A New Analytics Model for Large Scale Multidimensional Data Visualization

Jinson Zhang[1] and Mao Lin Huang[2, 1]

[1]School of Software, Faculty of Engineering & IT, University of Technology, Sydney
Sydney, Australia
[2]School of Computer Software, Tianjin University, China

Abstract. With the rise of Big Data, the challenge for modern multidimensional data analysis and visualization is how it grows very quickly in size and complexity. In this paper, we first present a classification method called the *5Ws Dimensions* which classifies multidimensional data into the 5Ws definitions. The 5Ws Dimensions can be applied to multiple datasets such as text datasets, audio datasets and video datasets. Second, we establish a *Pair-Density* model to analyze the data patterns to compare the multidimensional data on the 5Ws patterns. Third, we created two additional parallel axes by using pair-density for visualization. The attributes has been shrunk to reduce data over-crowding in pair-density parallel coordinates. This has achieved more than 80% clutter reduction without the loss of information. The experiment shows that our model can be efficiently used for Big Data analysis and visualization.

Keywords: Multidimensional data; Big Data; 5Ws dimension; Parallel coordinate; pair-density; shrunk attribute; Big Data visualization

1 INTRODUCTION

Big Data is considered to be structured or unstructured data that contains texts, images, audios, videos and other forms of data collected from multiple datasets, which grows rapidly in size and complexity. Big Data comes from everywhere in our life, and so is too big, too complex and moves too fast for us to analyze using traditional methods. For example, posting statuses or pictures on Facebook; uploading and watching videos on YouTube; sending and receiving messages through smart phones; broadcasting viruses over the Internet – all those activities collected by different datasets count as Big Data.

Based on Gartner's 3Vs definition [1], Big Data has three main characteristics: Volume, Velocity and Variety. The volume represents how datasets are extremely large and easily reach terabytes of information. The velocity describes how fast datasets are being produced. The variety illustrates the complexity of the datasets, including both structure and unstructured data which contains thousands of different attributes in multiple dimensions. Our approach establishes the analytic model for large scale multidimensional data.

Multidimensional data normally contains a large amount of noise data in different dimensions. Most current approaches try using different techniques to detach those noise data, including data reduction, data integration and data clustering [5] [7]. Data reduction shrinks the data size to separate the noise data; data integration merges multiple data dimensions into coherent data attributes; and data clustering classifies the data into different groups which eliminates the noise data.

Data clustering plays a main role in multidimensional data analysis, which classifies the data dimensions into different groups, such as social media data clustering [8], airline flight data clustering [9], and petrol data clustering [10]. The cluster methods vary depending on the data structure, such as k-means cluster method [11], hierarchical cluster method [12] and density cluster method [13].

In this paper, we have further developed our previous works [6] [20] to classify the multidimensional data into the 5Ws dimensions based on their data behaviours, and then introduce 5Ws patterns crossing multiple datasets. Second, we establish pair-density patterns for analyzing the multidimensional data; four pair-density patterns are introduced to measure the different topics and patterns. Third, we created two additional parallel axes by using pair-density patterns in parallel coordinate visualization.

The paper is organized as follows; Section 2 illustrates the 5Ws dimensions and its patterns. Section 3 demonstrates the Pair-Density model. Section 4 shows the results of implementation. Section 5 describes related works, and Section 6 summarises our achievements and future works.

2 5Ws DIMENSIONS

Multidimensional data contains texts, images, audios, videos and other forms of data, which occur every day in our lives. These include Facebook images, Twitter comments, YouTube videos or email contents. These multidimensional datasets grow very fast in size and complexity, which makes them hard to analyze using traditional database tools. Here, we analyze these data attributes and classify its behaviours into the 5Ws dimensions.

2.1 Multidimensional Data and Attributes

Assume that the first data incident, known as a data node, contains attributes

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

where *j* indicates the *jth* dimension, an attribute d_{1j} illustrates the *1st* data incident of the *jth* dimension. Therefore, the whole dataset can be illustrated as in (1) where $j=1,2,3,\dots m$ indicates the number of dimensions and $i=1,2,3,\dots n$ indicates the number of incidents. The total number of attributes $n\times m$ in the dataset can reach millions, even billions, in size.

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

For example, during the 2014 FIFA World Cup Final between Germany and Argentina, there were 280 million Facebook interactions including posts, comments and likes across 88 million Facebook users [2]. Assume those interactions contained 5 dimensions, the total attributes in the entire dataset was 280 × 5 = 1.4 billion.

2.2 5Ws Behaviour Pattern

We classify the multidimensional data into the 5Ws dimensions based on its behaviours. The 5Ws dimensions are defined in this paper as; When the data occurred, Where the data came from, What the data contained, How the data was transferred, Why the data occurred, and Who received the data. Therefore, the dataset *D* can be demonstrated through the 5Ws pattern as

