Data Behaviours Analytics Model for Big Data Visualisation

1 Introduction

Big Data comes from everywhere in our life, and so has extremely large size, complexity and is created very rapidly. For example, posting pictures and writing posts on Facebook, uploading and watching videos on YouTube, sending and receiving messages through smartphones, sending viruses to the victim's systems – all these activities collected by different datasets count as Big Data. According to Pingdom 2012 (Pingdom., 2013), there were 7 petabytes of photo content added on Facebook per month, 5 billion mobile phone users who used 1.3 exabytes of global mobile data traffic per month, 2.2 billion email users who sent 144 billion emails per day, and 4 billion hours of video watched on YouTube per month. This enormous volume of complex data which accumulates rapidly is what we mean when we talk about Big Data, and it is a growing issue in the world today.

Big Data has three main characteristics: Velocity, Volume and Variety, based on Gartner's 3Vs definition (Stamford., 2011). Velocity describes how fast the dataset is produced, volume describes how large the dataset is, usually reaching terabytes of information, while variety describes the different types of data within the dataset. To process such large volumes of data in 'real-time', researchers have tried analysing Big Data on parallel distribution systems, such as Hadoop Distributed File System (HDFS) and MapReduce (Hurwitz et al., 2013) to try and handle Big Data's issues with extremely high velocity and volume. HDFS and MapReduce takes a query over the entire dataset, divides it into many small fragments, and runs the fragments through parallel computing. This reduces data processing time, data sharing time or data clustering time for Big Data analysis (Wu, J. and Hong, B. 2014; Gupta et at., 2013; Malyshkin, V.E. 2014).

Variety describes how datasets contain both structured and unstructured data with hundreds, even thousands, of different attributes in multiple dimensions. This makes Big Data very difficult to analyse using traditional visualisation methods. Researchers have proposed different methods, such as dimension reduction and data clustering, to deal with Big Data's variety issues. Dimension reduction is the process of transferring high dimensional data into lower dimensions, and data clustering groups data with similar attributes into a classical structure for data analysis and visualisation (Wu et al., 2013; Esteves et al., 2014; Babaee et al., 2013).

Currently, Big Data visualisation has three main practices: data-type visualisation, special topic visualisation, and dataset visualisation. Data-type visualisation targets the particular type of data, such as text data visualisation (Afzal et al., 2012), video data visualisation (Meghdadi et al., 2012), or audio data visualisation (Lamboray et al., 2005). Special topic visualisation focuses on a particular topic during the visual algorithm progress, such as network traffic visualisation (Shi et al., 2013), diabetic retinopathy visualisation (Rocha et al., 2012), or word cloud visualisation (Chi et al., 2010). Dataset visualisation focuses on the particular dataset, such as weather dataset visualisation (Sanyal et al., 2010), social network dataset visualisation (Hadiak et al., 2011), or medical dataset visualisation (Wang et al., 2011). All of these visualisation types have their advantages, but are also faced with several flaws.

In this paper, we have further developed our previous works presented at I-SPAN 2014 (Zhang et al., 2014) by creating five densities and five parallel axes for Big Data analysis and visualisation. Firstly, we classified Big Data attributes into our 5Ws dimensions based on data characteristics and behaviours, and then illustrated these characteristics using the 5Ws parallel axes to indicate different dimensions. Secondly, we established our five dimensional densities based on these characteristics: sending density, content density, transferring density, purpose density and receiving density, in order to measure Big Data patterns for any form of data crossing multiple datasets. Thirdly, we created five additional parallel axes by using our previously defined five densities to illustrate the 5Ws patterns for Big Data visualisation. No previous work, to the best our knowledge, had addressed Big Data visualisation by using additional density axes and combining multiple datasets for any form of data, meaning that our research is exploring an entirely new method of data analysis.

The paper is organized as follows; Section 2 defines and illustrates the 5Ws dimensions and behaviours patterns. Section 3 introduces the five densities with additional axes and shrunk attributes in parallel coordinates. Section 4 shows the implementation of our model. Section 5 explains related works. Section 6 summarizes our conclusions and explores future works.

2 Behaviors pattern

2.1 5Ws dimensions and pattern

Big Data, which is collected from multiple datasets, contains text, image, video, audio, mobile or other forms of data and is rapidly growing in size and complexity. To establish a model that suits any form of data in multiple datasets, we classify data incidents into our 5Ws dimensions based on the 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. Those 5Ws dimensions can be illustrated by using six sets.

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

- A set $Q=\{q_1, q_2, q_3,\dots,\}$ represents who received the data

Therefore, each data incident can be defined as a node in the 5Ws pattern as

$$f\,(t, p, x, y, z, q)$$

where
$t \mid T\{\ \}$ is the time stamp for each data incidence.

$p \mid P\{\ \}$ represent where the data came from, such as dataset “Twitter”, “Facebook”, or “hacker” sending a virus spread into a network.

$x \mid X\{\ \}$ represents what the data contained, such as “like”, “comment” in Facebook posts, or “virus” sent by “hacker”.

$y \mid Y\{\ \}$ represents how the data was transferred, such as “by Internet” , “by Phone” or “by email”.

$z \mid Z\{\ \}$ represents why the data occurred, such as “posting photos”, “posting comment” or “spreading a virus”.

$q \mid Q\{\ \}$ represent who received the data, such as “follower” in Twitter, “Facebook friend” or “victim”.

Figure 1. Big Data 5Ws pattern

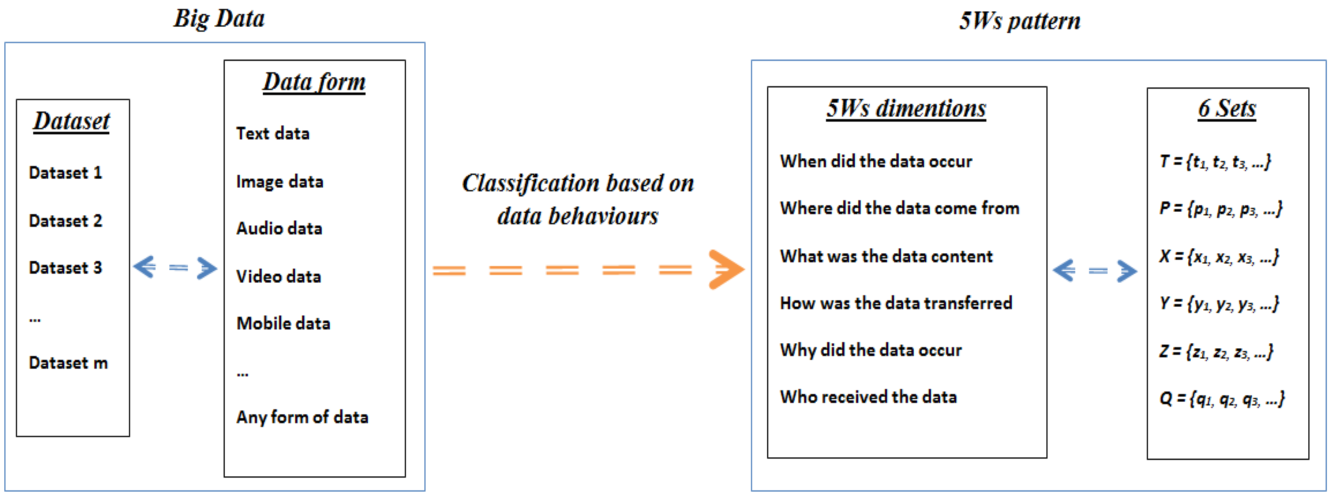


Fig 1 shows the Big Data classification in our 5Ws dimensions. All data incidents in the T time slot are represented as a set F, which contains number n incidences, and so can be defined as

$$F = \{f_1, f_2, f_3, \dots, f_n\} \tag{1}$$

F contains all the incident nodes in the 5Ws pattern within a certain time period. 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. Twitter users also sent 618,725 messages per minute at the moment of Germany’s victory (Lorenzetti, L 2014).

