plane.

4.3.2 Arc-Coordinate Geometry

The parallel coordinates plot, as a popular projective geometric visualization method, enjoys some elegant duality properties with usual Cartesian orthogonal coordinate

representation (Inselberg 1985; Wegman 1990). In fact, it is to be noticed that the axes in the traditional plot (PCP) are constructed by parallel line segments. Considering that the arc is longer than line segment when they start/end with the same points, we propose a property to show the extension rate of arc-coordinate plot to visualize more data items than PCP and it can keep better geometric structure of some circular data sets. Therefore, we obtain one to one mapping between the Cartesian coordinates and arc-coordinates in line with the transitivity of these bi-relations. According to the above assumption, we consider the first arc-based axis as our projection example to map the vertices in PCP to ACP.

Property: Comparing these two different visualization systems, the extension rate of the axis length from PCP to ACP is $\frac{\sqrt{2}\pi}{4}$.

Proof: From Fig.17, we easily notice that

$$|O_1M| = |O_1M'| = |x_0| = \frac{\sqrt{2}}{2} \text{ while } |MM'| = 2T = 1.$$

Hence, $\Delta O_1MM'$ is a right angled isosceles triangle. The length of arc MM' equals to one quarter of the perimeter of a circle exactly, i.e. $\frac{\pi x_0}{2} = \frac{\sqrt{2}\pi}{4}$.Therefore, we finish the proof of extension rate is $\frac{\sqrt{2}\pi}{4}$.

As we all know, there is only one straight line which passes through the points outside and inside the known line. To our visualization projection, there is one intersection

A_1'' when we draw a line between the arc axis and the line segment joining O_1 to A_1. In the similar manner, we can define the other intersections. It is to be noticed that it is a straightforward way defining the intersection A_1'' as the projection of the vertex A_1 from PCP to ACP. On the other hand, considering that the extension rate of the axis-length from PCP to ACP is $\frac{\sqrt{2}\pi}{4}$, we utilize increment of the arc length to project A_1'' to another vertex A_1' in the arc. In the following, we analyze the computation of increment of it in detail.

To simplify the computational complexity, we study the projection of vertices just in the positive semi-axis OX_1' of PCP and arc OM' in Fig. 17. Infact, the result of negative semi-axis is the same expression as the positive one. The slope of line O_1X_1' is $\frac{\sqrt{2}}{2}$, while the angle of OM' is just right the half of the right angle, $\frac{\pi}{4}$. The length of the arc is from $\frac{\sqrt{2}}{2}\arctan\frac{\sqrt{2}}{2}$ to $\frac{\sqrt{2}\pi}{8}$. The increment of it is $\frac{\pi}{4\arctan\frac{\sqrt{2}}{2}}$. To all vertices in the positive semi-axis, this increment is taken into account as our extension rate. In addition, due to the symmetric property of the axes in PCP and ACP, we can term this extension rate to the negative semi-axis. The explanation of this is illustrated in Fig. 18.

In conclusion, we propose the following function to project the point (i, y_0) in the $(i+1)-th$ PCP axis to the ACP one:

$$f:(i,y_0)\to\left(\frac{\cos\theta}{\sqrt{2}}+i-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\sin\theta\right),\ \text{where}\,\theta=\frac{\pi\arctan\left(\sqrt{2}y_0\right)}{4\arctan\left(\frac{\sqrt{2}}{2}\right)}. \qquad \textbf{eq. 1}$$

3

Here we adopt two steps for building the projection method:

In the first stage, we obtain the intersection coordinates between the line and arc by solving the following nonlinear system.

$$\begin{cases} y-y_0=\sqrt{2}y_0\left(x-i\right), \\ \left(x+\frac{\sqrt{2}}{2}-i\right)^2+y^2=\frac{1}{2}. \end{cases} \qquad \textbf{eq. 14}$$

The $\left(\frac{1}{\sqrt{2\left(2y_0^2+1\right)}}+i-\frac{\sqrt{2}}{2},\frac{y_0}{\sqrt{2y_0^2+1}}\right)$ coordinates can be obtained from the above nonlinear system.

In the second stage, we leverage the thought that extension rate $\frac{\pi}{4\arctan\left(\frac{\sqrt{2}}{2}\right)}$,as our extension factor, and multiplies the arc length which starts from the point $(i,0)$ in the horizontal axis and ends with the intersection coordinates. This makes the final projection vertex of the original point (i,y_0). To get the final coordinates, we have to link the arc length with the coordinate system. Therefore, we have the following system:

$$\begin{cases} y_0\cot\theta=x_0-i+\frac{\sqrt{2}}{2}, \\ \left(x_0+\frac{\sqrt{2}}{2}-i\right)^2+{y_0}^2=\frac{1}{2}. \end{cases} \qquad \textbf{eq. 15}$$

Finally, we get the result of projection $\left(\frac{\cos\theta}{\sqrt{2}}+i-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\sin\theta\right)$, where $\theta=\frac{\pi\arctan\left(\sqrt{2}y_0\right)}{4\arctan\left(\frac{\sqrt{2}}{2}\right)}$.