When, Where, What, How, Why, Who

$$D = \left\{\begin{matrix} d_{1T} & d_{1P} & d_{1X} & d_{1Y} & d_{1Z} & d_{1Q} \\ d_{2T} & d_{2P} & d_{2X} & d_{2Y} & d_{2Z} & d_{2Q} \\ d_{3T} & d_{3P} & d_{3X} & d_{3Y} & d_{3Z} & d_{3Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{iT} & d_{iP} & d_{iX} & d_{iY} & d_{iZ} & d_{iQ} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{nT} & d_{nP} & d_{nX} & d_{nY} & d_{nZ} & d_{nQ} \end{matrix}\right\} \tag{2}$$

where $T=\{t_1, t_2, t_3,\ldots,\}$ represents when the data occurred, $P=\{p_1, p_2, p_3,\ldots,\}$ represents where the data came from, $X=\{x_1, x_2, x_3,\ldots,\}$ represents what the data contained, $Y=\{y_1, y_2, y_3,\ldots,\}$ represents how the data was transferred, $Z=\{z_1, z_2, z_3,\ldots,\}$ represents why the data occurred and $Q=\{q_1, q_2, q_3,\ldots,\}$ represents who received the data. A data incident d_i can be illustrated as a node using the 5Ws pattern as $d_i\{t, p, x, y, z, q\}$. The dataset *D* therefore can be defined as

$$D = \{d_1, d_2, d_3, \ldots, d_n\} \tag{3}$$

We use a parallel axis to illustrate a dimension in the 5Ws behaviours pattern, in order to create the 5Ws parallel coordinates for visualization. Parallel coordinates are a popular information visualization tool for high-dimensional data introduced by Alfred Inselberg and Bernard Dimsdale [4]. Each parallel axis represents a dimensional data and polylines are drawn between independent axes at appropriate values. The data examined using the axes shows the data frequencies and the data relationships.

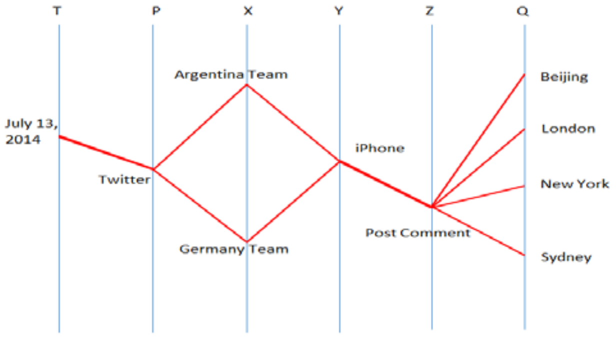

Fig. 1. Example of 5Ws parallel axes

Fig 1 shows an example using the 2014 FIFA World Cup Final between Germany and Argentina. Overall, Twitter users sent 618,725 messages per minute at the moment of Germany's victory [2]. Let us assume that the particular dataset contains the team names x_1 = *"Argentina Team"* and x_2 = *"Germany Team"*, which were posted through iPhone, and that countries which received the data were q_1 = *"Beijing"*, q_2 = *"London"*, q_3 = *"New York"* and q_4 = *"Sydney"*. These particular data incidents can be illustrated in the 5Ws parallel coordinates.

2.3 Dimension Clustering

The 5Ws dimensions can also be explored by clustering if necessary. For example, we want to explore the locations for who received the data by the countries and by the cities, Fig 1 has, then be changed as Fig 2 which shows clustering relationship between *Q1* and *Q2*.

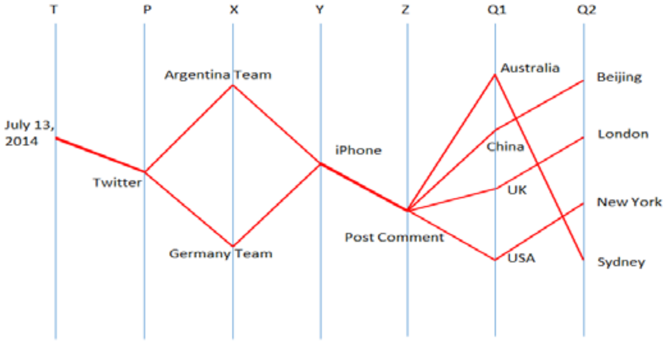

Fig. 2. Example of clustered 5Ws parallel axes

2.4 Shrunk Attributes

Each dimension contains hundreds, even thousands, of attributes, which can lead to the overcrowding of polylines in the pair-density parallel coordinates. To reduce this polyline cluttering, we create Shrunk Attributes (SA) to collect the attributes that are not displayed in each parallel axis, Fig 3 shown the example given.

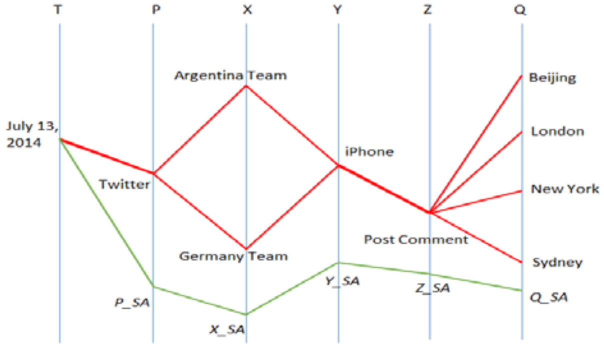

Fig. 3. Example of 5Ws parallel axes with SA

In Fig 3, *P_SA* collects the attributes that are not displayed in the *P* axis, *X_SA* for *X* axis, *Y_SA* for *Y* axis, *Z_SA* for *Z* axis and *Q_SA* for *Q* axis.

Fig 1, Fig 2 and Fig 3 have all clearly illustrated the 5Ws behaviours patterns, but it has also raised an important issue: how do we compare between these patterns on What the data contained, How the data was transferred and Why the data occurred? To solve this issue, we established pair density to measure the 5Ws behaviours pattern.

3 PAIR-DENSITY MODEL

In this section, four Pair-Densities have been established, which are Sending Density via Receiving Density; Sending Density via Purpose Density; Sending Density via Transferring Density and Sending Density via Content Density. We will use Sending Density via Receiving Density to demonstrate the pair-density model.

