Visual Analytics for BigData Variety and Its Behaviours

Article *in* Computer Science and Information Systems · December 2015

Impact Factor: 0.48 · DOI: 10.2298/CSIS141122050Z

3 authors, including:

Jinson Zhang

University of Technology Sydney

17 PUBLICATIONS **33** CITATIONS

Computer Science and Information Systems 12(4):1171–1191 DOI: 10.2298/CSIS141122050Z

Visual Analytics for BigData Variety and Its Behaviours

Jinson Zhang[1], Mao Lin Huang[2,1], Zhao-Peng Meng[2]

[1] School of Software, Faculty of FEIT,
University of Technology Sydney,Australia
{Jinson.Zhang, Mao.Huang}@uts.edu.au
[2] School of Computer Software, Tianjin University,
Tianjin, China
{mlhuang, mengzp}@tju.edu.cn

Abstract. BigData, defined as structured and unstructured data containing images, videos, texts, audio and other forms of data collected from multiple datasets, is too big, too complex and moves too fast to analyze using traditional methods. This has given rise to a few issues that must be addressed; 1) how to analyze BigData across multiple datasets, 2) how to classify the different data forms, 3) how to identify BigData patterns based on its behaviours, 4) how to visualize BigData attributes in order to gain a better understanding of data. It is therefore necessary to establish a new framework for BigData analysis and visualization. In this paper, we have extended our previous works for classifying the BigData attributes into the „5Ws‟ dimensions based on different data behaviours. Our approach not only classifies BigData attributes for different data forms across multiple datasets, but also establishes the „5Ws‟ densities to represent the characteristics of data flow patterns. We use additional non-dimensional parallel axes in parallel coordinates to display the „5Ws‟ sending and receiving densities, which provide more analytic features for BigData analysis. The experiment shows that our approach with parallel coordinate visualization can be efficiently used for BigData analysis and visualization.

Keywords: BigData, 5Ws dimensions, data visualization, parallel coordinate.

1. Introduction

Data visualization techniques have been widely used to analyze and explore data to allow users to interactively select and scale down the scope of view for a better understanding of data. With the rise of BigData, new analytical methods for BigData visualization have to be developed to handle very large and complex datasets in a short period of time. According to Gartner‟s 3Vs definition in 2011 [1], BigData has three main characteristics: Volume, Velocity and Variety. Dominik Klein et al [2] illustrated the 3Vs characteristics in a visual graph, shown in Fig 1.

Volume describes how BigData datasets are extremely large and easily reached terabytes, even yottabytes, of information. This large volume of data is not only a storage issue, but also a massive analysis issue if we continue to use traditional visual analytics techniques.

Velocity describes how fast the dataset is produced. According to Pingdom 2012 [3], in 2012 there were more than 1.7 million emails sent per second, 57,870 Google

searches per second, 7 petabytes of photo content added on Facebook every month, 4 billion hours of video watched on YouTube every month, and 5 billion mobile phone users used 1.3 Exabytes of global mobile data traffic per month.

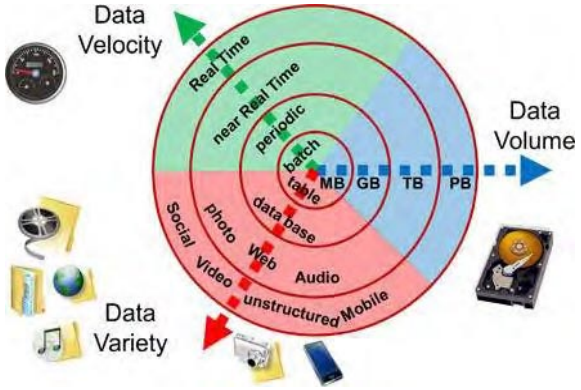

Fig. 1. Dominik Klein et al [2] illustrated BigData"s 3Vs characteristics

Variety describes how BigData datasets contain both structured and unstructured data, such as documents, emails, audio files, images and videos data, Internet click streams, mobile data or network log files. Hundreds, even thousands, of different attributes in multiple dimensions provide too much information for traditional visualization tools to handle.

Later researchers have since added Value and Veracity into BigData characters, which results in the 5Vs of BigData. Yuri Demchenko et al [4] created a graph to display these 5Vs characteristics, shown in Fig 2.

BigData comes from everywhere in our life, and so is too big, too complex and moves too fast for traditional visualization tools. For example, posting pictures and comments on Facebook; uploading and watching videos on YouTube; sending and receiving messages through smart phones and spreading a virus through the Internet all count as BigData collected by different datasets.

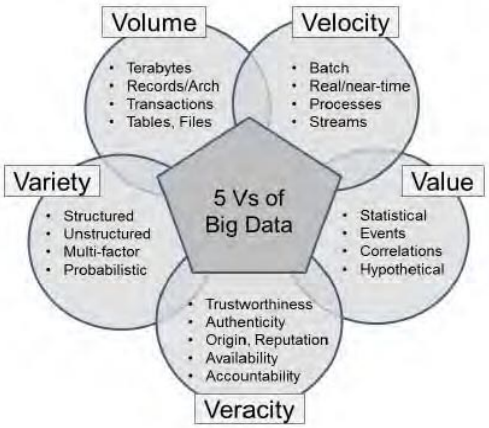

Fig. 2. Yuri Demchenko et al [4] demonstrated BigData"s 5Vs characteristics

The most current approaches for BigData visualization are practiced on a single form of dataset. Jinglan Zhang et al [5] analyzed the national bird‟s audio dataset, and used time-frequency, tags-linking and GeoFlow as the visualizing techniques for BigData audio data visualization. Seunggwoo Jeon et al [6] transformed unstructured email texts into a graph database. Yinglong Xia et al [7] visualized large scale graph data using "System G", the graph processing system they developed. Ryan Compton et al [8] visualized a Twitter dataset containing one hundred million users using a geolocation method.

To the best of our knowledge, no previous work has addressed BigData visualization by combining multiple datasets and different data forms. In this paper, we have classified BigData attributes into the „5Ws" dimensions based on the data behaviours. Each data incident contains these 5Ws dimensions, can be applied for multiple datasets across different data forms. This provides more analytical and measurement features for business, government and organizational needs.

Based on the data behaviors, we have extended our previous works [9] [33], and further developed our visual analytics method by using parallel coordinate visualization techniques for BigData analysis and visualization. Firstly, we analyzed the BigData attributes for multiple datasets and introduced the subsets measuring the BigData flow patterns. Secondly, we established the 5Ws sending density and receiving density to measure BigData flow patterns across multiple datasets. Thirdly, we created the 5Ws density parallel axes to illustrate the 5Ws patterns in parallel coordinates for BigData visualization.

The paper is organized as follows; Section 2 introduces our 5Ws dimension and density approach. Section 3 demonstrates the 5Ws density parallel coordinates visualization. Section 4 explains the implementation of this model. Section 5 describes related works, and Section 6summarises our approach.

