Big Data Behaviour Modelling and Visual Analytics

Jinson Zhang

A thesis submitted for the degree of

Doctor of Philosophy in Computing Sciences

Supervisor by: A/Professor Mao Lin Huang

May 2017

Faculty of Engineering & Information Technology

University of Technology Sydney

Australia

Certificate of Authorship/Originality

Data: **October 2016**

Author: **Jinson Zhang**

Title: **Big Data Behaviour Modelling and Visual Analytics**

Degree: **Doctor of Philosophy in Computer Sciences**

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

Firstly, I would like to thank my supervisor A/Professor Mao Lin Huang for giving me the opportunity to do my PhD within the Faculty of Engineering & Information Technology at the University of Technology Sydney, and for his enormous encouragement, advice, and academic guidance during my part-time research period. Dr. Mao Lin have given me not only academic advice and freedom in my research area, but had also provided me with teaching opportunities over the last 15 years across a broad range of different subjects, including in Internet Programming and Data Visualization.

Secondly, I would like to thank our Visual Analytics Team, consisting of Dr. Jie Liang, Dr. Jie Hua, Dr. Liang Fu Lu, and Dr. Wen Bo Wang, for their knowledge sharing and information exchange. Their focus and patience have encouraged me to finish my work, and I have greatly increased my skills and visualization techniques during our fortnightly team meetings. Their humor and trust have benefited me not only through academic partnerships, but also through friendships outside of research.

Thirdly, I would like to thank the Library for their support during my research period. As a full-time Library staff and part-time researcher, I have been given the ability to work flexibly in order in allowing me to attend conferences and other research/teaching activities. Without their support, I would not have been able to finish my research. The Library has also provided me with enormous online materials, and has helped me with referencing works.

Furthermore, I would like to thank the School of Software, Faculty of Engineering & Information Technology, and the Graduate Research School for their help in providing the Faculty Travel Fund and Vice-Chancellor's Conference Fund to attend national and

international conferences. This has been enormously beneficial for part-time PhD researchers such as myself.

Finally, I would like to take this opportunity to thank my wife Lisa and son Edmund for their enormous help and support. Having started my PhD as part-time researcher, their understanding and belief in my capability has provided enormous encouragement during my research period. They have provided continuous support and help in making this PhD thesis possible, and I am thankful for having my family behind me every step of the way.

Table Contents

List of Figures

List of Tables

List of Equations

Abstract

Big Data is composed of text, images, video, audio, mobile or other forms of data collected from multiple datasets, and is rapidly growing in both size and complexity. This has created a huge volume of multidimensional data within a very short time period. Big Data is therefore too big, too complex and moves too fast for us to analyze using traditional methods. Big Data behaviour is considered as a set of concepts and categories that descripts Big Data's acts towards others. The challenges facing Big Data analysis and visualization include: 1) how to classify Big Data across multiple datasets and different forms of data, 2) how to visualize structured and unstructured Big Data behaviour patterns for multidimensional data, 3) how to display Big Data behaviour patterns with very large volumes onto a normal-sized screen, 4) how to visualize Big Data behaviour patterns without the loss of information.

Big Data visualization normally requires optimized solutions through using different visual techniques for integrating display and exploration. To illustrate the huge amount of multidimensional data within a standard-size screen, visualization needs to find an efficient classification method for multiple datasets across any form of data. The current data interactive exploration has normally optimized data for visualization by excluding some pieces of information, resulting in missing information. Big Data visualization also suffers from visual cluttering and data overcrowding problems, whilst dealing with huge amounts of multidimensional data.

My approach includes two parts: Big Data behaviour modelling and Big Data visualization. I have firstly established the 5Ws dimensions for Big Data classification, based on data behaviour ontologies, that can be applied to multiple datasets and to any

form of data. Each data incident contains these 5Ws dimensions, which are posed as a set of concepts and categories that descripts Big Data acts for; **When** did the data occur, **Where** did the data come from, **What** did the data contain, **How** was the data transferred, **Why** did the data occur, and **Who** received the data. Secondly, I have introduced Pair-Density algorithms to measure Big Data behaviour patterns, which enables comparison and analysis between any two dimensions of behaviours. Two non-dimensional axes in parallel coordinates have then been created by using Pair-Density to measure and compare visual patterns for Big Data visualization. Finally, Shrunk Attributes has been deployed into Pair-Density parallel coordinates. This not only narrows down Big Data patterns for better understanding, but also dramatically reduces data cluttering and overcrowding in Big Data visualization.

Three different datasets with a combined total of more than 2.5 million data incidents have been implemented for measuring and visualizing different data patterns, including both numerical and non-numerical dimensions. The experimental results have shown that my new approach has significantly improved the accuracy of Big Data visualization, reduced data cluttering by more than 80% without the loss of information. The use of 5Ws dimensions and Pair-Density parallel coordinates therefore has large potential benefits and applications across both the business and research fields.

This thesis contains the research approach and implementation results obtained by the author during his Ph.D period. The majority of methods and results have been published in **Seventeen** research papers in journals and conference proceeding by May 2016.

Author's Seventeen published papers for PhD

1). J. Zhang, ML. Huang, (2016), "2D Approach Measuring Multidimensional Data Pattern in Big Data Visualization", IEEE first International Conference on Big Data Analysis (ICBDA2016), In Proceeding of 2016 IEEE International Conference on Big Data Analysis, IEEE Computer Society, pp. 194-199, March 2016, DOI: 1109/ICBDA.2016.7509823

2). LF Lu, ML. Huang, and J. Zhang, (2016), "Two Re-ordering Methods in Parallel Coordinates Plots", Journal of Visual Languages and Computing, Elsevier, Vol. 33, pp. 3-12, April 2016, DOI: 10.1016/j.jvlc.2015.12.001

3). J. Zhang and M.L. Huang, (2016), "Data Behaviours Model for Big Data Visual Analytics", International Journal of Big Data Intelligence, InderScience Publishers, Vol. 3, No. 1, pp. 1-17, Dec 2015, DOI: 10.1504/IJBDI.2016.073899

4). J. Zhang, M.L. Huang, and Z. Meng, (2015), "Visual Analytics for BigData Variety and Its Behaviours", International Journal of Computer Science and Information Systems, ComSIS Consortium, Vol. 12, No. 4, pp. 1171-1191, Nov 2015, DOI: 10.2298/CSIS141122050Z

5). WB. Wang, M.L. Huang, J. Zhang, and W. Lai, (2015), "Detecting Criminal Relationships Through SOM Visual Analytics", 19th International Conference on Information Visualization (IV2015), In Proceeding of 19th International Conference on Information Visualization, IEEE Computer Society, pp. 316-321, July 2015, DOI:10.1109/IV.2015.62

6). J. Zhang and M.L. Huang, (2015), "A New Analytics Model for Large Scale Multidimensional Data Visualization", Cloud Computing and Big Data (CloudCom-Asia 2015), Lecture Notes in Computer Science, Vol. 9106, pp. 55-71, June 2015, DOI: 10.1007/978-3-319-28430-9_5

7). J. Zhang, M.L. Huang, W.B. Wang, L.F. Lu, and Z.P. Meng, (2014), "Big Data Density Analytics using Parallel Coordinate Visualization", The 13th

International Symposium on Pervasive System, Algorithm, and Network (I-SPAN2014). In Proceeding of IEEE 17th International Conference on Computational Science and Engineering (CSE), IEEE Computer Society, pp. 1115-1120, Dec 2014, DOI:10.1109/CSE.2014.219

8). W.B. Wang, M.L. Huang, L.F. Lu, and J. Zhang, (2014), "Improving Performance of Forensics Investigation with Parallel Coordinates Visual Analytics", The 8th International Conference on Frontier of Computer Science and Technology (FCST2014), In Proceeding of IEEE 17th International Conference on Computational Science and Engineering (CSE), IEEE Computer Society, pp.1838-1843, Dec 2014, DOI:10.1109/CSE.2014.337

9). J. Zhang, M.L. Huang, and Z. Meng, (2014), "BigData Visualization: Parallel Coordinates using Density Approach", The 2nd International Conference on System and Informatics (ICSAI2014), In Proceeding of IEEE 2nd International Conference on System and Informatics, IEEE Computer Society, pp. 1056-1063, Nov 2014, DOI:10.1109/ICSAI.2014.7009441

10). J. Zhang and M.L. Huang, (2014), "Density approach: a new model for BigData analysis and visualization", Concurrency and Computation: Practice and Experience, Vol. 28, Issue. 3, pp 661-673, First publish online July 2014, DOI: 10.1002/cpe.3337

11). J. Zhang and M.L. Huang, (2013), "5Ws Model for BigData Analysis and Visualization", The 2nd International Conference on Big Data Science and Engineering (BDSE2013), In Proceeding of IEEE 16th International Conference on Computational Science and Engineering (CSE), IEEE Computer Society, pp. 1021-1028, Dec 2013, DOI:10.1109/CSE.2013.149

12). J. Zhang and M.L. Huang, (2013), "Detecting Flood Attack through New Density-Pattern Based Approach", The 15th International Conference on High Performance Computing and Communication (HPCC2013), In Proceeding of 15th IEEE International Conference on High Performance Computing and Communication, IEEE Computer Society, pp. 246-253, Nov 2013, DOI: 10.1109/HPCC.and.EUC.2013.44

13). M.L. Huang and J. Zhang, (2013), "Visual Analysis and Detection of Network Flood Attack through Two-Layer Density Approach", The 3rd International Conference on Computer Science and Network Technology (ICCSNT2013), In Proceeding of 3rd IEEE International Conference on Computer Science and Network Technology, IEEE Computer Society, pp. 625-629, Oct 2013, DOI: 10.1109/ICCSNT.2013.6967191

14). J. Zhang and M.L. Huang, (2013), "Visual Analytics Model for Intrusion Detection in Flood Attack", The 12th International Conference on Trust Security and Privacy in Computing and Communications (TrustCom2013), In Proceeding of 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, IEEE Computer Society, pp. 277-284, July 2013, DOI:10.1109/TrustCom.2013.38

15). J. Zhang, M.L. Huang, and D.B. Hoang, (2013), "Visual Analytics for Intrusion Detection in Spam Emails", International Journal of Grid and Utility Computing, InderScience Enterprises, Vol. 4, No. 2/3, pp. 178-186, Jan 2013, DOI:10.1504/IJGUC.2013.056254

16). J. Zhang, M.L, Huang, and D.B. Hoang, (2011), "Detecting DDoS Attack in Spam Emails using Density-Weight Model", The 2nd International Conference on Information Theory and Information Security, In Proceeding of 2nd IEEE International Conference on Information theory and Information Security, IEEE Press, Vol. II, pp. 344-352, Nov 2011, https://opus.lib.uts.edu.au/research/handle/10453/29567

17). M.L. Huang, J. Zhang, Q. Nguyen, and J. Wang, (2011), "Visual Clustering of Spam Emails for DDoS Analysis", The 15th International Conference on Information Visualization (IV2011), In Proceeding of 15th IEEE international Conference on Information Visualization, IEEE Computer Society, pp. 65-72, July 2011, DOI:10.1109/IV.2011.41

Chapter 1: Introduction

Big Data, according to Wikipedia (accessed June 2016), is "the term for data sets that are so large or complex that traditional data processing applications are inadequate." I consider Big Data to be structured and unstructured data containing text, image, audio, video, mobile and other forms of data, that is collected from multiple datasets and is rapidly growing in both size and complexity (Zhang, J., Huang, M.L., et al 2014). For example, Pingdom (2013) estimates that there are 2.2 billion email users who send 144 billion emails worldwide every day, seven petabytes of photo content added on Facebook every month, four billion hours of video watched on YouTube every month, and five billion mobile phone users who use 1.3 exabytes of global mobile data traffic per month. Such enormous data flows create thousands, even millions, of different attributes across multiple dimensions within datasets, which is far too much information for traditional analysis and visualization tools to handle.

1.1. Big Data Characteristics

Big Data comes from everywhere in our life. For example, posting pictures and writing comments on Facebook or Twitter; uploading and watching videos on YouTube; sending and receiving messages through smart phones; and sending viruses over the Internet all constitute Big Data since they all involve different formats of information that are collected by different datasets. According to Gartner's 3Vs definition (Stamford 2011), Big Data has three main characteristics: Volume, Variety and Velocity. Dominik Klein (Klein et al 2013) drew the model shown in Figure 1.1 to model these dimensions. The symbols PB, TB, GB and MB, in the Volume dimension represent Petabytes, Terabytes, Gigabytes and Megabytes respectively. The details of data size are explained in Table 1.1.

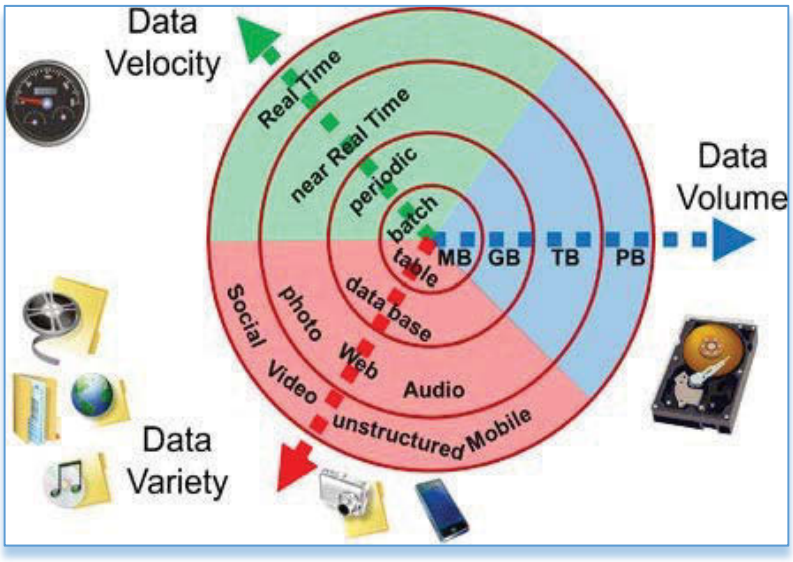

Figure 1.1 3Vs model of Big Data

Table 1.1. Dataset Volume Size

VALUE	ABBREVIATION	NAME
1000^1	KB	Kilobytes
1000^2	MB	Megabytes
1000^3	GB	Gigabytes
1000^3	GB	Gigabytes
1000^4	TB	Terabytes
1000^5	PB	Petabytes
1000^6	EB	Exabytes
1000^7	ZB	Zettabytes
1000^8	YB	Yottabytes

1.1.1. Volume

Volume describes the data size, scale or amount, and illustrates how datasets are extremely large. Big Data volumes can easily reach petabytes, even yottabytes, of

information. An excessively large volume of data is not only a storage issue, but also a massive analysis issue.

1.1.2. Variety

Variety illustrates the complexity of datasets, with datasets including both structured and unstructured data in different formats. Examples include documents, emails, audio files, images, videos, click streams, log files, mobile data, and financial transactions. Hundreds, even thousands, of different attributes in multiple dimensions create too many combinations and varieties for traditional database management tools to handle. This is because there is no primary key to link these datasets. For example, during the FIFA 2014 World Cup final between Germany and Argentina, there were 88 million Facebook users who collectively sent 280 million Facebook interactions including posts, comments or likes (Lorenzetti 2014). These interactions were communicated by hundreds of different data providers, using hundreds of different types of devices, and contained thousands of different comments that were posted and received in thousands of different cities around world. This creates a huge number of combinations and varieties in Big Data analytics.

1.1.3. Velocity

The velocity of a dataset is the speed of data generation, and illustrates how fast the datasets are being produced. Based on Royal Pingdom's Internet 2012 in numbers (Pingdom 2013), there were more than 57,870 Google searches launched per second, and 1.7 million emails sent worldwide every second. During the FIFA 2014 World Cup final, more than 10,312 Twitter messages were sent per second at the moment of Germany's victory (Lorenzetti 2014). Those fast growing datasets increase the difficulty for real-time Big Data visualization.

1.1.4. Value and Veracity

Later approaches have since added Value and Veracity into Big Data characters, therefore amending the 3Vs model into a 5Vs model. Yuri Demchenko (Demchenko et al 2013) drew the 5Vs model to illustrate the characters of Value and Veracity, which are shown in Figure 1.2.

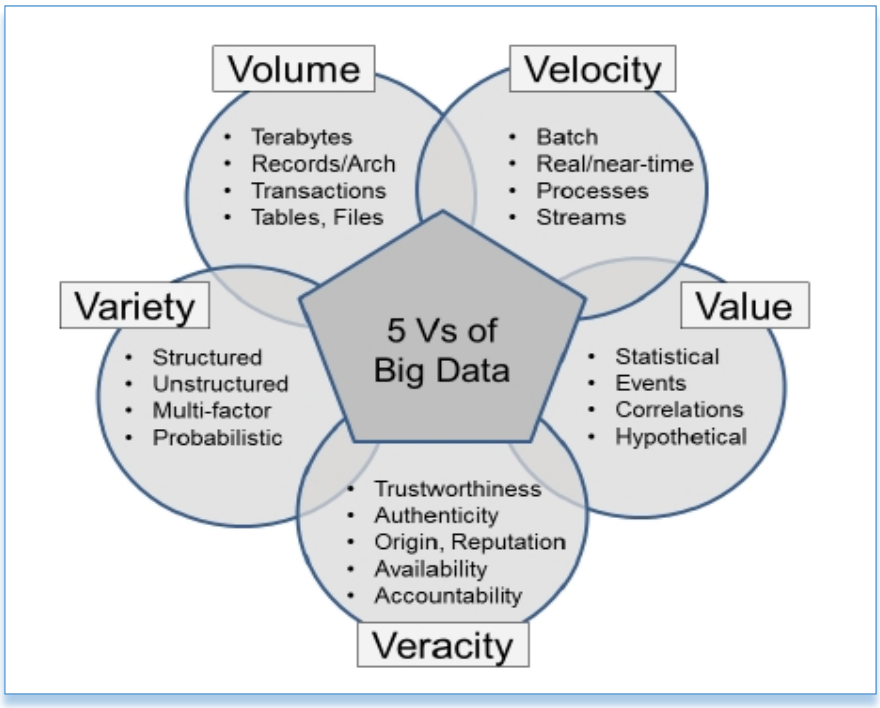

Figure 1.2 5Vs model of Big Data

The value of a dataset indicates the potential worth of the data, incorporating the high dollar value placed on data by government, organization, and business practices. For example, in 2012 the retailer giant Target created a Pregnancy Prediction model (Hill 2012), which analysed customers' consumption patterns from their shopping datasets to deduce whether a customer was pregnant long before they needed to start buying diapers. This Customer Relationship Management (CRM) system helps Target predict when goods needed to be increased, which stores they needed to be increased in, and which items needed to be added to which shop floor. Another example is Starwood Hotels & Resorts Worldwide Inc., which had created a Revenue Optimization System (Norton

2015) to analyse their huge datasets and determine how to push promotional offers on room prices in different seasons onto potential guests. This had increased Starwood's revenue from $128 million to $234 million a year later.

The veracity of a dataset represents the data consistency and trustworthiness. Veracity ensures that accurate data is trusted and protected during the whole data lifecycle, by collecting from trusted data sources, processing on trusted computer facilities, and storing data in protected data storages. It provides accountability for decision-making in government, organization and business practices.

1.2. Big Data Visual Analytics

Distributed computing and parallel processing techniques are widely used in industry for Big Data applications. Hadoop (High-availability distributed object oriented platform), the most popular open-source platform for reliable, scalable and distributed computing, is often referenced by Big Data researchers. Two main core frameworks in Hadoop are Hadoop Distributed File System (HDFS), which provides high-throughput access to application data, and MapReduce, which schedules jobs and tasks for parallel processing of large datasets. Both these core frameworks are used in industry for managing cluster distributed data centres, such as Facebook, Twitter, IBM, Adobe, Google, Yahoo, New York Times, LinkedIn, Amazon, Alibaba, and eBay (Hadoop 2014).

Big Data researchers have utilised HDFS and MapReduce for data processing, data sharing and data clustering because it has the ability to make a query over a dataset, divide the dataset into many small fragments, and run the fragments in parallel computing – thus reducing the time needed for data processing, data sharing and data clustering (Kraska 2013; Narayan et al 2012; Menon 2012). In addition, it provides a distributed system that stores data on the computer nodes, with high aggregate bandwidth across the

cluster for Big Data visualization. Currently, Big Data visual analytics has three main practices: dataset visualization, data-type visualization, and particular topic visualization.

1.2.1. Dataset Visualization

Big Data visualization is mostly practiced on a single dataset. Jibonananda Sanyal (Sanyal et al 2010) visualized a weather dataset by using a software tool called Noodles that was developed for meteorologists to help visualize ensemble uncertainty. They developed uncertainty ribbons, which quantify the uncertainty along a value's contour from the ensemble means, to visualize 2D uncertainty. Yu-Shuen Wang (Wang, Y.S et al 2011) visualized a scientific dataset by using feature-preserving and focus+context to reduce the dataset volume, and combined transfer function drives, continuous voxel repositioning and resampling techniques to demonstrate their 3D visualization. Steffen Hadlak (Hadlak et al 2011) visualized a social network dataset by developing a situ visualization method in their new classification approach, based on graph structure and temporal domain. Tamara L. Berg (Berg et al 2010) visualized an image dataset with text features using their recognition approach, which provides two classification methods: text classifier and visual classifier for the image recognition. The output combined these two strategies, preventing errors occurring if one classifier fails.

1.2.2. Data-Type Visualization

Big Data contains text, image, video, audio, mobile and other form of data, which normally uses different visual techniques to visualise different data types. Amir H. Meghdadi and Pourang Irani (Meghdadi 2013) visualized a video dataset by creating a Selective Video Summarization and Interaction Tool (sViSIT) that summarizes and visualizes both video content and trajectories. The tool extracts the motion paths of moving objects in surveillance video, and visualizes those motions in a space-time cube.

Jinglan Zhang (Zhang, J., Huang, K., et al 2013) visualized the audio data of a bird's audio dataset, and used time-frequency, tags-linking and GeoFlow as the visualization techniques for audio data visualization. Xiaotong Liu (Liu et al 2013) developed CompactMap to streamline the text data for visual analytics. Shehzad Afzal (Afzal et al 2012) visualized text data using automatic typographic maps that merge text and spatial data into visual lines. Their technique wraps text along paths to create lines, fills polygons with text to builds shapes, and colours in lines in visualization graphics.

1.2.3. Particular Topic Visualization

Most approaches in Big Data visual analytics focus on a particular topic during the visual algorithm progress, such as word cloud visualization (Cui et al 2010), spam email visualization (Zhang, J., Huang, M.L, Hoang, D 2013), diabetic retinopathy visualization (Rocha et al 2012), and flood DDoS attack visualization (Zhang, J., Huang, M.L 2013). Lei Shi (Shi et al 2013) visualized the network traffics by using the Structural Equivalence Grouping (SEG) method, which grouped multiple sub-nodes into one node during network traffic visualization.

1.3. Big Data Behaviours

In Oxford dictionary, the word "behaviour" is descripted as "the way in which one acts or conducts oneself, especially towards others" (Oxford Dictionary online 2016). In Wikipedia, it is consisted as actor, operation, interactions, and their properties. The behaviours are normally demonstrated as human behaviours, chemical behaviours, physical behaviours, health behaviours, or scientific behaviours.

Jorge Azorin-Lopez (Azorin-Lopex et al 2015) developed self-organizing activity description map for representing and classifying the customer behaviours from shopping

centre video clips. Yufei Chen (Chen, Y. and Shen, C. 2017) analyzed smartphone data to measure phone owner's movement patterns. Khandakar Tareq Alam (Alam et al 2016) studied human behaviours based on social network data, and categorised human acts such as "Happy", "Sad", and "Emotional" while questioning for "Journey Lovers", "Music Lovers", Movie Lovers", "Food Lovers", and "Sports Lovers". They used bar chart to illustrate behaviours pattern they found. Figure 1.3 shows their analyzing methods.

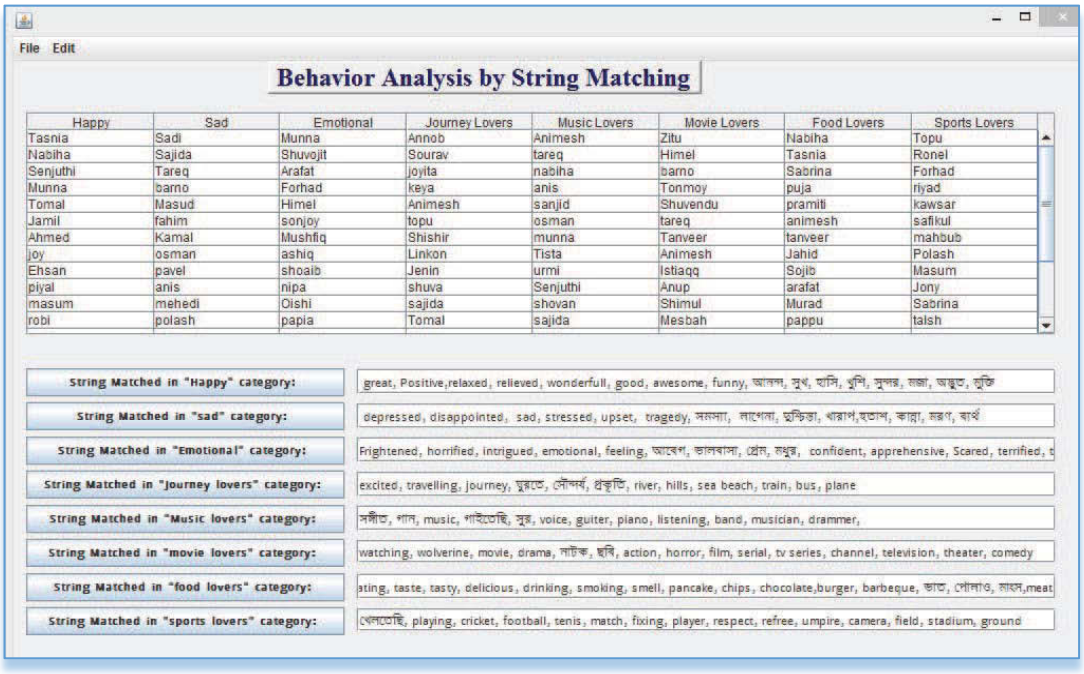

Figure 1.3 Behaviour analysis by string matching (Alam et al 2016)

Considering Big Data has combined the different datasets with variable data forms, I have defined Big Data behaviours as a set of concepts and categories that descripts Big Data's acts for; **When** did the data occur, **Where** did the data come from, **What** did the data contain, **How** was the data transferred, **Why** did the data occur, and **Who** received the data. It can also be descripted as **5W1H**, or **5Ws** which indicates the same concepts. I will discuss Big Data behaviours in the different visualization techniques in following sections.

1.4. Visualization Techniques

Big Data researchers have already attempted to reduce the dimensions in their visual approach when visualizing unstructured data. Seungwoo Jeon (Jeon et al 2013) transformed unstructured email texts into a graph database to visualise email texts. Richard K. Lomotey and Ralph Deters (Lomotey et al 2013) extracted topics and terms from unstructured text data by using the TouchR2 tool that they created for visualization. Woo Sik Seol (Seol et al 2013) proposed a reduction in the number of association rules to reduce multidimensional data attributes in their visualization approach.

Popular Big Data visualization techniques include: bubble charts/scatter plots to display data by using bubbles of different sizes and colours; word clouds and heat maps to represent statistical results by using different fonts and colours; Geography maps to illustrate data in geographic maps; histograms to display data aggregation; tree maps and data clustering to group and classify similar data attributes together; and parallel coordinates to demonstrate the relationships between multidimensional data.

I will discuss parallel coordinates technique in more detail later since Big Data contains huge multidimensional data, both structured and unstructured data, and because I have used parallel coordinates throughout my visual analytics in the following chapters.

1.4.1. Visualization Techniques for Low-Dimensional Data

The popular Big Data visualization techniques for low-dimensional data include bubble charts, scatter plots, word clouds, heat maps, Geography maps and histograms.

1.4.1.1. Bubble Charts/Scatter Plots

Scatter plots display data as a point with only two variables, normally indicated on both horizontal ‘x’ and vertical ‘y’ axes. A bubble chart is a chart that can display up

to four dimensions of data. The size and colour of the bubble demonstrate two dimensions, and the data's position in the 'x' and 'y' coordinates represent the other two dimensions. Recently, researchers have developed bubble plot in tree structures and clustering structures to group and classify similar data together.

The famous Big Data bubble chart was created by Hans Rosling (Rosling, H., 2009), and has become famous due to its use in a TED Talk. He compared the life expectancy, income per person and total population for every country, classified by region, over the last 200 years. He explained **Where** the population came from, **What** life expectancy were, **How** income per person moved, and **When** was happened. The graph is shown in Figure 1.4.

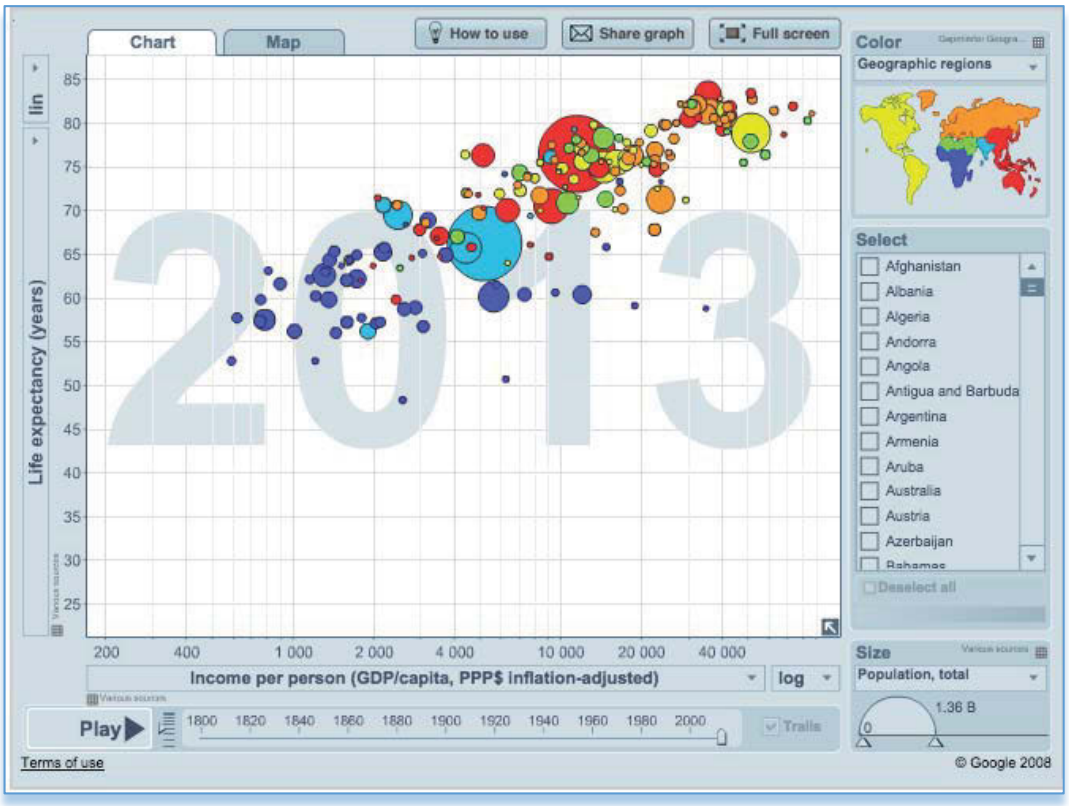

Figure 1.4 Hans Rosling's bubble chart

In Figure 1.4, there are four data dimensions illustrated in the graph corresponded for **Where**, **What**, **How** and **When**: life expectancy (years) is plotted in the 'y' axis;

average income per person is shown in the ‘x’ axis; population is illustrated through the bubble’s size; and the country’s geographic region is classified by different colours. The data’s time series of 200 years allows viewers to select a year, or alternative play a video speeding through many years worth of data in order to see the bubbles move over time. This makes the chart interactive, attractive and easy to understand. A particular country can be selected and tracked over time by clicking on the country’s bubble, or by selecting it in the right-hand control panel.