3.1 Sending Density via Receiving Density

Based on (2) and (3), the sending pattern, which measures where the data came from for a particular attribute *d{t, p, x, y, z}*, is defined as a subset of *D(t, p, x, y, z)*

$$D_{(t,\,p,\,x,\,y,\,z)} = \left\{ \begin{pmatrix} d_t & d_p & d_x & d_y & d_z & d_{1Q} \\ d_t & d_p & d_x & d_y & d_z & d_{2Q} \\ d_t & d_p & d_x & d_y & d_z & d_{3Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_t & d_p & d_x & d_y & d_z & d_{iQ} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_t & d_p & d_x & d_y & d_z & d_{mQ} \end{pmatrix} \right\} \qquad (4)$$

$$= \{d \in D \mid d\,(t, p, x, y, z, Q)\}$$

$Q=\{q_1, q_2, q_3,\ldots q_m\}$ represents who received the particular attribute $d\{t, p, x, y, z\}$. The subset $D_{(t, p, x, y, z)}$ collects all data that has the same attribute, regardless of who received it. For example, Fig 2 shows two sending patterns, {*x* = "*Argentina Team*"} and {*x* = "*Germany Team*"}, regardless of which country or city receiving them.

The Sending Density (*SD*) measures the sender's pattern during data transferal. Based on (3) and (4) for particular attributes {*t, p, x, y, z*}, the Sending Density is defined as $SD_{(t, p, x, y, z)}$.

$$SD_{(t, p, x, y, z)} = \frac{|D(t,p,x,y,z)|}{|D|} \times 100\% \tag{5}$$

The receiving pattern measures who received the data for particular attribute $d\{t, x, y, z, q\}$, which is defined as a subset of $D_{(t, x, y, z, q)}$

$$D_{(t, x, y, z, q)} = \begin{Bmatrix} d_t & d_{1P} & d_x & d_y & d_z & d_q \\ d_t & d_{2P} & d_x & d_y & d_z & d_q \\ d_t & d_{3P} & d_x & d_y & d_z & d_q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_t & d_{iP} & d_x & d_y & d_z & d_q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_t & d_{mP} & d_x & d_y & d_z & d_q \end{Bmatrix} \tag{6}$$

$$= \{ d \in D \mid d\,(t, P, x, y, z, q)\}$$

$P=\{p_1, p_2, p_3,\ldots p_m\}$ represents where the particular attribute $d\{t, x, y, z, q\}$ came from. The subset $D_{(t, x, y, z, q)}$ collects all data that has the same attribute no matter where the data came from. For example, Fig 2 shows eight receiving patterns; {*x* = "*Argentina Team*", *q* = "*Australia*"}, {*x* = "*Germany Team*", *q* = "*Australia*"}, {*x* = "*Argentina Team*", *q* = "*China*"}, {*x* = "*Germany Team*", *q* = "*China*"}, {*x* = "*Argentina Team*", *q* = "*UK*"}, {*x* = "*Germany Team*", *q* = "*UK*"}, {*x* = "*Argentina Team*", *q* = "*USA*"}, {*x* = "*Germany Team*", *q* = "*USA*"}.

The Receiving Density (*RD*) measures the receiver's pattern during data transferal. Based on (3) and (4) for particular attributes {*t, x, y, z. q*}, the Receiving Density is defined as $RD_{(t, x, y, z, q)}$

$$RD_{(t, x, y, z, q)} = \frac{|D(t,x,y,z,q)|}{|D|} \times 100\% \tag{7}$$

The dataset *D* which illustrates the incidents summary represents the volume and velocity of Big Data. The 5Ws density $SD_{(t, p, x, y, z)}$ and $RD_{(t, x, y, z, q)}$ demonstrates the variety of Big Data utilising the patterns for sending and receiving.

3.2 Noise Data

The noise data is defined in this paper as the unknown or undefined attribute in the 5Ws density algorithm methods. We define the unknown attributes in *P* dimension as *u_p*; in *X* dimension as *u_x*; in *Y* dimension as *u_y*; in *Z* dimension as *u_z*; and in *Q* dimension as *u_q*. A subset for any unknown attribute is defined as

$$D_{(u)} = \{\, d \in D \mid d\,(t, p, x, y, z, q),\ p=u\_p \vee x=u\_x \vee y=u\_y \vee z=u\_z \vee q=u\_q \,\} \quad (8)$$

If the subset $D_{(u)}$ collects all the unknown attributes in the 5Ws pattern, then this improves the accuracy for the density algorithms. The $SD_{(t,\ p,\ x,\ y,\ z)}$ and $RD_{(t,\ x,\ y,\ z,\ q)}$ should then be re-defined as

$$SD_{(t,\ p,\ x,\ y,\ z)} = \frac{|\,D(t,p,x,y,z)|}{|\,D\,|-|D(u)|} \times 100\% \quad (9)$$

$$RD_{(t,\ x,\ y,\ z,\ q)} = \frac{|\,D(t,x,y,z,q)|}{|\,D\,|-|D(u)|} \times 100\% \quad (10)$$

$SD_{(t,\ p,\ x,\ y,\ z)}$ and $RD_{(t,\ x,\ y,\ z,\ q)}$ now represents the sender's and receiver's known patterns, which significantly improves the accuracy for Big Data analysis because both densities have avoided noise data.

3.3 Pair-Density Parallel Axes

Here, we create two additional axes by using $SD_{(\,)}$ and $RD_{(\,)}$. The value of each axis is arranged by alphabetical order, which ranges from 0 to 9, A to Z and a to z.