2. 5Ws Dimension and Density

Each data incident, based on its behaviours, contains the 5Ws dimensions; What does the data contain, Why did the data occur, Where did the data come from, When did the data occur, Who received the data and How was the data transferred.

2.1. 5Ws Dimension

The 5Ws dimensions can be illustrated by using six data sets, each set demonstrating a dimension.

- A set $T=\{t_1, t_2, t_j, \ldots, t_m\}$ represents when the data occurred
- A set $X=\{x_1, x_2, x_j, \ldots, x_m\}$ represents what the data contained
- A set $Y=\{y_1, y_2, y_j, \ldots, y_m\}$ represents how the data was transferred
- A set $Z=\{z_1, z_2, z_j, \ldots, z_m\}$ represents why the data occurred
- A set $P=\{p_1, p_2, p_j, \ldots, p_m\}$ represents where the data came from
- A set $Q=\{q_1, q_2, q_j, \ldots, q_m\}$ represents who received the data

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

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

where

- $t \mid T\{\}$ represents the time stamp for each data incidence.
- $x \mid X\{\}$ represents what the data contained, such as "email" or "Facebook comment".
- $y \mid Y\{\}$ represents how the data was transferred, such as "by Internet", "by smart phone" or "by email".
- $z \mid Z\{\}$ represents why the data occurred, such as "sharing photos", "sending message to friends" or "spreading a virus".
- $p \mid P\{\}$ represent where the data came from, such as "Twitter", "Facebook" or "hacker".
- $q \mid Q\{\}$ represents who received the data, such as "friend", "bank account" or "victim".

Therefore, each data incident can be illustrated as a 5Ws pattern, which flows from p to q and contains patterns x, y, z in time slot t, shown in Fig 3.

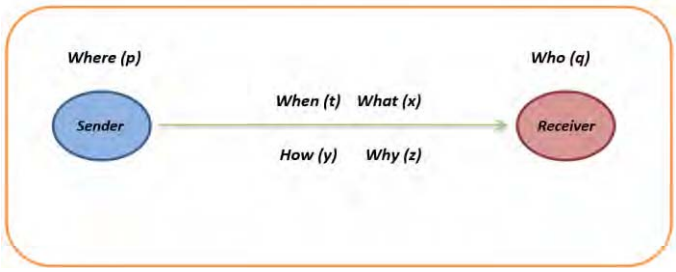

Fig. 3. 5Ws pattern that illustrated for each data incident

Suppose that in the time slot T there were n number of incidences, represented as a set Fsuch that

$$F = \{f_1, f_2, f_3, \ldots, f_n\} . \quad (1)$$

F therefore contains all incident nodes within a certain time period. For example, there were 9.66 million tweets during the Opening Ceremony of the London 2012 Olympic Games [3]. The twitter dataset for the Opening Ceremony is therefore $|F|$ = 9.66 million.

2.2. 5Ws Dimension crossing Multiple Datasets

5Ws data flow patterns are not only for one dataset, but can also be used to compare multiple datasets. For example, a Facebook dataset and an Internet banking transaction dataset are two different datasets. But similar attributes exist on both datasets, such as p = "users", y = "mobile connection". Therefore, the comparison between these two

datasets for p = "users" and y = "mobile connection" will export the ratio of Internet banking mobile users to Facebook mobile users.

Fig 4 shows the BigData classification across multiple datasets for different data forms using 5Ws dimensions. The left hand side shows multiple BigData datasets in different data forms. The right hand side illustrates these 5Ws dimensions by using the six datasets.

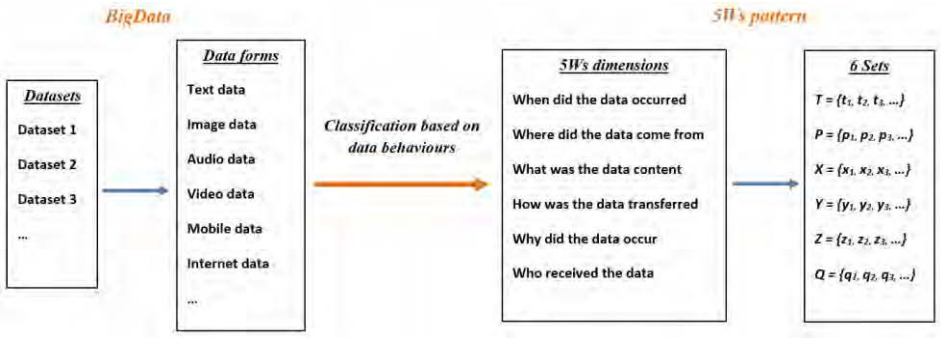

Fig. 4. BigData classification in 5Ws dimension crossing multiple datasets and data forms

Fig 5 shows a few examples of BigData in the 5Ws dimensions across multiple datasets and resources, such as social network datasets, finance datasets or email datasets.

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

Fig. 5. Example of BigData in 5Ws dimension crossing multiple datasets and resources

2.3. 5Ws Dimension Subsets

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

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

The subset $F_{(\alpha, \beta, \gamma, \delta, \varepsilon)}$ represents the particular incident pattern by the 5Ws data dimensions. For example, during the Opening Ceremony of the London 2012 Olympic Games, 9.66 million tweets were recorded containing multiple patterns such as α= "London" + "Olympics" + "Opening" + "Ceremony", β="sent or received", γ ="sharing opening ceremony" or "enjoying ceremony", δ = twitter, ε = users and t = 27-Jul-2012, 21:00 – 00:45.

The dataset $|F|$ illustrates the statistical results of both volume and velocity. The subset $F_{(\alpha, \beta, \gamma, \delta, \varepsilon)}$ demonstrates the variety for the particular incident pattern which we will focus on in this paper.

Each sender may send data to multiple receivers with the same or different α, β and γ. Each receiver may also receive data from multiple senders with the same or different α, β and γ. This results in a huge number of combinations and varieties between multiple senders and receivers. We therefore use the 5Ws data densities to measure the proportion of particular combinations relative to the entire set of variations.

2.4. Sending Density

We firstly establish sending density (SD) to measure a particular sender"s pattern during data transferal as a proportion of the entire dataset. Based on equation (2), the sending density for particular attributes $x=\alpha$, $y=\beta$, $z=\gamma$, and sender $p=\delta$, in time slot T, is defined as $SD_{(\alpha, \beta, \gamma, \delta)}$

$$SD_{(\alpha, \beta, \gamma, \delta)} = \frac{|F(\alpha,\beta,\gamma,\delta)|}{|F|} = \frac{1}{n}\sum_{i=1}^{n} f_{(i)}\left(t, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p_{(\delta)}, q\right) . \quad (3)$$

where $0 \le SD_{(\alpha, \beta, \gamma, \delta)} \le 1$.

$SD_{(\alpha, \beta, \gamma, \delta)}$ therefore represents the 5Ws dimensions for the particular sender"s pattern; the content was α, transferred by β, in $t \subset T$ time, for reason γ and sent by δ. A high value of $SD_{(\alpha, \beta, \gamma, \delta)}$ indicates that sender (δ) sent the most data compared to other senders.