1.4.1.2. Word Clouds

Word clouds, also called tag clouds, are visual representations of text data which use different font sizes, colours and spaces to display data frequency and category. A large font size indicates the attribute’s high frequency compared to other attributes. Word clouds have been widely used in text data visualization for Big Data analysis.

Weiwei Cui (Cui et al 2010) used overlapping and compacting principles in their force-directed model to reduce overlapping text by removing empty spaces between the words. The repulsive force, attractive force and spring force were introduced in their approach to avoid text overlapping, to ensure that the text is stable and meaningfully laid out, and to reduce empty space. Ming-Te Chi (Chi et al 2015) used rigid body dynamics with boundary constraints and overlap free techniques to arrange word tags in a compact layout graph. They illustrated **What** Apple products became popular in market, **When** those products launched. Figure 1.5 shows their word cloud visualization approach.

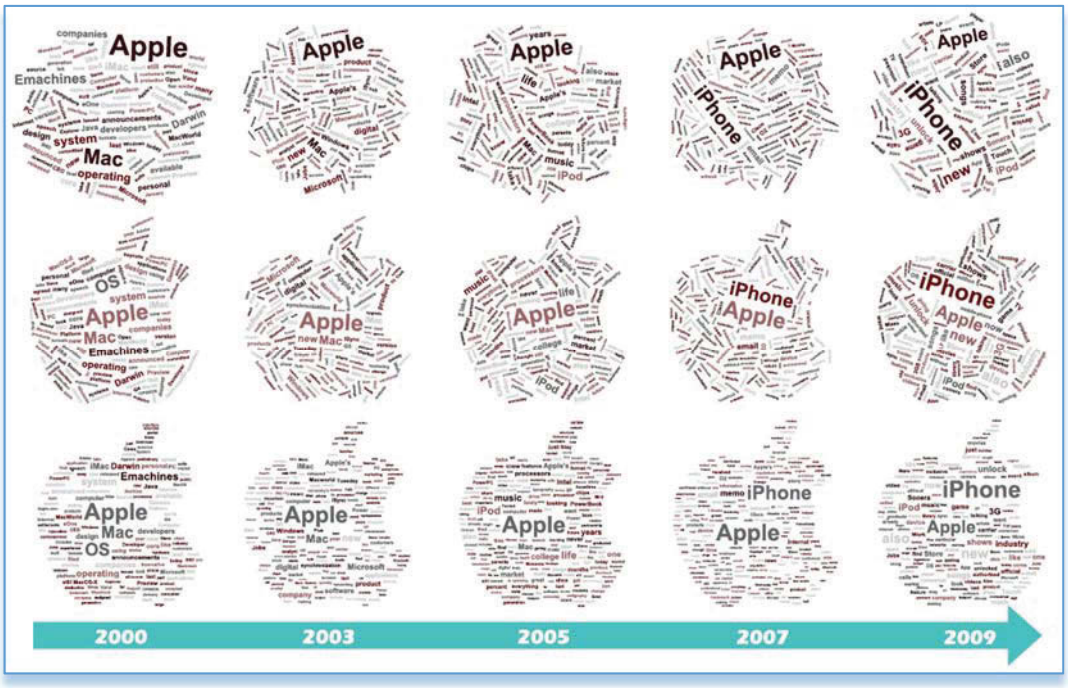

Figure 1.5 Example of word cloud

In Figure 1.5, the authors visualized the production dataset for Apple from 2000 to 2009 by using their word cloud approach which illustrated the company behaviours for **What** and **When**. The top row shows the normal layout for the word cloud in initial stage. The middle row applied the boundary constraints, which are shaped as the logo of Apple. The bottom row deployed overlap free technique to arrange the word tags. In 2000, Mac OS was the main product in the Apple brand. The iPod was launched in 2005, sparking the music-related tags. In 2007, the iPhone was released, and it became the company's biggest tag in 2009.

1.4.1.3. Heat Maps

Heat maps are used to represent two-dimensional data statistical values in a data matrix. The colour of each tile represents the data value within a range of numerical values. Recently, researchers have used geographical heat maps to illustrate weather data, and tree heat maps to catalogue the data classification.

Daniel Cheng (Cheng et al 2013) proposed Tile-Based Visual Analytics (TBVA) to explore one billion pieces of Twitter data. TBVA created tiled heat maps and tiled density strips for Big Data visualization. SAS Visual Analytics Explorer (Abousalh-Neto 2012) scales, visualizes and analyses massive datasets to find visual patterns. They created heat maps for Bank World's activity based on the latitude and longitude of each log entry. They demonstrated the ATM behaviours on **Where** the ATM located, **What** the activity value were. This graph is shown in Figure 1.6.

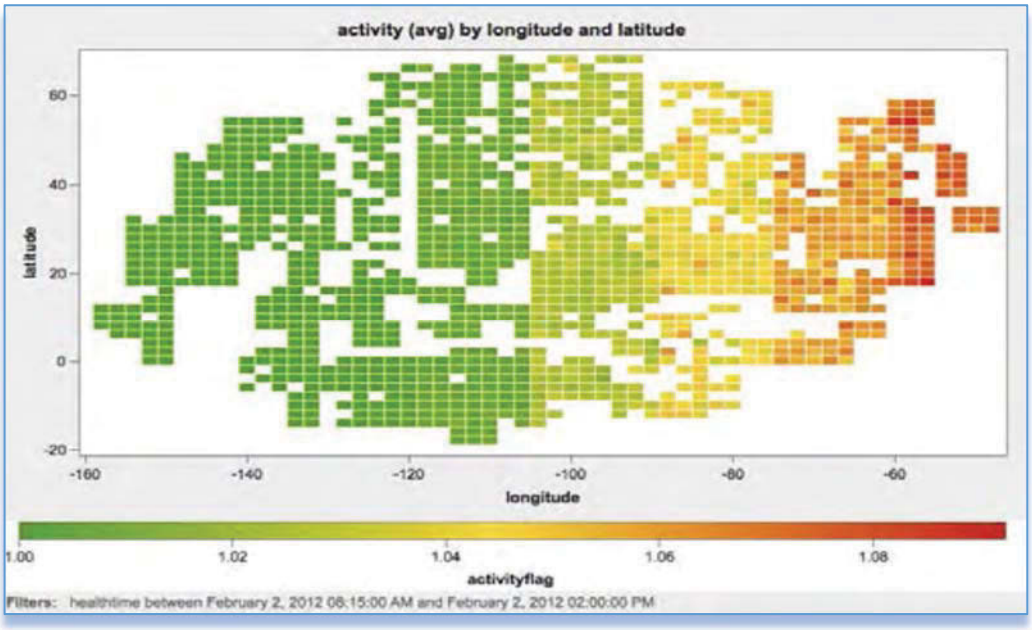

Figure 1.6 Heat map of Bank World's activity

In Figure 1.6, the authors display the average activity of Bank World including logs for ATM activity which demonstrated ATM behaviours for **What** and **Where**. The red colour symbolises a high value of activity while a green colour indicates a lower level of activity. An activity flag value above 1.06 illustrates the highest level of activity, which occurs around –60 in longitude. A longitude between -130 and -160 has the lowest level of activity, with an activity flag value below 1.02. The heat map also suggests that latitude value does not significantly influence the level of activity, as there is no pattern between activities with high latitude and long latitude.

1.4.1.4. Geography Maps

Geography maps have been a popular data visualization tool ever since Google Maps launched in 2005. Since then, data can be illustrated in much more detail, and 3D dimensions were introduced with the introduction of Street View. Geographical maps have been widely used for social network data visualization, security data visualization, and weather data visualization.

Jason Dykes (Dykes, J and Brunsdon, C 2007) introduced the geographically weighted visualization approach to explore spatial relationships on a range of scales by using gw-choropleth maps, multivariate gw-boxplots, gw-shading and scalograms methods. 'Gw' stands for Geographically Weight, which summarizes the statistics of geographic features such as population GW, income GW or property GW. Their approach provides many more features on a geography map.

Jibonananda Sanyal (Sanyal et al 2010) visualized a weather dataset by using a software tool called "Noodles" that was developed for operational meteorologists to visualize ensemble uncertainty. They explained the weather behaviours on **Where** the weather uncertainty was, and **What** the weather uncertainty metrics value. This is shown in Figure 1.7.

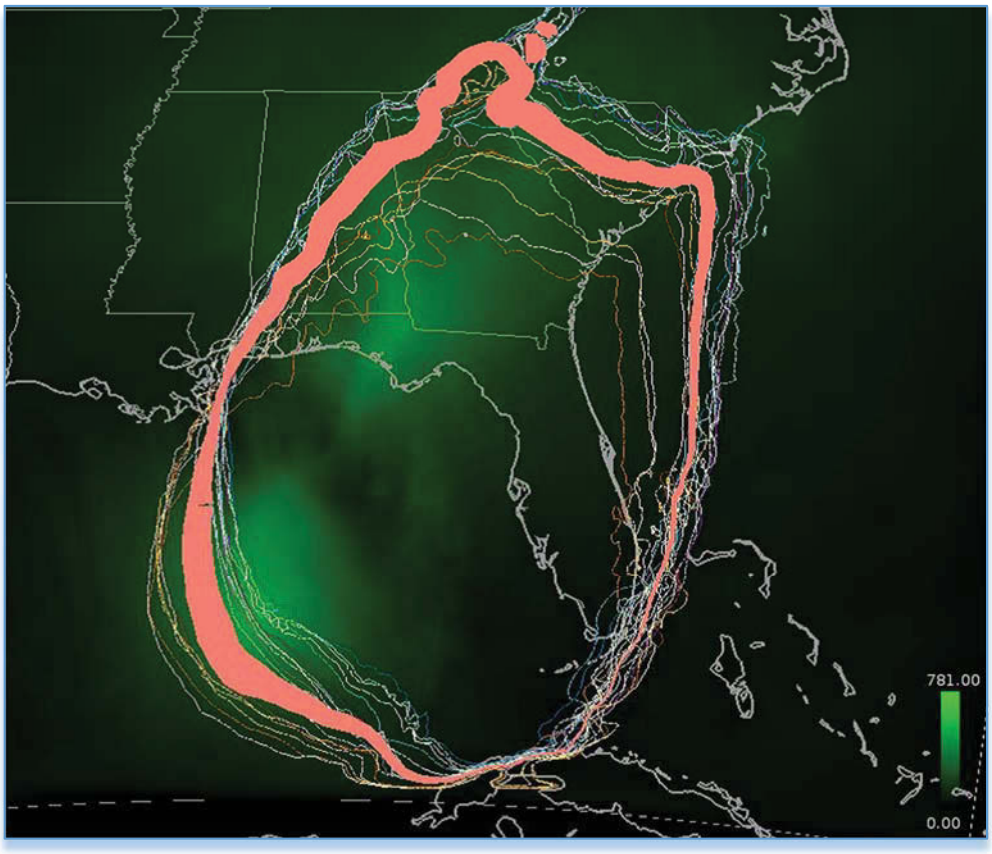

Figure 1.7 Uncertainty ribbon in geography map

In Figure 1.7, the authors analysed the weather behaviours for **Where** and **What**, and created the uncertainty ribbon, which quantifies the uncertainty along a value contour from the ensemble means, to visualise 2D weather uncertainty metrics in a geographical map.

1.4.1.5. Data Histograms

Data histogram is a popular visual tool for estimating the probability of a continuous numerical variable. The entire range of possible values is divided into small series intervals, which are illustrated on the horizontal axis. The vertical axis counts the number of values that fit into each series interval, and displays the height of each interval using rectangles. Histograms therefore visualise the density of data values, and provide estimations of variations in the data trend.

Petr Sereda (Sereda et al 2006) used LH histograms to visualize the complex properties of transfer function domains for the selection of boundaries between materials in their CT and MR volumetric data. Zhao Geng (Geng et all 2011) introduced angular histograms and attribute curves to visualize a real-world animal tracking dataset, and they combined parallel coordinates and histograms to visualize large and high dimensional data. Peter Torrione (Torrione et al 2014) explored the relationship between GPR (Ground Penetrating Radar) and HOG (Histogram of Oriented Gradients), and classified the sample GPR data in histogram chart. The GPR data behaviours they had focused on **When** the data occurred, and **How** to value the data down-sample. The histogram chart is shown in Figure 1.8.

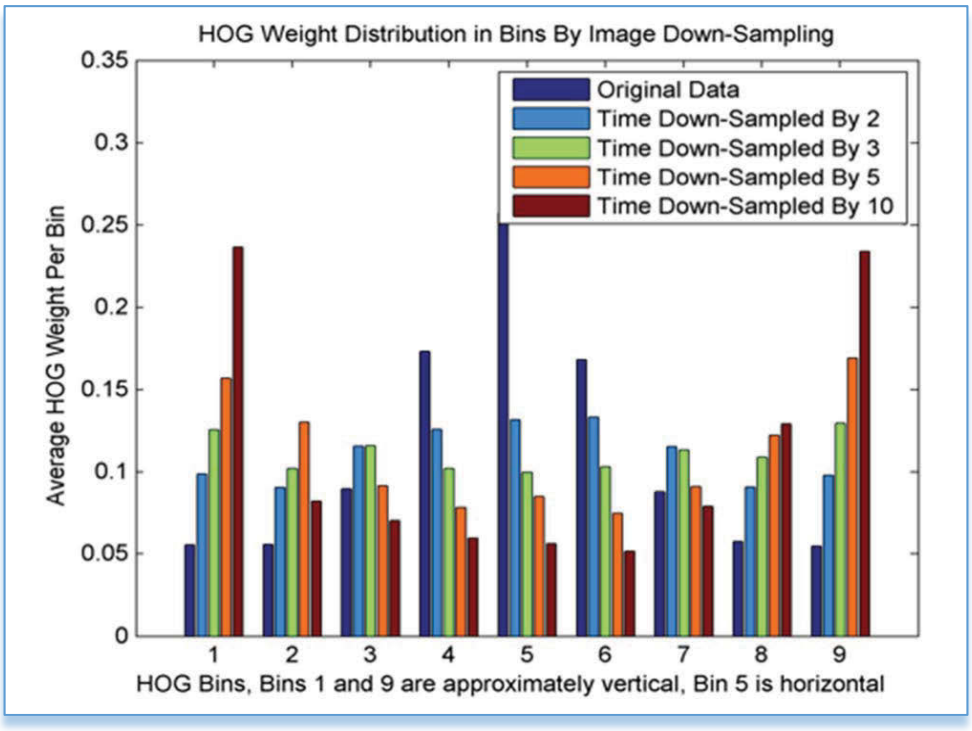

Figure 1.8 HOG histogram bin with different GPR downtime

In Figure 1.8, GPR data behaviours for **When** the data occurred, and **How** to value the data down-sample were illustrated in histogram chart. Bin 5 has highest value of

down-sample, and Bin 9 has lowest one. The ratio explained how they measured GPR data in their proposal.

1.4.2. Visualization Techniques for High-Dimensional Data

The popular Big Data visualization techniques for high-dimensional data include tree maps, data clustering and parallel coordinates.

1.4.2.1. Tree Maps

Tree maps are a popular Big Data visualization tool for displaying structured hierarchical data as a set of nested rectangles by using different sizes and colours to represent the dimensions. The leaf size, which is illustrated as a rectangle node represents the data value, while the colour categorises the data value and separates it from other leafs. In the visualization graph, the leaves are gathered together in order to further classify thousands of different leaves by their data attributes.

Jie Liang (Liang, J et al 2015) introduced various shapes to visualize large tree structural data which combined polygonal, angular and rectangular titling to create a tree map with a flexible layout. They studied the file behaviours which located in "Program Files" directory in Window 7 systems, and focused on **Who** the files are belong to, **What** files type were, **How** to value the file size. The result is shown in Figure 1.9.

Figure 1.9 Large file system with the different shapes in treemaps

In Figure 1.9, the file behaviours have been illustrated in treemap. Tags inside of treemap illustrated **Who** the files are belong to, the colour represented the different file extension for **What** files type were, and the size of sharps illustrated for **How** to value the file size. There are 145,000 files and folders illustrated in the flexible tree map, which combined irregular polygons, rotated rectangles and vertical-horizontal rectangles.

Roel Vliegen (Vliegen et al 2006) visualized business data and transferred it into tree maps by using several different tree map algorithms that combined the strengths of both business graphics and tree maps. They studied business calling data behaviours for **Who** did business connection, **How** they communicated, **What** volume they have made so far. Figure 1.10 shows their approach and results.

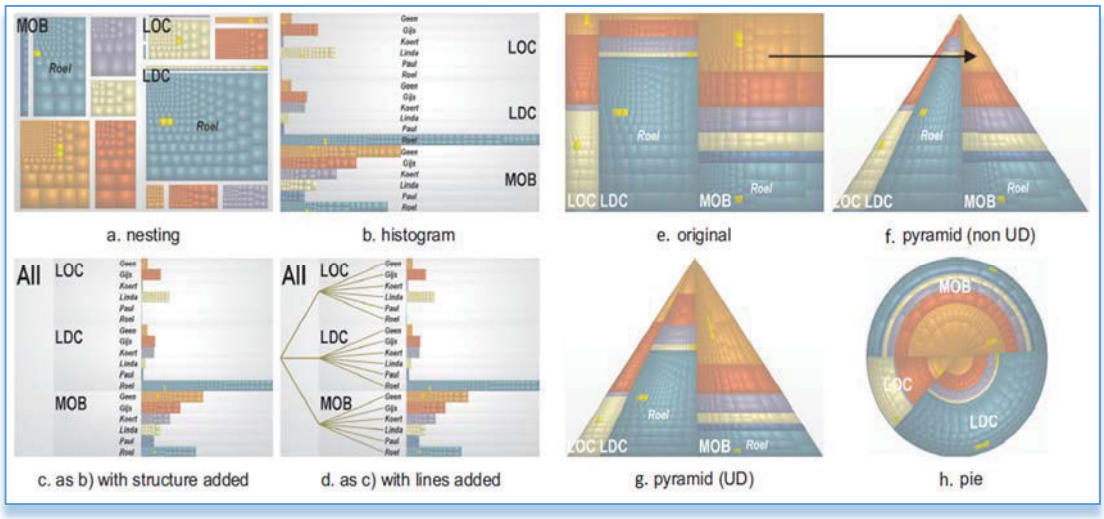

Figure 1.10 Matrix structure and transformation used in treemaps

In Figure 1.10, the business calling data behaviours have been illustrated by combined treemaps. The colour represents staff **Who** did business connection. Six different call methods; LDC (Long Distance Call), LOC (Local call), MOB (Mobile), INT (International call), SRV (Service call), and OTH (Other call) represented **How** they communicated. The value of sharps measured **What** volume they have made so far.

There are six persons who made calls: Gijs, Paul, Roel, Linda, Koert and Geen. UD means that the visualization density is uniform. Figure 1.8a) illustrates the nesting tree map for six persons in different groups, which can also be illustrated by a histogram graph as per Figure 1.8b). Figure 1.8c) explains the business structure, and Figure 1.8d) adds lines to the structure. From Figure 1.8e) to 1.8h), the authors have used different algorithm methods to transform the tree map into a pie chart, which is often used for business analysis.

1.4.2.2. Data Clustering

Data clustering is the process of organizing data into groups with similar elements for particular attributes, and has been widely used in Big Data visualization. Zhangye

Wang (Wang, Z., Zhou, J., et al 2013) clustered large-scale social data Microblog into user groups by using the attributes of user tag and user behaviour. The user behaviours they have focused on **When** user spend time on Microblog, **Who** user had followed, **How** user followed each other, **What** message user sent. Adding user tag, they created 6 clusters in their case study, which shown in Figure 1.11. The group outlined by the red rectangle having very high attribute values.

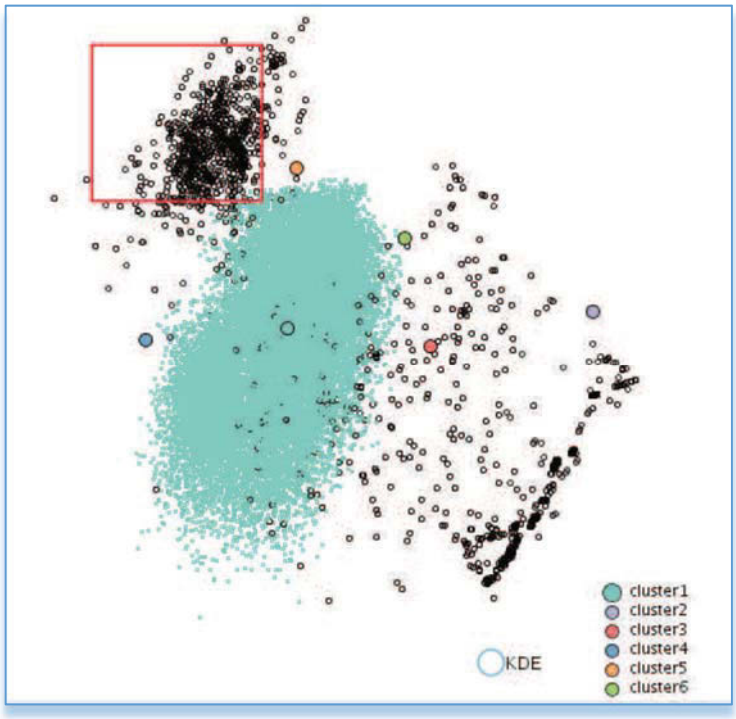

Figure 1.11 Group with high attribute value in red rectangle

Cheng-Long Ma (Ma 2012) used the K-means clustering method to find out the clustering centres for 3-D visualization. The distances of the three coordinate axes corresponded to the data in the original space, and they tested their model by using the Iris database. Zhenwen Wang (Wang, Z., Xiao, W., et al 2013) introduced ADraw for grouping the same attribute value nodes. They then created virtual nodes to group the same attribute value nodes together. Different virtual nodes are separated by different

colours in their visual analytics. They have tested four different datasets by using their ADraw model, and Figure 1.12 shows one dataset (blog network MSN) visualization.

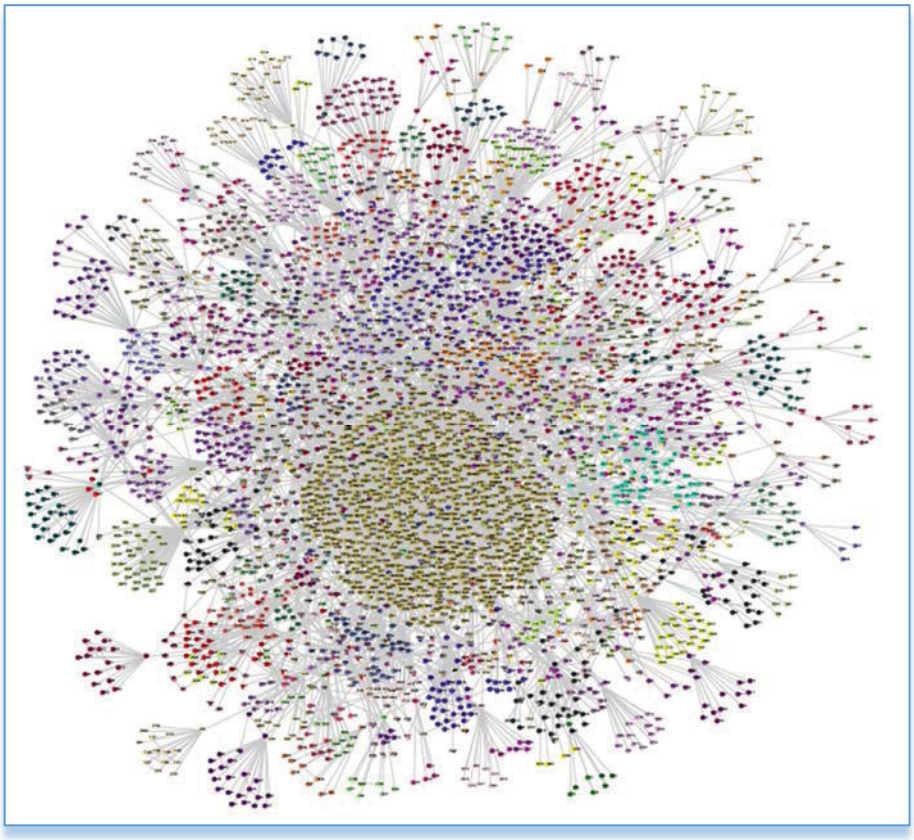

Figure 1.12 Blog network clustering by ADraw

In Figure 1.12, the authors focused on the Blog network behaviour for **Where** the data came from, **How** the data was connected, **Who** received the data, and **What** the data contained. They used the different colours as clustered nodes and combined with a tree structure to visualize blog networks and illustrate their complexity.

1.4.2.3. Parallel Coordinates

Parallel coordinates were introduced by Alfred Inselberg and Bernard Dimsdale (Inselberg et al 1990), who drew polylines between independent axes at appropriate values. Each axis represents a dimension, with the relationship between the axes joined by polylines show the data's frequencies, relationship and aggregation patterns. Parallel

coordinates are widely used for multidimensional data visualization. Parallel coordinate plots (PCP) are a simple but strong geometric high-dimensional data visualization method, and represents N-dimensional data in a 2-dimensional space with mathematical rigorousness. PCP, together with scatterplots and radar charts, have been widely used for visualizing multivariate datasets (Claessen et al 2011).

Parallel coordinate can demonstrate up to 100 dimensions in graph, so it is probably the best visualization tool for Big Data visual analytics. However, parallel coordinates have always suffered when visualizing Big Data, since the polylines clutter and crowd each other. This crowding of polylines reduces the ability for the user to identify data patterns from visual graphs, particularly when dealing with multiple datasets. Visual clustering, axes rotation or reordering, and reducing dimensions and feature selection are three common methods employed to reduce clutter in parallel coordinates.

1.4.2.3.1 Visual clustering

Clustering in parallel coordinates is the process of either revealing new cluster centers or expanding existing clusters. Clustering is normally bound to two-dimensional clusters, or to merging two-dimensional clusters into one cluster depending on the focus. Matej Novotny and Helwig Hauser (Novotny et al 2006) grouped data content into outliers, trends and focus, and set up three clustered parallel coordinates to reduce these cluttering issues. Kai Lun Chung and Wei Zhuo (Chung et al 2008) developed two visual analytic tools, selection graphs and relation graphs, to reduce visual clutter in parallel coordinates. A selection graph is a brushing tool which helps users highlight the regions selected, while a relation graph organizes clusters and provide interactions for users to explore the relationships between clusters. Julian Heinrich (Heinrich et al 2011) developed the BiCluster Viewer, which combines heatmaps and parallel coordinate plots to uncover

data patterns. The BiCluster Viewer contains many interactive features such as axis ordering, line coloring or zooming, which decrease data cluttering in visual graph. Hong Zhou et al (Zhou, H et al 2008) converted straight lines into curved lines to reduce visual clutter in clustered visualization. They also utilized a splatting framework (Zhou, H et al 2009) to detect clusters and reduce visual clutter. Figure 1.13 shows the principle of their curved lines.

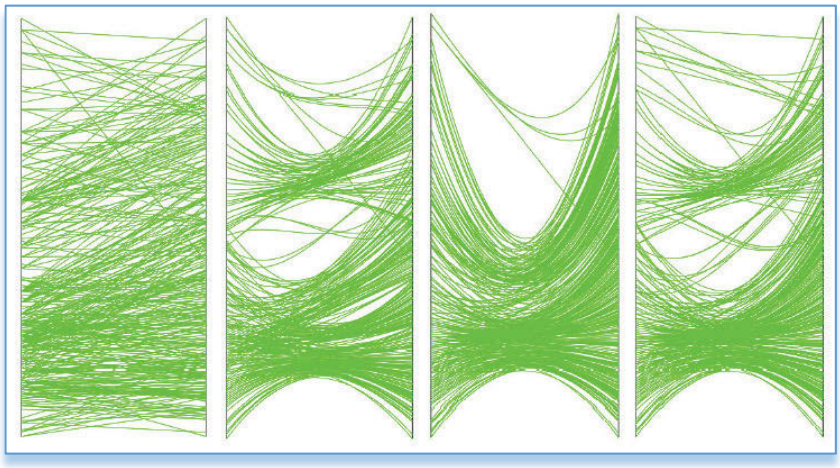

Figure 1.13 Curve edges to reduce the visual clutter

In Figure 1.13, the curved lines merged similar attributes from the left-hand side across to the right-hand side, thus reducing cluttering and leaving space to display more data. Clustering in parallel coordinates is a very useful technique to observe data pattern and structure with less clutter.

1.4.2.3.2 Axes rotation and reordering

The axes used in parallel coordinates are typically vertical and equally spaced, and data patterns and relationships can be illustrated as polylines between the axes. In order to explore particular relationships and patterns, these axes may need to be rotated or reordered. This not only demonstrates the relationship and pattern clearly, but also reduces the overlap and clutter of polylines.

Each vertical axis represents a dimension, characterised by numerical or non-numerical data. Importantly, each vertical axis does not have a standardised scale. For example, a numerical axis can have either an increasing or decreasing scale, depending on the initial layout of the graph, and on which particular patterns or relationships between axes are to be highlighted.

Liang Fu Lu (Lu, L.F et al 2012) proposed a similarity-based method to reorder the parallel axes which reduces clutter. The authors combined PCC (Pearson's Correlation Coefficient), SVD (Singular Value Decomposition), and NCC (Nonlinear Correlation Coefficient) visualization techniques to create a better visual representation. Aritra Dasgupta and Robert Kosara (Dasgupta et al 2010) proposed a model based on screen-space metrics to optimize the arrangement of axes. Koto Nohno (Nohno, K., et al 2014) introduced contractible parallel coordinates to reorder the parallel axes. This was done using spectral analysis of a weighted graph, which was composed by referring to data correlation among multiple dimensions, as shown in Figure 1.14.

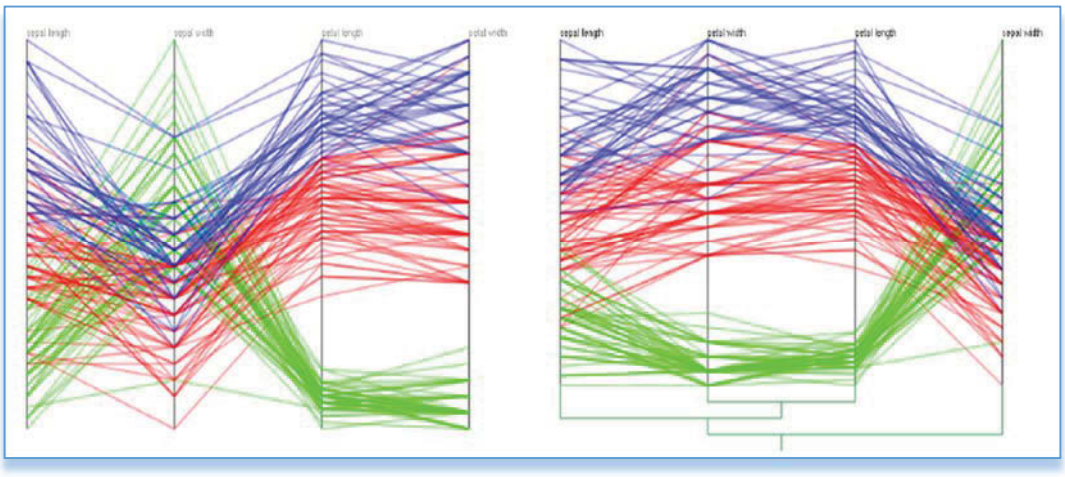

Figure 1.14 Axes re-ordering to reduce the visual clutter

In Figure 1.14, the left-hand graph is the original while the right-hand graph is the re-ordered graph. The first and third axes have remained the same in both graphs, while

the second and fourth axis have swapped positions between the left-hand and right-hand graphs. This reordering has clearly reduced the clustering and overlapping of polylines.