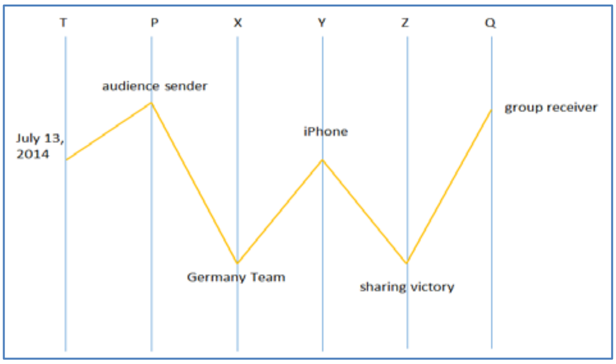
For a particular incident node where $p=\delta$, $x=\alpha$, $y=\beta$, $z=\gamma$ and $q=\varepsilon$, the incident node can then be represented as $f\,(t, p_{(\delta)}, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p_{(\varepsilon)})$. A subset $F_{(t,\,\delta,\,\alpha,\,\beta,\,\gamma,\,\varepsilon)}$ that contains all the particular incident nodes $f\,(t, p_{(\delta)}, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p_{(\varepsilon)})$ is therefore defined as

$$F_{(t,\,\delta,\,\alpha,\,\beta,\,\gamma,\,\varepsilon)} = \{\, f \in F \mid f\,(t, p, x, y, z, q),\ t,\ p=\delta,\ x=\alpha,\ y=\beta,\ z=\gamma,\ q=\varepsilon \,\} \tag{2}$$

The subset $F_{(t,\,\delta,\,\alpha,\,\beta,\,\gamma,\,\varepsilon)}$ represents the particular incident pattern by the 5Ws dimensions. For example, during 2014 FIFA World Cup final game, we assume 280 million Facebook posts contain a particular pattern as t = “13-Jul-2014”, δ = “audience sender”, α = “German team”, β =“iPhone”, γ =“sharing victory”, ε = “group receiver”, which shown in Fig 2.

In Fig 2, each parallel axis represents a 5Ws dimension, and is arranged as T, P, X, Y, Z, Q. The value of each axis is ordered by alphabetical order which ranges from 0 to 9, A to Z and a to z. The polyline shows a particular 5Ws pattern using parallel coordinate visualisation. The dataset $|F|$ = 280 million Facebook interactions illustrates the incidents summary representing the volume and velocity of Big Data characters. The subset $F_{(t,\,\delta,\,\alpha,\,\beta,\,\gamma,\,\varepsilon)}$ demonstrates the variety of this Big Data by representing 88 million Facebook senders and receivers though the clear and concise use of parallel axes and polylines.

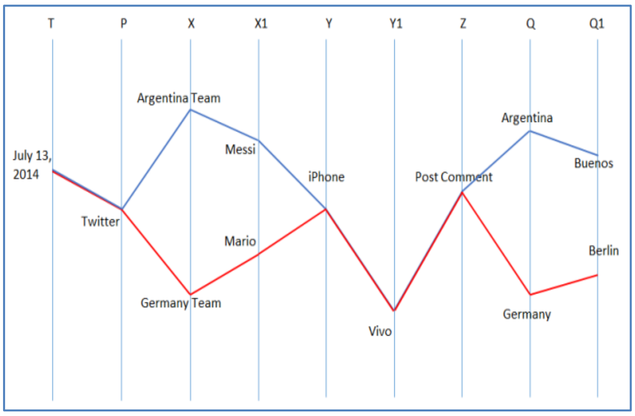
Figure 2. Example of 5Ws pattern in parallel coordinate



2.2 Dimensions clustering

Each dimension in the 5Ws patterns contains multiple classical relationships that can be clustered for that particular attribute. Suppose X1 is the clustered dimension for X, Y1 for Y and Q1 for Q. We will use Twitter messages during the 2014 FIFA World Cup Final to demonstrate the 5Ws dimensions clustering structure shown in Fig 3.

Figure 3. Example of 5Ws clustered dimensions



In Fig 3, dimension *T* indicates *t* = "July 13, 2014". Dimension *P* represents where the data came from, *p="Twitter"*. Dimension *X* represents what the data contained, *x="Argentina Team"* and "Germany Team". Dimension *Y* represents how the data was transferred, *y="iPhone"* means that the data was transferred by iPhone Apps. Dimension *Z* represents why the data occurred, *z="Post Comment"*. Dimension *Q* represents who received the data, *q="Argentina"* and *"Germany"* means the country received the data.

Three clustered dimensions are shown in Fig 3. X1 is the clustered dimension for X which represents the players for each team such as *"Messi"* for *"Argentina Team"*, *"Mario"* for *"Germany Team"*. Y1 is the clustered dimension for Y which represents the mobile communication providers such as *"Vivo"*. Q1 is the clustered dimension for Q which represents the city of the country such as *"Buenos"* for *"Argentina"*, *"Berlin"* for *"Germany"*. Fig 3 therefore clearly indicates the 5Ws dimensions with clustering structure for each of the particular topics.

Each dimension contains multiple attributes. For example, there were 32 teams competing in 2014 FIFA World Cup; each team has 23 players. Twitter users sent messages worldwide talking about these 32 teams, each with 23 players, which results in a huge number of combinations in 5Ws pattern. We can therefore establish the 5Ws dimensional densities to measure these varieties and patterns.

3 5ws densities and additional axes

3.1 5Ws densities

3.1.1 Sending density

Based on (2), a particular subset that contains all incidents, we can describe where the data come from for all incidents where $p=\delta$. This is defined as

$$F_{(\delta)} = \{\, f \in F \mid f\,(t, p, x, y, z, q),\ p=\delta \,\} \qquad (3)$$

We will introduce Sending Density (*SD*), which is used to measure the proportion of sender's pattern during data transferal for a particular attribute $p=\delta$. We define $SD_{(\delta)}$ as

$$SD_{(\delta)} = \frac{|\,F(\delta)\,|}{|\,F\,|} \times 100\% = \frac{\sum_{i=1}^{n} f_{(i)}(t,\ p_{(\delta)},\ x,\ y,\ z,\ q)}{\mathrm{n}} \times 100\% \qquad (4)$$

where $0 \le SD_{(\delta)} \le 1$ and *n* is the total number of incidents.

$SD_{(\delta)}$ collects all different patterns by different attributes in *x*, *y*, *z* and *q* as long as they all come from *δ*. For example, in Fig 3, $SD_{(Twitter)}$ collects two patterns that support different teams and are received by different countries, but which all are sent over Twitter. Assume *m* illustrates the total number of attributes in dimension *P*, so there are *m* different values of $SD_{(\ )}$ that demonstrates *m* number of different sending patterns. The sum of all $SD_{(\ )}$ should be one.

$$\sum_{j=1}^{m} SD_{(j)} = 1 \qquad (5)$$

$SD_{(\delta)}$ represents the 5Ws dimensions for the sender's pattern; in $t \subset T$ time, sent by *δ*, for any *t, x, y, z* and *q* as a proportion of total incidents n. A higher value of $SD_{(\delta)}$ indicates that sender (*δ*) sent more data compared to other senders.