4.3.3 Contribution-Based Layout

In the research field of matrix computation, singular value decomposition plays an important role for its revealing interesting and attractive algebraic properties, and conveying important geometrical and theoretical in-sights about transformations. The entries of each matrix obtained by the SVD algorithm have their special physical significances. Here we apply these significances of matrix to our method to measure the contribution of each dimension to the dataset.

For an $M\times M$ matrix D, the singular value decomposition of it is defined as the following form(Golub and Van Loan 1996): $D=U\Sigma V^*$, where U and V^* (V^* is the conjugate transpose of V)are $m\times m$ and $n\times n$ unitary matrices respectively. Σ is an $m\times n$ rectangular diagonal matrix with nonnegative real numbers (singular values of D)in order of decreasing magnitude on the diagonal.

There are many properties of SVD for the matrices, such as the singular values of the matrix D are the square roots of eigenvalues of matrix D^TD; the Euclidean norm of D is equal to the largest singular value and so on. Among these properties,

what impressed us most are that the columns of the matrices U and V form the orthonormal bases for the space spanned by the columns and rows of D.For example, in the literature(Krzysztof.Simek 2003), characteristic modes are defined to reconstruct the gene expression patterns based on this property. By combining and analyzing these properties, we can conclude the following property in perspective of the numerical properties for matrix:

Property: The entries of the first column of V in the singular value decomposition, which are denoted as $v_{1j}, j = 1,2,\cdots,n$, show the contributions of columns of D to the space spanned by them, $i.e. span\{d_1, d_2, \cdots, d_n\}$, d_i is the ith column of D.

By setting the contribution rate, as one of the simplest techniques, to retain as much characteristics of the whole data set as possible, we will get six attributes which retain up to 99.8% of the overall information. The order of the axis from left to right indicates the contribution rate of attributes, as shown in the subscript of axis in Fig. 20. Moreover, from the visualization shown in Fig. 20. we can easily find out two different attacks among 1113 data items. Therefore, we can visualize large volume of data using this approach effectively and can retain the main characteristics of data. From the perspective of data values, this method provides us effective and clear visualization structure of the data. It can help us take deeper insight into the dataset.

It is natural that we can compute the contribution rate of each dimension to the whole dataset using the following possible measure:

$$Ci = \frac{v_{1j}}{\sum_{j=1}^{n} v_{1j}} \times 100\% \qquad \textbf{eq. 16}$$

This approach not only provides us a new reordering method helping us take much more insights into the dataset but also gives rise to the following new method which can help in determination of the first dimension with the most contribution.

4.4 CASE STUDIES

This section presents examples of how ACP can be used to analyze multivariate data. We tested three different datasets on the novel method and compare it to conventional PCP, one describing random data set to illustrate that the rational geometric structure could be preserved in ACP. Another two about KDD Cup 1999 and Cars models for contribution-based and arc-based visualization. These two data sets we tested come from the literature.

4.4.1 Random and Car Datasets

At the first stage, we generated 50 data items with two dimensions randomly, which satisfy the equation of circle: $\left(x+\frac{\sqrt{2}}{2}\right)^2 + y^2 = \frac{1}{2}$. Using inverse mapping of our method, we project these data to one axis in parallel coordinates plane. As shown in

Fig.18, the density of points is different in two visualization plane. By using of $d = \frac{l}{n}$, where d , l , n behaves the density, length of each axis and the number of points with different values.

Therefore, we can obtain the mean densities 0.0222 , 0.0200 of two different graphs respectively.

From the readability perspectives, the ellipse in the graph presents the points illustrated by our method are sparser than the PCP. Without loss generality, the mean extension rate of points can be calculated by the following formula:

$R = \frac{l_{arc}}{l_{parllel}}$, in which l is the length of each axis in two different coordinates system. Moreover, the geometric property of the data can be displayed in the arc axis rather than in line segment.