2.5. Receiving Density

Similarly, the receiving density for $x=\alpha$, $y=\beta$, $z=\gamma$ and receiver $q=\varepsilon$, is defined as $RD_{(\alpha, \beta, \gamma, \varepsilon)}$

$$RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)} = \frac{|F(\alpha,\beta,\gamma,\varepsilon)|}{|F|} = \frac{1}{n}\sum_{i=1}^{n} f_{(i)}\left(t, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p, q_{(\varepsilon)}\right). \quad (4)$$

where $0 \le RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)} \le 1$

$RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$therefore represents the 5Ws dimensions for the particular receiver"s pattern; the contents was α, transferred byβ, in $t \subset T$ time, for reason γ and received by ε. A high value of $RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$ indicates that the receiver (ε) received the most data compared to other receivers.

2.6. Noise Data

Noise data is defined as the nodes with unknown or undefined attributes, such as x = *unknown_x*, y = *unknown_y*, z = *unknown_z*, p = *unknown_p* and q = *unknown_q*. A subset for any unknown nodes can be defined as

$$F_{(unknown)} = \{f \in F \mid f(t, x, y, z, p, q), x=unknown\_x, y=unknown\_y, z=unknown\_z, p=unknown\_p, q=unknown\_q\}. \quad (5)$$

Noise data should be excluded from the $SD_{(\alpha,\ \beta,\ \gamma,\ \delta)}$and $RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$ as it excludes immeasurable nodes. In equation (3), $SD_{(\ )}$would be then re-defined as

$$SD_{(\alpha,\ \beta,\ \gamma,\ \delta)} = \frac{|F(\alpha,\beta,\gamma,\delta)|}{|F| - |F(unknown\ )|}. \quad (6)$$

In equation (4), $RD_{(\ )}$would be then re-defined as

$$RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)} = \frac{|F(\alpha,\beta,\gamma,\varepsilon)|}{|F| - |F(unknown\ )|}. \quad (7)$$

By removing noise data, both $SD_{(\alpha,\ \beta,\ \gamma,\ \delta)}$and $RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$ represent the sender"s and receiver"s pattern with significantly greater accuracy for BigData analysis.

2.7. Comparing SD and RD

If $SD_{(\alpha,\ \beta,\ \gamma,\ \delta)} > RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$, this indicates that the data flow patterns from the sender (δ) are greater than the data flow to the receiver (ε), as shown as Fig 6a). Therefore, the receiver (ε) is only one target of the sender (δ).

If $SD_{(\alpha,\ \beta,\ \gamma,\ \delta)} < RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$, it indicates that the data flow patterns from the sender (δ) are less than the data flow to the receiver (ε), as shown as Fig 6b). Therefore, the sender (δ) is only one data source of the receiver (ε).

When $SD_{(\alpha,\ \beta,\ \gamma,\ \delta)} = RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$, it indicates the data flow densities are the same from the sender (δ) to the receiver (ε). This suggests that there is only one sender (δ) and one receiver (ε).

The 5Ws density can also be used for comparing between different attributes such as α_1 = "USA team" and α_2 = "UK team", or β_1 = "iPhone apps" and β_2 ="Android apps". This provides more analytic features and visual graphs for BigData visual analytics.

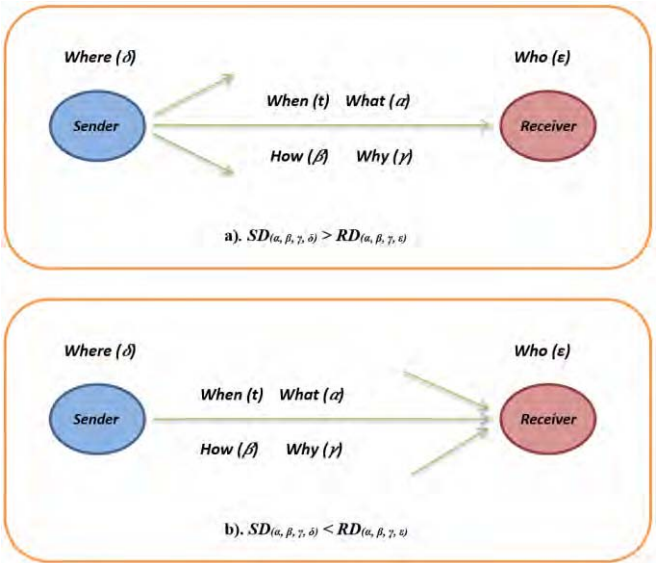

Fig. 6. Example of $SD_{(\alpha, \beta, \gamma, \delta)}$ via $RD_{(\alpha, \beta, \gamma, \varepsilon)}$

In summary, we have classified BigData into 5Ws dimensions based on data behaviours, created sending density $SD_{(\ )}$ and receiving density $RD_{(\ )}$ to measure data flow patterns, and removed noise data to increase the accuracy of BigData analysis.

3. 5Ws Density Parallel Coordinates

Parallel coordinates are a popular information visualization tools for multi-dimensional data [10]. It draws polylines between each axis at the appropriate values, and the graph explores data between the axes, showing data frequencies instead of individual data points. Large data can therefore be aggregated for visualization, up to a million pieces of data for 50 dimensions [11]. For example, Melanie Tory et al [12] proposed parallel coordinates for exploratory volume visualization, Aritra Dasgupta and Robert Kosara [13] adapted the parallel coordinates for the German credit dataset.

However, the parallel coordinates used to represent BigData dimensions have limitations because the size of the display screen is limited and polylines often clutter very easily. To implement parallel coordinates efficiently, we used the 5Ws dimensions and their densities as the parallel axes to visualize BigData. This not only reduced data cluttering, but also improved the measurement on the parallel coordinates for BigData visualization.

3.1. 5Ws Parallel Axes

Based on our BigData 5Ws flow pattern, we define the five parallel axes to represent the 5Ws data dimensions for parallel coordinate visualization.

- Left axis is P, which represents where the data came from
- Second axis is X, which represents what the data contained
- Third axis is Y, which represents how the data was transferred
- Fourth axis is Z, which represents why the data occurred
- Right axis is Q, which represents who received the data.

The value of each axis is ordered by alphabetical order, ranging from 0 to 9, *A* to *Z* and *a* to *z*.

We will use the 2012 London Olympic Opening Ceremony as an example to illustrate the 5Ws parallel axes, shown in Fig 7. Assume that during 2012 London Olympics Opening Ceremony, Facebook, Twitter, YouTube received huge data transactions through different smartphones, and containing different data contents from different locations around the world.

