2014 IEEE 17th International Conference on Computational Science and Engineering

Big Data Density Analytics using Parallel Coordinate Visualization

Jinson Zhang
Member IEEE
School of software, Faculty of Engineering & IT
University of Technology, Sydney
Sydney, Australia

Mao Lin Huang
School of Computer Software
Tianjin University
Tianjin, P.R. China
School of software, Faculty of Engineering & IT
University of Technology, Sydney
Sydney, Australia
Mao.Huang@uts.edu.au

Wen Bo Wang
School of software, Faculty of Engineering & IT
University of Technology, Sydney
Sydney, Australia

Liang Fu Lu
School of software, Faculty of Engineering & IT
University of Technology, Sydney
Sydney, Australia

Zhao-Peng Meng
School of Computer Software
Tianjin University, Tianjin, P.R. China

Abstract—**Parallel coordinate is a popular tool for visualizing high-dimensional data and analyzing multivariate data. With the rapid growth of data size and complexity, data clutter in parallel coordinates is a major issue for Big Data visualization. This has given rise to three problems; 1) how to rearrange the parallel axes without the loss of data patterns, 2) how to shrink data attributes on each axis without the loss of data trends, 3) how to visualize the structured and unstructured data patterns for Big Data analysis. In this paper, we introduce the 5Ws dimensions as the parallel axes and establish the 5Ws sending density and receiving density as additional axes for Big Data visualization. Our model not only demonstrates Big Data attributes and patterns, but also reduces data over-lapping by up to 80 percent without the loss of data patterns. Experiments show that this new model can be efficiently used for Big Data analysis and visualization.**

Keywords—Big Data; 5Ws dimension; shrunk attribute; parallel coordinates

I. INTRODUCTION

Big Data, structured and unstructured data, which contains text, image, video, audio and other forms of data, collected from multiple datasets, are rapidly growing in size and complexity. For example, based on Pingdom 2012 [1], there were 2.2 billion email users who sent 144 billion emails per day, 7 petabytes of photo content added on Facebook every month, and 5 billion mobile phone users who used 1.3 exabytes of global mobile data traffic per month. Hundreds, even thousands, of different attributes in multiple dimensions within datasets provide too much information for traditional analysis and visualization tools to handle.

Big Data is rapidly growing in size and complexity, which creates the need for multidimensional data for large volumes. Researchers have made some progress in reducing multidimensional data in their visual approaches. Woo Sik Seol et al [13] proposed the reduction of association rules to analyze Big Data set. Zhenwen Wang et al [14] 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 colors in the visualization. Zhangye Wang et al [16] clustered large-scale social data into users groups by using the information of user tag and user behaviour.

Currently, Big Data visualization has three main practices: special topic visualization, data-type visualization, and dataset visualization. Special topic visualization focuses on particular topics during the visual algorithm progress, such as network traffic visualization [9], spam email visualization [22], or word cloud visualization [10]. Data-type visualization targets on particular types of data, such as text data visualization [6], video data visualization [7], or audio data visualization [8]. Dataset visualization focuses on particular datasets, such as weather dataset visualization [3], social network dataset visualization [4], or medical dataset visualization [5].

In this work, we have further developed previous works [11][15] using the parallel coordinates to classify Big Data which can be applied across multiple datasets, different data-types and topics. Firstly, we analyzed the Big Data attributes and introduced the 5Ws subsets in parallel axes. Secondly, we established the 5Ws sending density and receiving density as additional parallel axes to measure Big Data flow patterns across multiple datasets for different data-types and topics.

The paper is organized as follows; Section II presents our 5Ws density parallel coordinate model. Section III demonstrates the implementation. Related works are explained in Section IV. Section V summarises our approach and future works.

II. 5WS DENSITY PARALLEL COORDINATE MODEL

A. *5Ws Dimension Model*

The 5Ws dimensions stand for; When did the data occur, Where did the data come from, What was the data content,

978-1-4799-7981-3/14 $31.00 © 2014 IEEE
DOI 10.1109/CSE.2014.219

How was the data transferred, Why did the data occur, and Who received the data. The 5Ws dimensions can therefore be illustrated by using six sets.