3.1.2 Content density

We can then define a particular subset that contains all incidents where the data content is $x=\alpha$, as

$$F_{(\alpha)} = \{\, f \in F \mid f\,(t, p, x, y, z, q),\ x=\alpha \,\} \qquad (6)$$

The Content Density (*CD*) is used to measure the data content's density patterns during data transferal for a particular attribute $x=\alpha$. It is defined as $CD_{(\alpha)}$

$$CD_{(\alpha)} = \frac{|\,F(\alpha)\,|}{|\,F\,|} \times 100\% = \frac{\sum_{i=1}^{n} f_{(i)}(t,\ p,\ x_{(\alpha)},\ y,\ z,\ q)}{\mathrm{n}} \times 100\% \qquad (7)$$

where $0 \le CD_{(\alpha)} \le 1$, *n* means total number of incidents and $\sum CD_{(\ )} = 1$.

$CD_{(\alpha)}$ represents the 5Ws dimensions for the data content's pattern; in $t \subset T$ time, contains *α*, for any *t, p, y, z* and q as a proportion of total incidents n. A higher value of $CD_{(\alpha)}$ indicates that that particular data content was more popular compared to other data contents.

3.1.3 Transferring density

Next, we shall also introduce a particular subset that contains all data incidents which are transferred by $y= \beta$, such that

$$F_{(\beta)} = \{ f \in F \mid f(t, p, x, y, z, q),\ y = \beta \} \tag{8}$$

Therefore, the Transferring Density (TD), to measure a particular attribute $y= \beta$ during data transferal, is defined as $TD_{(\beta)}$

$$TD_{(\beta)} = \frac{|F(\beta)|}{|F|} \times 100\% = \frac{\sum_{i=1}^{n} f_{(i)}\big(t,\ p,\ x,\ y_{(\beta),}\ z,\ q\big)}{\mathrm{n}} \times 100\% \tag{9}$$

where $0 \leq TD_{(\beta)} \leq 1$, $\sum TD_{(\,)} = 1$.

$TD_{(\beta)}$ represents the 5Ws dimensions for the data transferring pattern; in $t \subset T$ time, contains β, for any $t,\ p,\ x,\ z$ and q as a proportion of total incidents n. A higher value of $TD_{(\beta)}$ indicates that more data was transferred by $y= \beta$ compared to data transferred by other sources.

3.1.4 Purpose density

We will further define a particular subset that contains all incidents which occurred because of $z=\gamma$, as

$$F_{(\gamma)} = \{ f \in F \mid f(t, p, x, y, z, q),\ z = \gamma \} \tag{10}$$

The Purpose Density (PD) describes the purpose for data transferal in a particular attribute $z = \gamma$, and is defined as $PD_{(\beta)}$

$$PD_{(\gamma)} = \frac{|F(\gamma)|}{|F|} \times 100\% = \frac{\sum_{i=1}^{n} f_{(i)}\big(t,\ p,\ x,\ y,\ z_{(\gamma),}\ q\big)}{\mathrm{n}} \times 100\% \tag{11}$$

where $0 \leq PD_{(\gamma)} \leq 1$, $\sum PD_{(\,)} = 1$.

$PD_{(\gamma)}$ represents the 5Ws dimensions for the data purpose pattern; in $t \subset T$ time, contains γ, for any $t,\ p,\ x,\ y$ and q as a proportion of total incidents n. A higher value of $PD_{(\gamma)}$ illustrates that that particular purpose is the most likely purpose for the data occurring.

3.1.5 Receiving density

The particular subset that contains all data incidents by $q= \varepsilon$, is defined as

$$F_{(\varepsilon)} = \{ f \in F \mid f(t, p, x, y, z, q),\ q = \varepsilon \} \tag{12}$$

The Receiving Density (RD) illustrates the receiving pattern during data transferal for a particular attribute $q = \varepsilon$, and is defined as $RD_{(\varepsilon)}$

$$RD_{(\varepsilon)} = \frac{|F(\varepsilon)|}{|F|} \times 100\% = \frac{\sum_{i=1}^{n} f_{(i)}\big(t,\ p,\ x,\ y,\ z,\ q_{(\varepsilon)}\big)}{\mathrm{n}} \times 100\% \tag{13}$$

where $0 \leq RD_{(\varepsilon)} \leq 1$, $\sum RD_{(\,)} = 1$.

$RD_{(\varepsilon)}$ represents the 5Ws dimensions for the data receiving pattern; in $t \subset T$ time, contains ε, for any $t,\ p,\ x,\ y$ and z as a proportion of total incidents n. A higher value of $RD_{(\varepsilon)}$ demonstrates that $q= \varepsilon$ received more data compared to other receivers.

3.2 Noise data

Within each 5Ws dimension, there exists data incidents which have unknown or undefined attributes and properties due to either natural or artificial properties of the dataset. These incidents are known as noise data, which increases the processing time during the density algorithm analysis. We define the unknown attributes in the P dimension as *p_unknown*; in the X dimension as *x_unknown*; in the Y dimension as *y_unknown*; in the Z dimension as *z_unknown*; and in the Q dimension as *q_unknown*. Therefore, the noise data subset $F_{(p\_unknown)}$ collects all unknown attributes in the P dimension; $F_{(x\_unknown)}$ collects all noise data in the X dimension; $F_{(y\_unknown)}$ collects all noise data in the Y dimension; $F_{(z\_unknown)}$ collects all noise data in the Z dimension and $F_{(q\_unknown)}$ collects all noise data in the Q dimension.

Because the subset $F_{(unknown)}$ collects all unknown attributes in the 5Ws pattern, by removing noise data we therefore improve the accuracy for density algorithm. This is because we are removing irrelevant data, which saves processing time while maintaining accuracy and analysis quality. The densities $SD_{(\,)}$, $CD_{(\,)}$, $TD_{(\,)}$, $PD_{(\,)}$ and $RD_{(\,)}$ would then be re-defined as

$$SD_{(\delta)} = \frac{|F(\delta)|}{|F| - |F(p\_unknown)|} \tag{14}$$

$$CD_{(\alpha)} = \frac{|F(\alpha)|}{|F| - |F(x\_unknown)|} \tag{15}$$

$$TD_{(\beta)} = \frac{|F(\beta)|}{|F| - |F(y\_unknown)|} \tag{16}$$

$$PD_{(\gamma)} = \frac{|F(\gamma)|}{|F| - |F(z\_unknown)|} \tag{17}$$

$$RD_{(\varepsilon)} = \frac{|F(\varepsilon)|}{|F| - |F(q\_unknown)|} \tag{18}$$

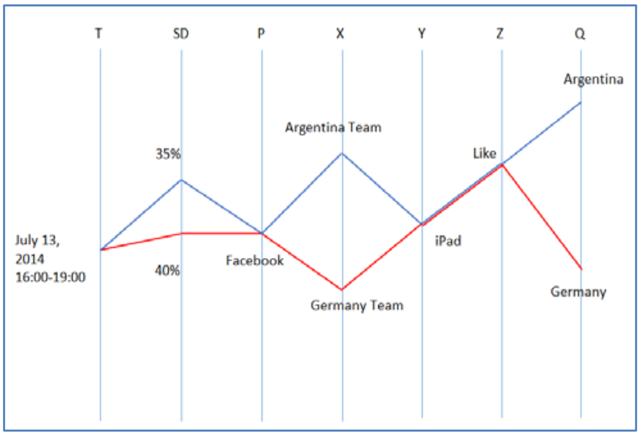
These five densities not only provide more analytical feature for Big Data visualisation, but also significantly improve the accuracy of Big Data analysis because all densities have avoided noise data, which hinders visualisation and analysis. The removal of noise data is a major way we can improve the efficiency of the analytical process.