Five Shrunk Attributes, SAs, *p_sa*, *x_sa*, *y_sa*, *z_sa and q_sa*, collect the attributes that are not illustrated in each axis. The Sending Density and Receiving Density for SA are defined as

$$SD_{(SA)} = \frac{|\,D(t,p\_sa,x\_sa,y\_sa,z\_sa)|}{|\,D\,|-|D(u)|} \times 100\% \quad (11)$$

$$RD_{(SA)} = \frac{|\,D(t,x\_sa,y\_sa,z\_sa,q\_sa)|}{|\,D\,|-|D(u)|} \times 100\% \quad (12)$$

We will use Facebook interactions during the 2014 FIFA World Cup Final between Germany and Argentina [2], as our example to show the pair-density parallel coordinates. Let us assume that $SD_{(\text{“Facebook”, “Germany Team”, “iPad”, “Like”})}$ = 40%, $SD_{(\text{“Facebook”, “Argentina Team”, “iPad”, “Like”})}$ = 35%, $RD_{(\text{“Germany Team”, “iPad”, “Like”, “Germany”})}$ = 20%, $RD_{(\text{“Argentina Team”, “iPad”, “Like”, “Argentina”})}$ = 18%, $SD_{(Others)}$ = 25% and $RD_{(Others)}$ = 62%. The pair-density parallel coordinate is shown in Fig 4.

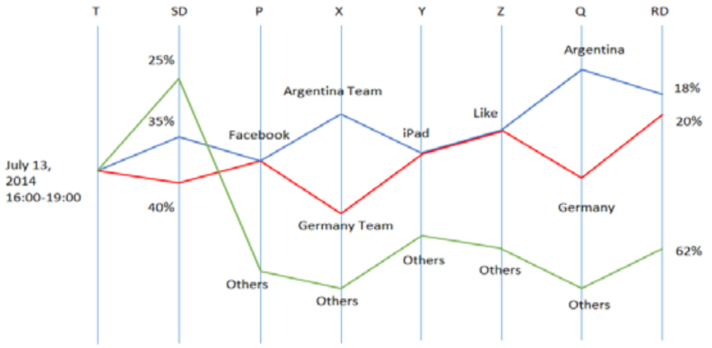

Fig. 4. Example of *SD-RD* parallel coordinates with SA

In Fig. 4, *"Others"* represents the SA that collect the other attributes for $p \neq$ *"Facebook"*, $x \neq$ *"Argentina Team"* or *"Germany Team"*, $y \neq$ *"iPad"*, $z \neq$ *"Like"*, $q \neq$ *"Argentina"* or *"Germany"*.

In Fig 4, 40% of Facebook senders supported the *"Germany Team"* compared to 35% of senders who supported the *"Argentina Team"*. 20% of Facebook receivers are located in *"Germany"* compared to 18% of receivers in *"Argentina"*. 62% of data goes to *"others"* countries and 25% of data came from sources other than *"Facebook"*.

The axes $SD_{()}$ and $RD_{()}$, which were closest to axes *P* and *Q*, have demonstrated senders and receivers patterns which significantly improves measurement for multidimensional data. The pair-density parallel axes, combined with the alphabetical axes and numerical axes, provide the most analytical method for Big Data analysis and visualization. It also explores the particular data patterns that enable multidimensional data analysis and visualization to be very efficient since it can contract or expand as required.

3.4 Clustering in Pair-Density Parallel Axes

The clustering axis in pair-density parallel coordinates can be assigned as several data types or topics. It will lead the values of $SD_{()}$ and $RD_{()}$ to change because a dimension in the 5Ws subset has been added. For example, after adding dimension *P1* as the clustered axis which contains attributes *P1{text, image, video, etc.}*, the subset has been changed to *{T, P, P1, X, Y, Z, Q}*. The value of $SD_{()}$ and $RD_{()}$ changes as well as a result. The clustered pair-density parallel coordinate is shown in Fig 5.

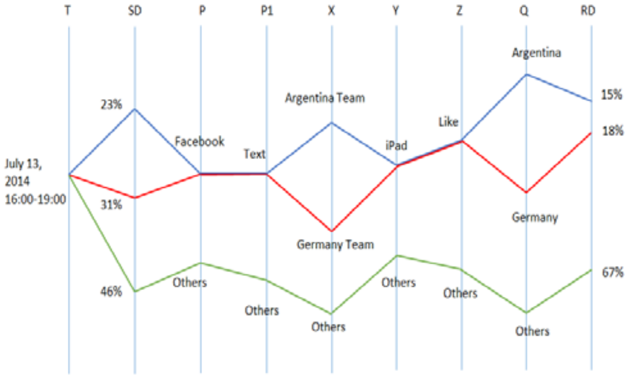

Fig. 5. Example of clustering in *SD-RD* parallel coordinates where *P1* is clustered axis *P*.

3.5 Other Pair-Densities

Sending Density via Purpose Density.

Based on (9) and (10), Sending Density *(SD)* and Purpose Density *(PD)*, which measures where the data came from and why the data occurred, are defined as

$$SD_{(t,\, p,\, x,\, y,\, q)} = \frac{|\, D(t,p,x,y,q)|}{|\, D\, | - |D(u)|} \times 100\% \qquad (13)$$

$$PD_{(t,\,x,\,y,\,z,\,q)} = \frac{|\,D(t,x,y,z,q)|}{|\,D\,|-|D(u)|} \times 100\% \tag{14}$$

SD$_{(\ )}$ measures where the data came from for particular attribute *{t, p, x, y, q}* regardless of why the data occurred. *PD*$_{(\ )}$ measures why the data occurred for particular attribute *{t, x, y, z, q}* regardless of where the data came from.