- A set $T=\{t_1, t_2, t_i, \ldots,\}$ represents when the data occurred
- A set $P=\{p_1, p_2, p_i, \ldots,\}$ represents where the data came from
- A set $X=\{x_1, x_2, x_i, \ldots,\}$ represents what the data contained
- A set $Y=\{y_1, y_2, y_i, \ldots,\}$ represents how the data was transferred
- A set $Z=\{z_1, z_2, z_i, \ldots,\}$ represents why the data occurred
- A set $Q=\{q_1, q_2, q_i, \ldots,\}$ represents who received the data

Therefore, each data incident can be mapped to a function

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

Where $t \mid T\{\ \}$ is the time stamp for each data incidence. $p \mid P\{\ \}$ represents where the data came from, such as “Twitter”, “Facebook” or “Sender”. $x \mid X\{\ \}$ represents what the data content was, such as “like”, “dislike” or “attack”. $y \mid Y\{\ \}$ 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 “sharing photos”, “finding friends” or “spreading a virus”. $q \mid Q\{\ \}$ represents who received the data, such as “friend”, “bank account” or “receiver”.

Suppose all data items are in time interval T, which are represented as a function set F with n number of items, i.e.

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

F therefore contains all data items within a certain time period. For example, there were 9.66 million tweets during the Opening Ceremony of the London 2012 Olympic Games [1]. Therefore, the scale of Twitter dataset for the Opening Ceremony is $|F| = 9.66$ million.

For a particular data item, assume $p=\delta$, $x=\alpha$, $y=\beta$, $z=\gamma$ and $q=\varepsilon$, the data pattern can then be represented as $f\,(t, p_{(\delta)}, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, q_{(\varepsilon)})$, shown as Fig 1.

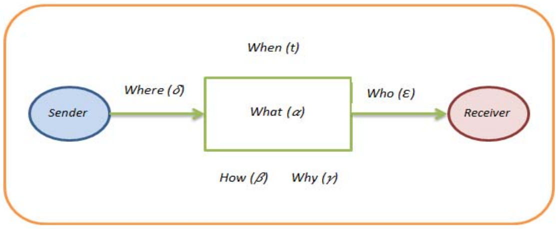

Figure 1. 5Ws pattern

A subset $F_{(\delta, \alpha, \beta, \gamma, \varepsilon)}$ which contains all the particular nodes $f\,(t, p_{(\delta)}, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, q_{(\varepsilon)})$ in the T time slot is therefore defined as

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

The subset $F_{(\delta, \alpha, \beta, \gamma, \varepsilon)}$, which represents the particular data pattern by the 5Ws dimensions, can be mapped using a polyline in the parallel coordinates. We use P as the first parallel axis which represents where the data came from, followed by X, Y and Z, and Q as the last parallel axis which represents who received the data, in order to gain a 5Ws parallel coordinate visualization.

Fig 2 shows an example of the 5Ws parallel coordinates for a particular data pattern where $p=\delta$, $x=\alpha$, $y=\beta$, $z=\gamma$ and $q=\varepsilon$. This 5Ws parallel coordinate visualization reduces the number of polylines because one polyline can demonstrate one 5Ws data pattern, which contains 5 data nodes in the subset.

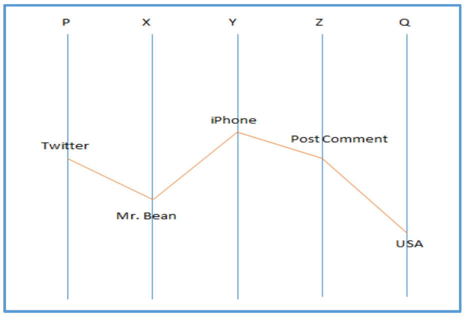

Figure 2. Example of 5Ws parallel coordinate

For example, if there were a lot of tweets during the London Olympics Opening Ceremony about Mr. Bean playing the piano using iPhone apps, *p=”Twitter”* describes where the data came from, *x=”Mr. Bean”* illustrates the data content, *y=”iPhone”* indicates how the data was transfered, *z=”Post Comment”* explains the reason for the data occurring and *q=”USA”* shows the country that received the data.

Each sender might send the data to multiple receivers with different attributes. Each receiver might also receive data from multiple senders with different attributes. This results in a huge number of combinations and varieties, and we use the 5Ws densities to measure these combinations and varieties.