3.3 Density parallel axes

Parallel coordinate visualisation is a popular information visualisation tool for multi-dimensional data. It establishes several parallel coordinate axes and then draws the polylines between each dimensional axis at the appropriate values, therefore allowing us to explore the statistical relationship between the dimensional axes. Based on our density algorithm, we created additional axes for each dimension by using its density to measure the 5Ws attribute patterns in parallel coordinate visualisation. This non-dimensional axis enables the measurement of attributes in any form of data for multiple datasets.

We secondly add additional axis next to each dimension in order to visualise its density. Fig 4 shows the dimensional axes with an additional $SD_{(\ )}$ axis located on the left hand side of *P* axis, which measures the density of the sender's patterns.

Figure 4. Example of $SD_{(\ )}$ parallel coordinates



In Fig 4, we assume $SD_{(\text{"Facebook", "Germany Team", "iPad", "Like"})}$ = 40%, and $SD_{(\text{"Facebook", "Argentina Team", "iPad", "Like"})}$ = 35%. The graph, which illustrates an example of $SD_{(\ )}$ density parallel coordinates during the 2014 FIFA World Cup Final, clearly indicates that 40% Facebook senders favour the "Germany Team" compared to 35% senders who support the "Argentina Team".

From Fig. 1, we can therefore see how our 5Ws density parallel axes, combined with the alphabetical axes and numerical axes, can provide another different analytical method for Big Data analysis and visualisation. The 5Ws model is more clear and concise, and the densities allow us for a deeper and more analytical examination of the dataset.

3.4 Shrunk attribute

When we have to examine very complex datasets, it becomes apparent that each dimension contains hundreds, even thousands, of attributes. This very quickly overcrowds the graph with crossing polylines and attributes in the 5Ws density parallel coordinates. To reduce this polylines cluttering, we create Shrunk Attributes (SA) to represents all attributes that we choose not to display in the density parallel axis. This could potentially be because we have decided to collect several minor or insignificant attributes into a single attribute known as 'Other'. Doing so will not only greatly reduce the number of attributes on each axis, but will also reduce polyline overcrowding. The densities to collect all SA attributes are defined as

$$SD_{(SA)} = \sum_{j=1}^{m} SD_{(j)}\left(t, p_{(sa)}, x, y, z, q\right) \tag{19}$$

$$CD_{(SA)} = \sum_{j=1}^{m} CD_{(j)}\left(t, p, x_{(sa)}, y, z, q\right) \tag{20}$$

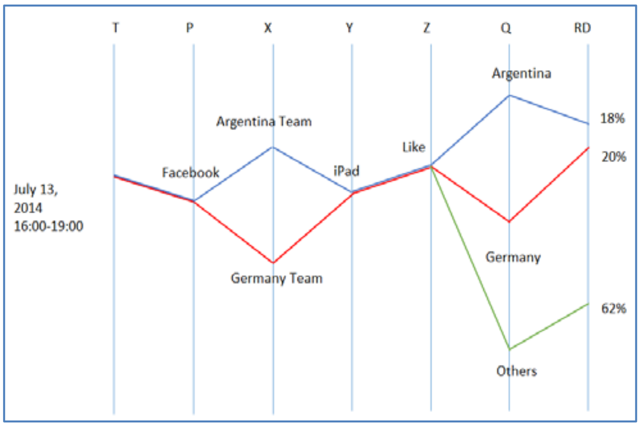
$$TD_{(SA)} = \sum_{j=1}^{m} TD_{(j)}\left(t, p, x, y_{(sa)}, z, q\right) \tag{21}$$

$$PD_{(SA)} = \sum_{j=1}^{m} PD_{(j)}\left(t, p, x, y, z_{(sa)}, q\right) \tag{22}$$

$$RD_{(SA)} = \sum_{j=1}^{m} RD_{(j)}\left(t, p, x, y, z, q_{(sa)}\right) \tag{23}$$

where, $j = 1 \rightarrow m$ means there were *m* attributes that were collected under a single attribute 'Other'. Fig 5 demonstrates an example of SA in $RD_{(\ )}$ parallel coordinate visualisation.

Figure 5. Example of SA in $RD_{(\ )}$ parallel coordinates



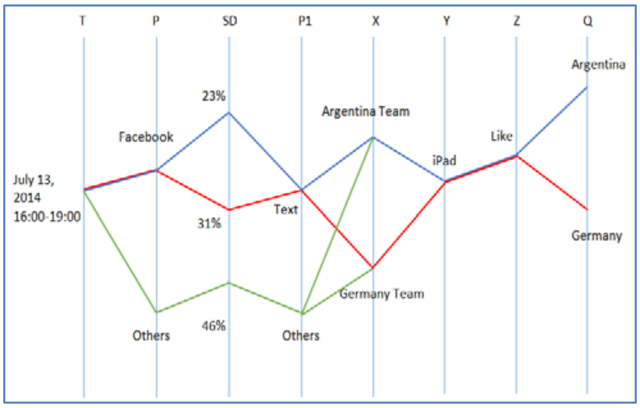
Here, we assign SA = "Others" in the Q dimension in order to collects the other attributes for $Q \neq$ *"Argentina" or* $Q \neq$ *"Germany"*. $RD_{(Others)}$ = 62% allows us to deduce that "Other" countries received 62% of data, Argentina receiving 18% and Germany receiving 20% of data.

As out graph clearly demonstrates, SA not only reduces the overcrowding of hundreds of attributes, but also summarizes all unseen attributes into a single attribute that keeps all information in track without any loss. This therefore reduces in increased clarity without any reduction in accuracy.

3.5 Clustering of density parallel axes

The clustering axis is an additional axis which stems from the 5Ws dimensions in the 5Ws density parallel coordinates, and can be used to explore the data types or topics. It will lead the value of densities to change because new elements in the subset have been added. Fig 6 shows the example.

Figure 6. Example of clustering in $SD_{()}$ parallel coordinates



In Fig 6, after adding dimension *P1{"text", "other"}* as the clustered *P* axis which represents the data-type for the dataset, the subset has been changed to *F{t, p, p1, x, y, z, q}*. This means that $SD_{()}$ has also changed. In Fig. 6, this shows that there are 54% of Facebook interactions that sent "likes" with text format using an "iPad", which contained the words "Argentina Team" or "Germany Team" and received in either Argentina or Germany countries.

In summary, we have classified Big Data into the 5Ws dimension based on data behaviours, and then established our five dimensional densities to measure the 5Ws patterns across multiple datasets. We have also created five additional axes in our parallel coordinate system for Big Data visualisation. Our approach has clearly demonstrated the data patterns for different datasets with any form of data attribute. The parallel axes with Shrunk Attribute and clustering views explores the particular data types and topics, while removing irrelevant information. This has successfully resulted in Big Data analysis and visualisation becoming much more efficient, easier to analyse and faster, while not leading to any loss of information as the SA can be expanded or narrowed as necessary.

4 Implementation

We have tested our approach by using six sample datasets from the ISCX2012 network dataset (Shiravi et al., 2012), an example of Big Data, which contains 906,782 incidents and 20 dimensions. The summary of these six sample datasets are shown in Table I. Unknown traffics in those six datasets are traded as unknown nodes which are calculated and illustrated in the graph. The applications describe how the data was transferred.