Sending Density via Transferring Density.

Sending Density *(SD)* and Transferring Density *(TD)*, which measures where the data came from and how the data was transferred, are defined as

$$SD_{(t,\,p,\,x,\,z,\,q)} = \frac{|\,D(t,p,x,z,q)|}{|\,D\,|-|D(u)|} \times 100\% \tag{15}$$

$$TD_{(t,\,x,\,y,\,z,\,q)} = \frac{|\,D(t,x,y,z,q)|}{|\,D\,|-|D(u)|} \times 100\% \tag{16}$$

SD$_{(\ )}$ measures where the data came from for particular attribute *{t, p, x, z, q}* regardless of how the data was transferred. *TD*$_{(\ )}$ measures how the data was transferred for particular attribute *{t, x, y, z, q}* regardless of where the data came from.

Sending Density via Content Density.

Sending Density *(SD)* and Content Density (*CD)*, which measures where the data came from and what the data contained, are defined as

$$SD_{(t,\,p,\,y,\,z,\,q)} = \frac{|\,D(t,p,y,z,q)|}{|\,D\,|-|D(u)|} \times 100\% \tag{17}$$

$$CD_{(t,\,x,\,y,\,z,\,q)} = \frac{|\,D(t,x,y,z,q)|}{|\,D\,|-|D(u)|} \times 100\% \tag{18}$$

SD$_{(\ )}$ measures where the data came from for particular attribute *{t, p, y, z, q}* regardless of what the data contained. *CD*$_{(\ )}$ measures what the data contained for particular attribute *{t, x, y, z, q}* regardless of where the data came from.

4 IMPLEMENTATION

We have tested our pair-density model by using six sample datasets from ISCX2012 network dataset [3], an example of Big Data with 20 data dimensions which contains 906,782 data incidents. The summary of these six sample datasets are shown in Table 1. Unknown traffics in those six datasets are traded as unknown nodes which are calculated and illustrated in the graph. We designed two stages to test our model for implementation. The first stage shows how 5Ws dimension works across 6 datasets. The second stage shows how the pair-density works with SA.

Table 1. Six sample datasets from ISCX2012

Dataset	Jun12	Jun13	Jun14	Jun15a	Jun15b	Jun15c
Network traffic node	133,193	275,528	171,380	101,482	94,911	130,288

Unknown TCP traffic	2	13,568	1,077	11,149	2	3
Unknown UDP traffic	254	414	6,172	36,149	20	36
Attacks	0	20,358	3,771	0	0	37,375
Source Ips	44	44	448	1,611	33	36
Destination Ips	2,610	2,645	7,959	15,067	2,164	1,656
Application Names	21	85	95	69	19	19

4.1 5Ws Parallel Coordinate

The first test stage is shown in Fig 6. The dimension *P* axis represents the source IPs which represented where the data came from. There were 1,948 attributes in the *P* axis. *P = "0.0.0.0"* indicates that the source address was invalid.

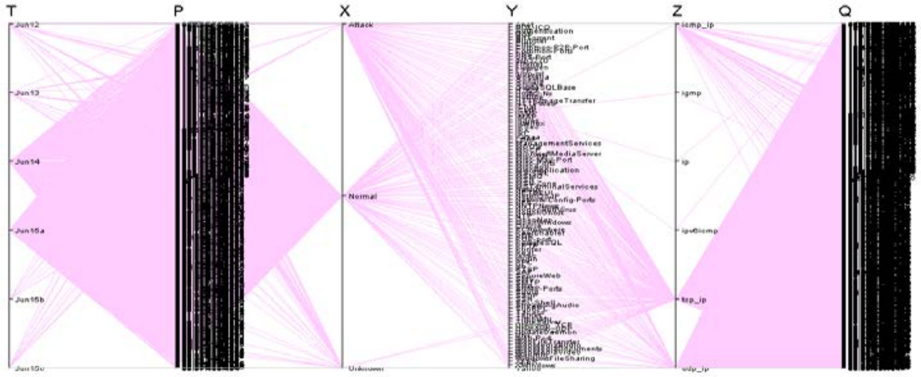

Fig. 6. 5Ws parallel coordinates without the pair-density algorithm where *P* axis contains 1,948 attributes, Y axis contains 105 attributes, and Q axis contains 24,372 attributes

The dimension *X* axis represents the data content, including *"Normal"* traffics, *"Attack"* traffics and *"Unknown"* traffics. The dimension *Y* axis represents the applications which describe how the data was transferred. There were 105 attributes in the *Y* axis. The dimension *Z* axis represents the protocol which illustrates why the data occurred. There were 6 attributes in the *Z* axis. The dimension *Q* axis represents the destination IPs which denotes who received the data. There were 24,374 attributes in the *Q* axis and Q = *"0.0.0.0"* means that the destination address was invalid. In total, 64,393 5Ws patterns are displayed in the graph from 906,782 data incidents. A lot of overlapping polylines and over-crowded attributes are shown in the *P* and *Q* axes as in Fig 6.