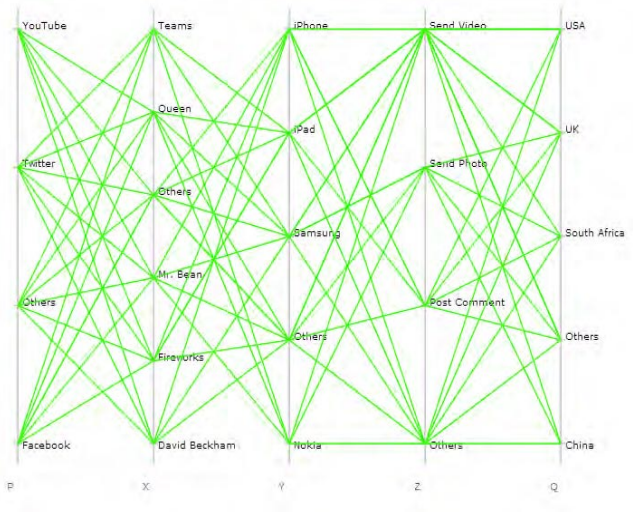

Fig. 7. Example of 5Ws axes in parallel coordinates

In Fig 7, P={Facebook, Twitter, YouTube} means that the data came from those datasets, X={David Beckham, Fireworks, Mr. Bean, Queen, Teams} represents what the data contained, Y={Nokia, Samsung, iPad, iPhone} indicates that the data was transferred via these smartphones, Z={Post Comment, Send Video, Send Photo} shows why the data occurred, and Q={China, South Africa, UK, USA} illustrates which countries received the data.

In each axis, the attribute {Others} represented all data attributes that are not represented by the nodes in each parallel axes. The 5Ws parallel coordinates illustrate the aggregation patterns between axes and so clearly show relationships between the senders and receivers.

3.2. Extension and Contraction of Attributes

The attributes inside {Others} can be extended in the 5Ws parallel axes by assigning particular attributes. For example, based on Fig 7, we can retrieve the "Message" dataset from the {Others} dataset and add it to the P axis. We can also add "007" attribute for X, add "LG" mobile phone for Y, and add country "Australia" for Q. The graph of the 5Ws parallel coordinates has therefore been extended, as per Fig 8.

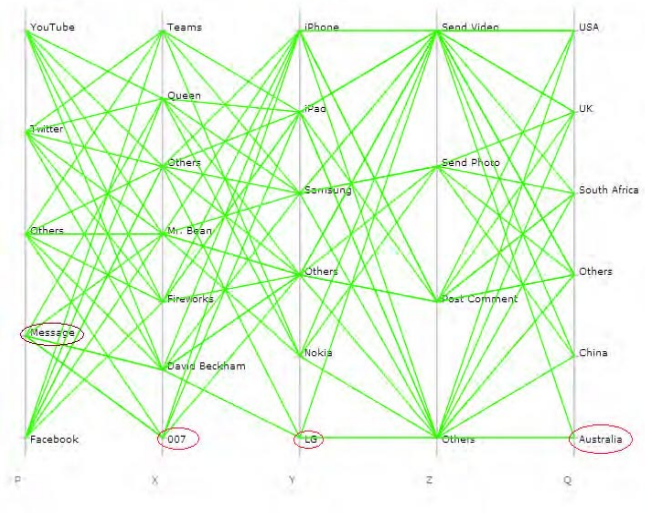

Fig. 8. Example of extended attributes in 5Ws axes in parallel coordinates

The extension of attributes in the visualization process in the 5Ws parallel axes occurs when moving from Fig 7 to Fig 8, while the contraction of attributes occurs when moving form Fig 8 to Fig 7. This extension and contraction enables BigData analysis and visualization to be very efficient as it can narrow down or extend particular attributes as required for each dimension.

3.3. Clustering 5Ws Parallel Axes

The 5Ws parallel axes can be clustered to provide a classical relationship for a particular dimension in BigData analysis and visualization. We will use Fig 8 as the example to demonstrate clustered parallel axes. We will add additional axes P1, X1, Y1, Z1 and Q1, which are defined as the clustered axes for each 5Ws axis, as shown in Fig 9.

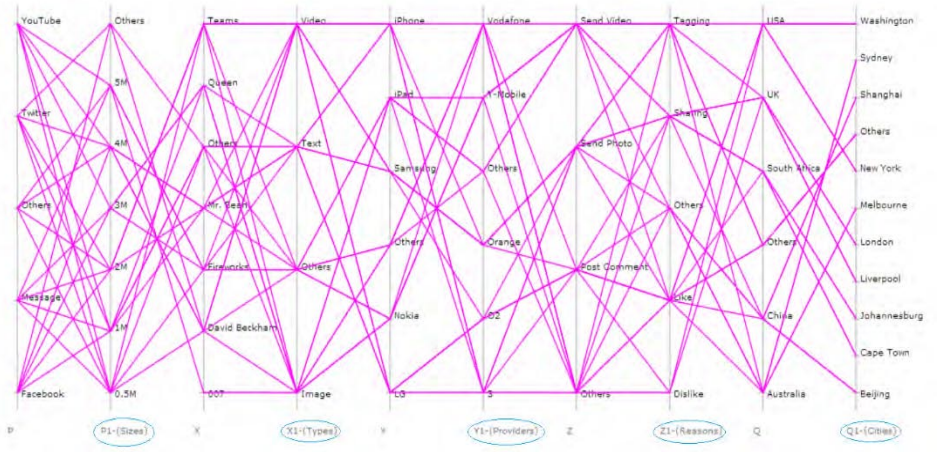

Fig. 9. Example of clustered 5Ws axes in parallel coordinates. P1 – the size of the data clustered from the sender P, measured in megabytes. X1 – the data types for the contents X, classified as "video", "audio", "text" and "other" forms of data. Y1 – the data providers used by the data-transferring devices in Y, including "Vodafone", "T-Mobile", "Orange", "O2", "3" and "others". Z1 – the clustered reasons why the data occurred in Z. Q1 – the cities within the country Q that received the data.

Fig 9 shows an example of a 5Ws clustered parallel axes. P1 indicates the sizes of the data coming from different senders P, such as "YouTube" or "Twitter". It demonstrates the aggregation patterns between datasets and data sizes. In X and X1, the relationships between data types and data contents are explored. The relationships between the data providers Y1 and data transferral mechanisms Y and the reasons of the data occurring Z and Z1 are also shown in the 5Ws clustered parallel coordinates.

3.4. Order of 5Ws Density Parallel Axes

We have now created density parallel axes such as the $SD_{(\alpha, \beta, \gamma, \delta)}$ and R $RD_{(\alpha, \beta, \gamma, \varepsilon)}$ axes to measure and visualize these 5Ws data flow patterns for particular data attributes in our 5Ws model. The SD is assigned as the first axis, and the RD as the last axis, as shown in Fig 10.

In Fig 10, each value in the SD and RD axes represents a subset for the particular attribute. Each polyline between different 5Ws dimensions demonstrates a particular flow pattern. This reduces line cluttering in the graph as one subset of attributes only has one polyline, and significantly reduces overlap for the parallel coordinates. The 5Ws density parallel axes, combined with the alphabetical axes and numerical axes, therefore provide a more analytical method for BigData visualization. Furthermore, no information has been lost during the analysis and visualization process.