B. 5Ws Sending and Receiving Density

The sending density (*SD*) is used to measure the sender’s pattern during data transferal. Based on (2), the sending density for particular attribute $p=\delta$, $x=\alpha$, $y=\beta$ and $z=\gamma$ in time slot T, is defined as $SD_{(\delta, \alpha, \beta, \gamma)}$

$$SD_{(\delta\,,\ \alpha,\ \beta,\ \gamma)} = \frac{|\,F(\delta,\alpha,\beta,\gamma)|}{|\,F\,|} \times 100\%$$
$$= \frac{\sum_{i=1}^{n} f_{(i)}(t,\ p_{(\delta)},\ x_{(\alpha)},\ y_{(\beta)},\ z_{(\gamma)},\ q)}{\mathrm{n}} \times 100\% \tag{3}$$

Where i is from 1 to n and indicates that there are n data items in time slot T. $f_{(i)}$ represents the i$^{\text{th}}$ incident.

$SD_{(\delta,\ \alpha,\ \beta,\ \gamma)}$ represents the 5Ws dimensions for the sender’s pattern; the content was α, transferred by β, in $t \subset T$ time, for reason γ and sent by δ. The data could have been sent to single or multiple receivers. A high value of $SD_{(\delta,\ \alpha,\ \beta,\ \gamma)}$ indicates that sender (δ) sent the most amount of data compared to other senders.

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

$$RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)} = \frac{|\,F(\alpha,\beta,\gamma,\varepsilon)|}{|\,F\,|} \ \times 100\%$$

$$= \frac{\sum_{i=1}^{n} f_{(i)}\left(t,\ p,\ x_{(\alpha)},\ y_{(\beta)},\ z_{(\gamma)},\ q_{(\varepsilon)}\right)}{\mathrm{n}} \times 100\% \tag{4}$$

$RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$ represents the 5Ws dimensions for the receiver's pattern; the contents was α, transferred by β, in $t \subset T$ time, for reason γ and received by ε. The data could have been sent by single or multiple senders. A high value of $RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$ indicates that the receiver (ε) received the most amount of data compared to other receivers.

C. Shrinkage and Extension of $SD_{(\,)}$ and $RD_{(\,)}$

Each dimension contains hundreds, even thousands of attributes that create huge levels of data cluttering in parallel coordinates. To reduce this over-crowded data without any loss of information, we establish Shrunk Attributes (SA) to collect the attributes that have not yet been illustrated in each dimension. Here, *sa-x* is defined as collection of the attributes for *X*-axis, *sa-y* for *Y*-axis, *sa-z* for *Z*-axis, *sa-p* for *P*-axis, and *sa-q* for *Q*-axis.

Therefore $SD_{(SA)}$ represents the sum of sending density for SA, which is defined as

$$SD_{(SA)} = \sum_{j=1}^{m} SD_{(j)}\left(p_{(sa-p)}, x_{(sa-x)}, y_{(sa-y)}, z_{(sa-z)}\right) \tag{5}$$

Where, *j is from 1 to m*, which means that there were *m* data items that have not yet been illustrated. $f_{(j)}$ represents the j^{th} incident of SA. Furthermore, $RD_{(SA)}$ represents the sum of receiving density for SA, which is also defined as

$$RD_{(SA)} = \sum_{j=1}^{m} RD_{(j)}\left(x_{(sa-x)}, y_{(sa-y)}, z_{(sa-z)}, q_{(sa-q)}\right) \tag{6}$$

The attributes hidden inside SA can also be extracted out for particular attributions in the 5Ws parallel axes. This extension and contraction enables Big Data analysis and visualization to be very efficient, since it can narrow down or extend the particular data-type as required.

D. 5Ws Density Parallel Axes

Here, we create two additional axes by proposing two densities $SD_{(\,)}$ and $RD_{(\,)}$, which each have a value for each data pattern, for parallel coordinate visualization in order to improve accuracy in parallel coordinate visualization. The values of $SD_{(\,)}$ and $RD_{(\,)}$ in both axes represent the data flow patterns shown as polylines among the 5Ws dimensions. This reduces data cluttering in the graph because one subset has only one polyline. 5Ws density parallel axes, combined with the alphabetical axes and numerical axes, have provided more analytical methods for Big Data visualization. No data patterns have been lost during the analysis and visualization process.

Furthermore, the values of $SD_{(SA)}$ and $RD_{(SA)}$ collected by multiple data patterns and shrunk it into one polyline in parallel coordinates have significantly reduced the visual clutter without loss any information. Fig 3 shows an example of 5Ws density parallel coordinates.