Table 1. Six Sample Datasets from ISCX2012

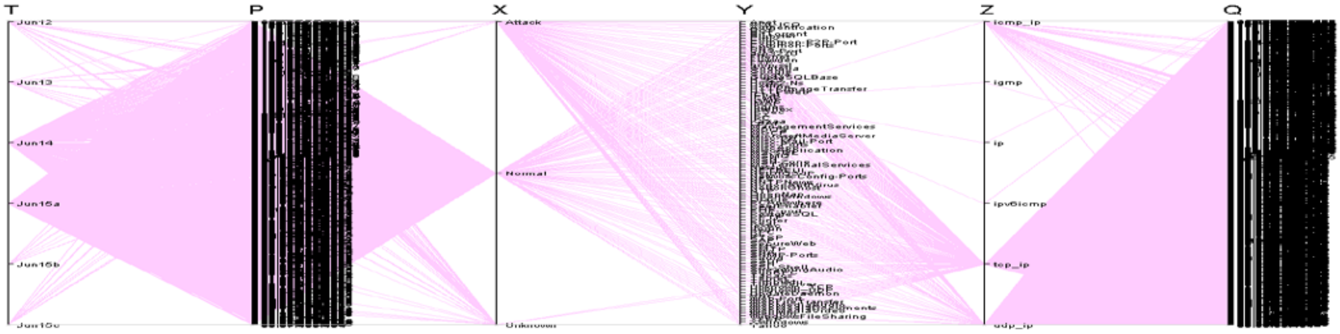
Dataset	TestbedJun12	TestbedJun13	TestbedJun14	TestbedJun15a	TestbedJun15b	TestbedJun15c
Network traffic nodes	133,193	275,528	171,380	101,482	94,911	130,288
Unknown TCP traffics	2	13,568	1,077	11,713	2	3
Unknown UDP traffics	254	414	6,172	36,149	20	36
Attacks	0	20,358	3,771	0	0	37,375
Source IPs	44	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

We designed two stages to test our model for implementation. In the first stage, we showed how the 5Ws dimensions worked across 6 datasets. The 5Ws pattern is illustrated in the 5Ws parallel coordinate as Fig 7, which contains a lot of polylines cluttering. The second stage shows the impact of applying the SA on the density algorithm. The five density patterns are implemented and applied with SA, and the results are shown in the Fig 8-13.

4.1 5Ws behaviours patterns

In the first test stage, shown in Fig 7, the *P* axis is chosen by the source IPs, which represents where the data came from. There were 1,948 attributes in *P* axis. The *X* axis is chosen by the data content, including "Normal" traffics, "Attack" traffics and "Unknown" traffics. The *Y* axis is chosen by the applications which describe how the data was transferred. There were 105 attributes in the *Y* axis. The *Z* axis is chosen by the protocol which illustrates why the data occurred. There were 6 attributes in the *Z* axis. The *Q* axis is chosen by the destination IPs which represents who received the data. There were 24,374 attributes in the *Q* axis. In total, Fig. 7 displays 64,393 5Ws patterns in the graph from a total of 906,782 data incidents.

Figure 7. Parallel coordinates in 5Ws patterns



In Fig 7, the P axis contains 1,948 attributes and the Q axis has 24,374 attributes. This causes a lot of overlapping polylines and over-crowded attributes in the graph. In the second stage, we use density algorithms and shrunk attributes to reduce cluttering and measure 5Ws patterns without losing clarity.

4.2 Densities with SA in parallel coordinate

In the second test stage, five densities with SA have been implemented in parallel coordinates, shown in Fig 8. SA has been applied in the *P*, *Y* and *Q* axes in order to reduce the over-crowding of attributes and cluttered polylines. For both the *P* and *Q* axes, we define SA for each subnet as “00x.xxx.xxx.xxx”, “0xx.xxx.xxx.xxx”, “1xx.xxx.xxx.xxx”, and “2xx.xxx.xxx.xxx” to represent where $SD_{(p)}$ < 1.0% and $RD_{(q)}$ < 1.0%. In other words, the attributes in *p* or *q* = “1xx.xxx.xxx.xxx” including all IPs with a range of {100-255. 1-255. 1-255. 1-255} with any density < 1.0%. For example, in *P* axis, if $SD_{(p=111.111.111.111)}$ < 1.0% and $SD_{(p=123.123.123.123)}$ < 1.0%, it will be summarized into one 5Ws pattern as $SD_{(p=1xx.xxx.xxx.xxx)}$ < 1.0%.

To narrow down security events for six datasets, we define the condition of SA on the *Y* axis as whether the attribute was transferring an “Attack” or not. If the attribute in the *Y* axis transferred an “Attack” or “Unknown traffics”, the attribute will stay in the *Y* axis as a separate attribute. If the attribute transferred a “Normal” incident, the attribute will then be shrunk into SA as “Other-Apps” in the Y axis. As a result, $TD_{(Other\text{-}Apps)}$ collects all attributes labelled as “Other-Apps” in the *Y* axis.

Fig 8 illustrates the second test result containing the 5Ws dimensions and five densities. Fig 9-13 represents each of the five individual densities and its 5Ws patterns.