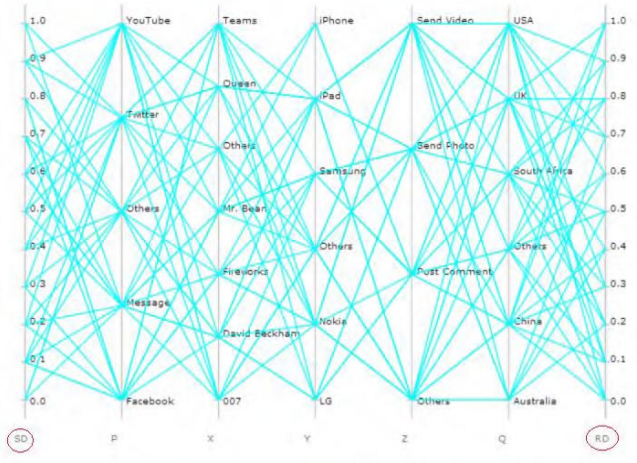

Fig. 10. Example of 5Ws density axes in parallel coordinates

3.5. Re-Ordering 5Ws Density Parallel Axes

The data with the highest contribution rate is always chosen as the first parallel axis because it more easily attracts the user"s attention [14]. Our 5Ws density parallel coordinate system has two density axes on either side of the graph that measure the data flow patterns. By re-ordering the 5Ws parallel axes, we can graphically illustrate the relationship between attributes by placing the desired axis close to the density axes.

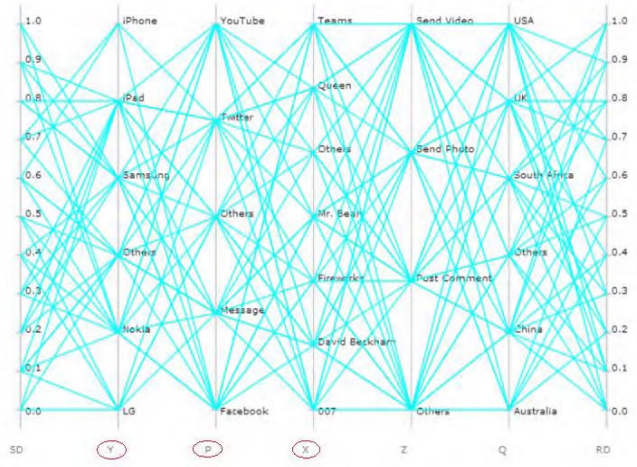

Fig. 11. Example of re-order of 5Ws density axes

For example, by having the Y axis next to the SD axis, we can easily demonstrate the data patterns between the sending density and the devices transferring the data. Fig 11 shows the 5Ws density parallel axes re-ordered on Y, P and X axes. This provides a

clear view of the visual structures and patterns between the SD and Y axes, creating better BigData visualisation and analysis which is much easier to read.

In summary, the 5Ws parallel axes clearly illustrate the 5Ws dimensions in order to provide a better view to understand data flow patterns. By creating two non-dimensional axes using 5Ws densities, we enhanced the measurement and comparison of BigData patterns. The 5Ws density parallel coordinates clearly demonstrate data patterns for different datasets and different data attributes. By extending and contracting attributes and utilising re-ordering and clustering of density parallel axes, the 5Ws parallel axes visualization system has significantly improved BigData analysis and visualization.

4. Implementation

The 5Ws density parallel coordinates model has been tested and evaluated by using the ISCX2012 dataset [15]. The summary of a dataset is shown as Table 1.

Table 1. ISCX2012 dataset – TestbedTueJun15c

Name	Amount
Network traffic nodes	130288
ICMP traffics	31
TCP traffics	119242
UDP traffics	11015
Unknown TCP traffics	3
Unknown UDP traffics	36
Attacks	37375
Source IPs	36
Source ports	23653
Destination IPs	1656
Destination ports	222
Application Names	19

The dataset contains 130288 incidents and 20 dimensions that indicates the total traffics are 130,288 ($|F|=130,288$) including 37,375 attacks, 36 sources IPs, 23,653 source ports, 1656 destinations IPs and 222 destination ports.

4.1. 5Ws Density Parallel Coordinates

The P axis represents the sources IPs, which represents where the data came from. 36 sources IPs have been assigned for the P axis. P= “0.0.0.0” means the source address is invalid.

The Q axis represents the destination IPs, which indicates who received the data. 1656 destinations IPs have been assigned for the Q axis. Q= “0.0.0.0” indicates the destination address is invalid. To save space on the Q axis, the axis has been shrunk by using the first 8 bits of the destination address only. For example, for normal traffics,

Q="111.222.222.222" will be shrink to "111.xxx.xxx.xxx". That means Q= "111.xxx.xxx.xxx" includes destination IPs for the range of {111. 1-255. 1-255. 1-255}.

The X axis represents the data content, including "Normal" traffics, "Attack" traffics and "Unknown" traffics. The Y axis represents the applications describing how the data was transferred, such as "FTP" transfer, or "HTTPWeb" transfer. The Z axis represents the protocol illustrating why the data occurred, including "UDP", "TCP" or "ICMP" connections.

SD is assigned as the first axis adjacent to the P axis, measuring the data patterns coming from the 36 sources IPs. SD ≈ 0.000 indicates SD <0.001, which means that $(|F_{(\alpha, \beta, \gamma, \delta)}| / |F|) < 0.001$, or $|F_{(\alpha, \beta, \gamma, \delta)}| < 0.001 \times |F|$. Using the data provided, this means that $|F_{(\alpha, \beta, \gamma, \delta)}| < 130$. The scale of the P axis means that sent data patterns < 130 will be drawn as SD (0.000).

RD is assigned as the last axis adjacent to Q, which illustrates the data patterns received by 1656 destination IPs. Received data patterns < 130 will be drawn as RD (0.000), and Q will be shrunk as ".xxx.xxx.xxx".

Fig 12 shows the test results using the ISCX2012 dataset. The SD patterns drawn to the P axis clearly illustrate the 36 sources IPs that are linked to the X axis, of which "Normal" and "Attack" are the major attributes. We can also see that the RD patterns linked to the Q axis are mostly connected by the "tcp_ip" protocol in the Z axis.