1.4.2.3.3 Reducing dimensions and feature selection

Dimension reduction is the process of reducing the number of axes in the parallel coordinates graph. Feature selection is focused on the data attributes that optimize the feature and reduced the number of attributes in the parallel coordinates axes. Both have reduced the data cluttering inside of graph.

Jim X Chen (2001) introduced quad-tree mapping to reduce the number of dimensions in parallel coordinates. Myung-Hoe Huh and Dong Yong Park (Huh et al 2008) introduced proportionate spacing between two adjacent axes, rather than the equal spacing used in conventional PCP parallel axes, in order to focus content between these two adjacent axes. Moreover, the curved lines used also possessed some statistical properties linking data points on adjacent axes. Tuan Nhon Dang et al (Dang et al 2010) proposed a visualization and interaction method to avoid overplotting and to preserve density information by stacking overlapping lines. Geoffrey Ellis and Alan Dix (Ellis et al 2006) developed three methods: raster algorithm, random algorithm and lines algorithm to measure occlusion in parallel coordinate plots and thus provide tractable measurement of the clutter. Xiaoru Yuan (Yuan et al 2009) scattered points in parallel coordinates to combine parallel coordinates and scatterplot scaling. This enlarged the spaces between the four axes, allowing for greater exploration of context and patterns, and reducing data crowding. Figure 1.15 illustrates their implementation graph.

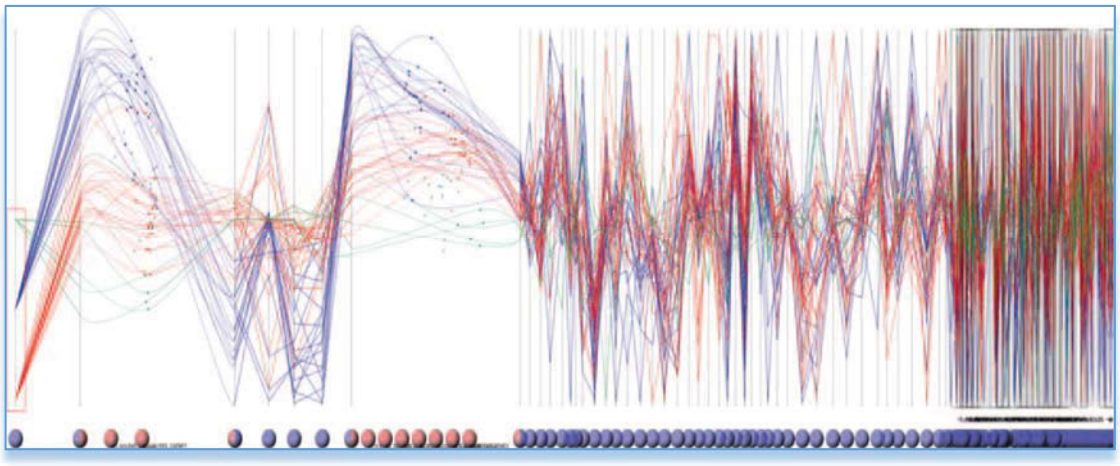

Figure 1.15 DNA Microarray data with Scattering Points in Parallel Coordinates

In Figure 1.15, up to 100 dimensions have been illustrated. The left-hand axes have been enlarged and focused while the right-hand axes have been shrunk and optimized. The graph illustrates clearly that visualization is affected by the selected axes, attributes and features, with data cluttering clearly reduced on the enlarged axes. Authors studied DNA Microarray data behaviours and focused on **How** classification needs to setup for possible early cancer diagnosis, **What** tissues were to identity the gene, and **When** the cancer diagnosis has to be exam again.

From above reviews, I found that the parallel coordinates can clearly illustrate Big Data behaviours for more than five dimensions, and I will use it throughout in this thesis.

1.5. Research Challenges and Motivations

The visualization techniques definitely help discover users the complex patterns and relationships that are hidden inside Big Data, but there are still limitations. Bubble charts, scatter plots, and data histograms have a maximum of up to four data dimensions. Word clouds can only illustrate text data, while heat maps and geography maps can only be used in two-dimensional (2D) graphs. Tree maps can deal with multidimensional data with clustering structures, but cannot compare data patterns using exact values. Parallel

coordinates and data clustering, therefore, are the best tools to visualize Big Data patterns, particularly for multidimensional data. However, these visualization methods suffer from data cluttering and overcrowded polylines.

The challenges faced by Big Data visual analytics is twofold. Firstly, how to classify Big Data patterns into categories that can be visualized for different datasets and data forms. Secondly, how to visualize Big Data patterns without the loss of information, particular while dealing with massive multidimensional data.

The challenges in classifying Big Data patterns for parallel coordinates using techniques that are suited across different datasets and data forms can be broken down into several steps.

- How to classify different data forms, such as text data, audio data or mobile data, in order to use visualization techniques
- How to measure Big Data patterns based on the data behaviours across multiple datasets
- How to value both structured and unstructured multidimensional data patterns that feed the parallel coordinates
- How to use parallel coordinates to identify Big Data patterns based on the data behaviours

Current Big Data visualization has three main foci: dataset visualization, data-type visualization, and particular topic visualization. Those three practices are based on single datasets of either only one data-type or one particular topic. My new approach classifies Big Data into 5Ws dimensions based on the data behaviour ontologies. This is suited for multiple datasets with any data-type or topic, since every data incident contains

these 5Ws dimensions. The 5Ws dimensions are posed as a set of concepts and categories that descripts Big Data's acts for; **When** did the data occur, **Where** did the data come from, **What** did the data contain, **How** was the data transferred, **Why** did the data occur, and **Who** received the data. To the best of my knowledge, no previous work had addressed Big Data visual analytics by using 5Ws dimensions that are suitable across multiple datasets for different data-types and multiple topics.

The challenges in visualizing Big Data patterns without the loss of information, particular when dealing with massive multidimensional data, can be summarized as the following.

- How to visualize Big Data patterns for different forms of data such as text data, mobile data, or audio data
- How to visualize structured and unstructured data patterns by using clustered parallel coordinates algorithm methods
- How to value Big Data attributes in order to gain a better understanding of data
- How to shrink and extend the data attributes on each dimension without losing information
- How to display multidimensional data on normal sized screens without losing data patterns

Parallel coordinates and data clustering are the best for visualizing Big Data patterns, particularly for multidimensional data. But data cluttering and overcrowded polylines are major issues, which cannot be resolved by visual clustering, axes rotation or reordering, and dimension reduction and feature selection since these methods inherently leave out information from the graph. My new approach has created Pair-Density to

measure and compare data patterns in parallel coordinates irrespective of their data forms. Pair-Density not only values all Big Data patterns, but also creates two additional non-dimensional axes to compare data patterns in parallel coordinates. Data cluttering and overcrowding is dramatically reduced using Pair-Density parallel coordinates. To the best of my knowledge, no previous work had created two more axes by using Pair-Density in parallel coordinates to measure multidimensional data patterns.

1.6. Author's Contributions in the Thesis

The contributions of this thesis are the introduction of two new methods for Big Data analysis and visualization. The first contribution is the creation of 5Ws dimensions for Big Data classification based on the data behaviours ontologies across multiple datasets for any form of data. 5Ws parallel coordinates have been introduced to visualize patterns between 5Ws dimensions. The second contribution is the establishment of Pair-Density parallel coordinates to measure and compare 5Ws patterns by creating two non-dimensional axes. Data cluttering and overcrowding of polylines have been significantly reduced using Pair-Density parallel coordinates.

Specific contributions of this thesis are:

- Establish 5Ws dimension to classify Big Data patterns based on the data behaviour ontologies suitable for different forms of data across multiple datasets

- Introduce Pair-Density to measure Big Data patterns to enable comparisons between any pair of dimensions for Big Data analytics

- Create two additional axes in the parallel coordinates by using Pair-Density to measure the visual patterns in Big Data visualization

- Introduce Shrunk Attributes (SA), which not only collect the value of elements not displayed in parallel coordinates, but also dramatically reduce data cluttering and overcrowding in Pair-Density parallel coordinates

This thesis contains the research approach and implementation results obtained by the author during his Ph.D period. The majority of methods and results have been published in **Seventeen** research papers in journals and conference proceeding by May 2016.

1.7. Thesis Organisation

The thesis is organised by following chapters:

Chapter 2 illustrates the data visualization of the 5Ws dimension classifications based on data behaviour ontologies. The approach can apply across multiple datasets for any form of data. Furthermore, 5Ws parallel coordinates is introduced to visualize the 5Ws patterns. Dimension clustering, Shrunk Attributes (SA) and noise data are also discussed in this chapter.

Chapter 3 illustrates the Pair-Density algorithm which compares any two dimensions for 5Ws patterns. Pair-Density parallel coordinates are also created by using Pair-Density as non-dimensional axes. In this chapter, ten different Pair-Density with different visual patterns are demonstrated and discussed.

Chapter 4 illustrates the case study, which contains three different cases implementing three different datasets. The first case study is based on US 2008 flight dataset (ASA, 2009), which has 29 data dimensions containing 1,048,575 flight incidents. The second case study is based on UTS Library 2009 email dataset, which has 17 data dimensions containing 585,300 email incidents. The third case study is based on

ISCX2012 network dataset (Shiravi et al 2012), which has 20 data dimensions containing 1,511,636 data incidents. The results of these three case studies show that our Big Data analysis and visualization techniques have significantly improved both the accuracy of data analytics and the visualization and interpretation of analysis.

Chapter 5 concludes this thesis by summarizing the research achievements, contributions and the possibility of future works.

Chapter 2: Data Behaviour Visual Analytics

Big Data collected from multiple datasets contains texts, images, audios, videos, mobile or other forms of data with multiple dimensions, and occurs every day through Facebook posts, Twitter comments, YouTube videos, Smartphone chats or email messages. I have analyzed these data attributes and classified its behaviours into the 5Ws dimensions, and then established the 5Ws patterns to measure its behaviours for Big Data analysis and visualization.

2.1. Multidimensional Data and Attributes

Assume that the first data incident d_1 contains m-dimension attributes. The data node can then be defined as:

$$d_{1m} = \{d_{11}, d_{12}, d_{13}, d_{1j}, \ldots, d_{1m}\}$$

Equation 2. 1

where j indicates the j^{th} dimension and attribute d_{1j} illustrates the 1^{st} data incident in the j^{th} dimension. Therefore, a whole dataset that has n incidents with m dimensions can be illustrated as

$$D = \begin{Bmatrix} d_{11} & d_{12} & d_{13} \cdots & d_{1j} \cdots & d_{1m} \\ d_{21} & d_{22} & d_{23} \cdots & d_{2j} \cdots & d_{2m} \\ d_{31} & d_{32} & d_{33} \cdots & d_{3j} \cdots & d_{3m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{i1} & d_{i2} & d_{i3} \cdots & d_{ij} \cdots & d_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & d_{n3} \cdots & d_{nj} \cdots & d_{nm} \end{Bmatrix}$$

Equation 2. 2

where $j=1,2,3,\ldots m$ indicates the number of dimensions and $i=1,2,3,\ldots n$ indicates the number of incidents.

The total number of attributes $n \times m$ in the dataset can reach millions, even billions, in number. For example, during the 2014 FIFA World Cup Final between Germany and Argentina, there were 280 million Facebook interactions including posts, comments and likes (Lorenzetti 2014). If we assume that these interactions contain only five dimensions (e.g. time, user, location, comment, device), the total number of attributes in the entire dataset would be 5 × 280 million = 1.4 billion attributes.

There will always be some duplicated attributes within each dimension. For example, assuming $j=2$ represents the sender's ages and $j=3$ represents the sender's location and attributes d_{13} and d_{33} both have the same location as "London", this means that $d_{13} = d_{33}$ = "London". The dataset D can then be illustrated as

$$D = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \boldsymbol{ages} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \boldsymbol{locations} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{Bmatrix} d_{11} & 23 & London & \cdots & d_{1j} & \cdots & d_{1m} \\ d_{21} & 24 & Sydney & \cdots & d_{2j} & \cdots & d_{2m} \\ d_{31} & 25 & London & \cdots & d_{3j} & \cdots & d_{3m} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ d_{i1} & 26 & Beijing & \cdots & d_{ij} & \cdots & d_{im} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ d_{n1} & 27 & New\ York & \cdots & d_{nj} & \cdots & d_{nm} \end{Bmatrix}$$

Equation 2. 3

Normal statistics for multidimensional datasets are calculated on a single dimension, such as the sum of total senders from the particular location "London", or the percentage of senders within the particular age range "23-25". The statistics result in dimension $j=3$ for attribute i = *"London"*, will be demonstrated as

$$Sum\_D_{(London,\ j=3)} = \sum_{i=1,j=3}^{n} (\ d_{(i,j)} | \ London)$$

Equation 2. 4

where $Sum\_D_{(London,\ j=3)}$ represents the statistical result for the attribute *"London"* at dimension $j=3$ in dataset D.

Based on the complexity of Big Data, the number of dimensions in Big Data datasets has rapidly increased, easily reaching hundreds, even thousands, of dimensions. This has raised issues of how to measure and visualize these very large volumes of multidimensional data across multiple datasets.

2.2. 5Ws Dimension and Behaviours Pattern

To establish a model suitable for any form of data across multiple datasets, I have classified the datasets into 5Ws dimensions based on data behaviours. Each data incident contains these 5Ws dimensions, which stand for; **When** did the data occur, **Where** did the data come from, **What** did the data contain, **How** was the data transferred, **Why** did the data occur, and **Who** received the data. Each of 5Ws dimension can be illustrated by using six sets.

- Set $T=\{t_1, t_2, t_3, \ldots,\}$ represents when the data occurred
- Set $P=\{p_1, p_2, p_3, \ldots,\}$ represents where the data came from
- Set $X=\{x_1, x_2, x_3, \ldots,\}$ represents what the data contained
- Set $Y=\{y_1, y_2, y_3, \ldots,\}$ represents how the data was transferred
- Set $Z=\{z_1, z_2, z_3, \ldots,\}$ represents why the data occurred
- Set $Q=\{q_1, q_2, q_3, \ldots,\}$ represents who received the data

Therefore, the dataset D can be denoted by the 5Ws dimensions as

$$D = \begin{Bmatrix} d_{11} & d_{12} & d_{13} \cdots & d_{1j} \cdots & d_{1m} \\ d_{21} & d_{22} & d_{23} \cdots & d_{2j} \cdots & d_{2m} \\ d_{31} & d_{32} & d_{33} \cdots & d_{3j} \cdots & d_{3m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{i1} & d_{i2} & d_{i3} \cdots & d_{ij} \cdots & d_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & d_{n3} \cdots & d_{nj} \cdots & d_{nm} \end{Bmatrix} = \begin{Bmatrix} d_{1T} & d_{1P} & d_{1X} & d_{1Y} & d_{1Z} & d_{1Q} \\ d_{2T} & d_{2P} & d_{2X} & d_{2Y} & d_{2Z} & d_{2Q} \\ d_{3T} & d_{3P} & d_{3X} & d_{3Y} & d_{3Z} & d_{3Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{iT} & d_{iP} & d_{iX} & d_{iY} & d_{iZ} & d_{iQ} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{nT} & d_{nP} & d_{nX} & d_{nY} & d_{nZ} & d_{nQ} \end{Bmatrix} \quad \textit{When, Where, What, How, Why, Who}$$

Equation 2. 5

The 5Ws dimensions have significantly simplified Big Data classification across multiple datasets for any form of data that feeds business, organization and government needs. Figure 2.1 illustrates Big Data classification for the 5Ws pattern.

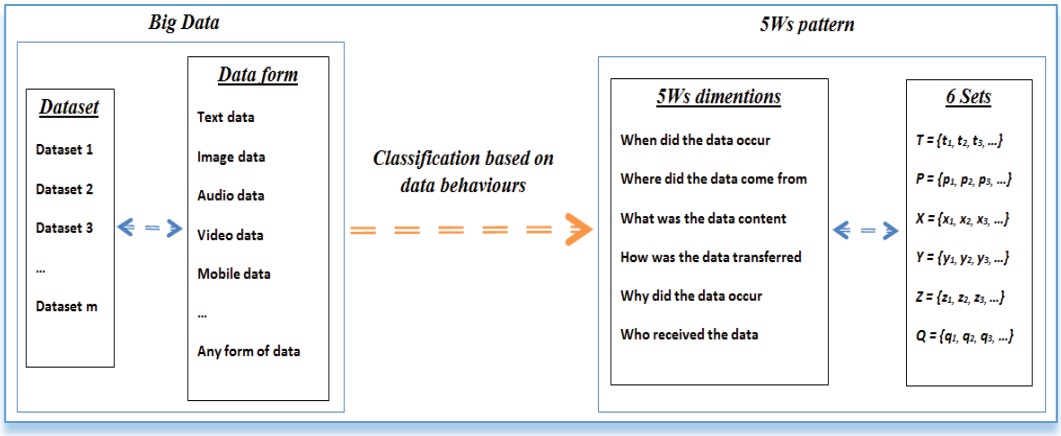

Figure 2.1 Big Data 5Ws pattern

According to the six datasets in the 5Ws dimensions, a data incident d_i can be defined as a node using the 5Ws pattern as $d_i\{t_i, p_i, x_i, y_i, z_i, q_i\}$. The dataset D with n incidents can therefore be illustrated as

$$D = \{d_1, d_2, d_3, \ldots, d_n\}$$

Equation 2. 6

2.3. 5Ws Parallel Coordinates

To visualize these 5Ws patterns, I have deployed parallel coordinates for 5Ws pattern visualization. Parallel coordinates are a popular information visualization tool for high-dimensional data, introduced by Alfred Inselberg and Bernard Dimsdale (Inselberg et al 1990). Each parallel axis represents a dimension and polylines are drawn between independent axes at appropriate values. The data examined using the axes shows data frequencies, data relationships and data aggregation patterns.

The value of each parallel axis can be any form of data, including numerical and text data. The order may be by alphabetical order, from 0 to 9 and A to Z. Figure 2.2 shows the 5Ws pattern parallel coordinates, using the example of the 2014 FIFA World Cup Final between Germany and Argentina. Overall, Twitter users sent 618,725 tweets per minute at the moment of Germany's victory (Lorenzetti 2014).

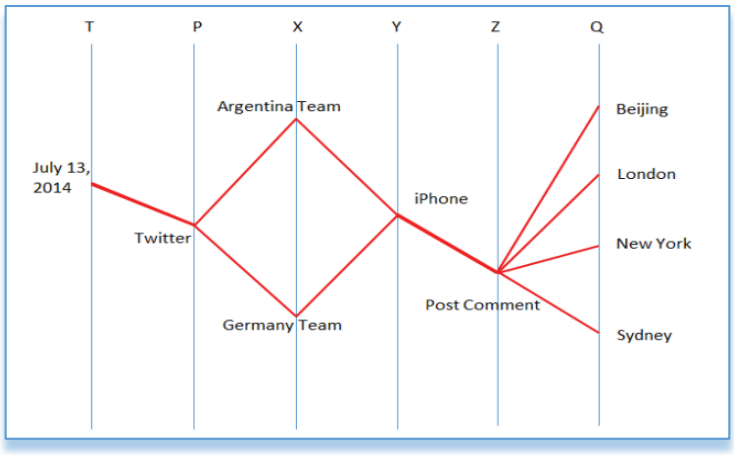

Figure 2.2 Example of 5Ws parallel coordinates

In Figure 2.2, assume that the particular data pattern illustrated contains the team names x_1 = "*Argentina Team*" and x_2 = "*Germany Team*", which were posted through iPhone only, and the countries whom received the data were q_1 = "*Beijing*", q_2 = "*London*",

q_3 = "*New York*" and q_4 = "*Sydney*". These particular data patterns can be illustrated in the 5Ws parallel coordinates, as per Figure 2.2.

5Ws patterns are not limited to just one dataset, but can also be used to compare multiple datasets. For example, a Facebook dataset and a bank transaction dataset are two different datasets. But similar attributes exist on both datasets, such as p = "users" and y = "mobile connection". By comparing these two datasets for p = "users" and y = "mobile connection", the user can find the ratio of internet banking mobile users via Facebook mobile users. Figure 2.3 shows an example of Big Data in the 5Ws data dimensions across multiple datasets and resources.

Big-Data	*What (X)*	*How (Y)*	*Why (Z)*	*When (T)*	*Where (P)*	*Who (Q)*
Social network						
- Facebook dataset	tag, text, photo, video	facebook user account	send to or receive from facebook	log on facebook	facebook user	facebook user
- Twitter dataset	text, photo	twitter user account	send to or receive from twitter	using email	twitter user	twitter user
Email network						
-Gmail dataset	text, attachment	gmail user account	send to or receive from email server	using email	email user	email user
Web logs						
- Google contents dataset	text, image, video	Google website	get information	online	Web site	anonymous user
Computer network						
- Traffic dataset	data exchange, attack	online network	send and receive data	anytime	send station	receive station
GPS						
- mobilephone tracks dataset	position	digital signal	data connection	mobilephone on	mobile phone	mobilephone station
Satellite data						
- Wether dataset	temperature, humidity	digital signal	position	anytime	weather sensor	satellite
Finance transactions						
- bank transaction dataset	amount, account	online banking	finance needs	anytime	bank account	bank account
Video streams						
- YouTube dataset	video	Internet	video sharing	online	Web	anonymous user
Smart phone						
- WeChat dataset	text, photo, audio, video	WeChat account	send to or receive from WeChat	online	WeChat user	WeChat user

Figure 2.3 Example of Big Data in 5Ws pattern crossing multiple datasets

2.4. 5Ws Dimension Clustering

Dimension clustering is the main method of reducing data overcrowding for Big Data analysis and visualization. The clustering of roles is based on the data's classification structure. The 5Ws pattern classification selects data behaviours through

collecting attributes in a tree structure, shown in Figure 2.4. The root is the 5Ws entity, and hundreds, even thousands, of different data nodes are grouped together through tree branches, with similar attributes linked to each root to show the classical relationships.

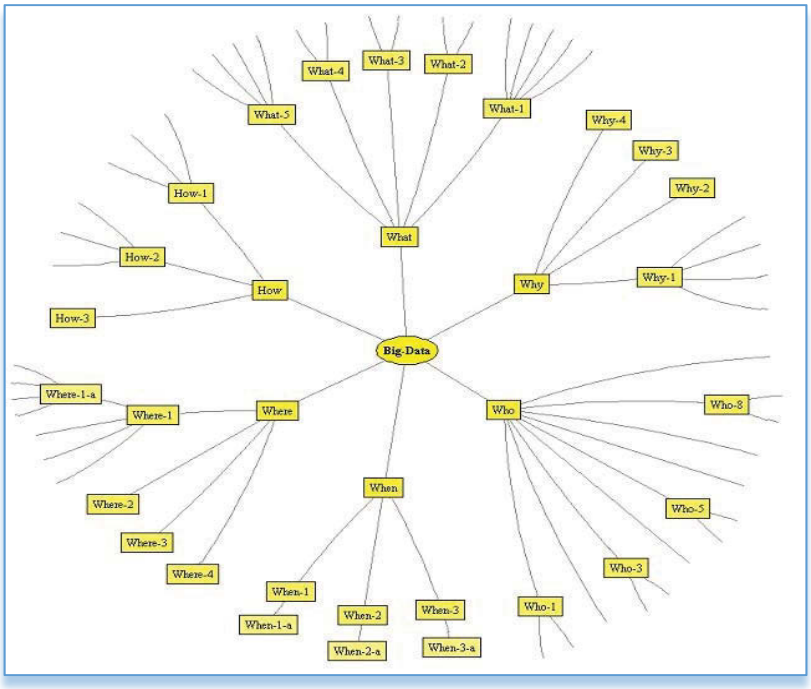

Figure 2.4 Tree structure of 5Ws pattern for Big Data

These classical relationships can also be easily explored by clustering. For example, say I want to explore the locations for who received the data by cities instead by countries. Q1 represents the "Country" and Q2 represents the "City" of who received the data. The dataset *D* will then be illustrated as

$$\boldsymbol{When, Where, What, How, Why, Country, City}$$

$$D = \begin{Bmatrix} d_{1T} & d_{1P} & d_{1X} & d_{1Y} & d_{1Z} & d_{1Q1} & d_{1Q2} \\ d_{2T} & d_{2P} & d_{2X} & d_{2Y} & d_{2Z} & d_{2Q1} & d_{2Q2} \\ d_{3T} & d_{3P} & d_{3X} & d_{3Y} & d_{3Z} & d_{3Q1} & d_{3Q2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{iT} & d_{iP} & d_{iX} & d_{iY} & d_{iZ} & d_{iQ1} & d_{iQ2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{nT} & d_{nP} & d_{nX} & d_{nY} & d_{nZ} & d_{nQ1} & d_{nQ2} \end{Bmatrix}$$

Equation 2. 7

Figure 2.5 shows an example of clustered 5Ws parallel coordinates on the *Q* axes. It clearly illustrates a clustering relationship between country (Q1) and city (Q2).

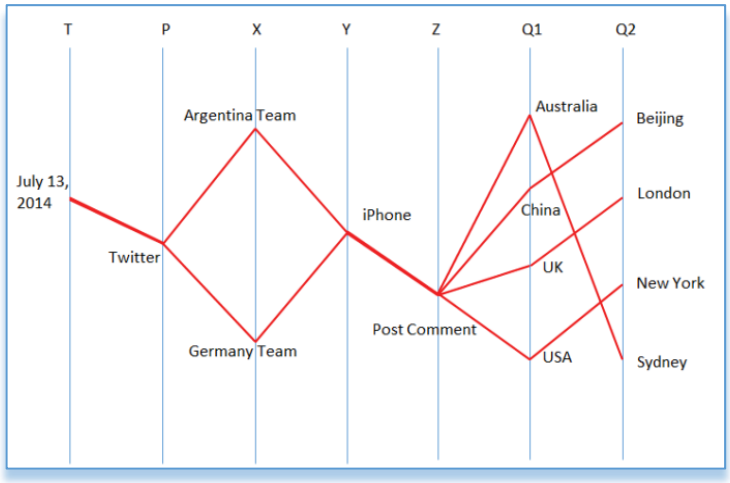

Figure 2.5 Example of clustered 5Ws parallel coordinates

2.5. 5Ws Shrunk Attributes

Each dimension contains hundreds, even thousands, of attributes, which leads to data overcrowding in Big Data visualization. Some approaches used the omission method to reduce the number of attributes in each dimension. However, this may result in the removal of important data patterns. To reduce data overcrowding without the loss of information in 5Ws patterns, I have defined Shrunk Attributes (SA) to collect the attributes that are not displayed in each dimension.

Assume *P_SA* collects the attributes that are not displayed in the *P* axis, *X_SA* collects the attributes that are not displayed in the *X* axis, *Y_SA* collects the attributes that are not displayed in the *Y* axis, *Z_SA* collects the attributes that are not displayed in the *Z* axis and *Q_SA* collects the attributes that are not displayed in the *Q* axis. The dataset *D*, with SA, will then be illustrated as

$$D = \begin{matrix} \textbf{\textit{When,}} & \textbf{\textit{Where,}} & \textbf{\textit{What,}} & \textbf{\textit{How,}} & \textbf{\textit{Why,}} & \textbf{\textit{Who}} \\ \left\{ \begin{matrix} d_{1T} \\ d_{2T} \\ d_{3T} \\ \vdots \\ d_{T\_SA} \end{matrix} \right. & \begin{matrix} d_{1P} \\ d_{2P} \\ d_{3P} \\ \vdots \\ d_{P\_SA} \end{matrix} & \begin{matrix} d_{1X} \\ d_{2X} \\ d_{3X} \\ \vdots \\ d_{X\_SA} \end{matrix} & \begin{matrix} d_{1Y} \\ d_{2Y} \\ d_{3Y} \\ \vdots \\ d_{Y\_SA} \end{matrix} & \begin{matrix} d_{1Z} \\ d_{2Z} \\ d_{3Z} \\ \vdots \\ d_{Z\_SA} \end{matrix} & \begin{matrix} d_{1Q} \\ d_{2Q} \\ d_{3Q} \\ \vdots \\ d_{Q\_SA} \end{matrix} \left. \right\} \end{matrix}$$

Equation 2. 8

Figure 2.6 shows an example of SA in 5Ws parallel coordinates. Each dimension has a SA to collect all the attributes which are not illustrated in the 5Ws parallel coordinates.

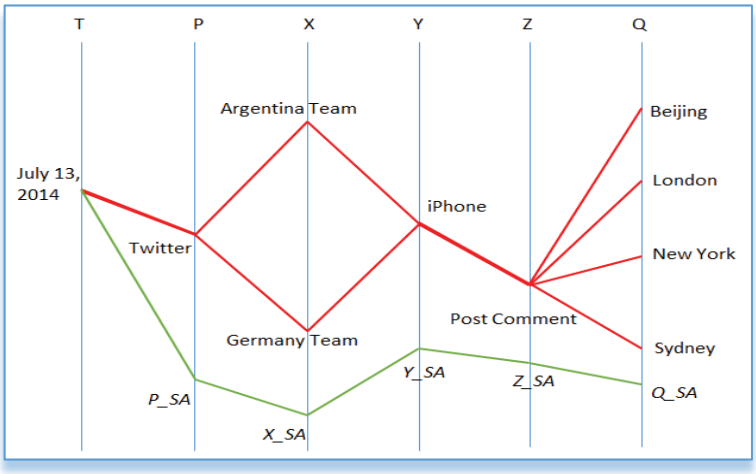

Figure 2.6 Example of SA in 5Ws parallel coordinates

2.6. Noise Attributes

Each dimension may contain noise data, which is defined as any unknown or undefined attributes in the 5Ws pattern. Noise data impacts the accuracy of measurement, and increasing the data algorithm time. I define *p_noise* as the noise data or unknown attributes in the *P* dimension; *x_noise* collects all noise data in the *X* dimension; *y_noise* collects all noise data in the *Y* dimension; *z_noise* collects all noise data in the *Z* dimension; and *q_noise* collects all noise data in the *Q* dimension.

Here, I assume that there is no noise data in the T dimension because the time stamp is sequential and hence it is unlikely that there is undefined time stamps. The dataset D that excludes noise data can then be illustrated as

$$D = \begin{matrix} \textbf{\textit{When,}} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \textbf{\textit{Where,}} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \textbf{\textit{What,}} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \textbf{\textit{How,}} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \textbf{\textit{Why,}} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \textbf{\textit{Who}} \\ \\ \\ \\ \\ \end{matrix} \left\{ \begin{matrix} d_{1T} & d_{1P} & d_{1X} & d_{1Y} & d_{1Z} & d_{1Q} \\ d_{2T} & d_{2P} & d_{2X} & d_{2Y} & d_{2Z} & d_{2Q} \\ d_{3T} & d_{3P} & d_{3X} & d_{3Y} & d_{3Z} & d_{3Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{nT} & d_{p\_noise} & d_{x\_noise} & d_{y\_noise} & d_{z\_noise} & d_{q\_noise} \end{matrix} \right\}$$

Equation 2. 9

Figure 2.7 shows an example of noise attributes in each dimension. It separates undefined or unknown attributes from other known and defined attributes. This significantly improves accuracy in Big Data measurement.