4.2 The $SD_{()}$ via $RD_{()}$ Parallel Coordinate with SA

SA has been implemented on two axes; *P* axis and *Q* axis in order to reduce the attributes over-crowding in Fig 7. We define SA for each subnet as *"00x.xxx.xxx.xxx"*, *"0xx.xxx.xxx.xxx"*, *"1xx.xxx.xxx.xxx"*, and *"2xx.xxx.xxx.xxx"* for where $SD_{(p)}$ < 1.0% or $RD_{(q)}$ < 1.0%. In another word, p or q = *"1xx.xxx.xxx.xxx"* including all IPs in the range of *{100-255. 1-255. 1-255. 1-255}* while $SD_{(p)}$ < 1.0% or $RD_{(q)}$ < 1.0%. For example, in *P* axis, if two attributes $SD_{(p=111.111.111.111)}$ < 1.0% and $SD_{(p=123.123.123.123)}$ <

1.0%, those two attributes will be shrunk into one attribute in the parallel coordinate as $SD_{(p=1xx.xxx.xxx.xxx)}$ < 1.0%.

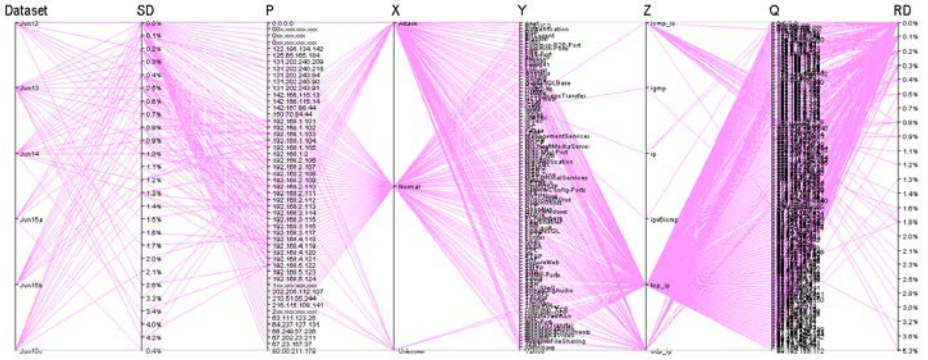

Fig. 7. $SD_{()}$ via $RD_{()}$ parallel coordinates with SA on *P* and *Q* axis where *P* axis remains 51 attributes and *Q* axis remains 200

In Fig 7 after implementing SA, axis *P* contained 51 items, down from 1,948 attributes and axis *Q* had 24,372 attributes shrunk down to 200. The cluttering polylines and over-crowded attributes have been significantly reduced from 64,393 to 8,030. Cluttering has therefore been reduced by over 85% without the loss of any information, which is a significant achievement. The attributes in each axis are represented as different topics such as *"Attack"* in *X* axis, or *"tcp_ip"* in *Z* axis. The data types can be extracted, such as *"http"* in *Y* axis. This provides comparisons between the different topics and types vital for business, government and organizational needs.

4.3 The $SD_{()}$ via $PD_{()}$ Parallel Coordinate with SA

SA has been applied on three axes; *P* axis, *Y* axis and *Q* axis, in $SD_{()}$ via $PD_{()}$ parallel coordinate shown in Fig 8. Axis *P* contains 29 attributes shrunk from 1948. Axis *Q* has been reduced to 52 attributes from 24,374. To analysis "Attack" patterns, SA has been assigned as *"Other-Apps"* for attribute not containing *"Attack"*. The attributes in *Y* axis shrunk to 81 from 105.

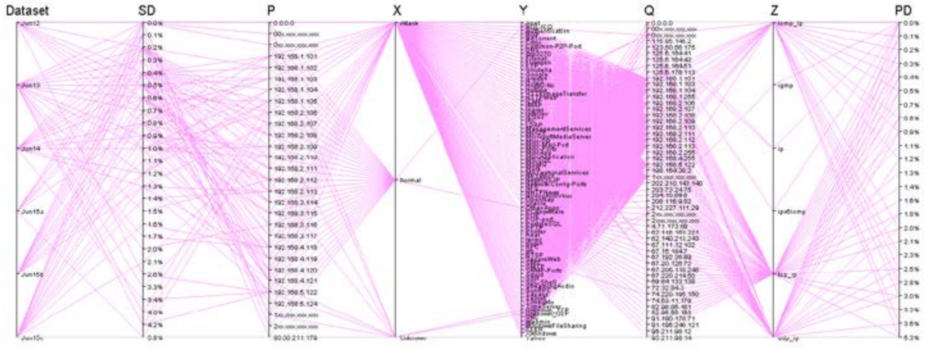

Fig. 8. $SD_{()}$ via $PD_{()}$ parallel coordinates with SA on *P*, *Y* and *Q* axis where *P* axis remains 29 attributes, *Y* axis has 81 attributes, and *Q* axis remains 52 attributes

In Fig 8, the pattern *"Attack"* has been clearly illustrated between the *X* and *Y* axis. *"Normal"* attribute in *X* axis points to *"Other-Apps"* on *Y* axis as a result of using SA on the *Y* dimension. The cluttering polylines and over-crowded attributes have been significantly reduced from 64,393 to 3,404 after implementing SA.

4.4 Reduction of Polylines Cluttering

We have measured the polylines from the original 20 dimensions to our 5Ws dimensions in the parallel coordinates, and found out that the 5Ws parallel coordinates has significantly reduced the cluttered polylines and over-crowded attributes by more than 78% shown in Fig 9. This is a significant boost in the analysis of Big Data as it provides ease of access and clarity to our analysis.