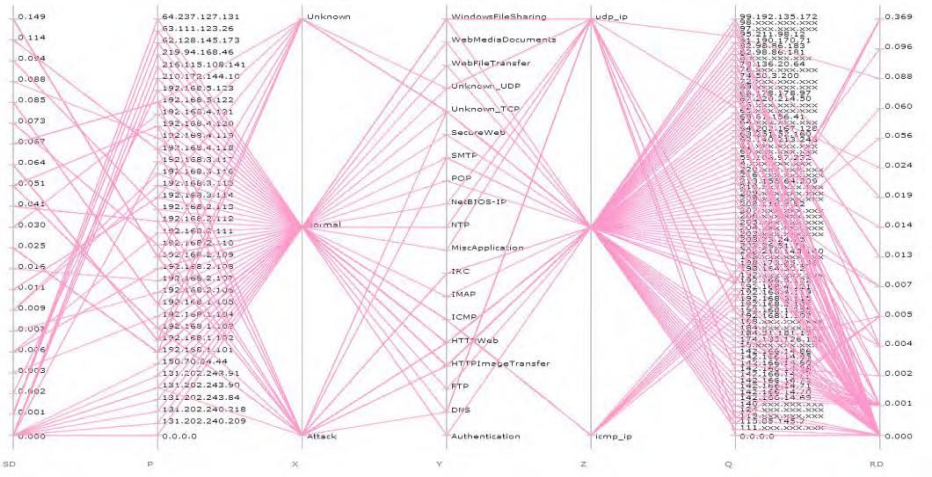

Fig. 12. 5Ws density parallel coordinates for dataset ISCX2012 TueJun15c

In Fig 12, it illustrates how the "Attack" in the X axis was sent from 10 source IPs in the P axis, and so most SD values for these attacks are above 0.050. The data has been transferred by 6 different methods; "DNS", "IRC", "SMTP", "HTTPImageTranfer", "Unknown_TCP" and "HTTPWeb". The X axis is positioned after the P axis and before the Y axis, as this clearly demonstrates that the SD patterns for the particular attribute X came from the P axis and were transferred by the methods on the Y axis.

Fig 12 also illustrates that the Y axis contains 19 applications describing how the data was transferred. 18 of 19 applications have transferred "Normal" data from the X axis, excluding IRC. 15 of 19 of applications use "tcp_ip" as the connecting method on the Z axis.

Our test therefore illustrates how the 5Ws density parallel axes visualization technique can be used effectively to process complex volumes of data, as well as effectively and efficiently illustrates this data is in easy to read and analyse graph. Therefore, the 5Ws density parallel axes visualization model effectively allows for much easier, faster and complete data analysis and visualization.

4.2. Re-Ordering 5Ws Density Parallel Axes

The SD patterns linked to the X axis are demonstrated in Fig 12. In the second part of our implementation, we have swapped the X and Z axes to re-order the 5Ws density parallel axes, shown in Fig 13.

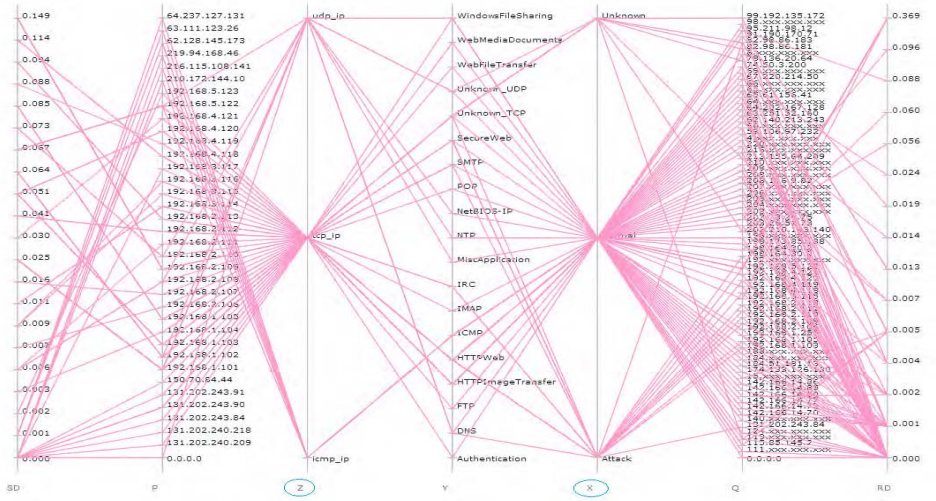

Fig. 13. Re-order of 5Ws density parallel coordinates for dataset ISCX2012 TueJun15c

In Fig 13, there are 11 destination IPs in the Q axis that received the data content "Attack". Most values of RD are below 0.001. 7 destination IPs received "Unknown" content, at the value of RD < 0.001, which was transferred by "Unknown_TCP" and "Unknown_UDP" in the Y axis.

4.3. Clustering 5Ws Density Parallel Axes

To study the patterns between the destination ports and the X axis, we previously defined the Q1axis for the clustered Q axis, which has 222 destination ports. To save space on the Q1 axis, the ports > 1000 have been shrunk unless RD > 0.001. The "Port 1xxx-1xxxx" represents Q1={1000 – 19999} and "Port 2xxxx-6xxxx" for Q1={20000 – 69999}. The clustered 5Ws density parallel coordinates is shown as Fig 14.

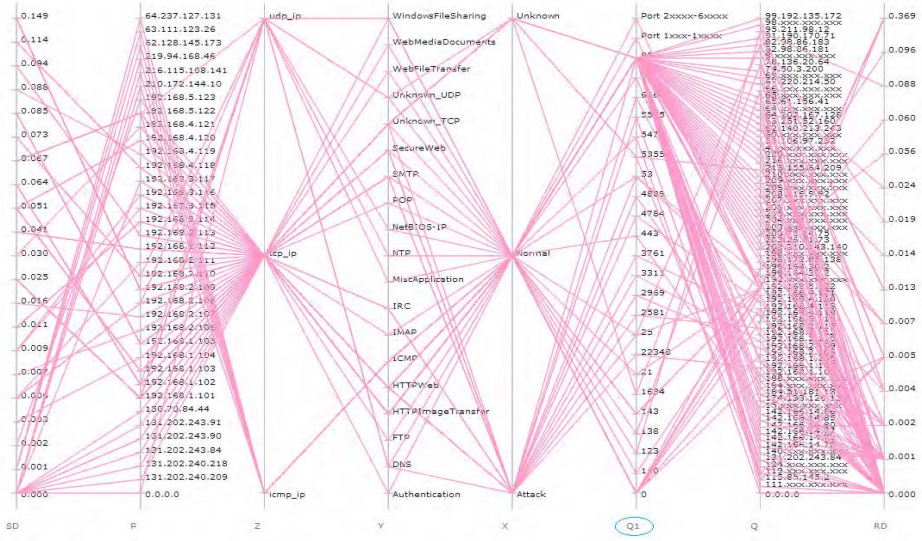

Fig. 14. Clustered 5Ws density parallel coordinates for dataset ISCX2012 TueJun15c

In Fig 14, we can deduce that there are 12 destination ports in the Q1 axis that received the data content "Attack", linked to 11 destination IPs in the Q axis. Furthermore, RD = 0.369 indicates that the destination IP "192.168.5.122" received "Attack" and "Normal" on port "80", and destination IP "82.98.86.183" on the Q axis received "Attack" and "Normal" on port "25" at value RD = 0.013. Three destination ports "80", "1694" and "5355" received the data contents "Unknown". The port "80" in the Q1 axis was connected to the most destination IPs on the Q axis. Therefore, we see the benefit of density parallel axes in visualizing and analysing BigData, as it is quicker, easier and clearer than conventional visualization techniques.

4.4. Reduction of Data over Plotting

Before we classified the data into 5Ws dimensions, the dataset has 130,288 data nodes within 20 dimensions. After applying 5Ws patterns, the data has been reduced to 4,178 nodes in 5Ws dimensions without the loss of any information. The details of the data nodes in the different dimensions (axes) are shown in Table 2.