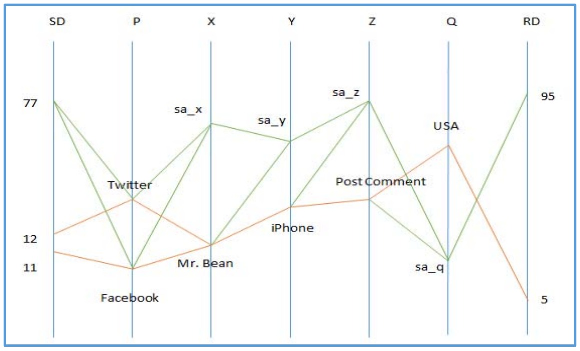

Figure 3. Example of 5Ws density parallel coordinates

For example, assume there were two datasets, Facebook and Twitters (*p="Twitter"* and *"Facebook"*), that post comment (*z="Post Comment"*) about "Mr. Bean" playing the piano (*x="Mr.Bean"*) through iPhone apps (*y="iPhone"*) during the Opening Ceremony of the London 2012 Olympic Games. At sender's end, $SD_{(Facebook,\ Mr.Bean,\ iPhone,\ Post\ Comment)} = 11$ and $SD_{(Twitter,\ Mr.Bean,\ iPhone,\ Post\ Comment)} = 12$ demonstrated the percentage of sending patterns, $SD_{(SA)} = 77$ illustrated the percentage of shrinking attributes *sa_x*, *sa_y* and *sa_z*. At receiver's end, $RD_{(Mr.Bean,\ iPhone,\ Post\ Comment,\ USA)} = 5$ indicated the percentage of receiving pattern, and $RD_{(sa\ q)} = 95$ summarized other countries receiving patterns.

It should be noticed that additional $SD_{(\,)}$ and $RD_{(\,)}$ axes bring extra features in the 5Ws density parallel coordinate visualization. Firstly, it enables the measurement of each data pattern, avoiding information loss. Secondly, the shrinkage and extension of attributes in each axis reduces the polylines over-crowding in graph without losing data patterns. Thirdly, it assigns values for different data-types, both structured and unstructured data patterns. Fourthly, multiple datasets can be analyzed and visualized at the same time. Finally, it provides the details of each data pattern that contains different topics to meet business, government and organization needs.

E. Clustering and Re-ordering

5Ws dimensions contain multiple classifications that provide a clustered view for particular topics and focuses. Each clustered axis, appearing in 5Ws density parallel coordinates, will change the value of $SD_{(\,)}$ and $RD_{(\,)}$. For example, if we assume that the Q axis has a clustered Q1 axis which demonstrate different states of countries such as Q1={California, Texas, sa_q1} and Q={USA, sa_q}, the value of $SD_{(\,)}$ and $RD_{(\,)}$ would be changed because the subset changed from $|F_{(P,\ X,\ Y,\ Z,\ Q)}|$ to $|F_{(P,\ X,\ Y,\ Z,\ Q,\ Q1)}|$. Example shows in Fig 4.

To illustrate the 5Ws data patterns in the parallel axes, we have ordered $SD_{(\,)}$ as the first axis and $RD_{(\,)}$ as the last axis. Fig 4 clearly demonstrates the data pattern from the sender to the receiver for the dimensions: how was the data transferred what the data contents were and why the data occurred. The 5Ws density parallel axes can be re-ordered to demonstrate the visual relationship with nearby axis, example shown as Fig 4.

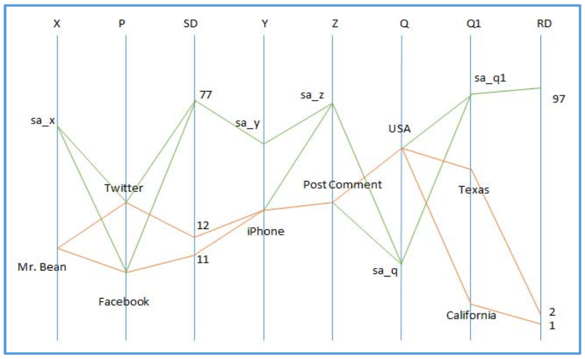

Figure 4. Example of clustering and re-ordering (comparing Fig 3)

If we assume $RD_{(Mr.Bean,\ iPhone,\ Post\ Comment,\ USA,\ California)} = 1$ $RD_{(Mr.Bean,\ iPhone,\ Post\ Comment,\ USA,\ Texas)} = 2$, $RD_{(sa\_q1)} = 97$ summarizes all the other receiving patterns. The re-ordered X axis and $SD_{(\,)}$ illustrates the close relationship between the Y axis and $SD_{(\,)}$.