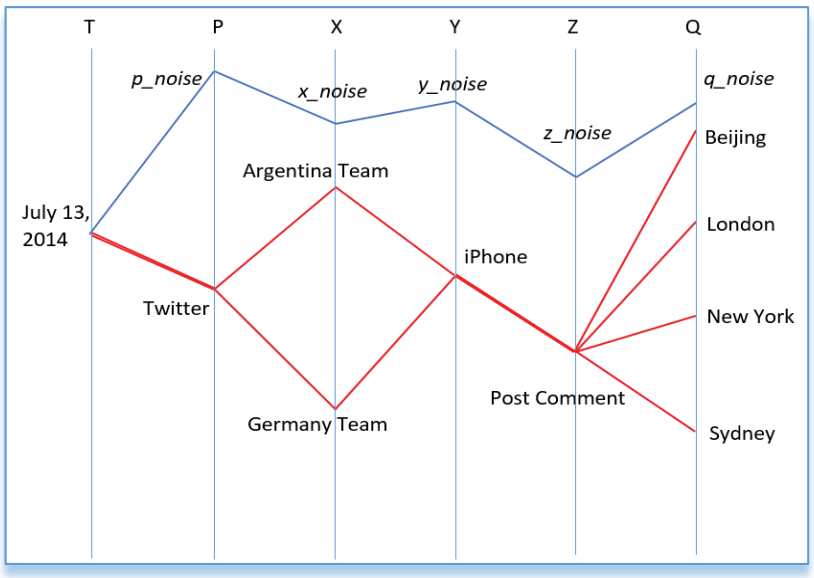

Figure 2.7 Example of noise data in 5Ws parallel coordinates

The 5Ws pattern creates a new approach for Big Data behavioural analysis and visualization. To the best of our knowledge, no previous work has addressed 5Ws dimension for Big Data behavioural analytics. Here, it also raises other challenges: how

do we compare Big Data patterns such as What the data contained via How the data was transferred, or Why the data occurred via Who received the data? To solve this issue, in the following chapter I have established 5Ws Pair-Density methods to analyze and visualize the 5Ws behavioural patterns.

Chapter 3: Pair-Density Parallel Coordinates

Almir Olivette Artero (Artero et al 2004) proposed their Parallel Coordinates Frequency and Density Plots. They created bi-dimensional frequency histograms for each pair of axes, and obtained information based on the data's frequency and relative density. The frequency is based on the graphic pixel value, which is calculated as

$$I_{(i,j)} = \frac{255 \times g(i,j)}{Max\ (G)}$$

Where I is the intensity of the pixel coordinates with a plot (i, j). G is a two-dimensional matrix whose value must be mapped across an adequate interval, normally [0, 255]. The plot's pixel resolution is $L \times W$, where $i = 1, \ldots, L$ defines the vertical resolution and $j = 1, \ldots W$ defines the horizontal resolution. $Max(G)$ is the largest value in matrix G. Lines with higher frequency are brighter than those with lower frequency, which illustrates the behaviour of attributes and aggregated patterns for data exploration. The density estimation is obtained by applying a smoothing filter, such as a square wave smoothing filter, which enables the user to highlight clusters.

This model, however, loses its efficiency in identifying multidimensional data patterns, since its approach to modelling frequency and density plots is based on the intensity of the pixels in the graph. This fails to calculate multi-dimensional data patterns. To measure the variation between the 5Ws patterns, I have created Pair-Density to value the 5Ws patterns between any two dimensions, and have also created Pair-Density parallel axes for visual analytics.

3.1. Pair-Density Algorithm

Aside from the T dimension, which is a time stamp using time serials, each data incident can be illustrated as a 5Ws node as *d(p,x, y, z, q)*. To compare any two dimensions, a pattern subset needs to be formed by collecting the other three dimensions together. For example, to compare P and Q dimensions, the pattern subset must contain dimensions X, Y and Z, such that the pattern can be defined as PAT{X, Y, Z}. The dataset D can then be illustrated as a pair of dimensions P and Q, with pattern subset PAT{X, Y, Z}.

$$D = \begin{matrix} P \\ \begin{Bmatrix} d_{1P} \\ d_{2P} \\ d_{3P} \\ \vdots \\ d_{iP} \\ \vdots \\ d_{nP} \end{Bmatrix} \end{matrix} \begin{matrix} \boldsymbol{PAT\{X,\ Y,\ Z\}} \\ \begin{Bmatrix} \boldsymbol{d_X} & \boldsymbol{d_Y} & \boldsymbol{d_Z} \\ \boldsymbol{d_X} & \boldsymbol{d_Y} & \boldsymbol{d_Z} \\ \boldsymbol{d_X} & \boldsymbol{d_Y} & \boldsymbol{d_Z} \\ \vdots & \vdots & \vdots \\ \boldsymbol{d_X} & \boldsymbol{d_Y} & \boldsymbol{d_Z} \\ \vdots & \vdots & \vdots \\ \boldsymbol{d_X} & \boldsymbol{d_Y} & \boldsymbol{d_Z} \end{Bmatrix} \end{matrix} \begin{matrix} Q \\ \begin{Bmatrix} d_{1Q} \\ d_{2Q} \\ d_{3Q} \\ \vdots \\ d_{iQ} \\ \vdots \\ d_{nQ} \end{Bmatrix} \end{matrix}$$

Equation 3. 1

A data incident $d_{(p,\ x,\ y,\ z,\ q)}$ can then be illustrated in a pair dimensions with particular pattern $pat_{(x,\ y,\ z)}$, which represented as $d_{(p,\ pat(x,\ y,\ z))}$ and $d_{(q,\ pat(x,\ y,\ z))}$. The data incident can be denoted pair dimensions as

$$d = \begin{cases} p,\ pat_{(x,y,z)} & -\ attribute\ in\ P\ dimension \\ q,\ pat_{(x,y,z)} & -\ attribute\ in\ Q\ dimension \end{cases}$$

Equation 3. 2

This enables the attributes comparison between P and Q dimensions, no matter of the different form of data (e.g. text, image or video). The pair subsets that contained all

incidents with particular pattern $pat_{(x, y, z)}$ for p attribute in P dimension, and for q attributes in Q dimension, can be illustrated as

$$\begin{cases} D(p, pat(x, y, z)) = \{d \in D | pat(x, y, z), p\} \\ D(q, pat(x, y, z)) = \{d \in D | pat(x, y, z), q\} \end{cases}$$

Equation 3. 3

where $\{d \in D|pat(x, y, z), p\}$ and $\{d \in D|pat(x, y, z), q\}$ represent the subsets that contained all attributes p and q with $pat_{(x, y, z)}$ in dataset D.

Accordingly, there are ten different pattern subsets that are formed by taking a Pair-Dimension. There are denoted as

$pat_{(x, y, z)} = \{pat \mid pat \in PAT(), x, y, z\}$, *(Pair-Dimension P and Q)*

$pat_{(x, y, p)} = \{pat \mid pat \in PAT(), x, y, p\}$, *(Pair-Dimension Z and Q)*

$pat_{(x, y, q)} = \{pat \mid pat \in PAT(), x, y, q\}$, *(Pair-Dimension P and Z)*

$pat_{(x, p, z)} = \{pat \mid pat \in PAT(), x, p, z\}$, *(Pair-Dimension Y and Q)*

$pat_{(x, q, z)} = \{pat \mid pat \in PAT(), x, q, z\}$, *(Pair-Dimension P and Y)*

$pat_{(p, y, z)} = \{pat \mid pat \in PAT(), p, y, z\}$, *(Pair-Dimension X and Q)*

$pat_{(q, y, z)} = \{pat \mid pat \in PAT(), q, y, z\}$, *(Pair-Dimension P and X)*

$pat_{(p, q, x)} = \{pat \mid pat \in PAT(), p, q, x\}$, *(Pair-Dimension Y and Z)*

$pat_{(p, q, y)} = \{pat \mid pat \in PAT(), p, q, y\}$, *(Pair-Dimension X and Z)*

$pat_{(p, q, z)} = \{pat \mid pat \in PAT(), p, q, z\}$, *(Pair-Dimension X and Y)*

Equation 3. 4

If the 5Ws pattern contains the clustering dimension, the particular attribute coming from the clustered dimension will be added to the pattern subset. For example, assume that the 5Ws pattern contains Z1 and Z2 dimensions, which are clustered from the Z dimension and Z{Z1, Z2}. If so, the particular pattern subset can then be illustrated as $pat_{(x, y, z1, z2)} = \{pat \mid pat \in PAT(), x, y, z_1, z_2\}$

3.2. Pattern and Pair-Density

Before establishing the 5Ws Pair-Density algorithms, I will first define the five densities used to measure each dimension in 5Ws patterns (i.e. P, X, Y, Z, and Q axes).

- Sending-Density (SD) measures the proportion of the sender's pattern (i.e. It collects the attributes in dimension P for where the data came from.)
- Content-Density (CD) measures the proportion of the content's pattern (i.e. It collects the attributes in dimension X for what the data contained.)
- Transferring-Density (TD) measures the proportion of the transmission pattern (i.e. It collects the attributes in dimension Y for how the data was transferred.)
- Purpose-Density (PD) measures the proportion of the purpose pattern (i.e. It collects the attributes in dimension Z for why the data occurred.)
- Receiving-Density (RD) measures the proportion of the receiver's pattern (i.e. It collects the attributes in dimension Q for who received the data.)

Those five densities provide the measurements and comparisons between 5Ws dimension, even they are in different form of data. I will now define the ten Pair-Densities

to measure the ten different patterns and define the ten different goals, as illustrated in Table 3.1.

Table 3.1. 5Ws Pair-Density and Patterns

Pair-Density	Pair-Dimension	Patterns	Target
$SD_{(p, pat(x, y, z))}$ via $RD_{(q, pat(x, y, z))}$	Dimension P via Q	*pat(x, y, z)*	Compare sender patterns via receiver patterns
$SD_{(p, pat(x, y, q))}$ via $PD_{(z, pat(x, y, q))}$	Dimension P via Z	*pat(x, y, q)*	Compare sender patterns via purpose/reason patterns
$SD_{(p, pat(x, z, q))}$ via $TD_{(y, pat(x, z, q))}$	Dimension P via Y	*pat(x, z, q)*	Compare sender patterns via transfer patterns
$SD_{(p, pat(y, z, q))}$ via $CD_{(x, pat(y, z, q))}$	Dimension P via X	*pat(y, z, q)*	Compare sender patterns via content patterns
$CD_{(x, pat(p, y, z))}$ via $RD_{(q, pat(p, y, z))}$	Dimension X via Q	*pat(p, y, z)*	Compare content patterns via receiver patterns
$CD_{(x, par(p, y, q)}$ via $PD_{(z, pat(p, y, q))}$	Dimension X via Z	*pat(p, y, q)*	Compare content patterns via purpose/reason patterns
$CD_{(x, pat(p, z, q))}$ via $TD_{(y, pat(p, z, q))}$	Dimension X via Y	*pat(p, z, q)*	Compare content patterns via transfer patterns
$TD_{(y, pat(p, x, z))}$ via $RD_{(q, pat(p, x, z))}$	Dimension Y via Q	*pat(p, x, z)*	Compare transfer patterns via receiver patterns
$TD_{(y, pat(p, x, q))}$ via $PD_{(z, pat(p, x, q))}$	Dimension Y via Z	*pat(p, x, q)*	Compare transfer patterns via purpose/reason patterns
$PD_{(z, pat(p, x, y))}$ via $RD_{(q, pat(p, x, y))}$	Dimension Z via Q	*pat(p, x, y)*	Compare purpose/reason patterns via receiver patterns

Each Pair-Density illustrates the particular pattern $pat_{( )}$ between two particular dimensions, which demonstrates the particular relationship in the 5Ws patterns. These Pair-Density and inside relationships can then be measured and compared. Here, I used

Equation 3.1 as an example to compare the relationship for Pair-Density $SD_{(p, pat(x, y, z))}$ via $RD_{(q, pat(x, y, z))}$. This is shown in Figure 3.1.

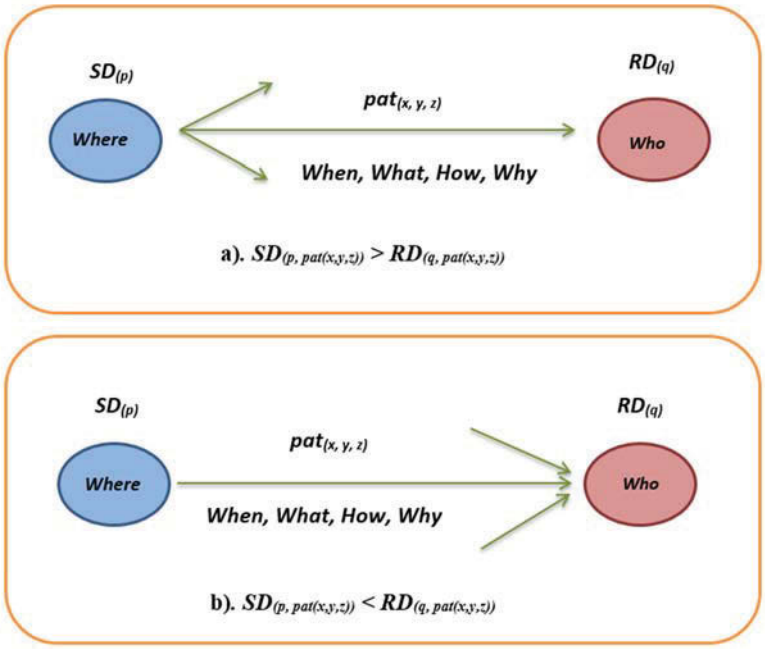

Figure 3.1 Example of relationship for $SD_{(p, pat(x, y, z))}$ via $RD_{(q, pat(x, y, z))}$

In Figure 3.1, graph a) shows $SD_{( )} > RD_{( )}$, which means that the sender p has sent $pat_{(x, y, z)}$ to multiple receivers, one of whom is q. Graph b) show $SD_{( )} < RD_{( )}$, which indicates that the receiver q has not only received data from p, but has also received data from other senders. The details of each Pair-Density, alongside with visual analytics, will be discussed in the following chapters.

3.3. $SD_{(p, pat(x, y, z))}$ via $RD_{(q, pat(x, y, z))}$ for *pat(x, y, z)*

The Pair-Density $SD_{(p, pat(x, y, z))}$ via $RD_{(q, pat(x, y, z))}$ measures the sender p and receiver q between attributes in the P and Q dimensions for the particular pattern *pat(x, y, z)*, which can be denoted as

$$\begin{cases} SD(p, pat(x,y,z)) = \dfrac{|D(p,\ pat(x,y,z))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(x,y,z), p\} \times 100\% \\ RD(q, pat(x,y,z)) = \dfrac{|D(q,\ pat(x,y,z))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(x,y,z), q\} \times 100\% \end{cases}$$

Equation 3. 5

$SD_{(p,\ pat(x,\ y,\ z))}$ measures the sender's pattern during data transferal for sender p and for the particular pattern $pat(x, y, z)$, $0 \leq SD_{(p,\ pat(x,\ y,\ z))} \leq 1$ and $\sum SD_{(p,\ pat(x,\ y,\ z))} = 1$. $RD_{(q,\ pat(x,\ y,\ z))}$ measures the receiver's pattern during data transferal for receiver q and for the particular pattern $pat(x, y, z)$, $0 \leq RD_{(q,\ pat(x,\ y,\ z))} \leq 1$ and $\sum RD_{(q,\ pat(x,\ y,\ z))} = 1$.

$SD_{(p,\ pat(x,\ y,\ z))}$ via $RD_{(q,\ pat(x,\ y,\ z))}$ both illustrate the same pattern $pat(x, y, z)$ but on different foci. $SD_{(p,\ pat(x,\ y,\ z))}$ represents the sender's pattern, sent by p irrespective of **Who** received the data. A high value of $SD_{(p,\ pat(x,\ y,\ z))}$ means that the sender p sent the data to more receivers compared to other senders. $RD_{(q,\ pat(x,\ y,\ z))}$ indicates the receiver's pattern, received by q irrespective of **Where** the data came from. A high value of $RD_{(q,\ pat(x,\ y,\ z))}$ means that the receiver q received more data from senders compared to other receivers.

I have created two additional parallel axes, $SD_{(\ )}$ and $RD_{(\ )}$, to illustrate the Pair-Density $SD_{(p,\ pat(x,\ y,\ z))}$ via $RD_{(q,\ pat(x,\ y,\ z))}$ in the 5Ws parallel coordinates, and to provide visual structure and measurement for the 5Ws pattern. The previous example has been used to demonstrate $SD_{(p,\ pat(x,\ y,\ z))}$ via $RD_{(q,\ pat(x,\ y,\ z))}$ in the 5Ws parallel coordinates, shown as Figure 3.2.

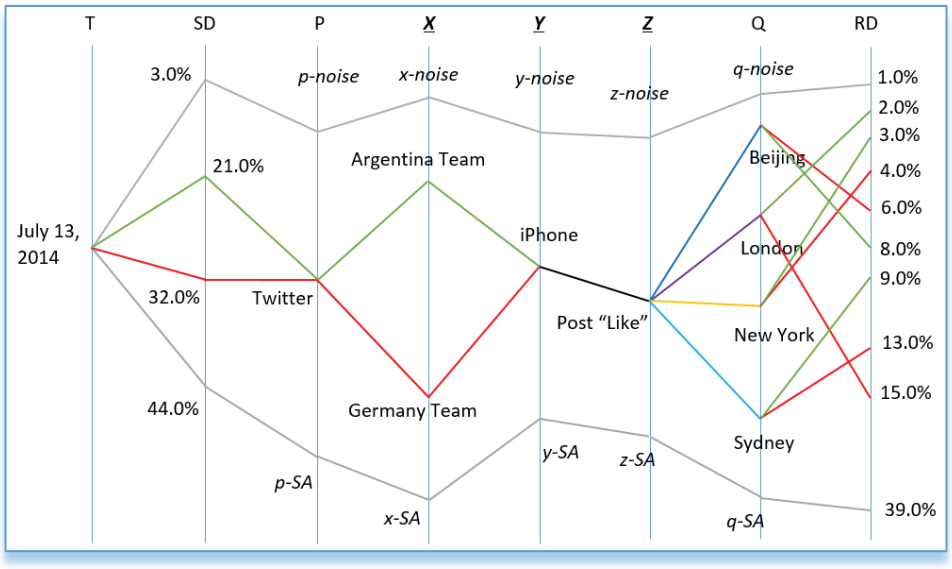

Figure 3.2 Example of 5Ws parallel coordinates with $SD_{(p,\ pat(x,\ y,\ z))}$ via $RD_{(q,\ pat(x,\ y,\ z))}$

In Figure 3.2, two $pat_{(x,\ y,\ z)}$ patterns have been illustrated in $SD_{(p,\ pat(x,\ y,\ z))}$ via $RD_{(q,\ pat(x,\ y,\ z))}$. One pattern $pat_{(Argentina\ Team,\ iPhone,\ Like)}$ contains tweets that post like messages for the Argentina Team by iPhone, and another pattern $pat_{(Germany\ Team,\ iPhone,\ Like)}$ posts like messages via Twitter for the Germany Team by iPhone as well. Both patterns are sent by Twitter and received by four different cities; Beijing, London, New York and Sydney. In the $SD_{(\ )}$ axis, Figure 3.2 illustrates that 32.0% of senders favour the Germany Team, compared to 21.0% that favour the Argentina Team. In the $RD_{(\ )}$ axis, 38.0% of received patterns favour the Germany Team (15.0%+13.0%+6.0%+4.0%=38.0%), compared to 22.0% of received patterns that favour the Argentina Team (9.0%+8.0%+3.0%+2.0%=22.0%). For Argentina, $SD_{(Argentina)}$ (21.0%) ≈ $RD_{(Argentina)}$ (22.0%), which suggests that the senders and receivers have a similar density. For Germany, $SD_{(Germany)}$ (32.0%) < $RD_{(Germany)}$ (38.0%), which indicates that some senders sent patterns to more than one receiver. For Shrunk Attributes, $SD_{(p\text{-}SA)}$ = 44.0% collects the other senders, such as Facebook, and $RD_{(q\text{-}SA)}$ = 39.0% aggregates data received by other cities.

In the $RD_{(\ )}$ axis, which corresponds with the Q dimension, Sydney, New York and London received more favourable tweets for the Germany Team compared to the Argentina Team, but this was reversed in Beijing, with more favourable tweets for the Argentina Team. The highest $RD_{(London,\ pat(Germany,\ iPhone,\ Like))}$ = 15.0% was received by London in favour of the Germany Team, and the lowest $RD_{(London,\ pat(Argentina,\ iPhone,\ Like))}$ = 2.0% was for the Argentina team, also received by London. The largest difference in $RD_{(\ )}$ occurred in London, with $RD_{(London,\ pat(Germany,\ iPhone,\ Like))}$ - $RD_{(London,\ pat(Argentina,\ iPhone,\ Like))}$ = 13.0%. This suggest that tweets were 7.5 times more likely to favour the Germany Team over the Argentina Team. The smallest difference in $RD_{(\ )}$ occurred in New York, with $RD_{(New\ York,\ pat(Germany,\ iPhone,\ Like))}$ - $RD_{(New\ York,\ pat(Argentina,\ iPhone,\ Like))}$ = 1.0%, which suggests a very similar likelihood of tweets favouring Germany and Argentina. New York's total $RD_{(\ )}$ = 7.0% (3.0%+4.0%) is also the lowest, with Sydney having the highest city with $RD_{(\ )}$ = 22.0% (9.0%+13.0%).

3.4. $SD_{(p,\ pat(x,\ y,\ q))}$ via $PD_{(z,\ pat(x,\ y,\ q))}$ with *pat(x, y, q)*

The Pair-Density $SD_{(p,\ pat(x,\ y,\ q))}$ via $PD_{(z,\ pat(x,\ y,\ q))}$ measures the pattern for senders and purposes (reasons) by comparing the attributes between the P and Z dimensions for the particular pattern *pat(x, y, q)*, which can be denoted as

$$\begin{cases} SD(p, pat(x,y,q)) = \dfrac{|D(p,\ pat(x,y,q))|}{|D|} \times 100\% = \dfrac{1}{n}\{d \in D | pat(x,y,q), p\} \times 100\% \\ PD(z, pat(x,y,q)) = \dfrac{|D(z,\ pat(x,y,q))|}{|D|} \times 100\% = \dfrac{1}{n}\{d \in D | pat(x,y,q), z\} \times 100\% \end{cases}$$

Equation 3. 6

$SD_{(p,\ pat(x,\ y,\ q))}$ measures the sender's pattern during data transferal for sender *p* and for the particular pattern *pat(x, y, q)*, $0 \leq SD_{(p,\ pat(x,\ y,\ q))} \leq 1$ and $\sum SD_{(p,\ pat(x,\ y,\ q))} = 1$. $PD_{(z,}$

$_{pat(x, y, q))}$ measures the purpose's pattern for reason z and for the particular pattern $pat(x, y, q)$, $0 \leq PD_{(z, pat(x, y, q))} \leq 1$ and $\sum PD_{(z, pat(x, y, q))} = 1$.

In $SD_{(p, pat(x, y, q))}$ via $PD_{(z, pat(x, y, q))}$, $SD_{(p, pat(x, y, q))}$ represents the sender's pattern for $pat(x, y, q)$, sent by p irrespective of **Why** the data occurred. A high value of $SD_{(p, pat(x, y, q))}$ means that the sender p sent more patterns than other senders. $PD_{(z, pat(x, y, q))}$ indicates the purpose's pattern with $pat(x, y, q)$, with the reason being irrespective of **Where** the data came from. A high value of $PD_{(z, pat(x, y, q))}$ means that the reason z has more data compared to other purposes.

By assigning $SD_{()}$ and $PD_{()}$ as two additional parallel axes, I will use the previous example to demonstrate the Pair-Density $SD_{(p, pat(x, y, q))}$ via $PD_{(z, pat(x, y, q))}$ in 5Ws parallel coordinates, shown as Figure 3.3.

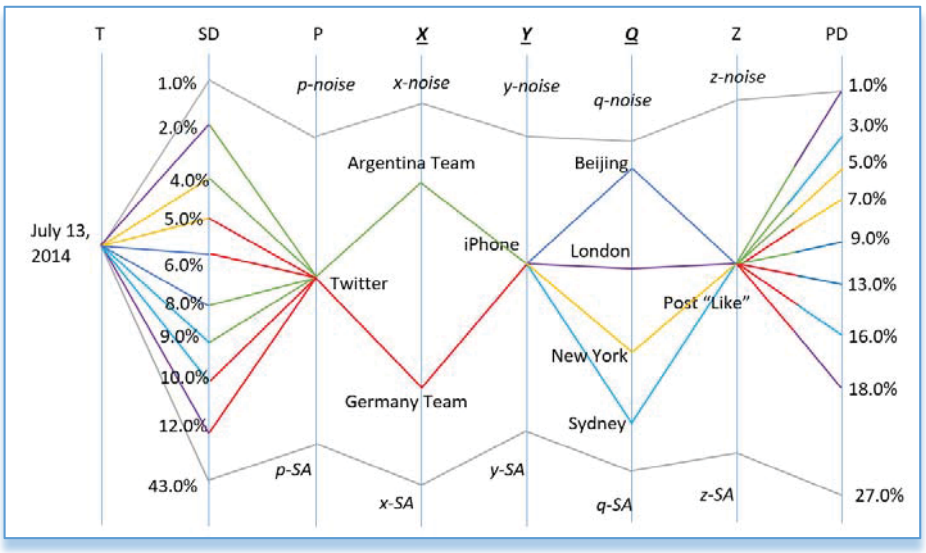

Figure 3.3 Example of 5Ws parallel coordinates with $SD_{(p, pat(x, y, q))}$ via $PD_{(z, pat(x, y, q))}$

In Figure 3.3， the three axes X, Y and Q have all been located in the middle of the graph to illustrate the pattern $pat(x, y, q)$, with the Pair-Densities axes located on both sides in order to clearly demonstrate the particular patterns. There are eight $pat_{(x, y, q)}$ patterns in this Pair-Density, with two attributes (Argentina Team and Germany Team)

in the X axis and four attributes (Beijing, London, New York and Sydney) in the Q axis. These eight patterns are all transferred by iPhone. To clearly illustrate the eight patterns, I have used 2×4 colour lines, including 2× colour lines to represent attributes in the X axis, and 4× colour lines to represent attributes in the Q axis. This then demonstrates patterns linked to both the $SD_{(\ )}$ and $PD_{(\ )}$ axes.

In the $SD_{(\ )}$ axis, the Germany Team has the two top densities (London and Sydney), compared to the $PD_{(\ )}$ axis, where it has the top three densities (London, Sydney and Beijing). The Argentina Team has the bottom two densities in the $SD_{(\ )}$ axis (London and New York), compared to the $PD_{(\ )}$ axis, where it has the top three densities (New York, Sydney and London). For Germany, the total $SD_{(Germany)}$ = 33% (12.0%+10.0%+6.0%+5.0%) is much less than the total $PD_{(Germany)}$ = 54.0% (18.0%+16.0%+13.0%+7.0%). This suggests that the like patterns were sent not only by Germany Team supporters, but also by the supporters of other teams, with a swing of more than 20% to the winning Germany Team. For Argentina, the total $SD_{(Argentina)}$ = 23% (9.0%+8.0%+4.0%+2.0%) is greater than the total $PD_{(Argentina)}$ = 18.0% (9.0%+5.0%+3.0%+1.0%). This indicates that the Argentina supporters sent more patterns, irrespective of whether they liked or disliked the match. The highest $SD_{(Germany)}$ = 12.0% is less than the highest $PD_{(Germany)}$ = 18.0%, which suggests that the supporters of other teams also thought that the Germany Team has the reasonable chance of winning the game. The highest $SD_{(Argentina)}$ = $PD_{(Argentina)}$ = 9.0%, which indicates that the Argentina Team supporters sent the same density of messages regardless of whether they liked the performance or not.

In the $PD_{(\ )}$ axis, the total like density $PD_{(like)}$ = 72.0%, $PD_{(z\text{-}noise)}$ = 1.0% and $PD_{(z\text{-}SA)}$ = 27.0% indicate that the main reason why peopled posted a message on Twitter was

because they liked the game. In the $SD_{()}$ axis, the Twitter density $SD_{(Twitter)}$ = 56.0% means that main method of posting messages during the game was Twitter. $SD_{(p\text{-}SA)}$ = 43.0% indicates the density for other transferral methods such as Facebook.

3.5. $SD_{(p,\ pat(x,\ z,\ q))}$ via $TD_{(y,\ pat(x,\ z,\ q))}$ with *pat(x, z, q)*

The Pair-Density $SD_{(p,\ pat(x,\ z,\ q))}$ via $TD_{(y,\ pat(x,\ z,\ q))}$ measures the sender's patterns and transfer patterns by comparing the attributes between the P and Y dimensions for the particular pattern *pat(x, z, q)*, which can be denoted as

$$\begin{cases} SD(p, pat(x,z,q)) = \dfrac{|D(p,\ pat(x,z,q))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(x,z,q), p\} \times 100\% \\ TD(y, pat(x,z,q)) = \dfrac{|D(y,\ pat(x,z,q))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(x,z,q), y\} \times 100\% \end{cases}$$

Equation 3. 7

$SD_{(p,\ pat(x,\ z,\ q))}$ measures the sender's pattern during data transferral for sender *p* and for the particular pattern *pat(x, z, q)*, $0 \leq SD_{(p,\ pat(x,\ z,\ q))} \leq 1$ and $\sum SD_{(p,\ pat(x,\ z,\ q))} = 1$. $TD_{(y,\ pat(x,\ z,\ q))}$ measures the transfer's pattern during data transferral for method *y* and for the particular pattern *pat(x, z, q)*, $0 \leq TD_{(y,\ pat(x,\ z,\ q))} \leq 1$ and $\sum TD_{(y,\ pat(x,\ z,\ q))} = 1$.

In $SD_{(p,\ pat(x,\ z,\ q))}$ via $TD_{(y,\ pat(x,\ z,\ q))}$, $SD_{(p,\ pat(x,\ z,\ q))}$ represents the sender's pattern for *pat(x, z, q)*, sent by *p* irrespective of **How** the data were transferred. A high value of $SD_{(p,\ pat(x,\ z,\ q))}$ means that the sender *p* sent more patterns compared to other senders, irrespective of transfer methods. $TD_{(y,\ pat(x,\ z,\ q))}$ indicates the transferring pattern with *pat(x, z, q)*, with the data being transferred by *y* irrespective of **Where** the data came from. A high value of $TD_{(y,\ pat(x,\ z,\ q))}$ indicates that the method *y* transferred more data compared to other methods.

After adding $SD_{(\ )}$ and $TD_{(\ )}$ as the two additional parallel axes, I have used the previous example to demonstrate the Pair-Density $SD_{(p,\ pat(x,\ z,\ q))}$ via $TD_{(y,\ pat(x,\ z,\ q))}$ in the 5Ws parallel coordinates, as shown as Figure 3.4.