Figure 8. Five densities in parallel coordinates with SA on P, Y, and Q axes

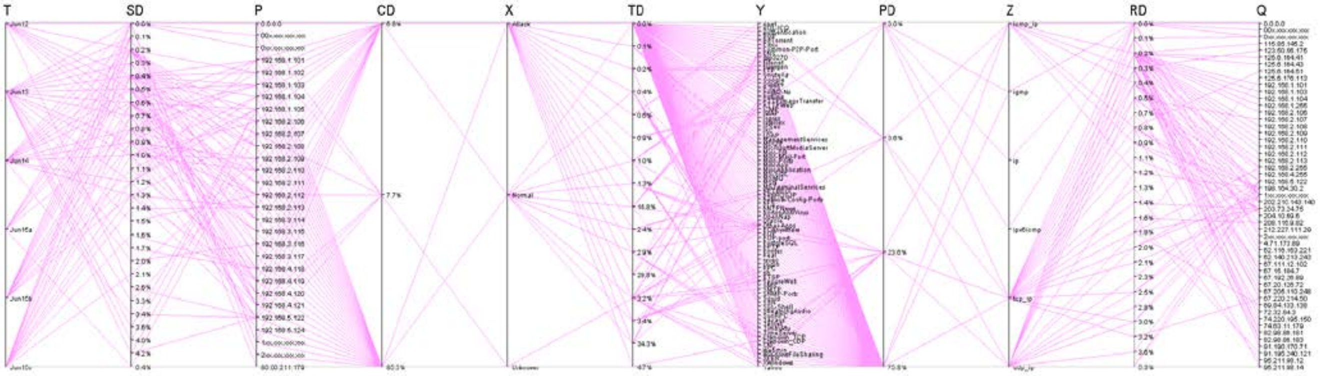


Figure 9. $SD_{(\,)}$ parallel coordinates with SA

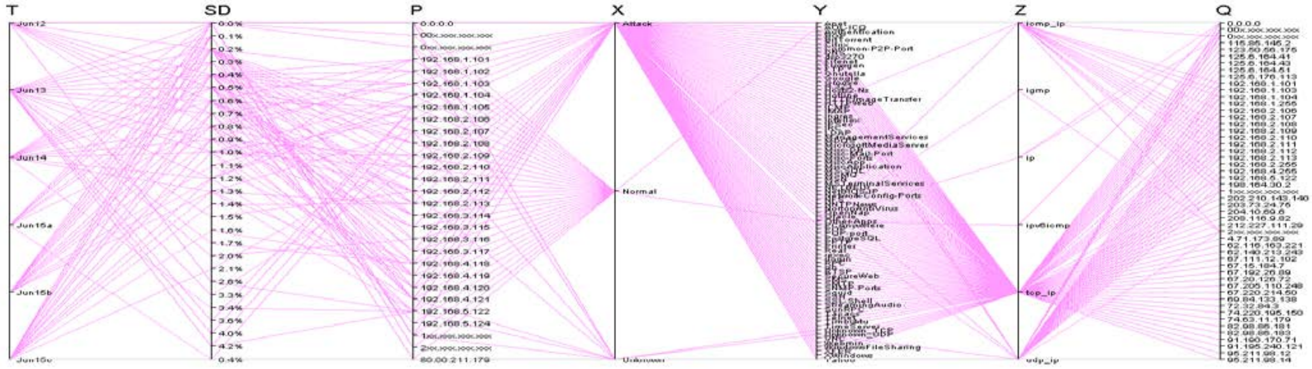


Figure 10. $CD_{(\,)}$ parallel coordinates with SA

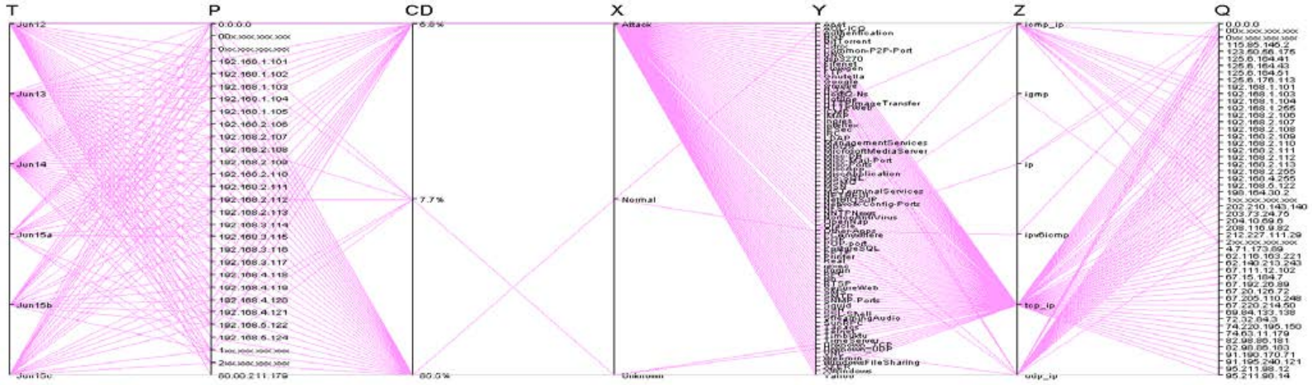


Figure 11. *TD*$_{(\,)}$ parallel coordinates with SA

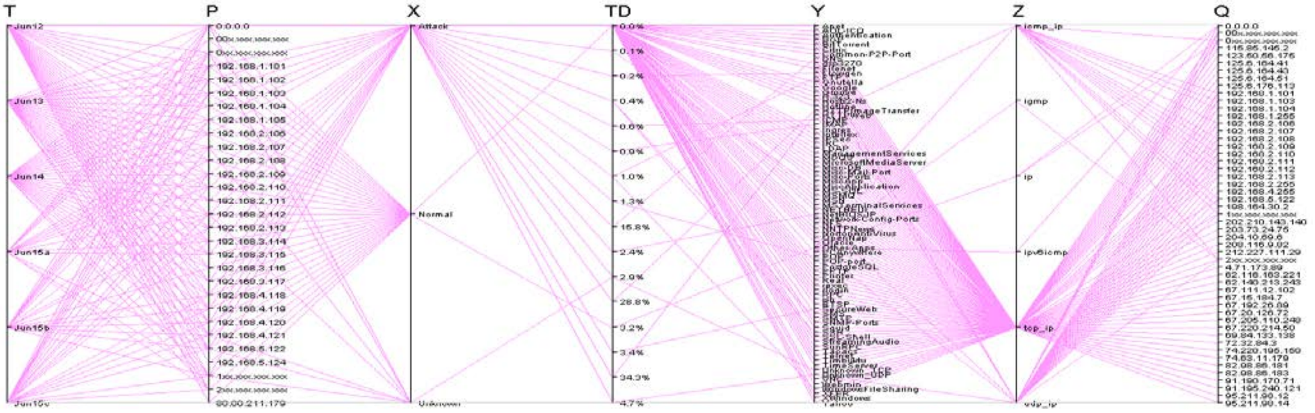


Figure 12. *PD*$_{(\,)}$ parallel coordinates with SA

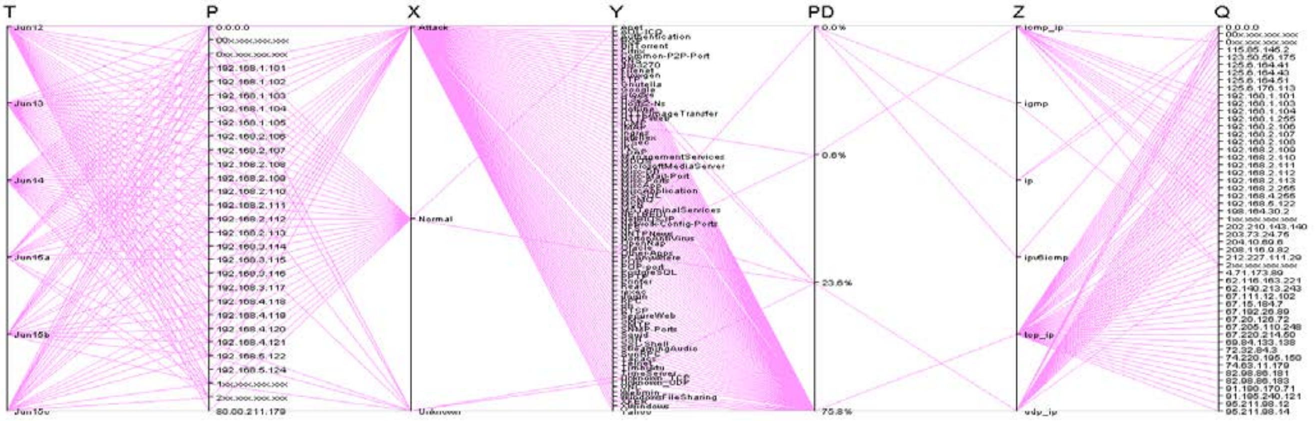
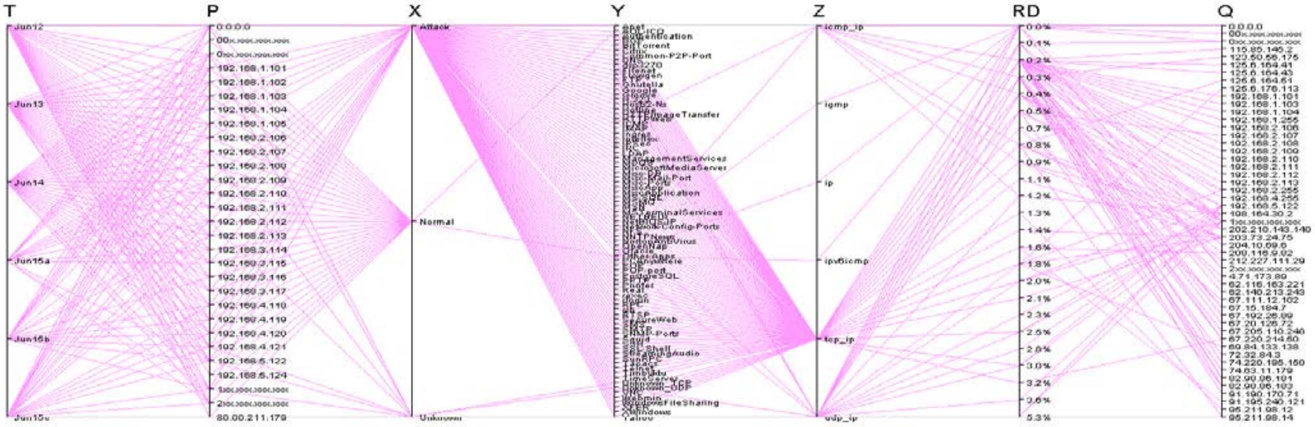


Figure 13. *RD*$_{(\,)}$ parallel coordinates with SA



In Fig 8, after we applied SA on the *P*, *Y* and *Q* axes, the number of attributes in each dimension were significantly reduced: *P* axis has dropped from 1,948 attributes to 29 attributes, *Y* axis has dropped from 105 attributes to 81 attributes, and *Q* axis has dropped from 24,374 attributes to 52 attributes. The total number of polylines has significantly reduced from 64,393 to 3,404 without the loss of information. This has therefore led to much greater clarity while maintaining accuracy.