The 5Ws density parallel coordinates with re-ordering and clustered views provide visual structures and patterns to illustrate the close relationship between the axes in a graphic layout. It clearly demonstrates Big Data patterns for different datasets, different topics and different data-types in visualization.

III. Implementation

Our 5Ws density parallel coordinate model has been tested by using three sample datasets from ISCX2012 network datasets [12], which is an example of Big Data. Three datasets contain 326,681 incidents, 20 dimensions, 37,375 attacks, 1,625 source IPs, 17,115 destination IPs, and 72 applications. The summary of the datasets are shown in Table I.

TABLE I. Three sample ISCX2012 datasets

Name	Jun15a	Jun15b	Jun15c
Network traffic	101,482	94,911	130,288
Attacks	0	0	37,375
Source IPs	1,611	33	36
Destination IPs	15,067	2,164	1,656
Applications	69	19	19

The P axis is defined as the source IP, has 1,625 attributes which represents where the data came from. P= "0.0.0.0" means that the source address is invalid. The X axis is defined as the data content, including "Normal" traffics, "Attack" traffics and "Unknown" traffics. The Y axis is chosen as the application, has 72 attributes that describes how the data was transferred, such as "FTP" transfer, or "HTTPWeb" transfer. The Z axis is chosen as the protocol that illustrates why the data occurred, including "UDP", "TCP" or "ICMP" connections. The Q axis is chosen as the destination IPs, has 17,115 attributes which indicates who received the data. Q= "0.0.0.0" indicates that the destination address is invalid. In total, 31,764 5Ws patterns are displayed in the graph from 326,681 incidents.

To avoid the over-crowded polylines with overlapping attributes in the P and Q axis, we have implemented our density algorithm with SA, shown as Fig 5.

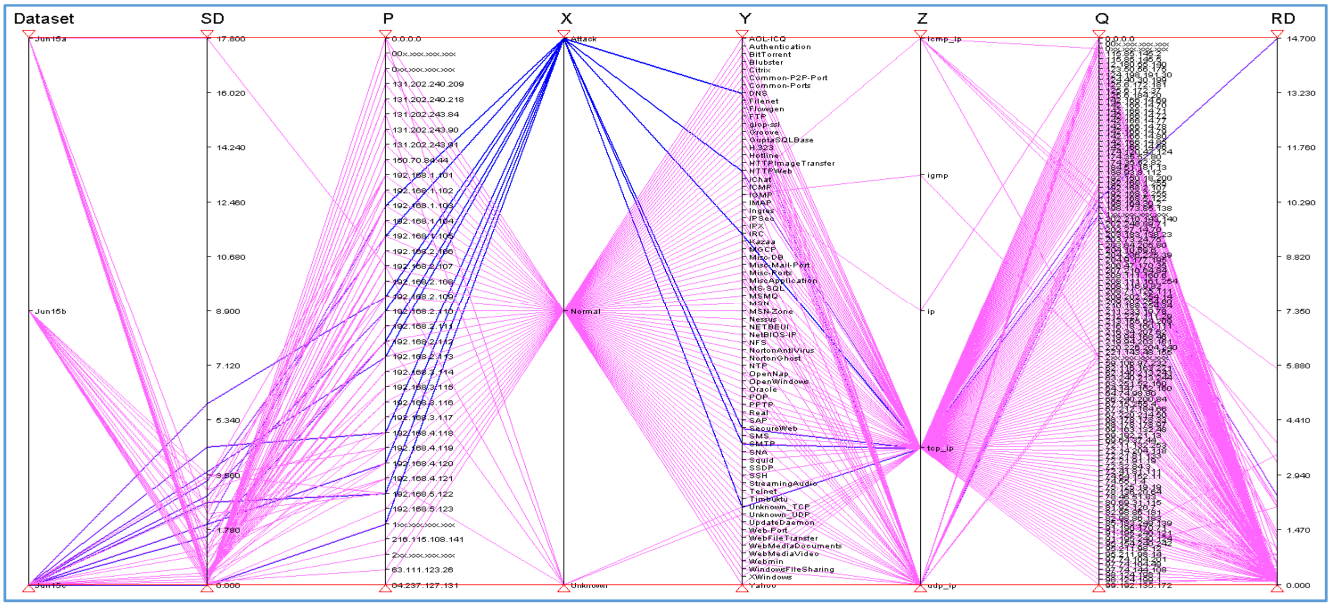

Figure 5. 5Ws density parallel coordinates for three datasets with SA in P-axis and Q-axis.