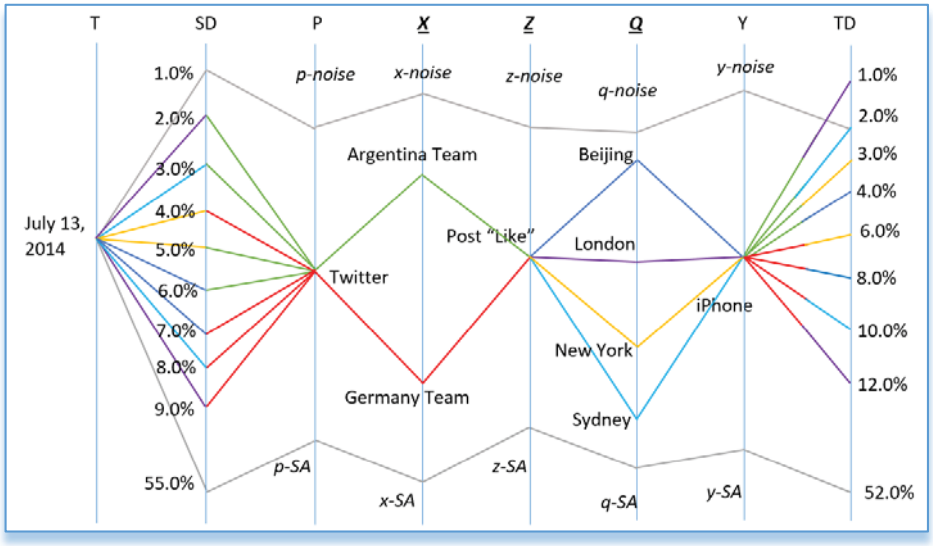

Figure 3.4 Example of 5Ws parallel coordinates with $SD_{(p,\ pat(x,\ z,\ q))}$ via $TD_{(y,\ pat(x,\ z,\ q))}$

In Figure 3.4, there are eight $pat_{(x,\ z,\ q)}$ patterns in this Pair-Density $SD_{(p,\ pat(x,\ z,\ q))}$ via $TD_{(y,\ pat(x,\ z,\ q))}$ that all have the same reason – the sender liked the game. In the $SD_{(\ )}$ axis, the total $SD_{(Germany)}$ = 28.0% (9.0%+8.0%+7.0%+4.0%) is greater than the total $SD_{(Argentina)}$ = 16.0% (6.0%+5.0%+3.0%+2.0%), indicating that 12.0% more senders liked Germany Team's performance compared to the Argentina Team's. $SD_{(London)}$ = 11.0% (9.0%+2.0%) and $SD_{(Sydney)}$ = 11.0% (8.0%+3.0%), which means that senders in both cities had the same sending density for this game. $SD_{(Beijing)}$ = 13.0% (7.0%+6.0%) is greater than $SD_{(New\ York)}$ = 9.0% (5.0%+4.0%), which suggests that Beijing has more soccer fans than New York. In the $TD_{(\ )}$ axis, the total $TD_{(Germany)}$ = 36.0% (12.0%+10.0%+8.0%+6.0%) is much greater than the total $TD_{(Argentina)}$ = 10.0% (4.0%+3.0%+2.0%+1.0%), which suggests that the Germany Team supporters used iPhones to transfer data more frequently compared to Argentina Team supporters, with a difference of more than 25%. $TD_{(London)}$ = 13.0%, and $TD_{(Sydney)}$ = $TD_{(Beijing)}$ = 12.0%

suggests that these three cities have a similar density in using an iPhone. $TD_{(New\ York)}$ = 9.0% indicates that New York has a lower ratio of iPhone usage compared to the other three cities.

Comparing the Pair-Density $SD_{(\ )}$ and $TD_{(\ )}$, $SD_{(Germany)}$ = 28.0% < $TD_{(Germany)}$ = 36.0%, which suggests that supporters of the other teams also used their iPhone to support the Germany Team. $SD_{(Argentina)}$ = 16.0% > $TD_{(Argentina)}$ = 10.0%, which indicates that Argentina's voters used other devices more frequently than they used their iPhone. $SD_{(London)}$ (9.0%+2.0% = 11.0%) ≈ $TD_{(London)}$ (12.0%+1.0% = 13.0%), $SD_{(Sydney)}$ (8.0%+3.0% = 11.0%) ≈ $TD_{(Sydney)}$ (10.0%+2.0% = 12.0%), $SD_{(Beijing)}$ (7.0%+6.0% = 13.0%) ≈ $TD_{(Beijing)}$ (8.0%+4.0% = 12.0%), and $SD_{(New\ York)}$ (5.0%+4.0% = 9.0%) = $TD_{(New\ York)}$ (6.0%+3.0% = 9.0%). This suggests that there is no significant difference between the Pair-Densities for these four cities. $SD_{(p\text{-}SA)}$ = 55.0% means that Twitter messages occurred in about 45% of this pattern, and $TD_{(y\text{-}SA)}$ = 52.0% indicates that other devices were used 4.0% more during the game.

3.6. $SD_{(p,\ pat(y,\ z,\ q))}$ via $CD_{(x,\ pat(y,\ z,\ q))}$ with *pat(y, z, q)*

The Pair-Density $SD_{(p,\ pat(y,\ z,\ q))}$ via $CD_{(x,\ pat(y,\ z,\ q))}$ measures the sender's patterns and content patterns by comparing the attributes between the P and X dimensions for the particular pattern *pat(y, z, q)*, which can be denoted as

$$\begin{cases} SD(p, pat(y,z,q)) = \dfrac{|D(p,\ pat(y,z,q))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(y,z,q), p\} \times 100\% \\ CD(x, pat(y,z,q)) = \dfrac{|D(x,\ pat(y,z,q))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(y,z,q), x\} \times 100\% \end{cases}$$

Equation 3. 8

$SD_{(p,\ pat(y,\ z,\ q))}$ measures the sender's pattern during data transferral for sender p and for the particular pattern $pat(y, z, q)$, $0 \leq SD_{(p,\ pat(y,\ z,\ q))} \leq 1$ and $\sum SD_{(p,\ pat(y,\ z,\ q))} = 1$. $CD_{(x,\ pat(y,\ z,\ q))}$ measures the content's pattern during data transferral for content y and for the particular pattern $pat(y, z, q)$, $0 \leq CD_{(x,\ pat(y,\ z,\ q))} \leq 1$ and $\sum CD_{(x,\ pat(y,\ z,\ q))} = 1$.

In $SD_{(p,\ pat(y,\ z,\ q))}$ via $CD_{(x,\ pat(y,\ z,\ q))}$, $SD_{(p,\ pat(y,\ z,\ q))}$ represents the sender's pattern for $pat(y, z, q)$, sent by p irrespective of **What** the data contained. A high value of $SD_{(p,\ pat(y,\ z,\ q))}$ means that the sender p sent more data compared to other senders, irrespective of data contents. $CD_{(x,\ pat(y,\ z,\ q))}$ indicates the content's pattern for $pat(y, z, q)$, with the data containing attribute x irrespective of **Where** the data came from. A high value of $CD_{(x,\ pat(y,\ z,\ q))}$ means more patterns have the attribute x compared to other contents inside the dataset, irrespective of who sent the data. Figure 3.5 shows an example of the 5Ws parallel coordinates for $SD_{(p,\ pat(y,\ z,\ q))}$ via $CD_{(x,\ pat(y,\ z,\ q))}$.

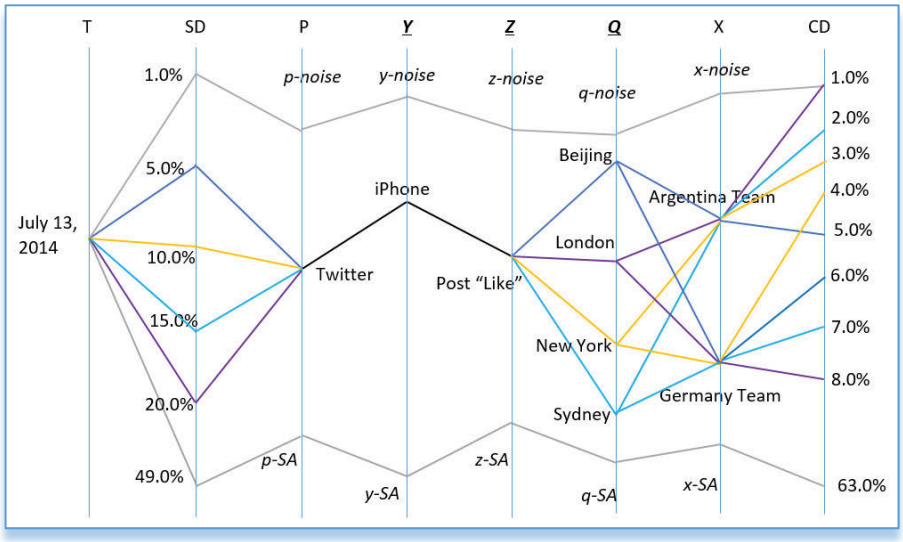

Figure 3.5 Example of 5Ws parallel coordinates with $SD_{(p,\ pat(y,\ z,\ q))}$ via $CD_{(x,\ pat(y,\ z,\ q))}$

In Figure 3.5, there are four $pat_{(y,\ z,\ q)}$ patterns in this Pair-Density $SD_{(p,\ pat(y,\ z,\ q))}$ via $CD_{(x,\ pat(y,\ z,\ q))}$, which include Beijing, London, New York and Sydney in the Q dimension. In the $SD_{(\ )}$ axis, the highest $SD_{(London)}$ = 20.0% suggests that London was the largest

sender of data. Sydney and New York follow with $SD_{(Sydney)}$ = 15.0% and $SD_{(New\ York)}$ = 10.0%, with Being having the lowest sending pattern during the game with $SD_{(Beijing)}$ = 5.0%. In the $CD_{(\ )}$ axis, the Germany Team had the three top densities, with the total $CD_{(Germany)}$ = 25.0% (8.0%+7.0%+6.0%+4.0%) being far more than the total $CD_{(Argentina)}$ = 11.0% (5.0%+3.0%+2.0%+1.0%). This suggests that 14.0% more data patterns mentioned the winning team compared to the losing team.

$SD_{(Beijing)}$ = 5.0% < $CD_{(Beijing)}$ = 11.0% (6.0%+5.0%), which indicates that senders from other countries also sent data patterns from Beijing. $SD_{(London)}$ (20.0%) > $CD_{(London)}$ (8.0%+1.0% =9.0%), and $SD_{(Sydney)}$ (15.0%) > $CD_{(Sydney)}$ (7.0%+2.0% =9.0%), which both illustrate that more senders posted sending patterns from London and Sydney. There is no significant difference between $SD_{(New\ York)}$ (10.0%) ≈ $CD_{(New\ York)}$ (4.0%+3.0%=7.0%). $SD_{(p\text{-}SA)}$ (49.0%) < $CD_{(x\text{-}SA)}$ (63.0%) indicates that 63% of supporters from other teams used Twitter during the game.

3.7. $CD_{(x,\ pat(p,\ y,\ z))}$ via $RD_{(q,\ pat(p,\ y,\ z))}$ with *pat(p, y, z)*

The Pair-Density $CD_{(x,\ pat(p,\ y,\ z))}$ via $RD_{(q,\ pat(p,\ y,\ z))}$ measures the content patterns and receiver patterns by comparing the attributes between the X and Q dimensions for the particular pattern *pat(p, y, z)*, which can be denoted as

$$\begin{cases} CD(x, pat(p,y,z)) = \dfrac{|D(x,\ pat(p,y,z))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,y,z), x\} \times 100\% \\ RD(q, pat(p,y,z)) = \dfrac{|D(q,\ pat(p,y,z))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,y,z), q\} \times 100\% \end{cases}$$

Equation 3. 9

$CD_{(x, pat(p, y, z))}$ measures the content's pattern during data transferral for the data content x and for the particular pattern $pat(p, y, z)$, $0 \leq CD_{(x, pat(p, y, z))} \leq 1$ and $\sum CD_{(x, pat(p, y, z))} = 1$. $RD_{(q, pat(p, y, z))}$ measures the receiver's pattern during data transferral for receiver q and for the particular pattern $pat(p, y, z)$, $0 \leq RD_{(q, pat(p, y, z))} \leq 1$ and $\sum RD_{(q, pat(p, y, z))} = 1$.

$CD_{(x, pat(p, y, z))}$ via $RD_{(q, pat(p, y, z))}$ illustrates the pattern $pat(p, y, z)$ for different foci. $CD_{(x, pat(p, y, z))}$ represents the content's pattern for $pat(p, y, z)$, with the data containing attribute x in the X dimension irrespective of **Who** received the data. A high value of $CD_{(x, pat(p, y, z))}$ means there are more patterns containing attribute x compared to other contents. $RD_{(q, pat(p, y, z))}$ indicates the receiver's pattern for $pat(p, y, z)$, with the data received by receiver q irrespective of **What** was the data contained. A high value of $RD_{(q, pat(p, y, z))}$ means that receiver q got more data patterns compared to other receivers. Figure 3.6 shows an example of the 5Ws parallel coordinates with $CD_{(x, pat(p, y, z))}$ via $RD_{(q, pat(p, y, z))}$.

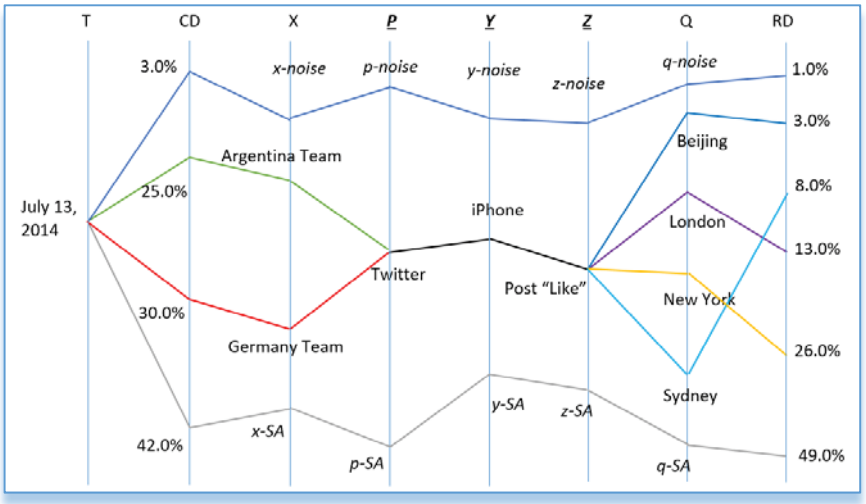

Figure 3.6 Example of 5Ws parallel coordinates with $CD_{(x, pat(p, y, z))}$ via $RD_{(q, pat(p, y, z))}$

In Figure 3.6, there is only one pattern $pat_{(Twitter, iPhone, Like)}$ in the Pair-Density. The values of $CD_{(x, pat(p, y, z))}$ has two attributes in the X dimension, and the values of $RD_{(q, pat(p, y, z))}$ demonstrate that there are four attributes in the Q dimension. In the $CD_{(\ )}$ axis,

$CD_{(Germany)} = 30.0\% > CD_{(Argentina)} = 25.0\%$, which indicates that 5.0% more data in pattern $pat_{(twitter, iPhone, Like)}$ favoured the Germany Team compared to the Argentina Team. In the $RD_{(\ )}$ axis, New York received the highest density at $RD_{(New\ York)} = 26.0\%$, which was much greater than the other three cities, with $RD_{(London)} = 13.0\%$, $RD_{(Sydney)} = 8.0\%$ and $RD_{(Beijing)} = 3.0\%$. This means that New York has the highest number of Twitter receivers who used iPhone compared to other cities, whilst Beijing had the lowest due to the low popularity of Twitter in China.

In the $CD_{(\ )}$ axis, the two teams have a total $CD_{(\ )} = 55.0\%$ across the four cities, compared to a total receiver density $RD_{(\ )} = 50.0\%$. $CD_{(x\text{-}SA)}$ (42.0%) < $RD_{(q\text{-}SA)}$ (49.0%), which means that other cities also received data patterns with similar data contents.

3.8. $CD_{(x, pat(p, y, q))}$ via $PD_{(z, pat(p, y, q))}$ with $pat(p, y, q)$

The Pair-Density $CD_{(x, pat(p, y, q))}$ via $PD_{(z, pat(p, y, q))}$ measures the content patterns and purpose patterns by comparing the attributes between the X and Z dimensions for the particular pattern $pat(p, y, q)$, which can be denoted as

$$\begin{cases} CD(x, pat(p,y,q)) = \dfrac{|D(x,\ pat(p,y,q))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,y,q), x\} \times 100\% \\ PD(z, pat(p,y,q)) = \dfrac{|D(z,\ pat(p,y,q))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,y,q), z\} \times 100\% \end{cases}$$

Equation 3. 1 0

$CD_{(x, pat(p, y, q))}$ measures the content pattern during data transferral for the data content x and for the particular pattern $pat(p, y, q)$, $0 \leq CD_{(x, pat(p, y, q))} \leq 1$ and $\sum CD_{(x, pat(p, y, q))} = 1$. $PD_{(z, pat(p, y, q))}$ measures the purpose pattern during data transferral for the reason z and for the particular pattern $pat(p, y, q)$, $0 \leq PD_{(z, pat(p, y, q))} \leq 1$ and $\sum PD_{(z, pat(p, y, q))} = 1$.

In $CD_{(x, pat(p, y, q))}$ via $PD_{(z, pat(p, y, q))}$, $CD_{(x, pat(p, y, q))}$ represents the content's pattern for *pat(p, y, q)* that contains attribute *x* in the X dimension, irrespective of **Why** the data occurred. A high value of $CD_{(x, pat(p, y, q))}$ means that more data had content *x* compared to other content attributes. $PD_{(z, pat(p, y, q))}$ indicates the purpose pattern for *pat(p, y, q)* that was sent for reason *z*, irrespective of **What** the data contained. A high value of $PD_{(z, pat(p, y, q))}$ means that reason *z* was responsible for more data compared to other reasons. Figure 3.7 shows an example of the 5Ws parallel coordinates with $CD_{(x, pat(p, y, q))}$ via $PD_{(z, pat(p, y, q))}$.

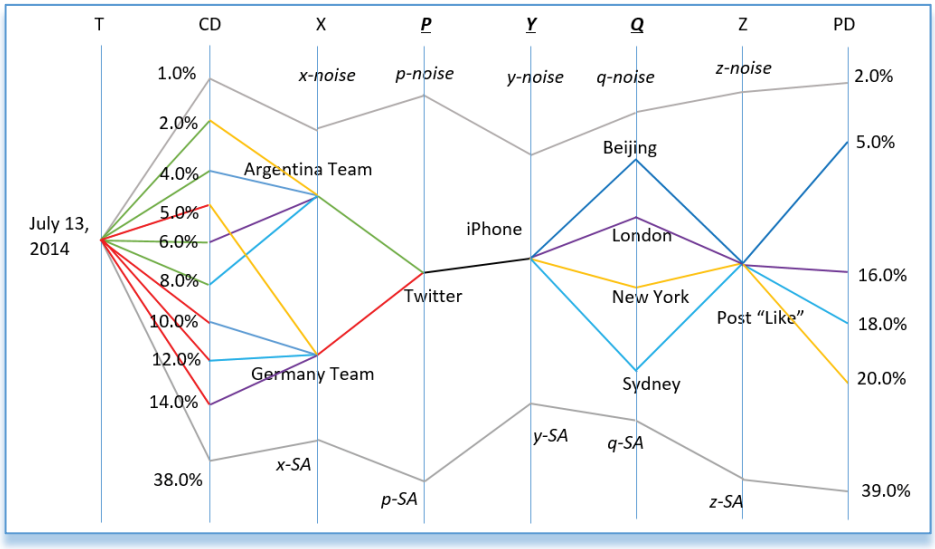

Figure 3.7 Example of 5Ws parallel coordinates with $CD_{(x, pat(p, y, q))}$ via $PD_{(z, pat(p, y, q))}$

In Figure 3.7, there are four $pat_{(p, y, q)}$ patterns between the P, Y and Q axes in this Pair-Density $CD_{(x, pat(p, y, q))}$ via $PD_{(z, pat(p, y, q))}$. In the $CD_{( )}$ axis, $CD_{(Beijing)}$ (10.0%+4.0% = 14.0%) > $PD_{(Beijing)}$ (5.0%), $CD_{(London)}$ (14.0%+6.0% = 20.0%) > $PD_{(London)}$ (16.0%), and $CD_{(Sydney)}$ (12.0%+8.0% = 20.0%) > $PD_{(Sydney)}$ (18.0%), which means that the content pattern occurred more frequently than the purpose pattern for these three cities. In other words, this indicates that these three cities have higher densities for both the Argentina and Germany Teams. $CD_{(New\ York)}$ (5.0%+2.0% = 7.0%) < $PD_{(New\ York)}$ (20.0%), which

indicates that the reason "Like" for New York also contained other teams within this pattern, with the difference being 13%.

In the $PD_{(\ )}$ axis, the lowest density is $PD_{(Beijing)}$ = 5.0%, which reflects that Twitter is not very popular in Beijing. The densities for the other three cities are all similar to each other, indicating that the popularity of Twitter and the FIFA final were similar among these three cities. In the $CD_{(\ )}$ axis, the Germany Team had the top three densities, which illustrates that most contents during the game were about the winning Germany Team. The highest $CD_{(Argentina)}$ is from Sydney, which indicates that the Argentina Team was much popular in Sydney compared to the other three cities. New York had the lowest density for both the Argentina and Germany Teams, which reflects the fact that soccer is not very popular in America. $CD_{(London)}$ (14.0%+6.0%) = $CD_{(Sydney)}$ (12.0%+8.0%) = 20.0%, which is greater than $CD_{(Beijing)}$ (10.0%+4.0%=14.0%) or $CD_{(New\ York)}$ (5.0%+2.0%=7.0%). This indicates that soccer is more popular in London and Sydney compared to Beijing and New York. $CD_{(x\text{-}SA)}$ (38.0%) ≈ $PD_{(z\text{-}SA)}$ (39.0%), suggesting that the other contents and reasons have similar densities.

3.9. $CD_{(x,\ pat(p,\ z,\ q))}$ via $TD_{(y,\ pat(p,\ z,\ q))}$ with *pat(p, z, q)*

The Pair-Density $CD_{(x,\ pat(p,\ z,\ q))}$ via $TD_{(y,\ pat(p,\ z,\ q))}$ measures the content patterns and transferring patterns by comparing the attributes between the X and Y dimensions for the particular pattern *pat(p, z, q)*, which can be denoted as

$$\begin{cases} CD(x, pat(p,z,q)) = \dfrac{|D(x,\ pat(p,z,q))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,z,q), x\} \times 100\% \\ TD(y, pat(p,z,q)) = \dfrac{|D(y,\ pat(p,z,q))|}{|\ D\ |} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,z,q), y\} \times 100\% \end{cases}$$

Equation 3. 1 1

$CD_{(x,\ pat(p,\ z,\ q))}$ measures the content pattern during data transferral for the data content x and for the particular pattern $pat(p, z, q)$, $0 \leq CD_{(x,\ pat(p,\ z,\ q))} \leq 1$ and $\sum CD_{(x,\ pat(p,\ z,\ q))} = 1$. $TD_{(y,\ pat(p,\ z,\ q))}$ measures the transferring pattern during data transferral for method y and for the particular pattern $pat(p, z, q)$, $0 \leq TD_{(y,\ pat(p,\ z,\ q))} \leq 1$ and $\sum TD_{(y,\ pat(p,\ z,\ q))} = 1$.

In $CD_{(x,\ pat(p,\ z,\ q))}$ via $TD_{(y,\ pat(p,\ z,\ q))}$, $CD_{(x,\ pat(p,\ z,\ q))}$ represents the content's pattern for $pat(p, z, q)$ that contains attribute x in the X dimension, irrespective of **How** the data was transferred. A high value of $CD_{(x,\ pat(p,\ z,\ q))}$ means that the content x occurred more frequently compared to other contents. $TD_{(y,\ pat(p,\ z,\ q))}$ indicates the pattern $pat(p, z, q)$ transferred by attribute y, irrespective of **What** the data contained. A high value of $TD_{(y,\ pat(p,\ z,\ q))}$ means that the method y has transferred more data compared to other methods. Figure 3.8 shows an example of the 5Ws parallel coordinates with $CD_{(x,\ pat(p,\ z,\ q))}$ via $TD_{(y,\ pat(p,\ z,\ q))}$.

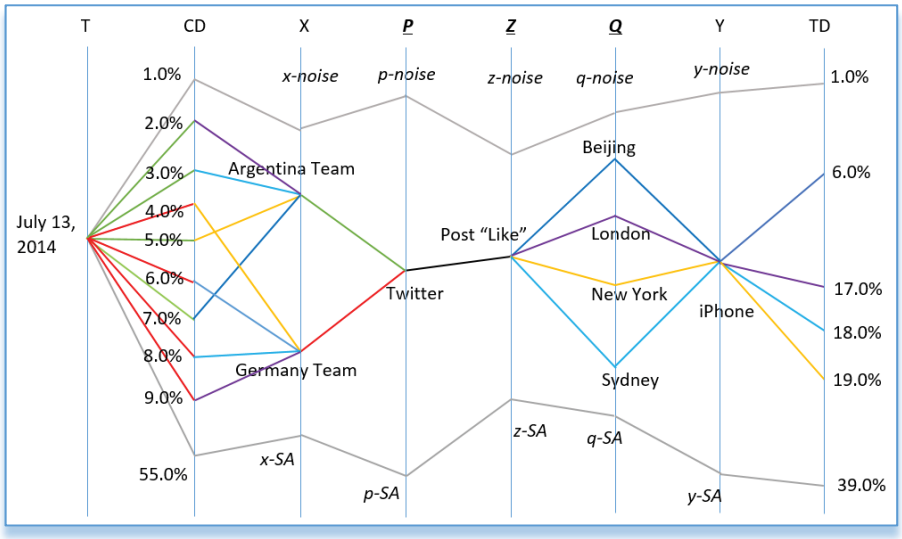

Figure 3.8 Example of 5Ws parallel coordinates with $CD_{(x,\ pat(p,\ z,\ q))}$ via $TD_{(y,\ pat(p,\ z,\ q))}$

In Figure 3.8, there are four $pat_{(p,\ z,\ q)}$ patterns between the P, Z and Q axes in this Pair-Density $CD_{(x,\ pat(p,\ z,\ q))}$ via $TD_{(y,\ pat(p,\ z,\ q))}$. In the $CD_{(\ )}$ axis, there are eight different values, which combine two attributes from the P axis and four attributes from the Q axis.

$CD_{(Beijing)}$ (7.0%+6.0%=13.0%) > $TD_{(Beijing)}$ (6.0%) suggests that in Beijing, the content pattern occurred more often than the transferring pattern. This indicates that messages in support of both the Argentina and Germany Team were often transferred by other devices. $CD_{(London)}$ (9.0%+2.0%=11.0%) < $TD_{(London)}$ (17.0%), and $CD_{(Sydney)}$ (8.0%+3.0%=11.0%) < $TD_{(Sydney)}$ (18.0%), which indicate that iPhone messages to London and Sydney contained many messages involving teams other than Germany and Argentina. New York had the lowest density $CD_{(New\ York)}$ = 9.0% (5.0%+4.0%), which suggests that there are not many football (soccer) fans compared to the other three cities. However, New York has the highest transferring density, with $TD_{(New\ York)}$ = 19.0%, which suggests that the iPhone is much more popular in New York compared to the other three cities. The difference in iPhone usage is about 10%.

In the $CD_{(\ )}$ axis, the Germany Team has the top two $CD_{(Germany)}$ densities, which were received by London and Sydney. This suggests that these two cities received the most data that contained references to the winning Germany Team. The highest $CD_{(Argentina)}$ is Beijing, which indicates that the Argentina Team is more popular in Beijing compared to the other three cities. The Argentina Team has the bottom two densities $CD_{(Argentina)}$, which were also received by London and Sydney. This suggests that these cities have few Argentina supporters. $CD_{(New\ York)}$ (9.0%) < $CD_{(\ Beijing)}$ (13.0%), which indicates that soccer is more popular in Beijing than in New York. $CD_{(London)}$ = $CD_{(Sydney)}$ = 11.0%, which suggests that there is the same density of football (soccer) fans in both cities. In the $TD_{(\ )}$ axis, the lowest density is $TD_{(Beijing)}$ = 6.0%, which means that Twitter iPhone users in Beijing were not very active during the game. The $TD_{(\ )}$ densities for the other three cities are close to each other, indicating that these three cities have similar iPhone usages. $CD_{(x\text{-}SA)}$ = 55.0% is much higher than $TD_{(y\text{-}SA)}$ = 39.0%, suggesting that

the supporters of other teams also used Twitter via iPhone, with a difference of about 16.0%.

3.10. $TD_{(y, pat(p, x, z))}$ via $RD_{(q, pat(p, x, z))}$ with *pat(p, x, z)*

The Pair-Density $TD_{(y, pat(p, x, z))}$ via $RD_{(q, pat(p, x, z))}$ measures the transferring patterns and receiving patterns by comparing the attributes between the Y and Q dimensions for the particular pattern *pat(p, x, z)*, which can be denoted as

$$\begin{cases} TD(y, pat(p,x,z)) = \dfrac{|D(y,\ pat(p,x,z))|}{|\,D\,|} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,x,z), y\} \times 100\% \\ RD(q, pat(p,x,z)) = \dfrac{|D(q,\ pat(p,x,z))|}{|\,D\,|} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,x,z), q\} \times 100\% \end{cases}$$

Equation 3. 1 2

$TD_{(y, pat(p, x, z))}$ measures the transferring pattern during data transferral for method y and for the particular pattern *pat(p, x, z)*, $0 \leq TD_{(y, pat(p, x, z))} \leq 1$ and $\sum TD_{(y, pat(p, x, z))} = 1$. $RD_{(q, pat(p, x, z))}$ measures the receiving pattern during data transferral for receiver q and for the particular pattern *pat(p, x, z)*, $0 \leq RD_{(q, pat(p, x, z))} \leq 1$ and $\sum RD_{(q, pat(p, x, z))} = 1$.

In $TD_{(y, pat(p, x, z))}$ via $RD_{(q, pat(p, x, z))}$, $TD_{(y, pat(p, x, z))}$ represents the transferring pattern for *pat(p, x, z)* that is transferred by y in the Y dimension, irrespective of **Who** received the data. A high value of $TD_{(y, pat(p, x, z))}$ means that the method y transferred more data compared to other methods. $RD_{(q, pat(p, x, z))}$ indicates the pattern *pat(p, x, z)* that is received by the attribute q, irrespective of **How** the data was transferred. A high value of $RD_{(q, pat(p, x, z))}$ means that the attribute q received more data compared to the other receivers. Figure 3.9 shows an example of the 5Ws parallel coordinates with $TD_{(y, pat(p, x, z))}$ via $RD_{(q, pat(p, x, z))}$.