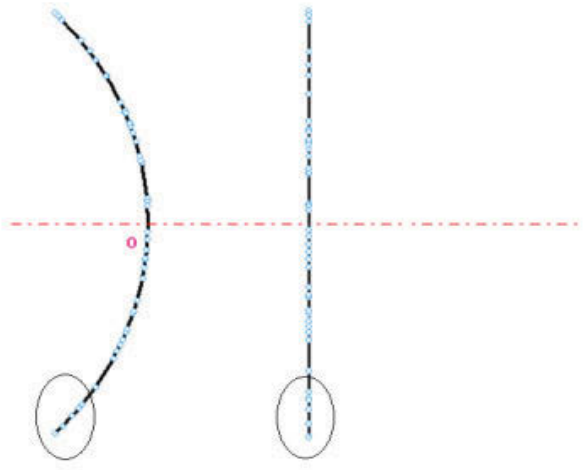


Figure 18. Random data represented in two different coordinates systems

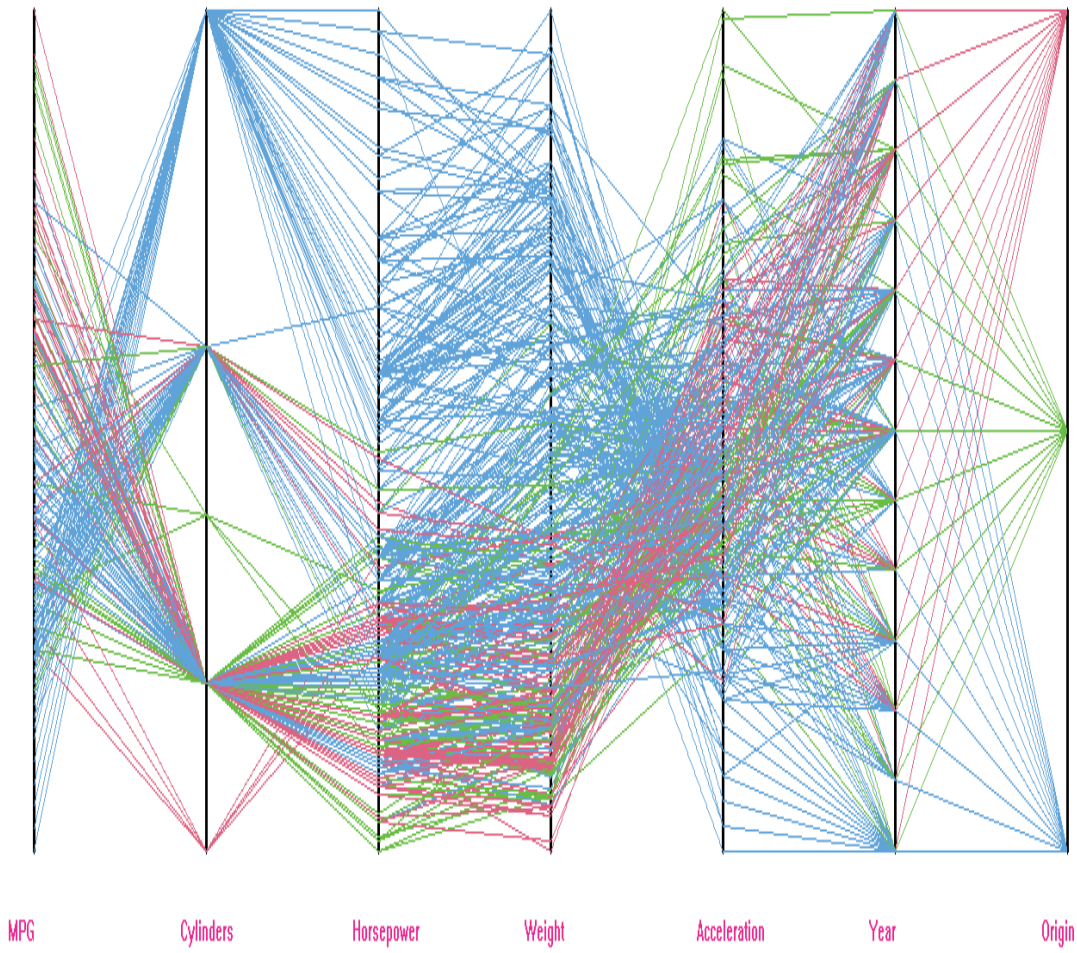
In Fig. 19, we visualize the Car dataset in both arced-axis parallel coordinates geometry and the traditional PCP. The comparison shows that our arced-axes parallel coordinate geometry can represent all datasets as the same quality as they are represented in the traditional vertical-line based parallel coordinate geometry. The arced-axes do not distort (or affect) the quality of visualization at all. In addition, our arced-axes approach could enlarge the mean density of points in the geometry that improves the readability of visualization.

To the three clusters in the dataset, they can be easily shown in the new system. From the comparison, it is to be noticed that our method can perform as same as PCP to this common dataset. Hence, our approach can not only do the same performance in visualizing the general datasets but also can do better in some circular datasets because of preserving the geometric structure of data.