In Fig 5, we defined SA for each subnet as "00x.xxx.xxx.xxx", "0xx.xxx.xxx.xxx", "1xx.xxx.xxx.xxx" and "2xx.xxx.xxx.xxx" that collected IPs within its subnet while $SD_{(p)} < 1.0\%$ or $RD_{(q)} < 1.0\%$. For example, in P axis, if $SD_{(p=123.123.123.123)} < 1.0\%$ and $SD_{(p=111.111.111.111)} < 1.0\%$, this will be summarized into $SD_{(p=1xx.xxx.xxx.xxx)} < 1.0\%$. In Fig 5, after applying SA, the P axis contains 37 attributes and the Q axis contains 107 attributes. The polylines have been reduced from 31,764 to 2,606 without the loss of data pattern, a significant improvement showing the success of the 5Ws SA visualisation model.

In Fig 5, an attribute “Attack” in the X-axis is chosen as the particular topic for visual analyzing (blue lines). The $SD_{(\ )}$ values are linked to 10 different source IPs in the P axis to illustrate these sending patterns. The attack has been transferred by 6 different methods in the Y axis; {“DNS”, “IRC”, “HTTPWeb” , “SMTP”, “SecurityWeb”, “Unknown_TCP”, }, and connected mostly by “tcp_ip” protocol in Z axis. The highest value of $RD_{(\ \ )}$ =14.7%, which is linked to 192.168.5.122 in Q axis, demonstrates that this receiver suffered the most “Attack” compared to others.

Three datasets with 326,681 incidents have been visualized by using only seven dimensions in Fig 5. Each polyline between seven dimensions demonstrates a particular data pattern that reduces the line cluttering in the graph. $SD_{(SA)}$ and $RD_{(SA)}$ collect multiple data patterns together and shrunk it into one polyline in the 5Ws density parallel coordinates. In our test, 326,681 polylines were reduced to 2,606 without the loss of data patterns, which has significantly reduced the polylines over-lapping and data processing time. Table II and Fig 6 illustrates the details of the polylines between the different axes with and without density algorithm.

TABLE II. The polylines for different axes between three datasets

Dimensions	Jun15a	Jun15b	Jun15c
20 dimensions (Original)	101,482	94,911	130,288
11 dimensions	75,030	90,778	113,482
9 dimensions	66,315	85,861	96,454
7 dimensions	64,605	82,987	94,189
5 dimensions **(5Ws)**	**21,561**	**6,016**	**4,187**
7 dimensions **(5Ws+density+SA)**	**614**	**1,227**	**765**

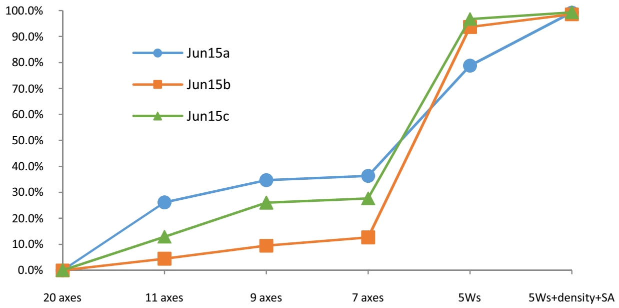

Figure 6. Reduction for different parallel axes between three datasets

In Table II, the datasets contain 20 axes (dimensions) are shown in the APPENDIX. The details for the different axes are: 11 axes {Source, SourcePort, TotalSourceBytes, TotalSourcePackets, ProtocolName, AppName, Tag, Destination, DestinationPort, TotalDestinationBytes, TotalDestinationPackets}; 9 axes {Source, SourcePort, TotalSourcePackets, ProtocolName, AppName, Tag, Destination, DestinationPort, TotalDestinationPackets}; 7 axes {Source, SourcePort, ProtocolName, AppName, Tag, Destination, DestinationPort}; 5 axes (5Ws) {Source, ProtocolName, AppName, Tag, Destination}; and 7 axes (5Ws+density+SA) {Sending Density, Source, ProtocolName, AppName, Tag, Destination, Receiving Density}.