Table 2. The data nodes for different axes

	20 (axes) Dimensions	11 (axes) Dimensions	9 (axes) Dimensions	5Ws (axes) Dimensions
Jun15c	130,288	113,482	96,454	4,187

The data cluttering hasbeen reduced more than 95% by using 5Ws pattern which is significantly improved. Fig 15 shows the percentage of cluttering reduction.

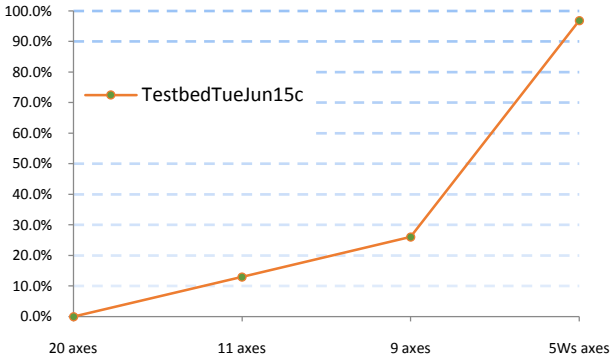

Fig. 15. Reduction in the different axes

To summarise, our experimental results in our 5Ws density parallel coordinates not only clearly illustrate the patterns of BigData in a clearly, easily accessible and accurate way, but also allows the user to shrink/extend any attribute for better display and analysis. The 5Ws density parallel axes re-ordering and clustering provide more details in analytics features and allow the user to clearly grasp links between dimensions. By introducing moveable axes, the 5Ws density parallel coordinates allows for a greater range of analysis and visualisation, improving the versatility and breadth of analysis while still maintaining depth and accuracy in the information it presents. The data cluttering has been significantly reduced up to 95% in our approach.

5. Related Works

Currently, BigData visualization has two main practices: dataset visualization and data form visualization. Dataset visualization focuses on a particular dataset during the visual algorithm progress, such as medical dataset visualization [16], social network dataset visualization [17], weather dataset visualization [18], or spam email dataset visualization [19]. On the other hand, data form visualization targets a particular form of data for visual analytics, such as the text data visualization [20], solar data visualization [21], the audio data visualization [22], the network data visualization [23], or the video data visualization [24].

BigData containing text, image, audio, video and other forms of data is too big and too complex to analyze using traditional methods. Researchers have tried using different visualization techniques to handle those challenges in their visual approaches. These include Bubble Plots to display visual area; Heatmaps for coloring; Histograms for aggregating; Clustering to group similar attributes, and Parallel Coordinates to illustrate multiple dimensional data.

Zhangye Wang et al [25] used Bubble Plots to cluster large-scale social data into user groups by using user tag and user behaviourinformation. Daniel Cheng et al [26] explored one billion pieces of Twitter for BigData visualization by using Heatmaps in their Tile-Based Visual Analytics (TBVA), which created tiled heat maps and tiled density strips data. Fangzhou Shen et al [27] used Histograms in their visual analytics to measure transfer functions. Zhenwen Wang et al [28] used a Clustering method to group

the same attribute value nodes, and then created virtual nodes to group the same attribute value nodes together. The different groups are separated by different colours in the visualization process. Cheng-Long Ma et al [29] used the K-means clustering method to find out the clustering centers for 3-D visualization. The distances between the three coordinate axes correspond to the data in the original space.

Parallel coordinates has been widely used for multidimensional data visualization since it was introduced by Alfred Inselberg and Bernard Dimsdale [10]. It draws polylines between independent axes at appropriate values, where each axis represents a dimension. The data explored between the axes shows the data frequencies, the data relationship and the data aggregation patterns. However, parallel coordinates faces a major problem when dealing with large datasets, as the polylines clutter and crowd each other. Xiaoru Yuan et al [30] scattered points in parallel coordinates to combine the parallel coordinates and scatterplot scaling, which reduced data crowding. Liang Fu Lu et al [14] proposed a similarity-based method to re-order the parallel axes and enable the reduction of the clutter. Geoffrey Ellis and Alan Dix [31] developed three methods: raster algorithm, random algorithm and lines algorithm for measuring occlusion in parallel coordinates plots to provide tractable measurement of the clutter. Matej Novotny and Helwig Hauser [32] grouped the data context into outliers, trends and focus, and set up three clustered parallel coordinates to reduce cluttering issues.

6. Conclusions

This work has focused on the variety of BigData and its behaviours. We have defined and studied the 5Ws dimensions for different data forms across multiple datasets. The Sending Density and Receiving Density have been established to measure BigData patterns without the loss of any information. These densities not only measure BigData patterns, but also enable comparisons between these dimensions. This provides more analytical features to feed business, government and organizational requirements.

The 5Ws parallel coordinates allows us to extend or shrinkthe data attributes whilstkeeping all data on each axis. This reduces data over-plotting and cluttering, therefore creating a simpler, easier to analyse graph that allows for faster and more accessible analysis. The clustering and reorderingof the 5Ws parallel axes enables a clearer and better understanding of BigData analysis and visualization, which has significantly improved the accuracy and ease of finding BigData patterns.

References

1. Stamford.: Gartner Says Solving „Big Data“ Challenge Involves More Than Just Managing Volumes of Data (2011). [Online]. Available:http://www.gartner.com/newsroom/id/1731916 (posted on June 27, 2011)
2. Pingdom,: Internet 2012 in numbers. (2012) [Online]. Available: http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/ (posted on Jan 16, 2013)
3. Klein, D., Tran-Gia, P., Hartmann, M.: Big Data. Informatik-Spektrum, Vol. 36, Issue 3, Springer-Verlag Berlin Heidelberg New York, 319-323.(2013)