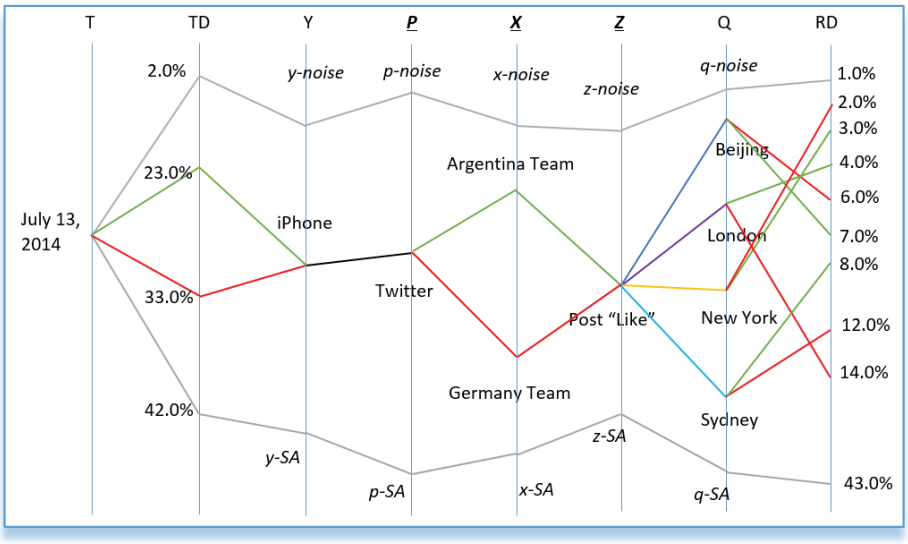

Figure 3.9 Example of 5Ws parallel coordinates with $TD_{(y,\ pat(p,\ x,\ z))}$ via $RD_{(q,\ pat(p,\ x,\ z))}$

In Figure 3.9, there are two $pat_{(p,\ x,\ z)}$ patterns between the P, X and Z axes in this Pair-Density $TD_{(y,\ pat(p,\ x,\ z))}$ via $RD_{(q,\ pat(p,\ x,\ z))}$. One pattern is for the Argentina Team and the other is for the Germany Team. $TD_{(Germany)}$ (33.0%) ≈ $RD_{(Germany)}$ (14.0%+12.0%+6.0%+2.0% = 34.0%), and $TD_{(Argentina)}$ (23.0%) ≈ $RD_{(Argentina)}$ (8.0%+7.0%+4.0%+3.0% = 22.0%), which indicate that both patterns have a similar transferring density and receiving density.

In the $TD_{(\ )}$ axis, $TD_{(Germany)}$ (33.0%) > $TD_{(Argentina\ )}$ (23.0%), which means that transferring patterns favoured the Germany Team by 10%. In the $RD_{(\ )}$ axis, $RD_{(Germany)}$ = 34.0% (14.0% + 12.0% + 6.0% + 2.0%) > $RD_{(Argentina)}$ = 22.0% (8.0% + 7.0% + 4.0% + 3.0%), which indicates that the receivers received 12.0% more messages about the winning Germany Team than the losing Argentina Team. The top two $RD_{(\ )}$ were received by London (14.0%) and Sydney (12.0%) for the Germany Team, which suggests that London and Sydney have the most Germany fans. The bottom two $RD_{(\ )}$ were received by New York for both the Germany and Argentina Teams, which indicates that there are not many football (soccer) fans in New York. Interesting, both Beijing and New York received more messages supporting Argentina than supporting Germany. $TD_{(y\text{-}SA)}$ (42.0%)

≈ $RD_{(q\text{-}SA)}$ (43.0%), which indicates that the density for other devices is similar to the density for other cities.

3.11. $TD_{(y, pat(p, x, q))}$ via $PD_{(z, pat(p, x, q))}$ with *pat(p, x, q)*

The Pair-Density $TD_{(y, pat(p, x, q))}$ via $PD_{(z, pat(p, x, q))}$ measures the transferring patterns and purpose patterns by comparing the attributes between the Y and Z dimensions for the particular pattern *pat(p, x, q)*, which can be denoted as

$$\begin{cases} TD(y, pat(p,x,q)) = \dfrac{|D(y,\ pat(p,x,q))|}{|\,D\,|} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,x,q), y\} \times 100\% \\ PD(z, pat(p,x,q)) = \dfrac{|D(z,\ pat(p,x,q))|}{|\,D\,|} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,x,q), z\} \times 100\% \end{cases}$$

Equation 3. 1 3

$TD_{(y, pat(p, x, q))}$ measures the transferring pattern during data transferral for method *y* with the particular pattern *pat(p, x, q)*, $0 \le TD_{(y, pat(p, x, q))} \le 1$ and $\sum TD_{(y, pat(p, x, q))} = 1$. $PD_{(z, pat(p, x, q))}$ measures the purpose pattern during data transferral for reason *z* in the particular pattern *pat(p, x, q)*, $0 \le PD_{(z, pat(p, x, q))} \le 1$ and $\sum PD_{(z, pat(p, x, q))} = 1$.

In $TD_{(y, pat(p, x, q))}$ via $PD_{(z, pat(p, x, q))}$, $TD_{(y, pat(p, x, q))}$ represents the transferring pattern for *pat(p, x, q)* which is transferred by *y* in the Y dimension, irrespective of **Why** the data occurred. A high value of $TD_{(y, pat(p, x, q))}$ means that the method *y* transferred more data compared to other methods. $PD_{(z, pat(p, x, q))}$ indicates the data pattern *pat(p, x, q)* for reason *z*, irrespective of **How** the data was transferred. A high value of $PD_{(z, pat(p, x, q))}$ means that the attribute *z* had more data patterns compared to other purposes. Figure 3.10 shows an example of 5Ws parallel coordinates with $TD_{(y, pat(p, x, q))}$ via $PD_{(z, pat(p, x, q))}$.

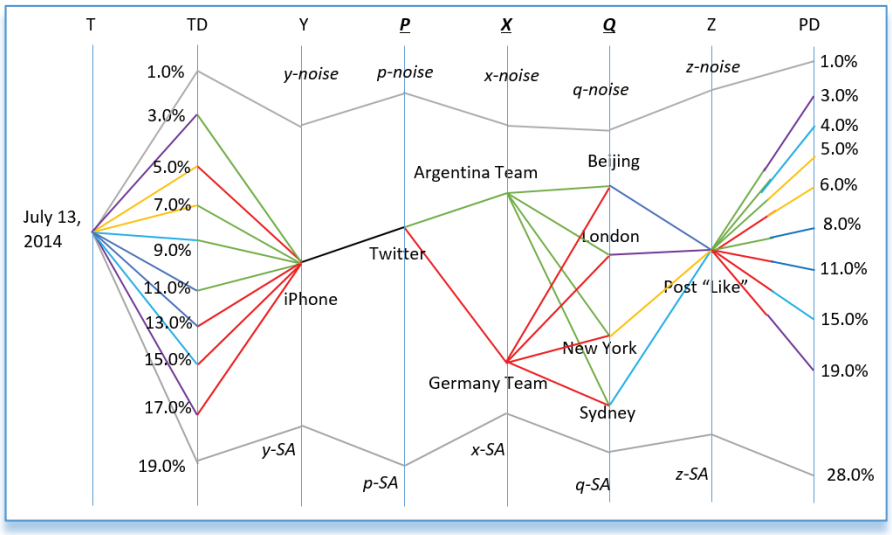

Figure 3.10 Example of 5Ws parallel coordinates with $TD_{(y,\ pat(p,\ x,\ q))}$ via $PD_{(z,\ pat(p,\ x,\ q))}$

In Figure 3.10, there are eight $pat_{(p,\ x,\ q)}$ patterns between the P, X and Q axes in this Pair-Density $TD_{(y,\ pat(p,\ x,\ q))}$ via $PD_{(z,\ pat(p,\ x,\ q))}$, which combine two attributes from the X dimension and four attributes from the Q dimension. Comparing $TD_{(\ )}$ axis and $PD_{(\ )}$ axis, the Germany Team has the three top densities for both densities at the same value, and total $TD_{(Germany)}$ (50.0%) ≈ $PD_{(Germany)}$ (51.0%). This all suggests that the transferring density and purpose density for pattern $pat_{(p,\ x,\ q)}$ are very similar. $TD_{(Argentina)}$ (11.0%+9.0%+7.0%+3.0% = 30.0%) > $PD_{(Argentina)}$ (8.0%+5.0%+4.0%+3.0% = 20.0%), which suggests that iPhones were used to post "Like" messages more than other devices. $TD_{(Beijing)}$ (13.0%+11.0% = 24.0%) > $PD_{(Beijing)}$ (11.0%+8.0% = 19.0%) and $TD_{(Sydney)}$ (15.0%+9.0% = 24.0%) > $PD_{(Sydney)}$ (15.0%+4.0% = 19.0%) both show that these two cities have high percentages of data sent through iPhone via Twitter. $TD_{(London)}$ (17.0%+3.0% = 20.0%) ≈ $PD_{(London)}$ (19.0%+3.0% = 22.0%) and $TD_{(New\ York)}$ (7.0%+5.0% = 12.0%) ≈ $PD_{(New\ York)}$ (6.0%+5.0% = 11.0%) both indicate that London and New York have similar densities for $TD_{(\ )}$ and $PD_{(\ )}$.

In the $TD_{(\ )}$ axis, $TD_{(iPhone,\ pat(Twitter,\ Argentina,\ New\ York))}$ (7.0%) > $TD_{(iPhone,\ pat(Twitter,\ Germany,\ New\ York))}$ (5.0%), suggesting that there are more Argentina supporters in New York

than Germany supporters. The other three cities all favoured the winning Germany team. The largest gap between Germany and Argentina support was in London, with a difference of $TD_{(iPhone, pat(Twitter, Germany, London))} - TD_{(iPhone, pat(Twitter, Argentina, London))} = 17.0\% - 3.0\% = 14.0\%$. In the $PD_{(\ )}$ axis, all cities favoured the winner team, with the largest gaps also occurring in London, where the difference was $PD_{(Like, pat(Twitter, Germany, London))} - PD_{(Like, pat(Twitter, Argentina, London))} = 19.0\% - 3.0\% = 16.0\%$. The $TD_{(y\text{-}SA)} = 19.0\%$ is less than $PD_{(z\text{-}SA)} = 28\%$, which indicates that other devices have also sent significant numbers of "Like" Twitter messages.

3.12. $PD_{(z, pat(p, x, y))}$ via $RD_{(q, pat(p, x, y))}$ with *pat(p, x, y)*

The Pair-Density $PD_{(z, pat(p, x, y))}$ via $RD_{(q, pat(p, x, y))}$ measures the purpose patterns and receiving patterns by comparing the attributes between the Z and Q dimensions for the particular pattern *pat(p, x, y)*, which can be denoted as

$$\begin{cases} PD(z, pat(p,x,y)) = \dfrac{|D(z,\ pat(p,x,y))|}{|D|} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,x,y), z\} \times 100\% \\ RD(q, pat(p,x,y)) = \dfrac{|D(q,\ pat(p,x,y))|}{|D|} \times 100\% = \dfrac{1}{n}\{d \in D | pat(p,x,y), q\} \times 100\% \end{cases}$$

Equation 3. 1 4

$PD_{(z, pat(p, x, y))}$ measures the purpose pattern during data transferral for reason z and for the particular pattern *pat(p, x, y)*, $0 \leq PD_{(z, pat(p, x, y))} \leq 1$ and $\sum PD_{(z, pat(p, x, y))} = 1$. $RD_{(q, pat(p, x, y))}$ measures the receiving pattern during data transferral for receiver q and for the particular pattern *pat(p, x, y)*, $0 \leq RD_{(q, pat(p, x, y))} \leq 1$ and $\sum RD_{(q, pat(p, x, y))} = 1$.

In $PD_{(z, pat(p, x, y))}$ via $RD_{(q, pat(p, x, y))}$, $PD_{(z, pat(p, x, y))}$ represents the purpose pattern for *pat(p, x, y)* at reason z in the Z dimension, irrespective of **Who** received the data. A high value of $PD_{(z, pat(p, x, y))}$ means that the purpose z has more data compared to the other

reasons. $RD_{(q,\ pat(p,\ x,\ y))}$ indicates the receiving pattern for the attribute q, irrespective of **Why** the data occurred. A high value of $RD_{(q,\ pat(p,\ x,\ y))}$ means that the attribute q received more data compared to the other receivers. Figure 3.11 shows an example of the 5Ws parallel coordinates with $PD_{(z,\ pat(p,\ x,\ y))}$ via $RD_{(q,\ pat(p,\ x,\ y))}$.

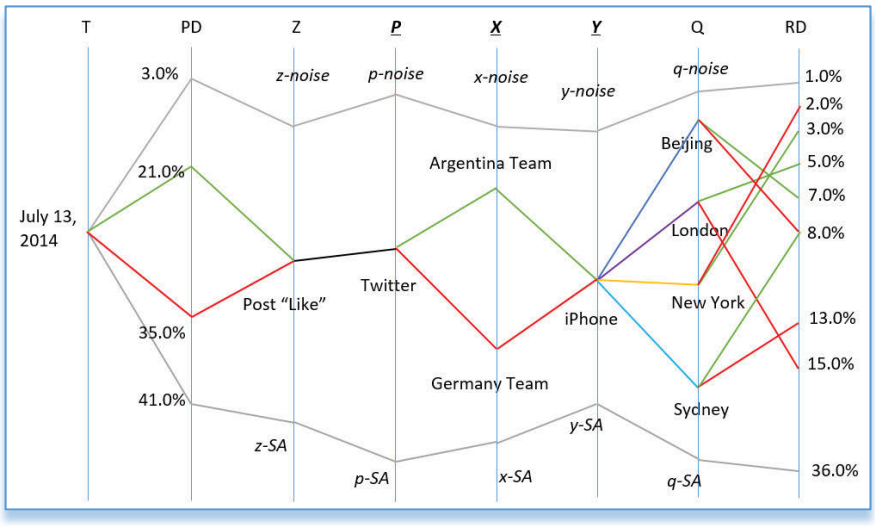

Figure 3.11 Example of 5Ws parallel coordinates with $PD_{(z,\ pat(p,\ x,\ y))}$ via $RD_{(q,\ pat(p,\ x,\ y))}$

In Figure 3.11, there are two patterns, $pat_{(Twitter,\ Argentina,\ iPhone)}$ and $pat_{(Twitter,\ Germany,\ iPhone)}$ in this Pair-Density $PD_{(z,\ pat(p,\ x,\ y))}$ via $RD_{(q,\ pat(p,\ x,\ y))}$. In comparing the $PD_{(\ )}$ axis and the $RD_{(\ )}$ axis, $PD_{(Germany)}$ (35.0%) < $RD_{(Germany)}$ (15.0%+13.0%+8.0%+2.0% = 38.0%) and $PD_{(Argentina)}$ (21.0%) < $RD_{(Argentina)}$ (8.0%+7.0%+5.0%+3.0% = 23.0%), which both indicate that the purpose pattern for both teams is less than the receiving patterns. Hence, these four cities received more twitters irrespective of the "Like" purpose or not.

In the $PD_{(\ )}$ axis, $PD_{(Germany)}$ (35.0%) > $PD_{(Argentina)}$ (21.0%), suggests that 14% more tweets were for the purpose of "liking" the Germany Team. In the $RD_{(\ )}$ axis, $RD_{(Germany)}$ = 38.0% > $RD_{(Argentina)}$ = 23.0%, which indicates that 15.0% more receivers received messages for the winning Germany Team compared to the losing Argentina Team. The top two $RD_{(\ )}$ favoured the winning Germany Team, which were received by London and Sydney. The bottom $RD_{(\ )}$ favoured the Argentina Team and was received by

New York, suggesting that football (soccer) and the Argentina Team are both not very popular in New York. Beijing, London and Sydney all received more data for the Germany Team than for the Argentina Team, suggesting that these cities had more supporters who favoured the winning Germany Team. However, New York was the opposite, with the Argentina Team receiving more messages or support than the Germany Team. In the $PD_{( )}$ axis, $PD_{(z\text{-}SA)} = 41.0\% > RD_{(q\text{-}SA)} = 36.0\%$, meaning that other reasons had a higher density on Twitter for the World Cup Final.

In conclusion, this chapter has discussed ten different Pair-Densities, corresponding with different data patterns in the 5Ws dimensions. Pair-Densities provide measurement and comparison for any pair of dimensions and patterns with any form of data. To the best of our knowledge, no previous work has addressed these comparisons. The following chapter describes the implementation of our model.

Chapter 4: Case Study

Our new approach has been tested using three different Big Data datasets, containing nearly three million data incidents combined. The first case study is based on the US 2008 flight dataset (ASA, 2009), which has 29 data dimensions containing 1,048,575 flight incidents. The second case study is based on the UTS Library 2009 email dataset, which has 17 dimensions containing 585,300 email incidents. The third case study is based on the ISCX2012 network dataset (Shiravi et al 2012), which has 20 data dimensions containing 1,511,636 data incidents. The implementation results show that our new approach has significantly improved both the accuracy and visualization of Big Data analytics.

4.1. Case One: Visual Estimate of US 2008 Flight Delay Patterns

The US 2008 flights dataset includes all flight incidents in the United States for January and February 2008, recording 1,048,575 incidents across 29 data dimensions. These dimensions include flight date, flight time, flight number, plan number, origin and destination airports, airline name, scheduled and actual departure time, scheduled and actual arrival time, and reasons for any delay. An example of a data entry for a flight incident can be described as:

"*At Thursday Jan-3 2008, FlightNum: 3920 (Airline: WN, plane TailNum: N464WN), departed at 18:29 (scheduled at 17:55) from IND airport, flew 77 minutes for 515 miles, and arrived at BWI at 19:50 (scheduled 19:25), resulting in a delay of 34 minutes due to CarrierDelay 2 minutes and LateAircraftDelay 32 minutes*".

Table 4.1 shows the 5Ws classifications for the US 2008 flight dataset. NAS delay stands for National Aviation System (NAS) delay. A plane is normally identified by its tail number, such as N612SW or N956AT.

Table 4.1. 5Ws classification for US 2008 flight dataset

5Ws Dimension	Dimensions in dataset	Details
When (T)	Year	2008
	Month	Jan - Feb
	Day of Month	1 – 31
	Day of Week	Mon - Sun
Where (P)	Origin Airport	286 airports
What (X)	Delay	Minutes
	Departure Delay	Minutes
	- Scheduled Departure	Minutes
	- Actual Departure	Minutes
	Arrival Delay	Minutes
	- Scheduled Arrival	Minutes
	- Actual Arrival	Minutes
	Carrier Delay (A)	Minutes
	Weather Delay (B)	Minutes
	NAS Delay (C)	Minutes
	Security Delay (D)	Minutes
	Late Aircraft Delay	Minutes
	Cancelled	Yes / No
	Cancellation Code	A / B / C / D
	Diverted	Yes / No

How (Y)	Flight No	7131 flights
	Plane No	4930 planes
	Airline	20
Why (Z)	Flight Distance	Miles
	Air Time	Minutes
	Actual Elapsed Time	Minutes
	Scheduled Elapsed Time	Minutes
Who (Q)	Destination Airport	287 airports

I selected Day of Week (T), Origin Airport (P), Delay (X), Airline (Y), Flight Distance (Z) and Destination Airport (Q) as the primary 5Ws dimensions in visual analysis. I then used clustered Pair-Density in the 5Ws parallel coordinates to narrow down the worst delay patterns.

4.1.1. 5Ws Pattern for US Flight Dataset

The details of the dataset in the 5Ws dimensions are shown in Table 4.2. 270,096 flight delays were recorded, with the worst delay being about 40 hours. To reduce the length of the X dimension axis and decrease data clutter, I have used SA to categorise delay times into 30 minute intervals, which are then scaled into units between 0 to 600 minutes. Delay times over 600 minutes were shrunk into a single attribute labelled ">600".

Table 4.2. 5Ws dimension for US 2008 flight dataset

5Ws Dimension	Dimensions in dataset	Details
When (T)	Day of Week	Mon - Sun
Where (P)	Origin Airport	286 airports

What (X)	Delay	270,096 flights
How (Y)	Airlines	20
Why (Z)	Flight Distance	Miles
Who (Q)	Destination Airport	287 airports

In the dataset, Flight Distance ranged from 24 to 4962 miles. This was also scaled using SA, with 100-mile intervals up to 3000 miles. Any distances over 3000 miles were shrunk into 1000-mile intervals.

The UniqueCarrier is the airline's code in the Y dimension. The dataset in the 5Ws parallel coordinates is shown in Figure 4.1.

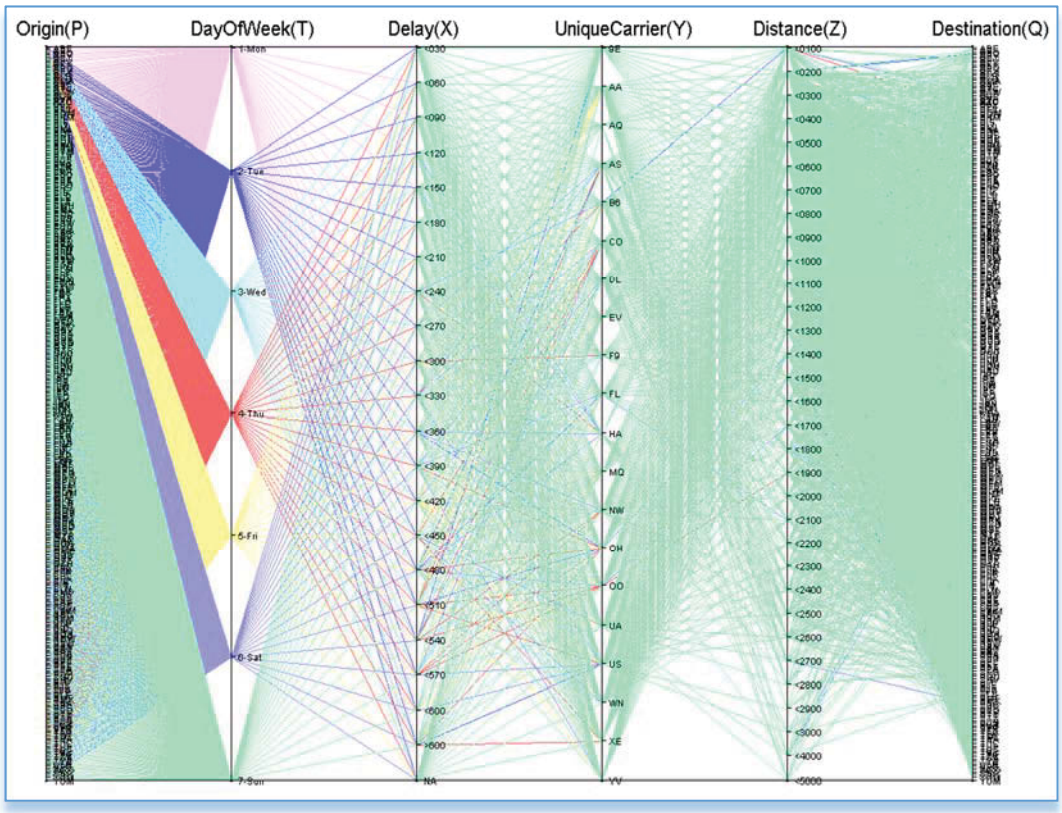

Figure 4.1 US 2008 flight patterns in 5Ws parallel coordinates

In Figure 4.1, attribute “NA” in the delay dimension (X dimension) means that the flight was not flagged as delayed, but instead departed or arrived on-time. There are 157,524 5Ws patterns illustrated in the graph, crossing from 286 origin airports (P) to 287 destination airports (Q). There are many overlapping polylines in the graph, with overcrowded attributes occurred in both the P and Q axes. This makes it hard to explore delay patterns in any sort of detail.

4.1.2. The Delay Pattern Combined Airline and Flight Distance

To investigate the worst delay patterns for delays over ten hours (X > 600 minutes), I have applied SA in the P and Q dimensions to collect all patterns $pat_{(x\leq 600,\, y,\, z)}$, including on-time flight patterns $pat_{(x=NA,\, y,\, z)}$, and then used the Pair-Density parallel coordinates to visualize the 5Ws dimensions. Figure 4.2 shows the outcomes of these delay patterns $pat_{(x>600,\, y,\, z)}$ after applying SA in the P and Q axes.

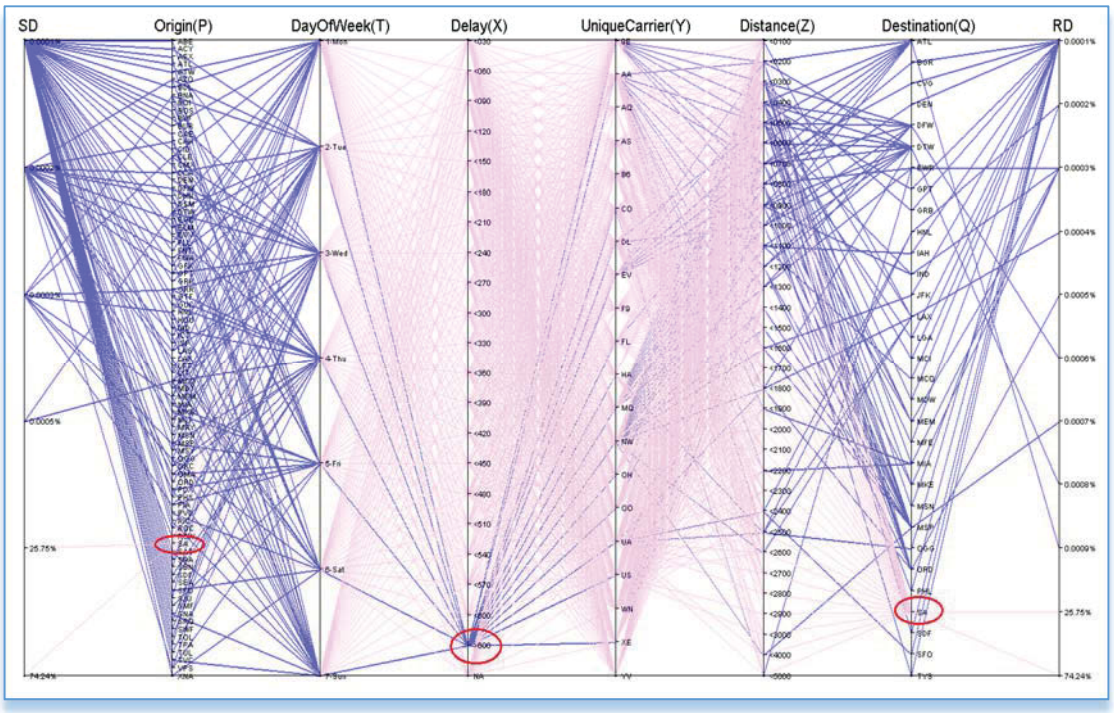

Figure 4.2 Delay pattern $pat_{(x>600,\, y,\, z)}$ in $SD_{(\,)}$ via $RD_{(\,)}$ parallel coordinates

In Figure 4.2, the overcrowding of polylines and attributes has been dramatically reduced. The SA attributes in the P and Q axes collected all patterns $pat_{(x\leq 600,\ y,\ z)}$ and $pat_{(x=NA,\ y,\ z)}$. This has left 132 delay patterns involving delays over ten hours, with $pat_{(x>600,\ y,\ z)}$ crossing 82 attributes in the P axis and 30 attributes in the Q axis. $SD_{(p\text{-}SA,\ pat(x\leq 600,\ y,\ z))} = RD_{(q\text{-}SA,\ pat(x\leq 600,\ y,\ z))} = 25.75\%$, which means that there is the same sender and receiver density for pattern $pat_{(x\leq 600,\ y,\ z)}$ in both the P and Q dimensions. $SD_{(p\text{-}SA,\ pat(x=NA,\ y,\ z))} = RD_{(q\text{-}SA,\ pat(x=NA,\ y,\ z))} = 74.24\%$ indicates that the density of on-time flights is the same in both axes.

The two highest values in the *SD* axis are $SD_{(p=DTW,\ pat(x>600,\ y,\ z))} = 0.0005\%$ and $SD_{(p=MCO,\ pat(x>600,\ y,\ z))} = 0.0005\%$, which indicates that the origin airports DTW and MCO have the highest frequency of flights that are delayed by over 600 minutes. The destination airport ATL has the highest value of $RD_{(\ )}$, where $RD_{(q=ATL,\ pat(x>600,\ y,\ z))} = 0.0009\%$, suggesting that this airport receives the most flights with major delays. The destination airport DTW has the second highest value of $RD_{(\ )}$, where $RD_{(q=DTW,\ pat(x>600,\ y,\ z))} = 0.0008\%$.

4.1.3. Clustered Delay Pattern in $SD_{(\ )}$ via $RD_{(\ )}$ Parallel Coordinates

To further investigate the delay patterns $pat_{(x>600,\ y,\ z)}$, which are highlighted in the previous graph, I have used the clustered 5Ws Pair-Density parallel coordinates to explore the worst delay patterns with delays over 600 minutes. Seven clustered X dimensions have been used in visualization: Departure Delay (X1); Carrier Delay (X2); Weather Delay (X3); NAS Delay (X4); Security Delay (X5); Late Aircraft Delay (X6); and Arrival Delay (X7). Figure 4.3 shows the result of the clustered parallel coordinates.

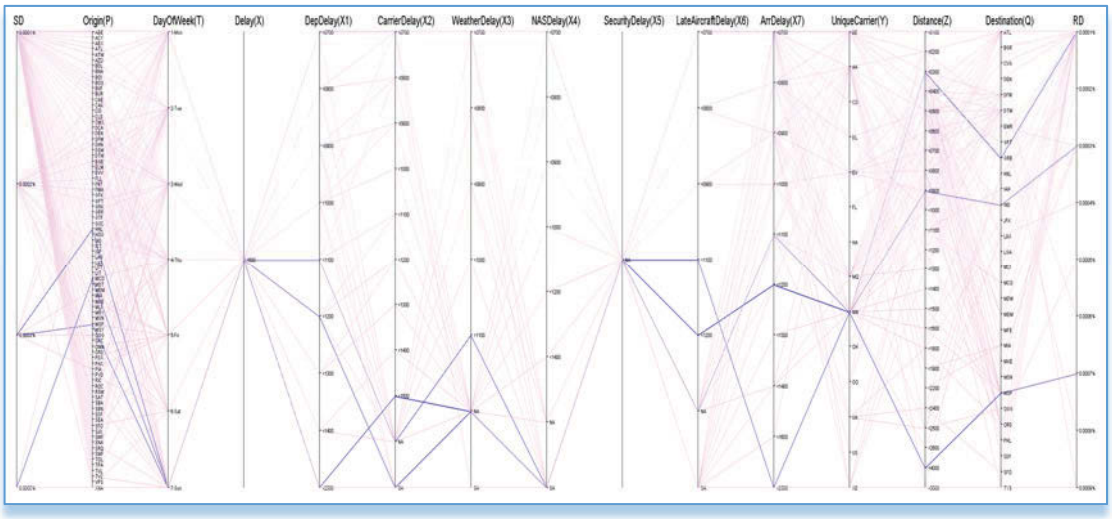

Figure 4.3 Clustered delay pattern $pat_{(x>600,\ y,\ z)}$ in $SD_{(\ )}$ via $RD_{(\ )}$ parallel coordinates

In Figure 4.3, X5 = NA means there is no delay in pattern $pat_{(x<600,\ y,\ z)}$ caused by Security issues. The worst delay pattern is identified as having the highest density of $SD_{(\ )}$ and $RD_{(\ )}$, which also has the highest value in the combination of clustered dimensions from X1 to X7.