In Fig 9, the highest value of *SD*$_{(\,)}$ amongst the six datasets, is 6.4%, which was sent from the source “192.168.2.107” in dataset “Jun15a” which contained “Normal” and “Unknown” data contents. There are 1,919 data incidents sent for which the value of *SD*$_{(\,)}$ was less than 1.0%. The six datasets with the highest values of *SD*$_{(\,)}$ are shown in Table 2.

Table 2. Higher value of *SD*$_{(\,)}$ with attribute in *P* axis

T	**Higher value of *SD*$_{(\,)}$**	*P*
Jun12	2.6%	192.168.5.122
Jun13	4.2%	192.168.2.106
Jun14	3.4%	192.168.5.122
Jun15a	6.4%	192.168.2.107
Jun15b	1.3%	192.168.5.122
Jun15c	2.1%	192.168.2.109

In Fig 10, the value of $CD_{(\,)}$ that contains “Attack” is 6.8%. There were sent from 20 attributes in the P axis, and by 80 attributes in the Y axis. Furthermore, we can also say that $CD_{(Normal)}$ = 85.5% and $CD_{(Unknown)}$ = 7.7%.

In Fig 11, there are 81 attributes in the Y axis. The highest value of $TD_{(\,)}$ is 34.3%. This indicates that most “Attack” data incidents were transferred by “HTTPWeb” from “tcp_ip” connection. There are 25 attributes in the Y axis transferring “Normal” content, which shrunk into the “Other-Apps” attribute as a result of SA, and $TD_{(Other\text{-}Apps)}$ = 3.2%.

In Fig 12, the highest value of $PD_{(\,)}$ is 75.8%. The value of $PD_{(\,)}$ and the corresponding attributes in the Z axis are shown in Table 3.

Table 3. Value of $PD_{(\,)}$

Attributes in *Z*	Value of $PD_{()}$	*P*
icmp_ip	0.6%	192.168.5.122
igmp	0.0%	192.168.2.106
ip	0.0%	192.168.5.122
ipv6icmp	0.0%	192.168.2.107
tcp_ip	75.8%	192.168.5.122
udp_ip	23.6%	192.168.2.109

In Fig 13, the highest value of $RD_{(\,)}$ is 5.3% which is connected by “udp_ip” in the Z axis. There are 24,324 attributes in the Q axis, who collectively received data of value $RD_{(\,)}$ < 1.0%.

4.3 Reduction of cluttering

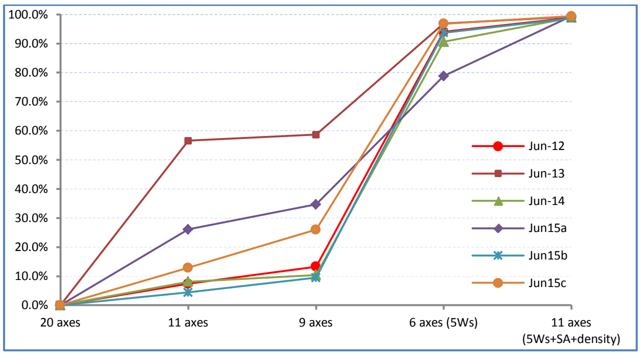
We have calculated the number of polylines in different axes for parallel coordinates, and have found out that the 5Ws density parallel coordinates with SA has significantly reduced the over-crowding of polylines. Table 4 illustrates the details of the number of polylines between different numbers of axes.

Table 4. Polylines for Different Axes

	20 Axes (raw data)	11 Axes	9 Axes	6 Axes (5Ws)	11 Axes (5Ws+SA +density)
Jun12	133,193	123,285	115,472	8,029	1,445
Jun13	275,528	119,580	113,761	8,408	2,113
Jun14	171,380	157,506	153,627	16,192	1,866
Jun15a	101,482	75,030	66,315	21,561	614
Jun15b	94,911	90,778	85,861	6,016	1,227
Jun15c	130,288	113,482	96,454	4,187	765

Fig. 14 demonstrates the percentage of these reductions in the number of polylines. Clearly, the 5Ws parallel coordinate system has reduced data over-crowding by more than 75% in our testing. Furthermore, the 5Ws density pattern with Shrunk Attributes has reduced the data cluttering by nearly 98% based on our density algorithms. This means that the 5Ws density pattern with Shrunk Attributes has very successfully reduced the data processing time for Big Data analysis enormously, and has greatly increased clarity and simplicity of use for Big Data visualisation. We can therefore see the usefulness of the 5Ws density pattern with Shrunk Attributes in Big Data analysis, visualisation and examination.

Figure 14. Reduction of cluttering



To summarise, our data behaviours model illustrates the 5Ws patterns and allows the user to shrink/extend any attribute in a parallel axis for better display and analysis. The five individual densities with additional parallel axes not only reduced data over-crowding, but also measured each data pattern in order to provide more analytical features for Big Data analysis and visualisation.

5 Related works

The most current approaches for Big Data visualisation are practiced on a single dataset. Jinglan (Zhang et al., 2013) analyzed the national bird’s audio dataset, and used time-frequency, tags-linking and GeoFlow as the visualizing techniques for audio data visualisation. Mu-Hsing (Kuo et al., 2014) analysed health datasets and extracted useful knowledge in order to design a pipelined framework of the guidelines in health using Big Data analysis. Xiaotong (Liu et al., 2013) developed a visual search engine based on CompactMap to stream the text data for visual analytics. Seunggwoo (Jeon et al., 2013) transformed unstructured email texts into a graph database and visualized them. Richard and Ralph (Lomotey, R.K. and Dters, R. 2013) extracted the topics and terms from unstructured data by using the TouchR2 tool they created. The tool algorithmically relied on bloom filtering and parallelization.

Researchers have tried to reduce multidimensional data in their visual approaches. Zhenwen (Wang et al., 2013) 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 visualisation. Zhangye (Wang et al., 2013) 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. Rui Maximo (Esteves et al., 2014) introduced competitive K-Means (CK-Means) to improve cluster analysis accuracy. Daniel (Cheng et al., 2013) 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 visualisation.

Parallel coordinates are a popular information visualisation tools for high-dimensional data, which were introduced by Alfred (Inselberg, A. and Dimsdale, B. 1990). Each parallel axis represents a dimension and the polylines were drawn between independent axes at appropriate values.