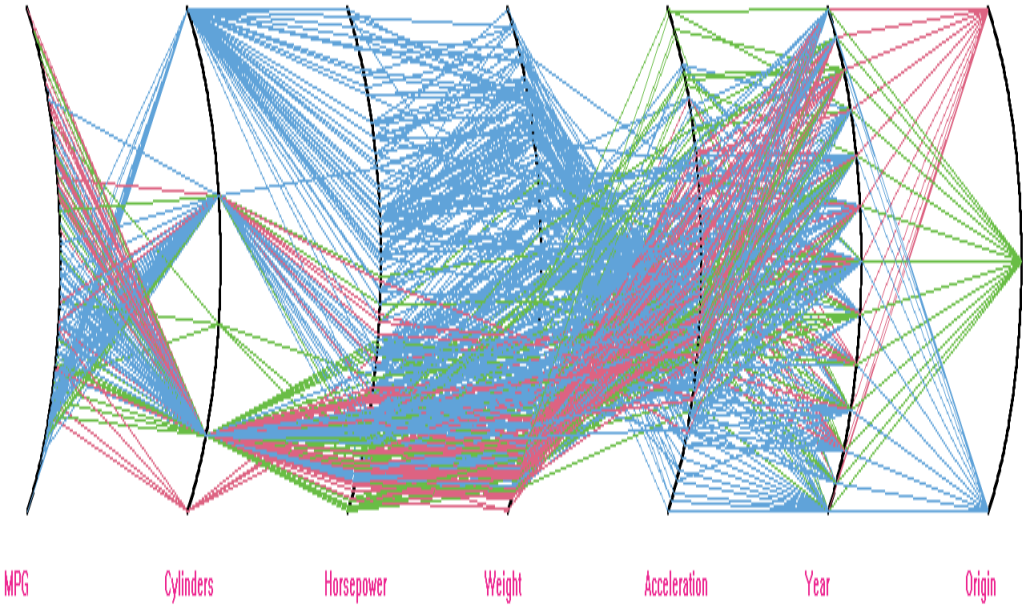
4.4.2 Case Study in Network Security Domain

In this section, we utilize KDD 1999 data sets to show the effectiveness of our contribution-based method in ACP. Firstly, part data from KDD 1999 consisting of 1113 data items with 42 attributes (including "normal" and "abnormal" labels) are analyzed in Fig.20.

To the whole 42 attributes, we use contribution-based method as a dimension reduction step for visualizing data set. By setting the contribution rate as one of the simplest techniques to retain as much characteristics of the whole data set as we can, we get the six attributes who retain the 99.8% of the overall information. It can be easily found that there are two different attacks in these 1113 data items: red lines and green lines represent Smurf and Neptune attacks respectively. From the polylines among the attributes "*dst _ host _ count*" and "*count*", we can find there is a big fluctuation between the normal and abnormal lines. As we all know, Smurf attack is represented by traffic-based features, such as *count* and *srv _ count* . While, Neptune attack quite distinct in src_bytes and count attributes. Hence, it is interesting that these patterns of attacks are totally presented in ACP.



(a)



(b)

Figure 19. Car dataset visualized in PCP and ACP respectively

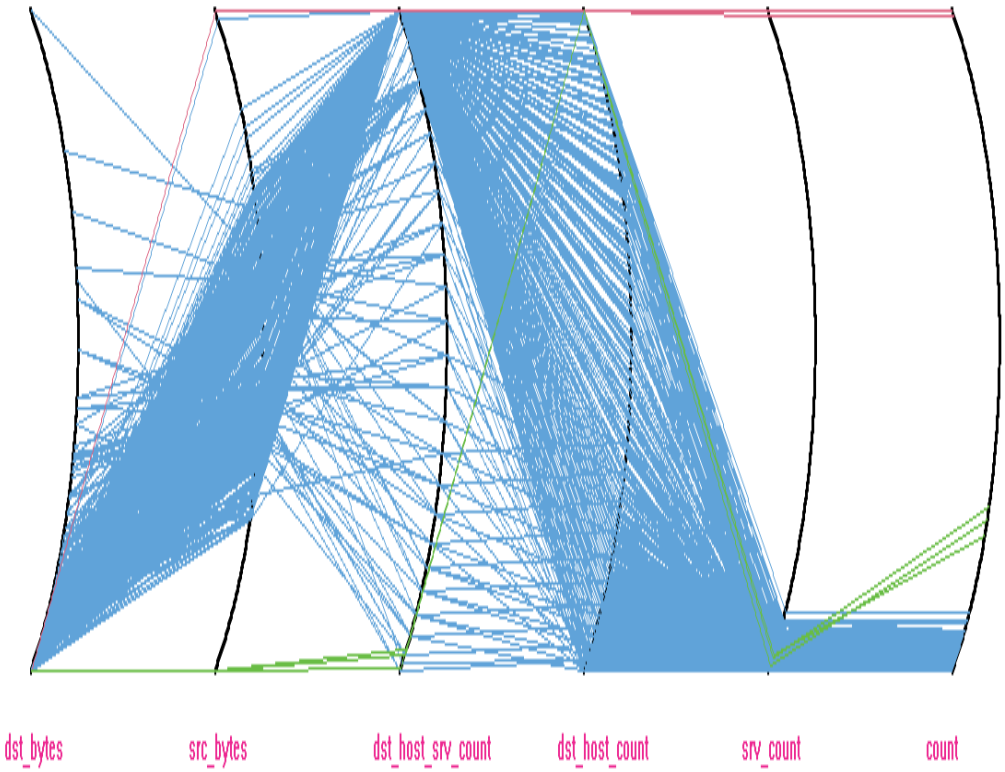


Figure 20. Detecting DDoS attacks using ACP:Red and green lines describe the Smurf and Neptune attacks respectively