The three worst delay patterns are highlighted in the graph: a) Flight from HNL to MSP with pattern $pat_{(x>600,\ x1\approx2500,\ x2\approx1500,\ x3\approx NA,\ x4=SA,\ x5=NA,\ x6\approx1100,\ x7\approx2500,\ y=NW,\ z\approx4000)}$; b) Flight from MCO to IND with pattern $pat_{(x>600,\ x1\approx1200,\ x2=SA,\ x3=NA,\ x4=SA,\ x5=NA,\ x6\approx1200,\ x7\approx1200,\ y=NW,\ z\approx900)}$; and c) Flight from MSP to GRB with pattern $pat_{(x>600,\ x1\approx1100,\ x2=NA,\ x3\approx1100,\ x4=SA,\ x5=NA,\ x6=NA,\ x7\approx1100,\ y=NW,\ z\approx300)}$. NW is the worst airline, with three data instances in the graph of worst delay patterns. Sunday also experiences higher levels of major delays. In the clustered $SD_{(\ )}$ via $RD_{(\ )}$ parallel coordinates, the graph also visually shows how flight distance does not significantly impact upon delay patterns.

4.1.4. Airline Flight Pattern between Origin and Destination Airport

To study airline flight patterns between the origin and destination airports, described under the 5Ws system as the pattern $pat_{(p,\ y,\ q)}$, I used the algorithm of $CD_{(x,\ pat(p,}$

$_{y,\ q))}$ via $PD_{(z,\ pat(p,\ y,\ q))}$ to compare the delay and flight distance dimensions. Figure 4.4 illustrates the $CD_{(x,\ pat(p,\ y,\ q))}$ via $PD_{(z,\ pat(p,\ y,\ q))}$ parallel coordinates for airline flight patterns.

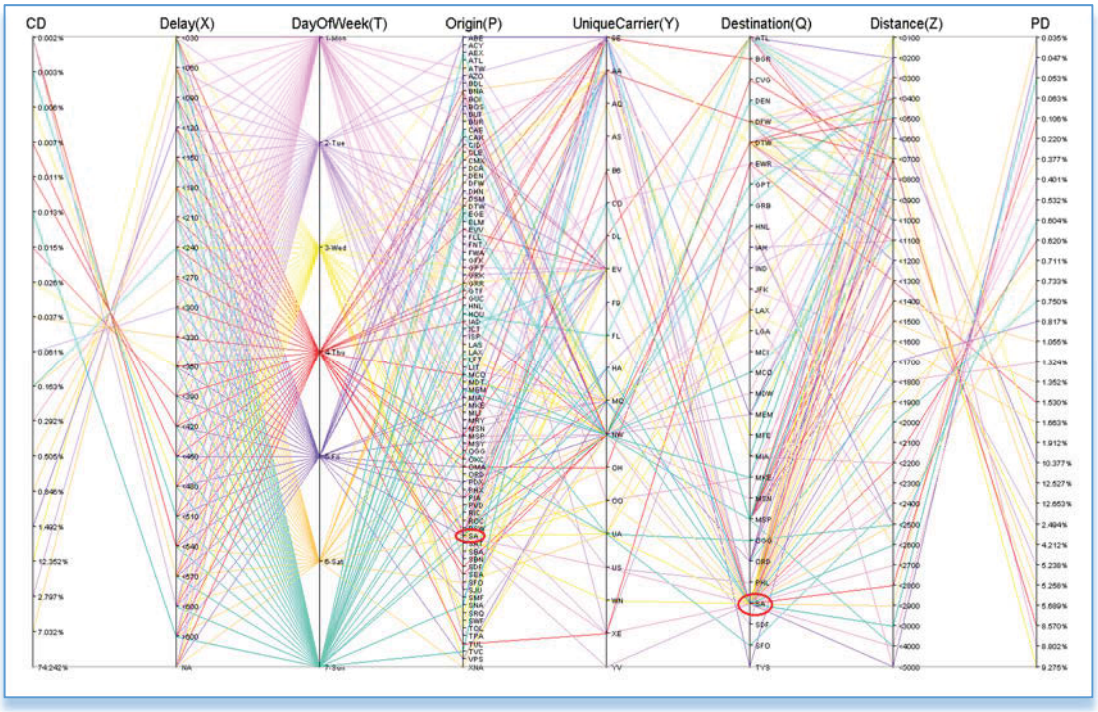

Figure 4.4 Airline flight pattern $pat_{(p,\ y,\ q)}$ in $CD_{(\ )}$ via $PD_{(\ )}$ parallel coordinates

In Figure 4.4, SA has been used in the P and Q axes to collect all patterns for attribute $x \leq 600$, including on-time flight patterns where $x = NA$. 140 airline flight patterns are illustrated in the graph, which combines 83 origin airports in the P axis with 31 destination airports in the Q axis.

In the delay (X) dimension, the highest value is $CD_{(x<030,\ pat(p,\ y,\ q))} = 12.352\%$, which means that the biggest proportion of flight delays are less than 30 minutes. The second highest value is $CD_{(x<060,\ pat(p,\ y,\ q))} = 7.032\%$, meaning that nearly 20% of flight delays are less than one hour. The density of delays over 600 minutes is $CD_{(x>600,\ pat(p,\ y,\ q))} = 0.013\%$.

In the distance (Z) dimension, the highest value is $PD_{(z<0400,\ pat(p,\ y,\ q))} = 12.652\%$, and the second highest value is $PD_{(z<0300,\ pat(p,\ y,\ q))} = 12.527\%$. This means that more than

25% of flight patterns are between 200 and 400 miles. Four airlines (CO, DL, UA and AA) have the longest flights patterns, all exceeding 4000 miles.

4.2. Case Two: Visual Analysis of 2009 Spam Email

Over a period of five months in 2009, the UTS Library email system recorded 585,300 emails involving 17 dimensions, including email date and time, sender name and email address, message head, message body and attachment, message transport name and method, attack type, recipient name and email address and email system reaction. An example of an email incident is shown in Figure 4.5.

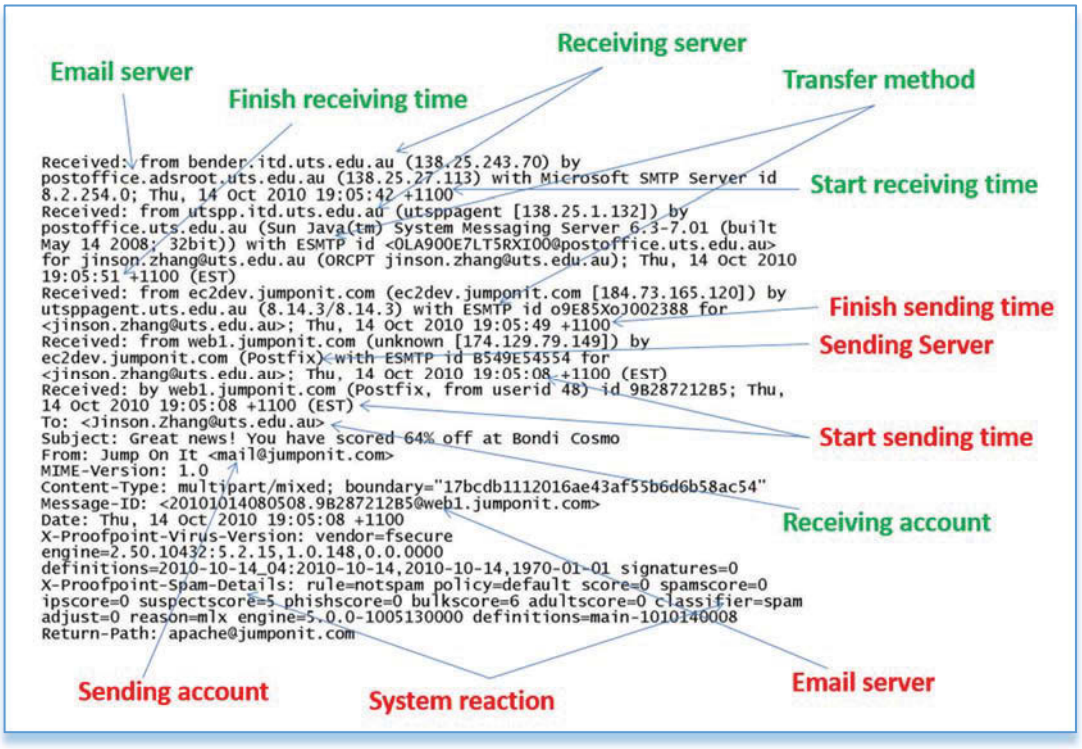

Figure 4.5 Example of an email incident

The 2009 UTS Library Email dataset contains 53,028 spam emails and 4,100 virus attacks, which were detected by our email system using Microsoft Forefront software to identify viruses and protect the email system. The junk-email rule in our email system is

based on the individual user. The 5Ws classification for the 2009 UTS Email dataset is shown in Table 4.3.

Table 4.3. 5Ws classification for 2009 email dataset

5Ws Dimension	Dimensions in dataset	Details
When (T)	Year	2009
	Month	July - Nov
	Day of Month	1 – 31
	Time	0:00 – 24:00
Where (P)	SenderName	423,947
	SenderAddress	363,582
	SenderIP	331,470
What (X)	MessageHead	517,232
	MessageBody	561,894
	Attachment	Y / N
How (Y)	TransmitMethod	Inbound / Outbound
	ScanJob	RealTime / Transport
	SystemAction	Transmit / Remove
Why (Z)	Incident	585,300
	Spam	53,028
	- Virus	4,100
Who (Q)	RecipientName	163,532
	RecipientAddress	159,874
	RecipientIP	138,568
	CcName	7,144
	CcAddress	6,617

	BccName	45,897
	BccAdress	40,821

I selected Time (T), SenderIP (P), MessageHead (X), TransmitMethod (Y), Virus (Z), and RecipientAddress (Q) as the primary 5Ws dimensions for the email virus analysis, and then studied the relationship between the virus and the attachments by using $SD_{()}$ via $PD_{()}$ parallel coordinates which to explore the patterns $pat_{(x, y, q)}$. Clustered Pair-Density parallel coordinates have also been used in visual analytics.

4.2.1. 5Ws Pattern for Virus Email

The details of the 5Ws pattern for virus emails is shown in Table 4.4. There are 25 attributes in the Y dimension to represent transmision methods, 988 attributes in the X dimension, and some virus attacks containing no message head. P = 2,450 > Q = 87, which indicates some recipients suffered from repeat attacks.

Table 4.4. 5Ws dimension for 2009 virus email

5Ws Dimension	Dimensions in dataset	Details
When (T)	Time	0:00 – 24:00
Where (P)	SenderIP	2,450
What (X)	Head	970
How (Y)	TransmitMethod	25
Why (Z)	Virus	4,100
Who (Q)	Recipient	87

Figure 4.6 shows the email virus patterns in the 5Ws parallel coordinates. In the P axis, a few SenderIP nodes have sent many virus attacks with high densities. In the Y axis, attribute "SMTPMessages\InboundAndOurbound" transmitted a large number of spam emails. In the Z axis, three different viruses are linked to more than five different transmission methods, which means that these three viruses are more dangerous than other viruses since it can be transmitted via multiple methods, making it hard to detect.

Figure 4.6 2009 email virus pattern in 5Ws parallel coordinates

In Figure 4.6, there exists a lot of overlapping between polylines, with overcrowded attributes occurring in the T, P, X and Z axes.

4.2.2. Virus Pattern $pat_{(x, y, z)}$ in $SD_{( )}$ via $RD_{( )}$ Parallel Coordinates

To reduce the cluttering of attributes and polylines, SA has been applied in the T, P, X and Z axes. In the T dimension, SA has scaled the time series into 25 hourly intervals. For the P dimension, which has 2,450 SenderIPs, SA has selected the first 8-bits of the network address, and shrunk the remaining 24-bits in the sending virus pattern $pat_{(x, y, z)}$ if the attribute occurred fewer than three times in the dataset. allowing similar network

addresses to be grouped together. For example, if sender $p = 123.123.123.123$ has sent two virus patterns $pat_{(x, y, z)}$, $\{d \in D \mid pat\ (x, y, z), p=123.123.123.123\}$=2, then the attribute p will be represented as $p = 123.xxx.xxx.xxx$, therefore shrinking the last 24-bits of the network address. After applying SA in the P dimension, the number of attributes in the P axis has reduced from 2,450 to 101, significantly reducing data overcrowding.

In the X dimension, SA has been applied to the pattern $pat_{(x, y, z)}$ if the data attribute occurred fewer than three times in the dataset. For example, if $\{d \in D \mid pat\ (x, y, z)\} = 2$, the pattern will be represented as $pat_{(x=SA, y, z)}$. The X dimension also has some repeating messages with similar characters in the subject, such as "DHL Tracking Number PMT2S0NQ" or "eBay: security issues <message id:5001020732>". SA has collected these repeated messages and shrunk them into the attributes "DHL.SA" and "eBay.SA", due to their similarity with the same SourceIP and Virus. After applying SA in the X dimension, the number of attributes in the X axis has reduced from 970 to 76, significantly reducing data overcrowding.

In the Z dimension, viruses that occur fewer than three times will be shrunk by SA into the attribute "VIRUS=SA". This has reduced the number of attributes in the Z axis from 4,100 to 88. The parallel coordinates for $SD_{()}$ via $RD_{()}$ with pattern $pat_{(x, y, z)}$ are shown in Figure 4.7.

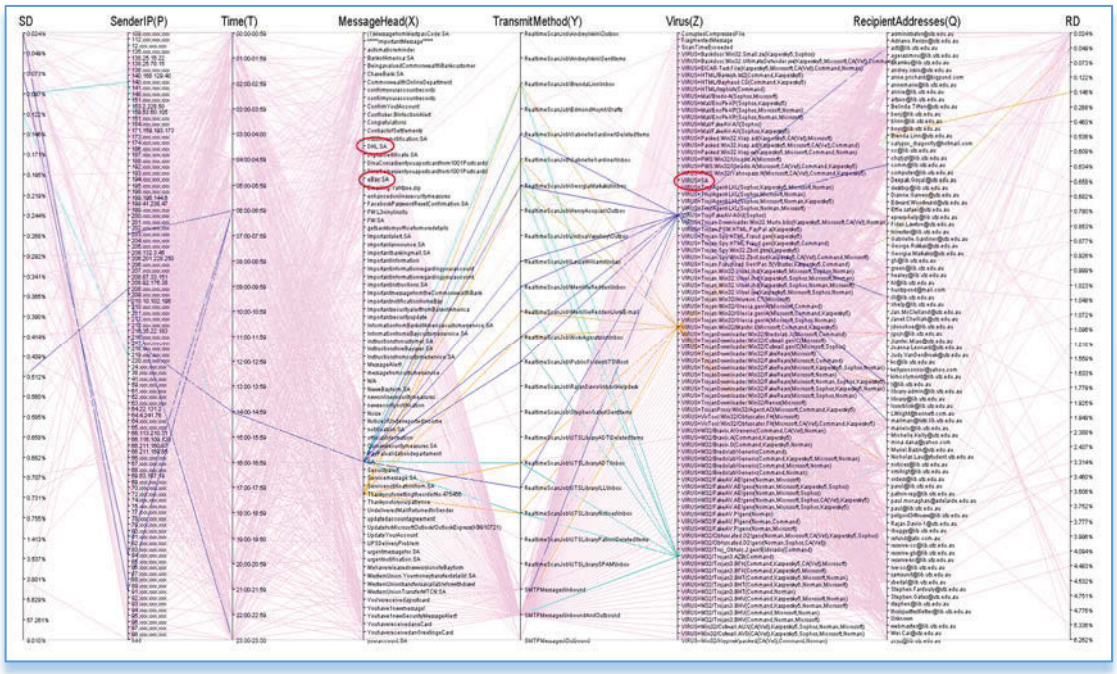

Figure 4.7 Virus pattern $pat_{(x=virus,\ y,\ z)}$ in $SD_{(\ )}$ via $RD_{(\ )}$ parallel coordinates with SA

In Figure 4.7, the overcrowding of attributes and overlapping of polylines have been dramatically reduced. The three viruses that are sent from multiple transmission methods have been highlighted, which helps explain the virus pattern structures for both the sending and receiving densities.

The highest value $SD_{(\ )}$ = 57.261% was sent from SenderIP p = *"138.25.78.15"*, with pattern $pat_{(x=\text{"Thank you for setting the order No:478456"},\ y=\text{"SMTPMessages\Inbound And Outbound"},\ z=\text{"VIRUS=W32/Trojan3BFK(...)"})}$. The $RD_{(\ )}$ values for the same pattern are between 0.05 and 6.26%, and are received by multiple RecipientAddress. This indicates that this sender sent multiple widespread attacks to the email system. The highest value $RD_{(\ )}$ = 6.262% was received by multiple RecipientAddress with multiple patterns. That means that no single victim suffered especially heavy attacks in the email dataset.

4.2.3. Clustered Transferring Pattern $pat_{(p, y, q)}$

The pattern $pat_{(p, y, q)}$, which describes the virus transferring pattern between the sender and recipient, can be explored by using Pair-Density $CD_{(x, pat(p, y, q))}$ via $PD_{(z, pat(p, y, q))}$. The relationship between the attachment and the virus for the transferring pattern $pat_{(p, y, q)}$ can be illustrated by using the clustered X dimension, which is shown in Figure 4.8.

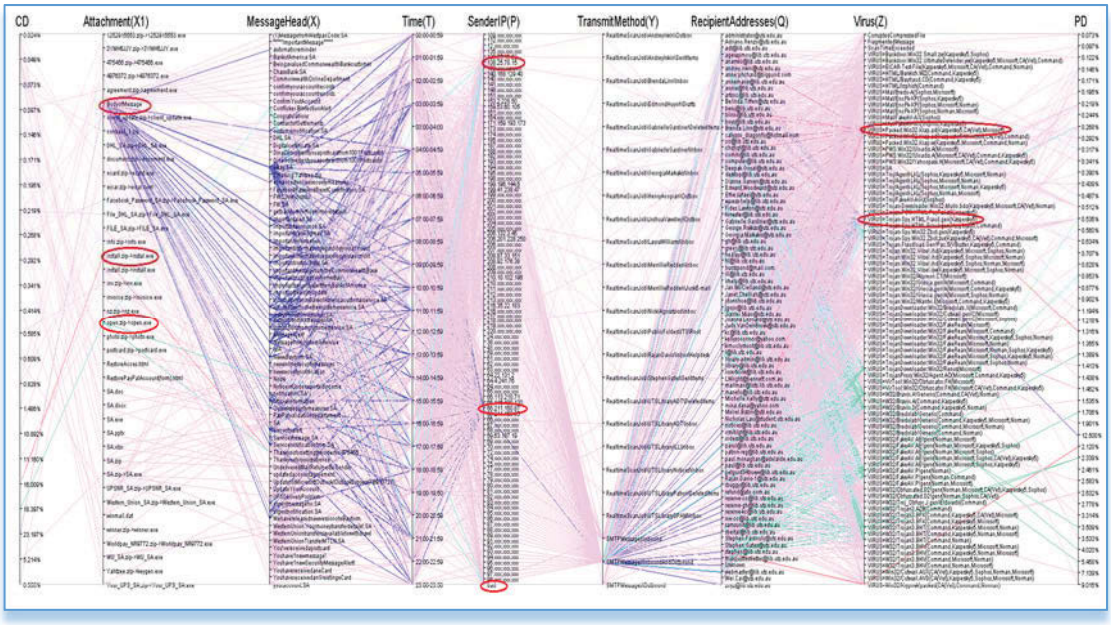

Figure 4.8 Virus transferring pattern $pat_{(p, y, q))}$ in $CD_{( )}$ via $PD_{( )}$ parallel coordinates

In Figure 4.8, dimension X1 represents the attachment with the virus, which is clustered from dimension X. The highest value of $CD_{(x1="BodyofMessage",\ pat\ (p="66.211.160.87",\ y="SMTPMessage\InboundAndOutbound",\ q))}$ = 23.197% is linked to multiple attributes in the X dimension over a long period of time. The second highest $CD_{(x1="open.exe",\ pat\ (p="138.25.78.15",\ y="SMTPMessage\InboundAndOutbound",\ q))}$ = 18.397%, and the third highest $CD_{(x1="install.exe",\ pat\ (p="138.25.78.15",\ y="SMTPMessage\InboundAndOutbound",\ q))}$ = 16.009%. These three figures indicates that these three attachments carried more than 50% of the virus attacks. In the Z dimension, the highest value of $PD_{(z="VIRUS=Trojan-Spy.HTML.Fraud.gen(Kaspersky5)",\ pat\ (p="66.211.160.87",\ y="SMTPMessage\InboundAndOutbound",\ q))}$ = 12.500%, and the second highest value

of $PD_{(z=\text{“VIRUS=HTML/Irsphish(Command)”},\ pat\ (p=\text{“199.196.144.6”},\ y=\text{“SMTPMessage\InboundAndOutbound”},\ q))}$ = 9.016%.

The highest value of $CD_{(BodyofMessage)}$ = 23.197% is larger than the highest value of $PD_{(VIRUS=Trojan\text{-}Spy....)}$ = 12.500%. Because both have the same pattern $pat_{(p=\text{“66.211.160.78”},\ y=\text{“SMTPMessage\InboundAndOutbound”},\ q)}$, this means that the attachment contained different kinds of viruses sent from the same attacker using the same transmission method. The attachments are linked to multiple attributes in the X dimension, indicating that the same attachment was hidden inside multiple messages. This makes it difficult for the email system administrator to prevent the virus reaching the victim, since the variations of the email subject name cannot be easily filtered out.

4.3. Case Three: Visual Detect DDoS Attacks in ISCX2012 Dataset

ISCX2012 dataset contains 1,511,636 network data incidents in 12 subsets, and has 20 data dimensions including SourceIP, SourcePort, DestinationIP, DestinationPort, StartData, StopData, TrafficTag, and ConnectMethod. An example of an ISCX2012 network data incident can be described as:

“On 13/June/2010 10:51:45-10:51:52 PM, SourceIP: 142.167.88.44 from SourcePort: 61506 sent an attack to DestinationIP: 192.168.5.122 on DestinationPort: 80, transferred by ProtocolName: tcp_ip, connected by AppName: HTTPWeb, TotalSourceBytes: 6372, TotalSourcePackets: 9, TotalDestinationBytes: 1014, TotalDestinationPackets: 9, SourcePayloadAsBase64: R0VUIC9PdkNnaS9s……., SourcePayloadUTF: Get/OvCgi/Toolbar.exe?XOM……, DestinationPayloadAsBase64: SFRUUC8×LjEgNDA……., DestinationPayloadAsUTF: HTTP/1.1404NotFound....., Direction: R2L, SourceTCPFlagDescription: F.S.P.A, DestinationTCPFlagDescription: F.S.P.A”

I have analyzed six subsets dated Jun12-15, containing 906,782 data incidents. Table 4.5 shows the 5Ws classifications for the ISCX2012 dataset.

Table 4.5. 5Ws dimension for ISCX2012 network dataset

5Ws Dimension	Dimensions in dataset	Details
When (T)	Year	2010
	Month	June
	Day of Month	12 – 17
	Time	0:00 – 24:00
Where (P)	SourceIP	1,948
	SourcePort	61,234
	SourceTCPFlag	25
What (X)	Network traffic	906,782
	NormalTraffic	775,868
	AttackTraffic	61,504
	UnknownTraffic	69,410
	TotalSourcePackets	0 – 996
	TotalDestinationPackets	0 – 996
How (Y)	ConnectMethod	105
	SourcePayloadAsUTF	
	DestinationPayloadAsUTF	
Why (Z)	TransferProtocol	6
	Direction	4
Who (Q)	DestinationIP	24,374
	DestinationPort	18,632
	DestinationTCPFlag	27

In Table 4.5, SourceTCPFlag and DestinationTCPFlag include the control bits NS, CWR, ECE, URG, ACK, PSH, RST, SYN and FIN in the TCP communication. SourcePayloadAsUTF means that the requesting source string is in UTF-8 unicode format, and DestinationPayloadAsUTF represents the responding connection. TransferProtocol includes tcp_ip, udp_ip, icmp_ip, igmp, ip and ipv6icmp. Direction means that the network traffic flow direction includes L2R (local host to remote host), L2L (local host to local host), R2L (remote host to local host) and R2R (remote host to remote host).

4.3.1. 5Ws Pattern for ISCX2012 Dataset

The primary 5Ws pattern that I have chosen for the ISCX2012 network traffic is shown in Table 4.6. The T axis has six subsets based on the time serials above. The X dimension represents the network traffic tag, which is labelled as either “Normal”, “Attack” or “Unknown” traffics. The Y dimension contains 105 attributes that describe the request apps, such as “HTTPWeb”, “DNS”, or “POP”. The Z axis has six transfer protocols.

Table 4.6. 5Ws pattern for ISCX2012 network traffic

5Ws Dimension	Dimensions in dataset	Details
When (T)	Date of subset	Jun12-15c
Where (P)	SourceIP	1,948
What (X)	TrafficTag	3
How (Y)	ConnectMethod	105
Why (Z)	TransferProtocol	6
Who (Q)	DestinationIP	24,374

Figure 4.9 illustrates the 5Ws pattern for the ISCX2012 dataset, which contains 64,393 different patterns. There are many overcrowded polylines and attributes, especially in the P, Y and Q axes. This makes it hard to detect particular attack patterns.

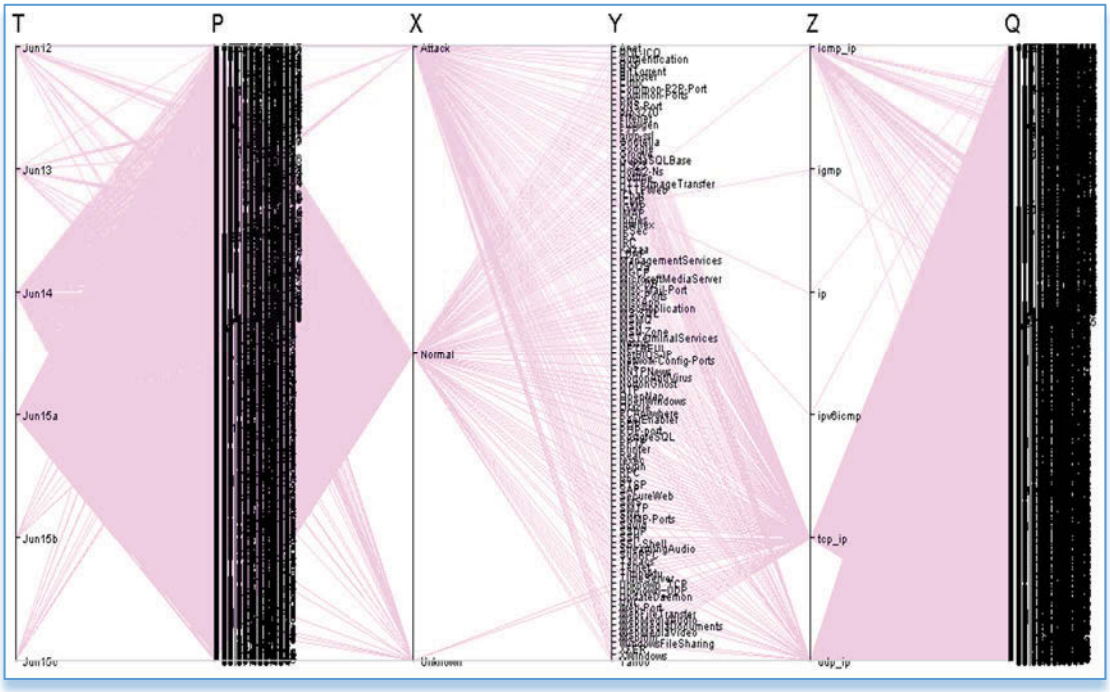

Figure 4.9 ISCX2012 network pattern in 5Ws parallel coordinates

4.3.2. Network Pattern $pat_{(x, y, z)}$ in $SD_{(\ )}$ via $RD_{(\ )}$ Parallel Coordinates

SA has been implemented on the P and Q axes in order to reduce the overcrowding of attributes. I defined the SA for each subset when $SD_{(p)} \leq 0.02\%$ or $RD_{(q)} \leq 0.02\%$as "00x.xxx.xxx.xxx", "0xx.xxx.xxx.xxx", "1xx.xxx.xxx.xxx", and "2xx.xxx.xxx.xxx". In other words, p or q = "1xx.xxx.xxx.xxx" including all IPs in the range of {100-255. 1-255. 1-255. 1-255} that also satisfy the conditions $SD_{(p)} \leq 0.02\%$ or $RD_{(q)} \leq 0.02\%$. For example, if two attributes in the P axis have $SD_{(p=111.111.111.111)} \leq 0.02\%$ and $SD_{(p=123.123.123.123)} \leq 0.02\%$. then these two attributes will be shrunk into one SA attribute as $SD_{(p=1xx.xxx.xxx.xxx)} \leq 0.02\%$. Figure 4.10 shows the network pattern $pat_{(x, y, z)}$ in the Pair-Density $SD_{(\ )}$ via $RD_{(\ )}$ parallel coordinates after SA has been applied to the P and Q axes.

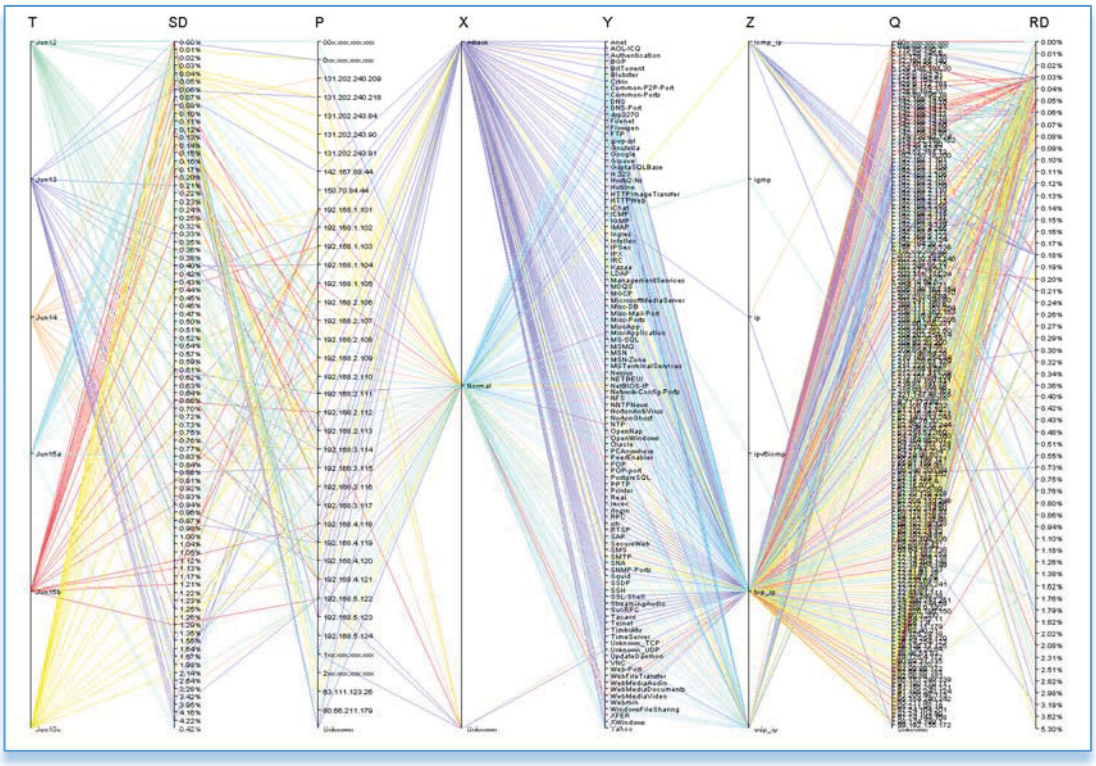

Figure 4.10 Network pattern $pat_{(x, y, z)}$ in $SD_{( )}$ via $RD_{( )}$ parallel coordinates

In Figure 4.10, after the implementation of SA, the number of attributes in the P axis has fallen from 1,948 to 51 items, while the number of attributes in the Q axis has fallen from 24,372 attributes to 200 attributes. The number of cluttered polylines and overcrowded attributes have been significantly reduced from 64,393 to 8,030. Cluttering has therefore been reduced by over 85% without any significant loss of information – which is a significant achievement.