In Fig 6, data cluttering has been reduced by around 80% in the 5Ws pattern: the“Jun15a” dataset reduced data cluttering by 78.8%; “Jun15b” dataset reduced data cluttering by 93.7% and the “Jun15c” dataset reduced data cluttering by 96.8%. Furthermore, the 5Ws density parallel coordinates with SA has reduced data over-crowding by more than 98% based on our test results, and so has significantly reduced data cluttering in parallel coordinates for Big Data analysis and visualization.

IV. Related Work

Parallel coordinate plots, as one of the most popular methods, was first proposed by Inselberg [23] and Wegman suggested it as a tool for high dimensional data analysis [24]. Coordinates of n-dimensional data can be represented in parallel axes in a 2-dimensional plane and connected by linear segments. As pointed out in the literature [25], many methods have been proposed to provide insight into multivariate data using interactive visualization techniques. Parallel coordinate plots (PCP), as a simple but strong geometric high-dimensional data visualization method and represents N-dimensional data in a 2-dimensional space with mathematical rigorousness.

Visual clustering, axis reordering and context focusing are common methods to reduce clutters in parallel coordinates. Dasgupta et al. [26] proposed a model based on screen-space metrics to pick the axes layout by optimizing arranges of axes. Huh et al. presented a proportionate spacing between two adjacent axes rather than the equal spacing in conventional PCP parallel axes. Moreover, the curves possessing some statistical property linking data points on adjacent axes are described in literature [27] as well. Zhou et al. [28] converted the straight-line edges into curves to reduce the visual clutter in clustered visualization. They also utilized the splatting framework [21] to detect clusters and reduce visual clutter. With the aim for avoiding over-plotting and preserving density information,

Kai Lun Chung and Wei Zhuo [20] developed two visual analytic tools: selection graphs and relation graphs, to reduce visual clutter in parallel coordinates. The selection graph is a brushing tool which helps users highlight the regions selected. The relation graph organizes clusters and provide interactions for users to explore the relationships between clusters. Julian Heinrich et al [19] developed BiCluster Viewer that combines heatmaps and parallel coordinates plots to uncover data patterns. The BiCluster Viewer contains many interactive features such as axis ordering, line coloring, or zooming that decrease data cluttering in visual graph. Matej Novotny and Helwig Hauser [18] grouped the data context into outliers, and then trended and focused the context in clustered parallel coordinates to reduce the cluttering issues. Xiaoru Yuan et al [17] scattered points in parallel coordinates to combine parallel coordinates and scatterplot scaling, which reduced data crowding. Users can reorder the coordinates by dragging the axes in their interface for better visual viewer.

No previous work, to the best of our knowledge, has created two additional axes by using the densities for the parallel coordinate visualization. We analyzed the data pattern first in order to obtain the values of SD and RD, and then visualized the data nodes. These data patterns reduced data over-lapping and crowding. 5Ws density parallel coordinates has significantly reduced data cluttering for Big Data analysis and visualization.

V. Conclusion & Future Work

The 5Ws density parallel coordinate model, a novel approach for Big Data analysis and visualization, has been introduced in this work. This model not only keeps the original data patterns during the analysis and visualization, but also explores multiple datasets for different data types and topics. The two extra parallel axes, the sending density and receiving density, have been created to measure the data pattern while still avoiding information loss, and to reduce the visual clutter from the over-crowded graph. The Shrunk Attributes applied in each dimension axis enables the attributes to be narrowed down or extended for a better graph view.

For future works, we plan to develop our 5Ws density parallel coordinate model in three focus areas. Firstly, we plan to implement our model on more practical datasets, such as finance datasets and social media datasets. Secondly, we plan to develop a 5Ws treemap to explore Big Data behaviours. Thirdly, we aim to study the combination of 5Ws parallel coordinates and treemaps.