4.5 SUMMARY

In this chapter, we propose a novel approach to extend the parallel axes in parallel coordinates plane theoretically. At the first stage, an arc-based parallel coordinates visualization method, termed as arc coordinate plots (ACP), is developed to extend the axes in parallel coordinate plots. Because the length of arc is longer than the line segments, the density of points displayed in each axis of our method could be enlarged. Moreover, ACP can preserve much more geometric structures of the data, such as the circular data. At the second stage, we leverage singular value decomposition algorithm to provide a new way of looking into the dimensions within datasets. We propose the contribution-based visualization method and a formula for contribution rate of each dimension. At last, the experimental evaluations demonstrate the effectiveness and rationale of our approaches especially the applications in security domain.

We plan to implement some real circular datasets in cloud computing to the experiments to demonstrate our method. And try to prove the property in the section 4 theoretically. Moreover, we consider combing some interaction techniques with our approach and strengthen our visualization system.

Chapter 5. CONCLUSION AND FUTURE WORK

We leverage a layered directed graph drawing algorithm into parallel coordinates for visualization of uncertainty. In the first stage, a nonlinear equation system is deployed to obtain the initial positions for all uncertain values. In the second stage, a multi-objective optimization algorithm is adapted to relocate positions for the dummy vertices. In the final stage, the penalty minimization method finalizes ordering of all vertices.

The case studies showed the clutter reduction among polylines and demonstrated the effectiveness of the method with better visual structure in parallel coordinates visualization. These experiments also illustrated that the number of edge crossings, uncertain values and the attributes of multi-dimensional data could play important roles in affecting visualization performance.

In order to advance technology towards information visualization of uncertainty, formal evaluation needs to be conducted in the future. Based on the evaluation results, we will not only work on visual quality but also further modify the algorithm in order to reduce complexity of computing time for larger datasets.

We proposed a new method to improve the readability and understandability of parallel coordinates visualization theoretically. At the first stage, we propose a new

way of looking into the dimensions within datasets based on the singular value decomposition algorithm. At the second stage, we present a method, named similarity-based reordering method, for calculating the similarity between the two dimensions based on the nonlinear correlation coefficient and singular value decomposition algorithms rather than the traditional Pearson's correlation coefficient, and then visualize the optimal dimension order according to the similarity in parallel coordinates. We have conducted the experimental evaluations to demonstrate the effectiveness and rationale of our approaches: NCC reordering method enlarges the mean crossing angles of the whole data set and reduces the amount of polylines between some neighboring dimensions.

During the process of calculation for nonlinear correlation coefficient, the more exact choice of rank grids will do much more help in the speed of calculation. Therefore, we consider this problem to be our first future work. And then we will apply our methods with interactive techniques to more real-world datasets and help users analyze the datasets using visualization.

We propose a novel approach to extend the parallel axes in parallel coordinates plane theoretically. At the first stage, an arc-based parallel coordinates visualization method, termed as arc coordinate plots (ACP), is developed to extend the axes in parallel coordinate plots. Because the length of arc is longer than the line segments, the density of points displayed in each axis of our method could be enlarged. Moreover,

ACP can preserve much more geometric structures of the data, such as the circular data. At the second stage, we leverage singular value decomposition algorithm to provide a new way of looking into the dimensions within datasets. We propose the contribution-based visualization method and a formula for contribution rate of each dimension. At last, the experimental evaluations demonstrate the effectiveness and rationale of our approaches especially the applications in security domain.

As to the future research work, firstly, to the uncertainty problems in information visualization, we will not only work on visual quality but also further modify the algorithm in order to reduce complexity of computing time for larger datasets.

Secondly, during the process of calculation for nonlinear correlation coefficient, the more exact choice of rank grids will do much more help in the speed of calculation.

Thirdly, we plan to implement some real circular datasets in cloud computing to the experiments to demonstrate our methods, and try to prove the property in the Chapter 4 theoretically.

Moreover, we consider combing some interaction techniques with our approaches and strengthen our visualization system. And then we will apply interactive techniques to more real-world datasets and help users analyze the datasets using parallel coordinates visualization.

At last, based on the approaches we achieved, we will try to implement all systems in real time and use them to reveal the patterns in real time such as in detecting the DDoS attacks and so on. Beside, formal evaluation needs to be conducted to evaluate the approaches we proposed in the future as well.