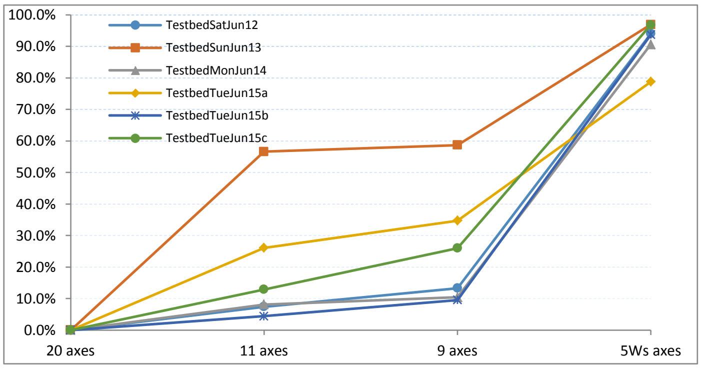

Fig. 9. Reduction for different axes between six datasets

Fig 10 shows the reduction of different SA between datasets. It has not only reduced polylines over-crowding in graphs, but also significantly reduced the data processing time for Big Data analysis and visualization – another major advantage of our model.

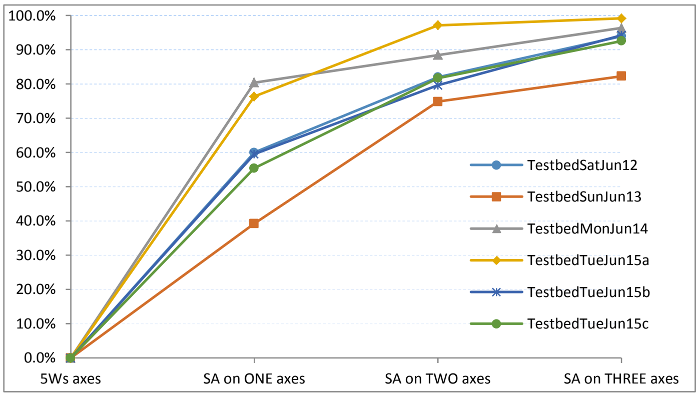

Fig. 10. Reduction for different SA between six datasets

5 RELATED WORKS

The multidimensional data analysis requires tools to explore the relationship between these dimensions. One powerful visual tool that explores multidimensional data is the parallel coordinate which is widely used for multidimensional data visualization. However, it has a problem while deals with large scale multidimensional data: the polylines clutter and over-crowd each other.

Xiaoru Yuan et al [14] scattered points in parallel coordinates to combine the parallel coordinates and scatterplot scaling, which reduced data over-crowding. Matej Novotny and Helwig Hauser [15] grouped the data context into outliers, trends and focus, and set up three clustered parallel coordinates to reduce the data cluttering issues. Yi Chen et al [23] used the parallel coordinates and enhanced ring (PCER) to explore the statistical results for students' scores to reveal any trend. Geoffrey Ellis and Alan Dix [16] developed three methods: raster algorithm, random algorithm and lines algorithm for measuring occlusion in parallel coordinates plots to provide tractable measurement of the clutter.

Most approaches for multidimensional data analysis and visualization are practiced on a single dataset such as text dataset, audio dataset, and image dataset. Xiaotong Liu et al [17] developed a visual search engine based on CompactMap to stream the text data for visual analytics. Seunggwoo Jeon et al [18] transformed unstructured email texts into a graph database and visualized them. Richard K. Lomotey and Ralph Deters [19] extracted the topics and terms from unstructured data by using the TouchR2 tool they created.

Researchers have tried to reduce multidimensional data in their visual approaches. Zhenwen Wang et al [21] introduced ADraw for grouping the same attribute value nodes. Then they created virtual nodes to group the same attribute value nodes together. The different groups are separated by different colours in the visualization. Zhangye Wang et al [8] clustered large-scale social data into users groups by using the information of user tag and user behaviour. The K-means algorithm has been deployed in their approach. Daniel Cheng et al [22] proposed the 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. Quan Li et al [24] proposed PatternTrack to detect visual patterns for multidimensional data, and mapped all dimension axes in concentric circles to integrate three level concentric groups: data values, patterns and gradient circles.

To the best of our knowledge, no previous work has used the 5Ws dimensions to classify the multidimensional data behaviours, nor has any work created two additional axes by using the pair-density in the parallel coordinate visualization. Common visualization methods trade each data as a node in visual graphics, and then find visual patterns to analyze the data. We have classified the data dimension first to obtain the 5Ws patterns, and then visualized those data patterns. Our method has significantly reduced the data processing time and the data cluttering for Big Data analysis and visualization.

6 CONCLUSIONS & FUTURE WORKS

Pair-density model, a novel approach for multidimensional data analysis and visualization, has been introduced in this work. We have demonstrated the 5Ws patterns across multiple datasets, established the pair-density for Big Data analysis, and created two additional axes in pair-density parallel coordinates to reduce data over-crowding without the loss of any information. The shrunk attributes applied in each dimension axis enables the attributes to be contracted or expanded for better visualisation as necessary. The dimension clustering in pair-density parallel axes provides a clear view of visual structures and patterns for better understanding of Big Data.

The pair-sending not only measures multidimensional data patterns, but also provides comparisons for multiple datasets between different topics and data types. This provides more analytical features for Big Data analysis. Our model has reduced data over-crowding by at least 75% in our testing. Even more, the pair-density pattern with shrunk attributes has reduced data cluttering by nearly 98% based on our density algorithm. It has also significantly reduced the data processing time for Big Data analysis.

In the future, we plan to develop our pair-density model and deploy it in more areas and different datasets such as financial datasets and Facebook datasets. The combination of pair-density parallel coordinates and Treemaps is our next stage for Big Data analysis and visualization.