The data was explored between the axes, showing the data frequencies, the data relationship and the data aggregation patterns. However, it has a severe problem when dealing with large datasets: the polylines clutter and over-crowd each other. Xiaoru (Yuan et al., 2009) scattered points in parallel coordinates to combine the parallel coordinates and scatterplot scaling, which reduced data overcrowding. Matej (Novotny, M. and Hauser, H. 2006) grouped the data context into outliers, trends and focus, and set up three clustered parallel coordinates to reduce the data cluttering issues. Geoffrey (Ellis, G. and Dix, A. 2006) developed three methods: raster algorithm, random algorithm and lines algorithm for measuring occlusion in parallel coordinates plots to provide tractable measurement of the clutter.

To the best our knowledge, no previous work had created 5Ws dimensions and additional density axes in Big Data parallel coordinate visualisation. Common visualisation methods trade each data as a node in visual graphics, and then find visual patterns to analyze the data. We have classified the data dimensions first to obtain the 5Ws patterns and its densities, and then visualized those data patterns. We have found that our visualisation and analytical method has significantly reduced the data cluttering and the data processing time for Big Data analysis and visualisation, while significantly increasing clarity and ease of use without any loss of information or accuracy. Our model has therefore achieved all of our goals in terms of efficiently and effectively presenting Big Data visualisation.

6 Conclusions and future works

The data behaviours model, a novel approach for Big Data analysis and visualisation, has been introduced in this work. We have demonstrated the 5Ws patterns across multiple datasets, established the multiple densities for any form of data and created additional axes in 5Ws 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 in each dimension to be narrowed down or extended for a better visualisation aesthetic. The 5Ws density parallel axes provide a clear view of visual structures and patterns for a better understanding and analysis of Big Data.

The 5Ws densities not only measure Big Data patterns, but also provide 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 more than 75% in our testing. Furthermore, additional 5Ws density axes with shrunk attributes have reduced the data cluttering by almost 98% based on our density algorithm. This has therefore significantly reduced the data processing time for Big Data analysis while improving the quality of Big Data visualisation.

In the future, we plan to develop our model to deploy in more areas and on different datasets such as finance datasets and health datasets. The combination of 5Ws density parallel coordinates and 5Ws treemaps is our next stage for Big Data analysis and visualisation.

References

Afzal, S., Maciejewski, R., Jang, Y., Elmqvist, N. and Ebert, D.S. (2012) ‘Spatial Text Visualization Using Automatic Typographic Maps’, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 18, No. 12, pp. 2556-2564.

Babaee, M., Datcu, M. and Rigoll, G. (2013) ‘Assessment of Dimensionality Reduction Based on Communication Channel Model: Application to Immersive Information Visualization’, in *Proceeding of IEEE International Conference on Big Data, IEEE Big Data ‘13*, pp. 1-6, IEEE Computer Society, Silicon Valley, CA, USA.

Cheng, E., Schretlen, P., Kronenfeild, N., Bozowsky, N. and Wright, W. (2013) ‘Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis’, in *Proceeding of IEEE International Conference on Big Data, IEEE Big Data ‘13*, pp. 2-4, IEEE Computer Society, Silicon Valley, CA, USA.

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

Ellis, G. and 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.

Esteves, R.M., Hacker, T. and Rong, C. (2014) ‘A new approach for accurate distributed cluster analysis for Big Data: competitive K-Means’, *Int. J. Big Data Intelligence*, Vol. 1, Nos. 1/2, pp. 50-64.

Gupta, U. and Fegaras, L., (2013) ‘Map-Based Graph Analysis on MapReduce’, in *Proceeding of IEEE International Conference on Big Data, IEEE Big Data ‘13*, pp. 24-30, IEEE Computer Society, Silicon Valley, CA, USA.

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

Hurwitz, J., Nugent, A., Halper, F. and Kaufman, M., (2013) ‘Big Data for Dummies’, *John Wiley & Sons. Inc.*

Inselberg, A. and Dimnsdale, B. (1990) ‘Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry’, *in Proceeding of First IEEE Conference on Visualization*, pp. 361-378, IEEE Computer Society, San Francisco, CA, USA.

Jeon, S., Khosiawan, Y. and Hong, B. (2013) ‘Making a Graph Database from Unstructured Text’, in *Proceeding of 16t IEEE International Conference on Computational Science and Engineering (CSE)*, pp. 981-988, IEEE Computer Society, Sydney, Australia.

Kuo, M., Sahama, R., Kushniruk, A.W., Borycki, E.M. and Grunwell, D.K. (2014) ‘Health big data analytics: current perspectives, challenges and potential solutions’, *Int. J. Big Data Intelligence*, Vol. 1, Nos. 1/2, pp. 114-126.

Lamboray, E., Wurmlin, S. and Gross, M. (2005) ‘Data Streaming in Telepresence Environments’, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 11, No. 6, pp. 637-648.

Liu, X., Hu, Y., North, S. and Shen, HW. (2013) ‘CompactMap: A Mental Map preserving Visual Interface for Streaming Text Data’, in *Proceeding of IEEE International Conference on Big Data, IEEE Big Data ‘13*, pp. 48-55, IEEE Computer Society, Silicon Valley, CA, USA.

Lomotey, R.K. and Dters, 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, IEEE Computer Society, Sydney, Australia.

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

Malyshkin, V.E. (2014) 'Peculiarities of numerical algorithms parallel implementation for exa-flops multicomputers', *Int. J. Big Data Intelligence*, Vol. 1, Nos. 1/2, pp. 65-73.

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

Novotny, M. and 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.

Pingdom, (2013) 'Internet 2012 in numbers', [online] http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/ (accessed July 2014).

Rocha, A., Carvalho, T., Jelinek, H.F., Goldenstein, S. and 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.

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

Shi, L., Liao, Q., Sun, X., Chen, Y. and Lin, C. (2013) 'Scalable Network Traffic Visualization Using Compressed Graphs', in *Proceeding of IEEE International Conference on Big Data, IEEE Big Data '13*, pp. 606-612, IEEE Computer Society, Silicon Valley, CA, USA.

Shiravi, A., Shiravi, H., Tavallaee, M. and Ghorbani, A.A. (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', [online] http://www.gartner.com/newsroom/id/1731916/ (accessed July 2014).

Wang, Y.S., Wang, C., Lee, T.Y. and 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. and 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 Big Data '13*, pp. 25-32, IEEE Computer Society, Silicon Valley, CA, USA.

Wang, Z., Zhou, J., Chen, W., Chen, C., Liao, J. and Maciejewski, R. (2013) 'A Novel Visual analytics Approach for Clustering Large-Scale Social Data', in *Proceeding of IEEE International Conference on Big Data, IEEE Big Data '13*, pp. 79-86, IEEE Computer Society, Silicon Valley, CA, USA.

Wu, C., Yang, H., Zhu, J., Zhang, J., King, I. and Lyu, M.R. (2013) 'Sparse Poisson Coding for High Dimensional Document Clustering', in *Proceeding of IEEE International Conference on Big Data, IEEE Big Data '13*, pp. 512-517, IEEE Computer Society, Silicon Valley, CA, USA.

Wu, J. and Hong, B. (2014) 'Multi-source streaming-based data access for MapReduce systems', *Int. J. Big Data Intelligence*, Vol. 1, Nos. 1/2, pp. 36-49.

Yuan, X., Guo, P., Xiao, H., Zhou, H. and 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, M.L., Wang, W.B., Lu, L.F., and Meng, ZP. (2014) 'Big Data Density Analytics Using Parallel Coordinate Visualization', in Proceeding of *17th IEEE International Conference on Computational Science and Engineering (CSE)*, pp. 1115-1120, IEEE Computer Society, Chengdu, China.

Zhang, J., Huang, K., Cottman-Fields, M., Truskinger, A., Roe, P., Duan, S., Dong, X., Towsey, M. and 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, Dec 2013, IEEE Computer Society, Sydney, Australia.