PUBLICATION LIST

1．Liang Fu Lu, Mao Lin Huang, Jinsong Zhang. Two Axes Re-ordering Methods in Parallel Coordinates Plots, *Journal of Visual Languages and Computing* 33, 2016:3-12. (75%)

2． Mao Lin Huang, Liang Fu Lu, Xuyun Zhang: Using arced axes in parallel coordinates geometry for high dimensional Big Data visual analytics in cloud computing. *Computing* 97(4), 2015: 425-437. (70%)

3． Jinson Zhang, Mao Lin Huang, Wenbo Wang, Liang Fu Lu, Zhao-Peng Meng: Big Data Density Analytics Using Parallel Coordinate Visualization. *Proceeding of IEEE Internationa Conference on Computational Science and Engineering (CSE2014):*1115-1120. (30%)

4． Wenbo Wang, Mao Lin Huang, Liang Fu Lu, Jinson Zhang: Improving Performance of Forensics Investigation with Parallel Coordinates Visual Analytics. *Proceeding of IEEE International Conference on Computational Science and Engineering (CSE2014):* 1838-1843. (25%)

5． Mohammed A. Ambusaidi, Zhiyuan Tan, Xiangjian He, Priyadarsi Nanda, Liang Fu Lu, Aruna Jamdagni: Intrusion detection method based on nonlinear correlation measure. *International Journal of Internet Protocol Technology* 8(2/3), 2014: 77-86. (20%)

6． Liang Fu Lu, Mao Lin Huang, Tze-Haw Huang: A New Axes Re-ordering Method in Parallel Coordinates Visualization. *Proceeding of Internationa Conference on Machine Learning and Applications (ICMLA (2)) 2012*: 252-257. (80%)

7. Liang Fu Lu, Mao Lin Huang, Yi-Wen Chen, Jie Liang, Quang Vinh Nguyen: Clutter Reduction in Multi-dimensional Visualization of Incomplete Data Using Sugiyama Algorithm. *Proceeding of Information Visualisation (IV 2012):* 93-99. (85%)

Note: The percent at the end of each publication paper represents the contribution of the candidate to the whole paper.

APPENDIX

To make the thesis easy to be read, the following abbreviations are the terms we used in the thesis:

PCP -Parallel coordinate plots;

SBAA -Similarity-Based Attribute Arrangement;

PCC-Pearson's Correlation Coefficient;

NCC- Nonlinear Correlation Coefficient;

SVD- Singular Value Decomposition;

NS-Neighbouring Sequence;

ACP -Arc Coordinate Plots.

REFERENCES

Albuquerque, G., M. Eisemann, et al. (2010). Improving the visual analysis of high-dimensional datasets using quality measures. Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST), 2010:19-26.

Ankerst, M., S. Berchtold, et al. (1998). Similarity clustering of dimensions for an enhanced visualization of multidimensional data. Proceedings of IEEE Symposium on Information Visualization, 1998:52-60.

A. Astel, K. A., M. Biziuk, J. Namieśnik2 (2006). "Clasification of Drinking Water Samples Using the Chernoff's Faces Visualization Approach." Polish Journal of Environmental Studies, 15(5): 691-697.

Artero, A. O., M. C. F. de Oliveira, et al. (2004). Uncovering Clusters in Crowded Parallel Coordinates Visualizations. Proceedings of IEEE Symposium on Information Visualization, 2004. INFOVIS 2004:81-88.

Artero, A. O., M. C. F. d. Oliveira, et al. (2006). Enhanced High Dimensional Data Visualization through Dimension Reduction and Attribute Arrangement. Proceedings

of the conference on Information Visualization, IEEE Computer Society: 707-712.

Becker, R. A. and W. S. Cleveland (1987). "Brushing Scatterplots." Technometrics 29(2): 127-142.

Bertini, E., A. Tatu, et al. (2011). "Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization." Visualization and Computer Graphics, IEEE Transactions on 17(12): 2203-2212.

Card, S. K., J. D. Mackinlay, et al. (1999). Readings in information visualization: using vision to think, Morgan Kaufmann.

Claessen, J. H. T. and J. J. van Wijk (2011). "Flexible Linked Axes for Multivariate Data Visualization." IEEE Transactions on Visualization and Computer Graphics, 17(12): 2310-2316.

Cyntrica Eaton, C. P., Terence Drizd (2005). Visualizing missing data: Classification and empirical study. Proceedings of INTERACT 2005, Springer: 861-872.

Dasgupta, A. and R. Kosara (2010). "Pargnostics: Screen-Space Metrics for Parallel Coordinates." IEEE Transactions on Visualization and Computer Graphics, 16(6): 1017-1026.