In Figure 4.10, the majority of traffics are tagged as “normal” by the “tcp_ip” protocol. The highest value of $SD_{(p=192.168.2.107)}$ = 6.42% containing the “normal” traffic came from the “tcp_ip” and “udp_ip” protocols. Similarly, the highest value of $RD_{(q=192.168.5.122)}$ = 5.30% has both “normal” and “attack” traffics.

4.3.3. Attack Pattern $pat_{(x="attack", y, z)}$ between Attacker and Victim

To explore the network attack pattern, I have applied SA on the *P*, *Y*, *Z* and *Q* axes for attribute x = “normal”. This has narrowed down the pattern $pat_{(x="attack", y, z)}$ in visual analytics. Figure 4.11 shows the attack pattern $pat_{(x="attack", y, z)}$ in the Pair-Density $SD_{()}$ via $RD_{()}$ parallel coordinates.

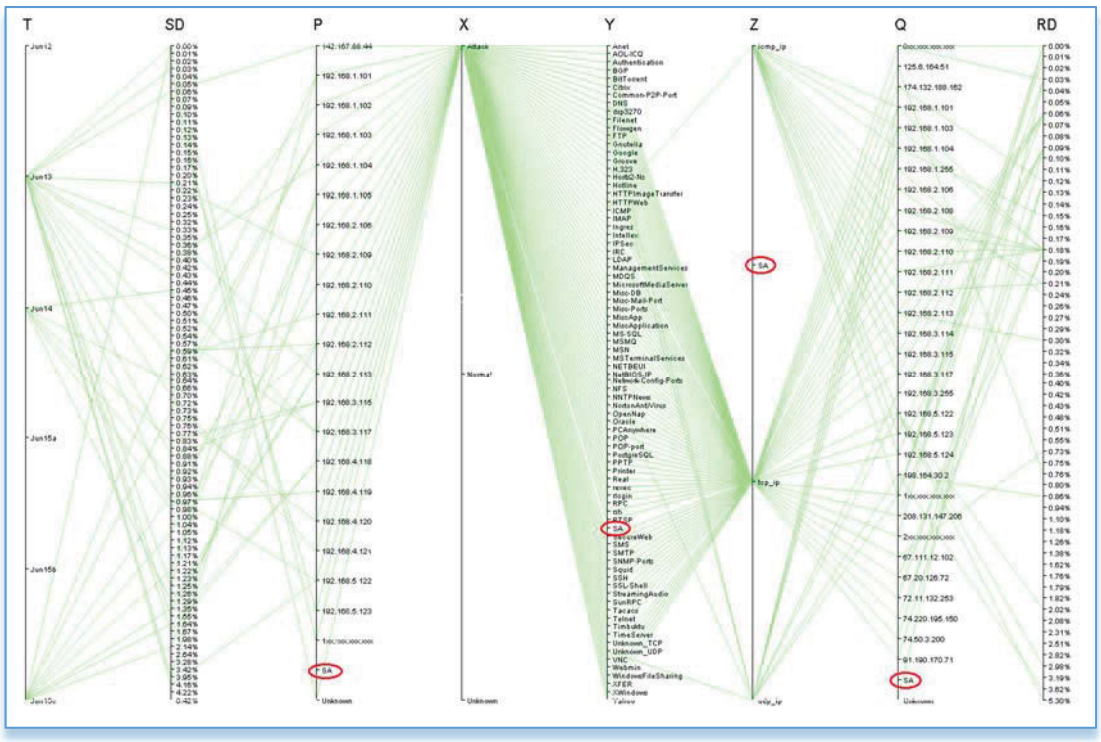

Figure 4.11 Attack pattern $pat_{(x="attack", y, z)}$ between attackers and victims

In Figure 4.11, there are 22 hackers that have sent multiple attacks to 32 victims by using three protocols “icmp_ip”, “tcp_ip” and “udp_ip” in the transaction. A total of 80 connection methods have been used for these network attacks, which are widespread in the network traffics. Among these attacks, the most dangerous network attack is the DDoS (Distributed Denial of Service) attack, and this needs to be addressed in visual analytics.

The main characteristic of DDoS attacks is that the victim receives high density attacks from multiple attackers using different connection methods, which thereby makes it difficult to prevent an attack due to the complexity of the attackers and the connection methods (Zhang, J., Huang, M.L 2013). Figure 4.12 shows the DDoS attacks in Pair-Density $SD_{()}$ via $RD_{()}$ parallel coordinates.

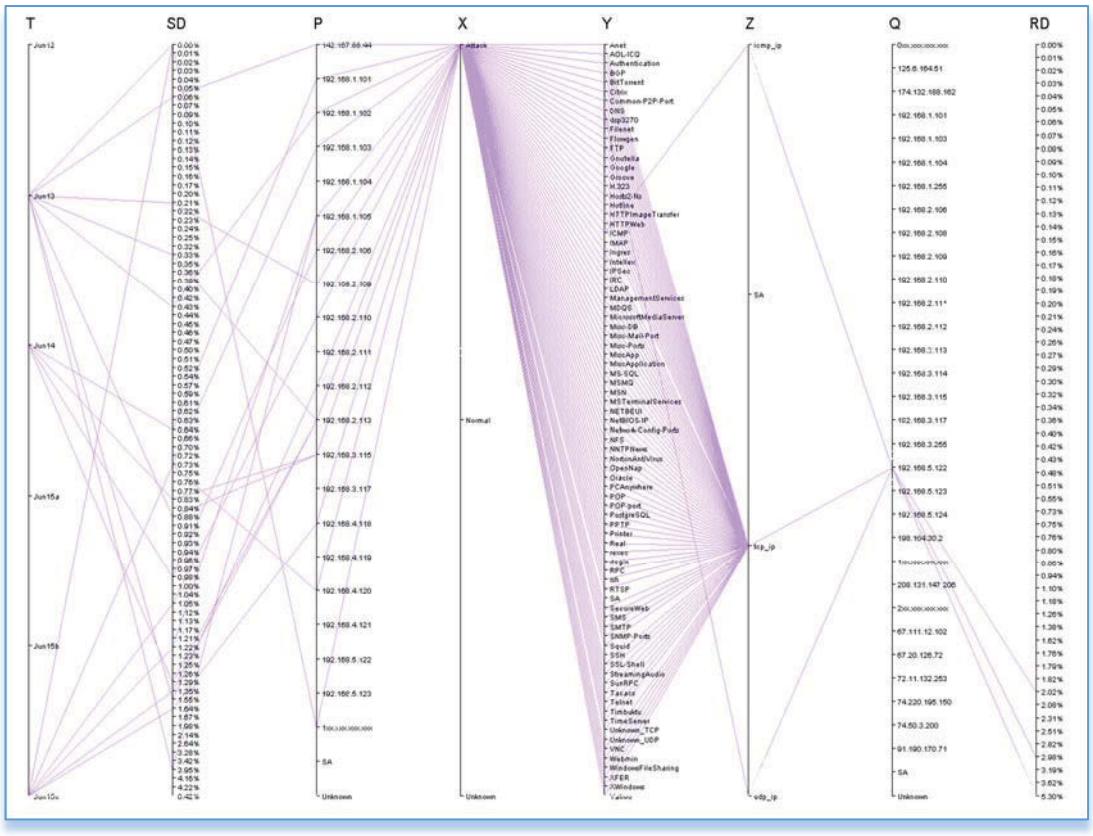

Figure 4.12 DDoS attack pattern in $SD_{()}$ via $RD_{()}$ parallel coordinates

In Figure 4.12, the victim (*q=192.168.5.122*) suffered a high density of DDoS attacks, reaching up to 10.30%. These DDoS attacks were sent from 14 different attackers using 78 different connection methods, such as FTP, HTTP, LDAP, Oracle, POP, SMTP, Telnet and WindowsFileSharing. To prevent the system from collapsing under the DDoS attack, the system administrator has to closely monitor these service connections. Fortunately, the DDoS attack patterns can be quickly and easily recognized in the $SD_{()}$

via $RD(\ )$ parallel coordinates, which are clearly visualized in order to enable the administrator to manage and control the key attributes, and to reduce system damage.

4.4. Reduction of Data Overcrowding

In above three case studies, I have deployed 5Ws algorithm and Pair-Density model to analyze the relationships between the particular pair dimensions, for uncovering the worst delay patterns in case one – US 2008 flights dataset; for finding the most dangerous spreading patterns in case two – UTS 2009 Library email dataset; for detecting the DDoS attacking patterns in case three – ISCX2012 dataset.

The Pair-Density model not only illustrates the visual pattern for multidimensional data, but also significantly reduces data cluttering by using SA in the Pair-Density parallel coordinates. Table 4.7 shows the details of the data pattern between the three different case studies.

Table 4.7. Data pattern for three cases

	Case one (2008 Flights)	**Case two** (2009 Email)	**Case three** (2012 Network)
Original	1,048,575	585,300	906,782
5Ws	157,524	4,104	64,393
Pair-Density(withSA)	132	3,765	8,030

In Table 4.7, the three cases have a total of 2,540,657 original data patterns. This has been reduced to a total of 226,021 5Ws data patterns, and further reduced to 11,927 patterns in the 5Ws Pair-Density parallel coordinates. These reductions in data patterns have not resulted in the loss of any information, since other patterns have merely been

shrunk into the SA attributes, and are represented in the Pair-Density axes. Figure 4.13 and Figure 4.14 display the reduction in data cluttering and overcrowding of polylines.

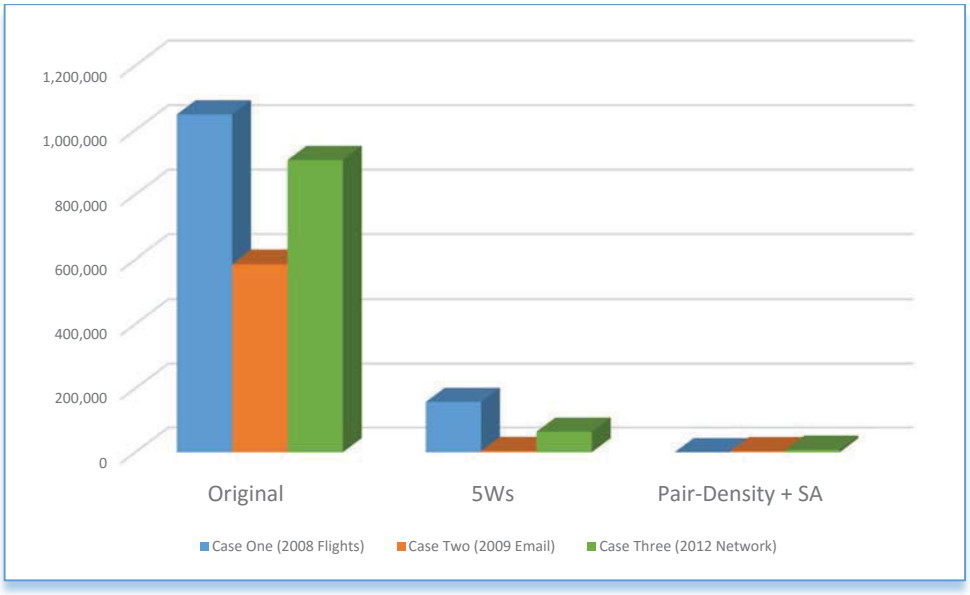

Figure 4.13 Reduction of data cluttering (a)

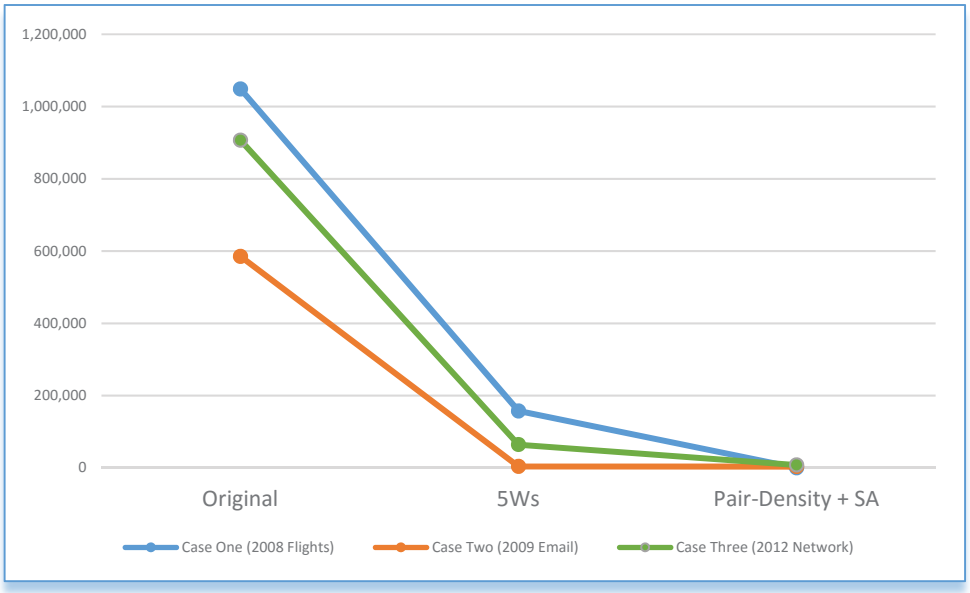

Figure 4.14 Reduction of data cluttering (b)

In Figure 4.13 and Figure 4.14, the 5Ws pattern has reduced data overcrowding in my three case studies by more than 80%. Furthermore, the 5Ws Pair-Density with SA has reduced data cluttering by more than 95% without the loss of any information. The

model that I propose has significantly reduced the data processing time for Big Data analysis, and significantly reduced the data overcrowding for Big Data visualization.

Chapter 5: Conclusions and Future Work

5.1. Conclusions

Big Data is composed of text, images, video, audio, mobile or other forms of data collected from multiple datasets, and is rapidly growing in both size and complexity. This has created a huge volume of multidimensional data within a very short time period, and therefore it is too big, too complex and moves too fast for us to analyze using traditional methods. Big Data characteristics include Volume, Variety and Velocity (3V) when it came up in the early ages. Later approaches have since added Value and Veracity into Big Data characters, therefore amending the 3Vs model into a 5Vs model.

This thesis presented a novel approach to Big Data visual analytics. The 5Ws dimensions has been established to classify Big Data patterns based on the data behaviours ontologies. Big Data behaviours are as a set of concepts and categories that descripts Big Data's acts for; **When** did the data occur, **Where** did the data come from, **What** did the data contain, **How** was the data transferred, **Why** did the data occur, and **Who** received the data. It can also be descripted as **5W1H**, or **5Ws** which indicates the same concepts.

The 5Ws dimensions are suitable for multiple datasets across any form of data, and has been visualized using the 5Ws parallel coordinates. This has narrowed down Big Data patterns to give us a better understanding of data behaviour and its relationship. To the best of my knowledge, no previous work had addressed Big Data visual analytics by using 5Ws dimensions.

Pair-Densities have also been created to measure different data patterns, and illustrated in the Pair-Density parallel coordinates in order to provide accurate

measurement and comparison for Big Data visualization. Pair-Density measures and compares data patterns in parallel coordinates irrespective of their data forms. Pair-Density not only values all Big Data patterns, but also creates two additional non-dimensional axes to compare data patterns in parallel coordinates. Data cluttering and overcrowding is dramatically reduced using Pair-Density parallel coordinates. To the best of my knowledge, no previous work had created two more axes by using Pair-Density in parallel coordinates to measure multidimensional data patterns.

Shrunk Attributes (SA) have been introduced which not only collect the value of elements not displayed in parallel coordinates, but also dramatically reduce data cluttering and overcrowding in Pair-Density parallel coordinates. Combined with Shrunk Attributes (SA) and dimension clustering, the 5Ws dimensions and Pair-Density parallel coordinates model has significantly improved the accuracy and accessibility of Big Data analysis and visualization, since noise data, data cluttering and overcrowded attributes have all been minimised through our process.

More than 2.5 million data incidents have been implemented from three different case studies, each containing different forms of data incidents, such as flight data incidents, spam email data incidents and network traffic data incidents. The flight delay patterns, spam email virus attack patterns and DDoS network attack patterns have all been explored and visualized using many graphs.

Through these three case studies, it is clear that our new approach can be used for many different datasets and for any form of data. This is especially the case for unstructured data, since Pair-Density parallel coordinates are able to measure non-numerical dimensions, enabling comparisons between any pair of dimensions.

Big Data visualizations normally use optimization techniques to explore data nodes, meaning that some data will inevitably be lost in the graph. Using our new approach, no information is lost in the 5Ws dimensions and Pair-Density parallel coordinates algorithm, since all minor attributes are clustered in the parallel axes through SA and dimension clustering. This has been verified through the three case studies. Pair-Densities not only measure the particular data patterns, but also value the noise data and SA patterns, irrespective of whether the data has been explored or not. This achievement is very significant due to the complexity of analysing huge volumes of Big Data.

Big Data visualization faced huge data cluttering and overcrowding issues, which reduced the ability to demonstrate clear views of particular patterns to feed business, organizational and government needs. In this new approach, data cluttering and overcrowding have been significantly reduced by more than 80% in the three case studies conducted, without any loss of information.

This thesis contains the research approach and implementation results obtained by the author during his Ph.D period. The majority of methods and results have been published in **Seventeen** research papers in journals and conference proceeding by May 2016.

5.2. Future Works

For future work, there are two areas that I will continue to study and practice. Firstly, I will further research the combination of treemaps with the 5Ws dimensions and Big Data behaviour patterns, and research the development of 5D software to illustrate Big Data visual movements. Ultimately, this may make it possible to create motion clips for Big Data research and industry engagement.

Secondly, I will further study Pair-Densities for hospital medical treatment research, since there are many pair-relationship patterns within hospital systems. Examples of medical pair-relationship patterns include the relationship between doctor and patient; disease and antibiotic; and patient and antibiotic. This will bring Big Data techniques into the hospital treatment system, and may provide huge benefits for both patient and hospital by reducing the hospital's cost of antibiotics management and patient treatment control, whilst also providing more efficient treatment to a greater number of patients. This will benefit the patients and our societies.

Bibliography

Abousalh-Neto, N.A. Kazgan, S (2012), “Big Data Exploration through Visual Analytics”, In Proceeding of IEEE Symposium on Visual Analytics Science and Technology, pp. 285-286

Afzal, S., Maciejewski, R., Jang, Y., Elmqvist, N., Ebert, D.S (2012), “Spatial Text Visualization using Automatic Typographic Maps”, IEEE Transaction on Visualization and Computer Graphics, Vol. 18, No. 12, pp. 2556-2564

Alam, K,T., Hossain, S.M.M., Arefin, M. S. (2016), “Feveloping Framework for Analyzing Social Networks to Identify Human Behaviours”, In Proceeding of 2nd International Conference on Electrical , Computer & Telecommunication Engineering (ICECTE), pp. 1-4, Dec 2016

Artero, A., de Oliveira, M., Levkowitz, H (2004), “Uncovering Clusters in Crowded Parallel Coordinates Visualization”, In Proceeding of IEEE Symposium of Information Visualization, pp. 81-88

ASA, (2009), “Statistical Computing Statistical Graphics”, Data expo 09, [online] posted on 2009, http://stat-computing.org/dataexpo/2009/the-data.html, accessed on Feb 2013

Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Garcia-Rodriguez, J., Orts-Excolano, S. (2015), “Self-Organizing Activity Description Map to Represent and Classify Human Behaviour”, In Proceeding of 2015 International Joint Conference on Neural Network (IJCNN), pp. 1-7, July 2015

Berg, T.L., Sorokin, A., Wang, G., Forsyth, D.A., Hoiem, D., Endres, I., Farhadi, A (2010), “It’s All About the Data”, Proceeding of the IEEE, Vol. 98, No. 8, pp.1434-1452, Aug 2010

Chen, J.X., and Wang, S. (2001), “Data visualization: parallel coordinates and dimension reduction”, Computing in Science and Engineering, Vol. 3, No. 5, pp 110-113

Chen, Y., and Shen, C., (2017), “Performance Analysis of Smartphone-Sensor Beharior for Human Activity Recognition”, IEEE Access, Vol. 5, pp. 3095-3110

Cheng, D., Schretlen, P., Kronenfeild, N., Bozowsky, N., Wright, W (2013), “Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis”, In Proceeding of IEEE International Conference on Big Data (IEEE BigData 2013), pp. 2-4

Chi, M.T., Lin, S.S., Chen, S.Y., Lin, C.H., and Lee, T.Y. (2015), “Morphable Word Clouds for Time-Varying Text Data Visualization”, IEEE Transactions on Visualization and Computer Graphics, Vol. 21, No. 12, pp. 1415-1426

Chung, K.L., Zhuo, W (2008), “Graph-Based Visual Analytic Tools for Parallel Coordinates”, ISVC 2008, Lecture Notes in Computer Science, Vol. 5359, pp. 990-999

Claessen, J.H., van Wijk, J.J (2011), “Flexible linked axes for multivariate data visualization”, IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No. 12, pp. 2310-2316

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M.X., Qu, H (2010), “Context Preserving, Dynamic Word Cloud Visualization”, IEEE Computer Graphics and Applications, Vol. 30, No. 6, pp.42-53

Dang, T.N., Wilkinson, L (2010), “A Stacking graphic elements to avoid over-plotting”, IEEE Transactions on Visualization and Computer Graphics, Vol. 16 No. 6, pp. 1044-1052

Dasgupta, A., Kosara, R (2010), “Pargnostics: Screen-Space Metrics for Parallel Coordinates”, IEEE Transactions on Visualization and Computer Graphics, Vol. 16, No. 6, pp. 1017-1026

Demchenko, Y., Grosso, P., Laat, C.D., Membrey, P (2013), “Addressing Big Data Issues in Scientific Data Infrastructure”, In Proceeding of International Conference on Collaboration Technologies and Systems (CTS), pp. 48-55

Dykes, J., Brunsdon, C (2007), “Geographically Weighted Visualization: Interactive Graphics for Scale-Varying Exploratory Analysis”, IEEE Transactions on Visualization and Computer Graphics, Vol. 13, No. 6, pp. 1161-1168

Ellis, G., Dix, A (2006), “Enabling Automatic Clutter Reduction in Parallel Coordinates Plots”, IEEE Transactions on Visualization and Computer Graphics, Vol. 12, No 5, pp 717-724

Geng, Z., Peng, Z.M., Laramee, R.S., Walker, R., Roberts, J.C (2011), “Angular Histograms: Frequency-Based Visualization for Large High Dimensional Data”, IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No 12, pp 2572-2580

Hadlak, S., Schulz, H.J., Schumann, H (2011), “In Situ Exploration of Large Dynamic Networks”, IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No. 12, pp. 2334- 2343

Hadoop (2014), [online] http://hadoop.apache.org, accessed on Aug 2014

Heinrich, J., Seifert, R., Burch, M., Weiskopf, D (2011), “BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data”, ISVC 2011, Lecture Notes in Computer Science, Vol. 6938, pp. 641-652

Hill, K (2012), “How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did”, Forbes, [online] posted on Feb 16, 2012, http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/, accessed on Aug 2014

Huh, M.H., Park, D.Y (2008), “Enhancing parallel coordinate plots”, Journal of the Korean Statistical Society, Vol. 37, No. 2, pp. 129 -133

Inselberg, A., Dimnsdale, B (1990), “Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry”, In Proceeding of First IEEE Conference on Visualization, pp. 361-378

Jeon, S., Khosiawan, Y., Hong, B (2013), “Making a Graph Database from Unstructured Text”, In Proceeding of 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 981-988

Klein, D., Tran-Gia, P., Martmann, M (2013), “Big Data”, Informatik-Spektrum, Vol. 36, issue 3, pp. 319-323

Kraska, T (2013), “Finding the Needle in the Big Data Systems Haystack”, Internet Computing, IEEE, Vol. 17, No. 1, pp. 84-86

Liang, J., Nguyen, Q.V., Simoff, S., Huang, M.L (2015), “Divide and Conquer treemaps: Visualizing large trees with various shapes”, Journal of Visual Languages and Computing, Vol. 31, pp. 104-127

Liu, X., Hu, Y., North, S., Shen, H.W (2013), “CompactMap: A Mental Map Preserving Visual Interface for Streaming Text Data”, In Proceeding of IEEE International Conference on Big Data (IEEE BigData 2013), pp. 48-55

Lomotey, R.K., Deters, R (2013), “Topics and Terms Mining in Unstructured Data Stores”, In Proceeding of 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 854-861

Lorenzetti, L (2014), “World Cup scores big on Twitter and Facebook”, Fortune, [online] posted on July 14, 2014, http://fortune.com/2014/07/14/world-cup-scores-big-on-twitter-and-facebook/, accessed on July 2014

Lu, L.F., Huang, M.L., Huang, T.H (2012), “A New Axes Re-ordering Method in Parallel Coordinates Visualization”, In Proceeding of 11th International Conference on Machine Learning and Applications, pp 252-257

Ma, C.L., Shang, X.F., Yuan, Y.B (2012), “A Three-Dimensional Display for Big Data Sets”, In Proceeding of International Conference on Machine Learning and Cybernetics, pp. 1541-1545

Meghdadi A.H., Irani, P (2013), “Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization”, IEEE Transaction on Visualization and computer Graphics, Vol. 19, No. 12, pp. 2119-2128

Menon, A (2012), “Big Data @ Facebook”, In Proceeding of Workshop on Management of Big Data System (MBDS ‘12), pp. 31-32

Narayan, S., Bailey, S., Daga, A (2012), “Hadoop Acceleration in a OpenFlow-Based Cluster”, In Proceeding of High Performance Computing, Networking, Storage and Analysis, SC Companion 2012, pp. 535-538

Nohno, K., Wu, H.Y., Watanabe, K., Takahashi, S., Fujishiro, I (2014), "Spectral-Based Contractible Parallel Coordinates", In Proceeding of 18th IEEE international Conference on Information Visualization, pp. 7-12

Norton, S (2015), "Starwood Hotels Using Big Data to Boost Revenue", The Wall Street Journal, [online] posted on Feb 10, 2015, http://blogs.wsj.com/cio/2015/02/10/starwood-hotels-using-big-data-to-boost-revenue/, accessed on Mar 2015

Novotny, M., Hauser, H (2006), "Outlier-preserving Focus+Context Visualization in Parallel Coordinates", IEEE Transactions on Visualization and Computer Graphics, Vol. 12, No 5, pp 893-900

Oxford Dictionary, [online] https://en.oxforddictionaries.com/definition/behaviour/, accessed on Oct 2016

Pingdom (2013), "Internet 2012 in numbers", Pingdom, [online] posted on Jan 16, 2013, http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/, accessed on Oct 2013

Rocha, A., Carvalho, T., Jelinek, H.F., Goldenstein, S., Wainer, J (2012), "Points of Interest and Visual Dictionaries for Automatic Retinal Lesion Detection", IEEE Transactions on Biomedical Engineering, Vol. 59, No. 8, pp. 2244-2253

Rosling, H (2009), "200 years that changed the world", Gapminder, [online] posted on May 8, 2009, http://www.gapminder.org/video/200-years-changed-the-world/, accessed on Feb 2011

Sanyal, J., Zhang, S., Dyer, J., Mercer, A., Amburn, P., Moorhead, R.J (2010), "Noodles: A Tool for Visualization of Numerical Weather Model Ensemble Uncertainty", IEEE transactions on Visualization and Computer Graphics, Vol. 16, No. 6, pp. 1421-1430

Seol, W.S., Jeong, H.W., Lee, B., Youn, H.Y (2013), "Reduction of Association Rules for Big Data Sets in Socially-Aware Computing", In Proceeding of 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 949-956

Sereda, P., Bartroli, A.V, Serlie, I.W.O., Gerritsen, F.A (2006), "Visualization of Boundaries in Volumetric Data Sets Using LH Histograms", IEEE transactions on Visualization and Computer Graphics, Vol. 12, No. 2, pp. 208-218

Shi, L., Liao, Q., Sun, X., Chen, Y., Lin, C (2013), "Scalable Network Traffic Visualization using Compressed Graphs", In Proceeding of IEEE International Conference on Big Data (IEEE BigData 2013), pp. 606-612

Shiravi, A., Shiravi, H., Tavallaee, M., and Ghorbani, AA (2012), "Toward developing a systematic approach to generate benchmark datasets for intrusion detection", Computers & Security, Vol. 31, No. 3, pp. 357-374

Stamford (2011), "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data", Gartner, [online] posted on June 27, 2011, http://www.gartner.com/newsroom/id/1731916/, accessed on July 2013

Torrione, P.A., Morton, K.D., Sakaguchi, R., Collins, L.M. (2014), "Histograms of Oriented Gradients for Landmine Detection in Ground-Penetrating Radar Data", IEEE Transactions on Geoscience and Remote Sensing, Vol. 52, No. 3, pp.1539-1550

Vliegen, R., van Vijk, J.J., Van der Linden, E.J. (2006), "Visualizing Business Data with Generalized Treemaps", IEEE Transactions on Visualization and Computer Graphics, Vol. 12, No. 5, pp. 789-796

Wang, Y.S., Wang, C., Lee, T.Y., Ma, K.L (2011), "Feature-Preserving Volume Data Reduction and Focus+Context Visualization", IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No. 2, pp. 171-181

Wang, Z., Xiao, W., Ge, B., Xu, H (2013), "ADraw: A novel social network visualization tool with attribute-based layout and coloring". In Proceeding of IEEE International Conference on Big Data (IEEE BigData 2013), pp. 25-32

Wang, Z., Zhou, J., Chen, W., Chen, C., Liao, J., Maciejewski, R (2013), "A Novel Visual analytics Approach for Clustering Large-Scale Social Data", In Proceeding of IEEE International Conference on Big Data (IEEE BigData 2013), pp. 79-86

Yuan, X., Guo, P., Xiao, H., Zhou, H., Qu, H (2009), “Scattering Points in Parallel Coordinates”, IEEE Transactions on Visualization and Computer Graphics, Vol. 15, No 6, pp 1001-1008

Zhang, J., Huang, K., Cottman-Fields, M., Truskinger, A., Roe, P., Duan, S., Dong, X., Towsey, M., Wimmer, J (2013), “Managing and Analysing Big Audio Data for Environmental Monitoring”, In Proceeding of 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 997-1004

Zhang, J., Huang, M.L (2013), “Visual Analytics Model for Intrusion Detection in Flood Attack”, In Proceeding of 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom2013), pp. 277-284

Zhang, J., Huang, M.L, Hoang, D (2013), “Visual analytics for intrusion detection in spam emails”, International Journal of Grid and Utility Computing, Vol. 4, No. 2/3, pp. 178-186

Zhang, J., Huang, M.L., Wang, W.B., Lu, L.F., Meng, Z.P (2014), “Big Data Density Analytics using Parallel Coordinate Visualization”, In Proceeding of IEEE 17th International Conference on Computational Science and Engineering (CSE), pp. 1115-1120

Zhou, H., Cui, W., Qu, H., Wu, Y., Yuan, X., Zhuo, W (2009), “Splatting the lines in parallel coordinates”, Computer Graphics Forum, Vol. 28, No. 3, pp. 759-766

Zhou, H., Yuan, X., Qu, H., Cui, W., Chen, B (2008), “Visual clustering in parallel coordinates”, Computer Graphics Forum, Vol. 27, No. 3, pp. 1047-1054