4. Demchenko, Y., Grosso, P., Laat, C.D., Membrey, P.: Addressing Big Data Issues in Scientific Data Infrastructure. In Proceeding of International Conference on Collaboration Technologies and Systems (CTS), 48-55.(2013)
5. Zhang, J., Huang, K., Cottman-Fields, M., Truskinger, A., Roe, P., Duan, S., Dong, X., Towsey, M., Wimmer, J.: Managing and Analysing Big Audio Data for Environmental Monitoring.In Proceeding of 16th IEEE International Conference on Computational Science and Engineering (CSE), 997-1004.(2013)
6. Jeon, S., Khosiawan, Y., Hong, B.: Making a Graph Database from Unstructured Text.In Proceeding of 16th IEEE International Conference on Computational Science and Engineering (CSE), 981-988.(2013)
7. Xia, Y., Tanase, I.G., Nai, L., Tan, W., Liu, Y., Crawford, J., Lin, C-Y.: Explore Efficient Data Organization for Large Scale Graph Analytics and Storage.In Proceeding of IEEE International Conference on Big Data (IEEE BigData), 942-951.(2014)
8. Compton, R., Jurgens, D., Allen, D.: Geotagging One Hundred Million Twitter Account with Total Variation Minimization.In Proceeding of IEEE International Conference on Big Data (IEEE BigData), 393-401.(2014)
9. Zhang, J., Huang, M.L., Meng, Z.P.: BigData Visualization: Parallel Coordinates using Density Approach. In Proceeding of 2nd International Conference on Systems and Informatics (ICSAI), 1056-1063. (2014)
10. Inselberg, A., Dimnsdale, B.: Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry. In Proceeding of First IEEE Conference on Visualization, 361-378.(1990)
11. Blaas, J., Botha, C.P., Post, F.H.:Extensions of Parallel Coordinates for Interactive Exploration of Large Multi-Timepoint Data Sets. IEEE Transactions on Visualization and Computer Graphics, Vol. 14, No 6, IEEE Computer Society,1436-1443. (2008)
12. Tory, M., Potts, S., Moller, T.: A Parallel Coordinates Style Interface for Exploratory Volume Visualization. IEEE Transactions on Visualization and Computer Graphics, Vol. 11, No 1, IEEE Computer Society,71-80.(2005)
13. Dasgupta A., Kosara, R.: Adaptive Privacy-Preserving Visualization using Parallel Coordinates. IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No 12, IEEE Computer Society,2241-2248.(2011)
14. Lu, L.F., Huang, M.L., Huang, T.H.:A New Axes Re-ordering Method in Parallel Coordinates Visualization. In Proceeding of 11th International Conference on Machine Learning and Applications, 252-257.(2012)
15. Shiravi, A., Shiravi, H., Tavallaee, M., Ghorbani, A.A.: Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security, Vol 31, No. 3, Elsevier, 357-374.(2012)
16. Wang, Y.S., Wang, C., Lee, T.Y.,Ma, K.L.:Feature-Preserving Volume Data Reduction and Focus+Context Visualization. IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No 2, IEEE Computer Society, 171-181, (2011)
17. Hadiak, S., Schulz, H.J.,Schumann, H.:In Situ Exploration of Large Dynamic Networks. IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No 12, IEEE Computer Society, 2334-2343.(2011)
18. Sanyal, J., Zhang, S., Dyer, J., Mercer, A., Amburn, P., Moorhead, R.J.:Noodles: A Tool for Visualization on Numerical Weather Model Ensemble Uncertainty. IEEE Transactions on Visualization and Computer Graphics, Vol. 16, No 6, IEEE Computer Society, 1421-1430.(2010)
19. Zhang, J., Huang M.L., Hoang, D.: Visual analytics for intrusion detection in spam emails. International Journal of Grid and Utility Computing, Vol 4, No 2/3, InderScience Plublishers, 178-186.(2013)
20. Afzal, S., Maciejewski, R., Jang, Y., Elmqvist, N., Ebert, D.S.:Spatial Text Visualization Using Automatic Typographic Maps. IEEE Transactions on Visualization and Computer Graphics, Vol. 18, No 12, IEEE Computer Society, 2556-2564.(2012)

21. Velx S., Csillaghy, A.:A computer vision approach to mining big solar data.In Proceeding of IEEE International Conference on Big Data (IEEE BigData), 27-35.(2014)
22. Lamboray, E., Wurmlin, S., Gross, M.:Data Streaming in Telepresence Environments. IEEE Transactions on Visualization and Computer Graphics, Vol. 11, No 6, IEEE Computer Society, 637-648.(2005)
23. Shi, L., Liao, Q., Sun, X., Chen,Y., Lin, C.:Scalable Network Traffic Visualization Using Compressed Graphs.In Proceeding of IEEE International Conference on Big Data (IEEE BigData), 606-612.(2013)
24. Zhao, X., Ma, H., Zhang, H., Tang, Y., Fu, G.:Metadata Extraction and Correction for Large-Scale Traffic Surveillance Videos. In Proceeding of IEEE International Conference on Big Data (IEEE BigData), 412-420.(2014)
25. Wang, Z., Zhou, J., Chen, W., Chen, C., Liao, J., Maciejewski, R.:A Novel Visual analytics Approach for Clustering Large-Scale Social Data.In Proceeding of IEEE International Conference on Big Data (IEEE BigData), 79-86.(2013)
26. Cheng, D., Schretlen, P., Kronenfeild, N., Bozowsky, N., Wright, W.:Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis.In Proceeding of IEEE International Conference on Big Data (IEEE BigData), 2-4, (2013)
27. Shen, F., Wang, K., Yuan, Y., Wang, L.: Visualization using Histogram Based Transfer Functions for 3D Cardiac Volume Data Set. In Proceeding of International Conference on Information and Automation, 977-980.(2012)
28. Wang, Z., Xiao, W., Ge, B., Xu, H.: ADraw: A novel social network visualization tool with attribute-based layout and coloring.In Proceeding of IEEE International Conference on Big Data (IEEE BigData), 25-32.(2013)
29. Ma, C.L., Shang, X.F., Yuan, Y.B.:A Three-Dimensional Display for Big Data Sets. In Proceeding of International Conference on Machine Learning and Cybernetics, 1541-1545.(2012)
30. Yuan, X., Guo, P., Xiao, H., Zhou, H., Qu, H.:Scattering Points in Parallel Coordinates. IEEE Transactions on Visualization and Computer Graphics, Vol. 15, No 6, IEEE Computer Society, 1001-1008.(2009)
31. Ellis G., Dix, A.: Enabling Automatic Clutter Reduction in Parallel Coordinates Plots. IEEE Transactions on Visualization and Computer Graphics, Vol. 12, No 5, IEEE Computer Society, 717-724.(2006)
32. Novotny, M., Hauser, H.: Outlier-preserving Focus+Context Visualization in Parallel Coordinates. IEEE Transactions on Visualization and Computer Graphics, Vol. 12, No 5, IEEE Computer Society, 893-900.(2006)
33. Zhang, J., Huang, M.L.: Density approach: a new model for BigData analysis and visualization. Concurrency and Computations Practice and Experience, online at 1st July 2014, DOI: 10.1002/cpe.3337. (2014)

Jinson Zhang has been working in IT industry in full-time base more than 20 years, and a part-time researcher in GBDTC (Global Big Data Technologies Centre) of Faculty of Engineering and Information Technology at University of Technology Sydney. His research interests include Big Data, Information Visual Analytics, Cloud Computing, Machine Learning, Network Security and System Security.

Mao Lin Huang is Associate Professor in Faculty of Engineering and Information Technology at University of Technology Sydney, and professor in School of Computer Software at Tianjin University. His research is focusing on Visual Data Analytics, Visual Network Intrusion Detection, Information Visualization and Social Network Visualization. He is an expert and an internationally recognized researcher in the above

areas. Dr. Huang has published more than 170 research papers in the high quality journals and international conferences since 1997.

Zhao-Peng Meng is Professor and the Dean in the School of Computer Software at Tianjin University. Dr. Meng has hold 5 patents as the first inventor, and participated in more than 20 research projects including National Natural Science Fund. His research interests including Big Data, CSCV-based Collaborative System, Internet of Things, Human-Computer Interaction, Artificial Intelligence, Computer Network and Applications, and Distributed Multimedia.

Received: November 22, 2014; Accepted: June 30, 2015.