Dimsdale, B. (1984). "Conic transformations and projectivities." IBM Los Angeles Scientific Center. Rep. G320-2753.

Drmota, M. and W. Szpankowski (2004). "Precise minimax redundancy and regret." IEEE Transactions on Information Theory, 50(11): 2686-2707.

Ellis, G. and A. Dix (2006). "Enabling Automatic Clutter Reduction in Parallel Coordinate Plots." IEEE Transactions on Visualization and Computer Graphics, 12(5): 717-724.

Ellis, G. and A. Dix (2007). "A Taxonomy of Clutter Reduction for Information Visualisation." IEEE Transactions on Visualization and Computer Graphics, 13(6): 1216-1223.

Feng, D., L. Kwock, et al. (2010). "Matching visual saliency to confidence in plots of uncertain data." IEEE Transactions on Visualization and Computer Graphics 16(6): 980-989.

Ferdosi, B. J. and J. B. T. M. Roerdink (2011). "Visualizing High-Dimensional Structures by Dimension Ordering and Filtering using Subspace Analysis." Computer Graphics Forum 30(3): 1121-1130.

Friendly, M. and E. Kwan (2003). "Effect ordering for data displays." Computational Statistics & Data Analysis 43(4): 509-539.

Gershon, N., S. Card, et al. (1998). Information visualization tutorial. CHI 98 Cconference Summary on Human Factors in Computing Systems. Los Angeles, California, USA, ACM: 109-110.

Golub, G. H. and C. F. V. Loan (1983). Matrix Computations. Baltimore, John Hopkins University Press.

Golub, G. H. and C. F. Van Loan (1996). "Matrix computations. 1996." Johns Hopkins University, Press, Baltimore, MD, USA: 374-426.

Guo, D. (2003). "Coordinating computational and visual approaches for interactive feature selection and multivariate clustering." Information Visualization 2(4): 232-246.

Hahsler, M., K. Hornik, et al. (2008). "Getting things in order : an introduction to the {R} package seriation." Journal of Statistical Software.

Hauser, H., F. Ledermann, et al. (2002). Angular brushing of extended parallel

coordinates. Proceedings of IEEE Symposium on Information Visualization, 2002. INFOVIS 2002:127-130.

Hong Zhou, X. Y., Huamin Qu, Weiwei Cui, Baoquan Chen (2008). "Visual Clustering in Parallel Coordinates." Computer Graphics Forum 27(3): 1047-1054.

Huang, W. and M. Huang (2014). "Exploring the relative importance of number of edge crossings and sinze of crossing angle:a quantitative perspective." Advanced Intelligence 3(1):25–42.

Huh, M.-H. and D. Y. Park (2008). "Enhancing parallel coordinate plots." Journal of the Korean Statistical Society 37(2): 129-133.

Hurley, C. B. and R. W. Oldford (2010). "Pairwise Display of High-Dimensional Information via Eulerian Tours and Hamiltonian Decompositions." Journal of Computational and Graphical Statistics 19(4): 861-886.

Inselberg, A. (1985). "The plane with parallel coordinates." The Visual Computer 1(2): 69-91.

Irvine, U. (Machine Learning Repository). from http://archive.ics.uci.edu/ml/.

J. Yang, M. O. W., E.A. Rundensteiner and S. Huang (2003). Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets. Joint EUROGRAPHICS - IEEE TCVG Symposium on Visualization (2003). S. H. G.-P. Bonneau, C. D. Hansen: 19-28.

J.Bertin (1983). Semiology of graphics, University of Wisconsin Press.

Jakobsen, M. R. and K. Hornbaek (2013). "Interactive Visualizations on Large and Small Displays: The Interrelation of Display Size, Information Space, and Scale." IEEE Transactions on Visualization and Computer Graphics, 19(12): 2336-2345.

Johansson, S. and J. Johansson (2009). "Interactive Dimensionality Reduction Through User-defined Combinations of Quality Metrics." IEEE Transactions on Visualization and Computer Graphics 15(6): 993-1000.

Johnson, C. R. and A. R. Sanderson (2003). "A next step: Visualizing errors and uncertainty." IEEE Computer Graphics and Applications 23(5): 6-10.

Keim, D. A. (2000). "Designing pixel-oriented visualization techniques: theory and applications." IEEE Transactions on Visualization and Computer Graphics 6(1): 59-78.

Keim, D. A. (2002). "Information visualization and visual data mining." IEEE Transactions on Visualization and Computer Graphics 8(1): 1-8.

Krzysztof.Simek (2003). "Properties of a Singular Value Decomposition Based Dynamical Model of Gene Expression Data." International Journal of Applied Mathematics and Computer Science 13(3): 337-345.

Laguna, M., R. Mart, et al. (1997). "Arc crossing minimization in hierarchical digraphs with tabu search." Computers and Operations Research 24(12): 1175-1186.