REFERENCES

1. Stamford, "Gartner Says Solving 'Big Data' Challenge Involcves More Than Just Managing Volumes of Data", posted on June 27, 2011, http://www.gartner.com/newsroom/id/1731916 /
2. L. Lorenzetti, "World Cup scores big on Twitter and Facebook", Fortune, posted on July 14, 2014, http://fortune.com/2014/07/14/world-cup-scores-big-on-twitter-and-facebook/
3. A. Shiravi, H. Shiravi, M. Tavallaee, and A.A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," Computers & Security, vol. 31, no. 3, pp 357-374, May 2012
4. A. Inselberg and B. Dimnsdale, "Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry", In Proc. First IEEE Conference on Visualization, pp. 361-378, Oct 1990.
5. K. Qu, N. Lin, Y. Lu, and D.G. Payan, "Multidimensional Data Integration and Relationship Inference", IEEE Intelligent Systems, vol. 17, no 2, pp.21-27, Mar/April 2002
6. J. Zhang and M.L. Huang, "Density approach: a new model for BigData analysis and visualization", Concurrency and Computation: Practice and Experience, published online in Wiley online Library, July 2014, DOI:10.1002/cpe.337
7. X.F. Yin, "Multidimensional Data Clustering Based on Fast Kernel Density Estimation", In Proc. 2013 International Conference on Machine Learning and Cybernetics, pp. 311-315, July 2013
8. Z. Wang, J. Zhou, W. Chen, C. Chen, J. Liao, and R. Maciejewski, "A Novel Visual Analytics Approach for Clustering Large-Scale Social Data", In Proc. 2013 IEEE International Conference on Big Data (IEEE Big Data 2013), pp. 79-86, Oct 2013

9. L. Li, M. Gariel, R.J. Hansman, and R. Palacios, “Anomaly Detection in Onboard-Recored Flight Data using Cluster Analysis”, In Proc. 2011 IEEE/AIAA 30th Digital Avionics Systems Conference (DASC), pp. 4A4-1 – 4A4-11, Oct 2011
10. S.L. Nimmagadda, and H. Dreher, “Petro-data cluster mining – knowledge building analysis of complex petroleum system”, In Proc. 2009 IEEE International Conference on Industrial Technology (ICIT2009), pp. 1-8, Feb 2009
11. J. We, and W. Yu, “Optimization and Improvement based on K-Means Cluster algorithm”, In Proc. 2009 2nd International Symposium on Knowledge Acquisition nd Modeling (KAM09), vol. 3, pp. 335-339, Nov 2009
12. B. Nie, J. Du, H. Liu, G. Xu, Z. Wang, Y. He, and B. Li, “Crowds’ classification using hierarchical cluster, rough sets, principal component analysis and its combination”, In Proc. 2009 International Forum on Computer Science-Technology and Application (IFCSTA09), pp. 287-290, Dec 2009
13. J. Zhang and M.L. Huang, “Detecting Flood Attacks through New Density-Pattern Based Approach”, In Proc. 2013 IEEE 15th International Conference on High Performance Computing and Communications (HPCC2013), pp. 246-253, Nov 2013
14. X. Yuan, P. Guo, H. Xiao, H. Zhou, and H. Qu, “Scattering Points in Parallel Coordinates”, IEEE Transactions on Visualization and Computer Graphics, vol. 15, no 6, pp 1001-1008, Nov/Dec 2009
15. M. Novotny, and H. Hauser, “Outlier-preserving Focus+Context Visualization in Parallel Coordinates”, IEEE Transactions on Visualization and Computer Graphics, vol. 12, no 5, pp 893-900, Sep/Oct 2006
16. G. Ellis and A. Dix, “Enabling Automatic Clutter Reduction in Parallel Coordinates Plots”, IEEE Transactions on Visualization and Computer Graphics, vol. 12, no 5, pp 717-724, Sep/Oct 2006
17. X. Liu, Y. Hu, S. North, and HW. Shen, “CompactMap: A Mental Map preserving Visual Interface for Streaming Text Data”, In Proc. 2013 IEEE International Conference on Big Data (IEEE Big Data 2013), pp. 48-55, Oct 2013
18. S. Jeon, Y. Khosiawan, and B. Hong, “Making a Graph Database from Unstructured Text”, In Proc. 2013 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 981-988, Dec 2013
19. R.K. Lomotey and R. Dters, “Topics and Terms Mining in Unstructured Data Stores”, In Proc. 2013 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 854-861, Dec 2013
20. J. Zhang and M.L. Huang, “5Ws Model for BigData Analysis and Visualization”, In Proc. 2013 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 1021-1028, Dec 2013
21. Z. Wang, W. Xiao, B. Ge, and H. Xu, “ADraw: A novel social network visualization tool with attribute-based layout and coloring”, In Proc. 2013 IEEE International Conference on Big Data (IEEE Big Data 2013), pp. 25-32, Oct 2013
22. D. Cheng, P. Schretlen, N. Kronenfeild, N. Bozowsky and W. Wright, “Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis”, In Proc. 2013 IEEE International Conference on Big Data (IEEE Big Data 2013), pp. 2-4, Oct 2013
23. Y. Chen, X. Cheng, and H. Chen, “A Multidimensional Data Visualization Method Based On Parallel Coordinates and Enhanced Ring”, In Proc. 2011 International Conference on Computer Science and Network Technology (ICCSNT), vol. 4, pp. 2224-2229, Dec 2011
24. Q. Li, L. Chen. H. Liao, and J. Yong, “PatternTrack: A Visual Pattern Detection Technique for Multidimensional Data”, In Proc. 2012 International Conference on Computer Science and Service System (CSSS), pp. 1360-1365, Aug 2012