Martin Theus, H. H., Bernd Siegl & Antony Unwin (1997). "MANET: Extensions to Interactive Statistical Graphics for Missing Values." New Techniques and Technologies for Statistics II: 247-259.

Matsuda, H. (2000). "Physical nature of higher-order mutual information: Intrinsic correlations and frustration." Physical Review E 62(3): 3096-3102.

Pang, A. T., C. M. Wittenbrink, et al. (1997). "Approaches to uncertainty visualization." Visual Computer 13(8): 370-390.

Pascale, K., P. Bruno, et al. (2006). "Minimizing crossings in hierarchical digraphs with a hybridized genetic algorithm." Journal of Heuristics 12(1-2): 23.

Peihong, G., X. He, et al. (2010). Interactive local clustering operations for high dimensional data in parallel coordinates. Proceedings of the IEEE Pacific Visualization Symposium (PacificVis), 2010:97-104.

Peng, W., M. O. Ward, et al. (2004). Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering. Proceedings of the IEEE Symposium on Information Visualization, IEEE Computer Society: 89-96.

Popov, S. (2006). Nonlinear Visualization of Incomplete Data Sets Computer Science – Theory and Applications. D. Grigoriev, J. Harrison and E. Hirsch, Springer Berlin / Heidelberg. 3967: 524-533.

Rao, R. and S. Card (1994). The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. Proceedings of the SIGCHI conference on Human factors in computing systems, Boston, MA, USA, ACM.

Repository, U. M. L.

Rodgers, J. and A. Nicewander (1988). "Thirteen Ways to Look at the Correlation

Coefficient." The American Statistician 42(1): 59-66.

Skeels, M., B. Lee, et al. (2010). "Revealing uncertainty for information visualization." Information Visualization 9(1): 70-81.

Sugiyama, K., S. Tagawa, et al. (1981). "Methods for Visual Understanding of Hierarchical System Structures." IEEE Transactions on Systems, Man and Cybernetics 11(2): 109-125.

Swayne, D. F. and A. Buja (1998). "Missing Data in Interactive High-Dimensional Data Visualization." SSRN eLibrary 13(1): 15–26.

Tatu, A., G. Albuquerque, et al. (2011). "Automated Analytical Methods to Support Visual Exploration of High-Dimensional Data." Visualization and Computer Graphics, IEEE Transactions on 17(5): 584-597.

Tatu, A., G. Albuquerque, et al. (2009). Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. Proceedings of IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009:59-66.

Tominski, C., J. Abello, et al. (2004). Axes-based visualizations with radial layouts.

Proceedings of the 2004 ACM symposium on Applied computing, ACM:1242-1247.

Tuan Nhon, D., L. Wilkinson, et al. (2010). "Stacking Graphic Elements to Avoid Over-Plotting." IEEE Transactions on Visualization and Computer Graphics 16(6): 1044-1052.

Wang, Q., Y. Shen, et al. (2005). "A nonlinear correlation measure for multivariable data set." Physica D: Nonlinear Phenomena 200(3–4): 287-295.

Wegman, E. J. (1990). "Hyperdimensional data analysis using parallel coordinates." Journal of the American Statistical Association 85(411): 664-675.

Wei-xiang, S. and H. Jing-wei (2001). "Edge Crossing Minimization Algorithm for Hierarchical Graphs Based on Genetic Algorithms." Wuhan University Journal of Natural Sciences 6(1-2): 555-559.

Wijk, J. J. v. (2002). "Image based flow visualization." ACM Trans. Graph. 21(3): 745-754.

Xiaoru, Y., G. Peihong, et al. (2009). "Scattering Points in Parallel Coordinates." IEEE Transactions on Visualization and Computer Graphics 15(6): 1001-1008.

Xuyun, Z., L. Chang, et al. (2013). "A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud." IEEE Transactions on Parallel and Distributed Systems 24(6): 1192-1202.

Xuyun, Z., L. T. Yang, et al. (2014). "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud." IEEE Transactions on Parallel and Distributed Systems 25(2): 363-373.

Zhang, X., C. Liu, et al. (2013). "An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud." Journal of Computer and System Sciences 79(5): 542-555.

Zheng Rong, Y. and M. Zwolinski (2001). "Mutual information theory for adaptive mixture models." IEEE Transactions on Pattern Analysis and Machine Intelligence 23(4): 396-403.

Zhiyuan, S., W. Qiang, et al. (2011). Effects of statistical distribution on nonlinear correlation coefficient. IEEE Instrumentation and Measurement Technology Conference (I2MTC), 2011.

Zhou, H., W. Cui, et al. (2009). "Splatting the Lines in Parallel Coordinates."

Computer Graphics Forum 28(3): 759-766.

Zhuang Chu Qiang, W. Y. S. (1992). Mathematical Statistics with Applications. Guangzhou, South China science and technology university press.