References

[1] Pingdom, “Internet 2012 in numbers”, posted on Jan 16, 2013, http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

[2] A. Inselberg and B. Dimnsdale, “Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry”, In Proc. First IEEE Conference on Visualization, pp. 361-378, Oct 1990

[3] J. Sanyal, S. zhang, J. Dyer, A. Mercer, P. Amburn, and R.J. Moorhead, “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, Nov/Dec 2010

[4] S. Hadiak, H.J Schulz, and H. Schumann, “In Situ Exploration of Large Dynamic Networks”, IEEE Transactions on Visualization and Computer Graphics, vol. 17, no 12, pp 2334-2343, Dec 2011

[5] Y.S. Wang, C. Wang, T.Y. Lee, and K.L. Ma, “Feature-Preserving Volume Data Reduction and Focus+Context Visualization”, IEEE Transactions on Visualization and Computer Graphics, vol. 17, no 2, pp 171-181, Feb 2011

[6] S. Afzal, R. Maciejewski, Y. Jang, N. Elmqvist, and D.S. Ebert, “Spatial Text Visualization Using Automatic Typographic Maps”, IEEE Transactions on Visualization and Computer Graphics, vol. 18, no 12, pp 2556-2564, Dec 2012

[7] A.H. Meghdadi, and P. Irani, “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, Dec 2013

[8] E. Lamboray, S. Wurmlin, and M. Gross, “Data Streaming in Telepresence Environments”, IEEE Transactions on Visualization and Computer Graphics, vol. 11, no 6, pp 637-648, Nov/Dec 2005

[9] L. Shi, Q. Liao, X. Sun, Y. Chen and C. Lin, “Scalable Network Traffic Visualization Using Compressed Graphs”, In Proc. 2013 IEEE International Conference on Big Data (IEEE BigData 2013), pp. 606-612, Oct 2013

[10] W. Cui, Y. Wu, S. Liu, F. Wei, M.X. Zhou, and H. QU, “Context-Preserving, Dynamic Word Cloud Visualization”, IEEE Computer Graphics and Applications, vol. 30, no 6, pp. 42-53, Nov/Dec 2010

[11] J. Zhang and M.L Huang, “5Ws Model for Big Data Analysis and Visualization”, In Proc. 2013 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 1021-1028, Dec 2013

[12] A. Shiravi, H. Shiravi, M. Tavallaee, and A.A. Ghorbani, “Toward developing a systematic approach to generate benchmark datasets for intrusion detection,” Computers & Security, vol. 31, no. 3, pp 357-374, May 2012

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

[14] Z. Wang, W. Xiao, B. Ge, and H. Xu, “ADraw: A novel social network visualization tool with attribute-based layout and coloring”, In Proc. 2013 IEEE International Conference on Big Data (IEEE BigData 2013), pp. 25-32, Oct 2013

[15] J. Zhang and M.L. Huang, “Density approach: a new model for BigData analysis and visualization”, Concurrency and Computation: Practice and Experience. publish online July 2014, DOI:10.1002/cpe.3337

[16] Z. Wang, J. Zhou, W. Chen, C. Chen, J. Liao and R. Maciejewski, “A Novel Visual analytics Approach for Clustering Large-Scale Social Data”, In Proc. 2013 IEEE International Conference on Big Data (IEEE BigData 2013), pp. 79-86, Oct 2013

[17] X. Yuan, P. Guo, H. Xiao, H. Zhou, and H. Qu, “Scattering Points in Parallel Coordinates”, IEEE Transactions on Visualization and Computer Graphics, vol. 15, no 6, pp 1001-1008, Nov/Dec 2009

[18] M. Novotny, and H. Hauser, “Outlier-preserving Focus+Context Visualization in Parallel Coordinates”, IEEE Transactions on Visualization and Computer Graphics, vol. 12, no 5, pp 893-900, Sep/Oct 2006

[19] J. Heinrich, R. Seifert, M. Burch, and D. Weiskopf, “BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data”, ISVC 2011, Lecture Notes in Computer Science, vol. 6938, pp641-652, 2011

[20] K.L. Chung and W. Zhuo, “Graph-Based Visual Analytic Tools for Parallel Coordinates”, ISVC 2008, Lecture Notes in Computer Science, vol. 5359, pp 990-999, 2008

[21] H. Zhou, W. Cui, H. Qu, Y. Wu, X. Yuan, W. Zhuo, “Splatting the lines in parallel coordinates”, Computer Graphics Forum, 28(3), pp. 759-766, 2009

[22] J. Zhang, M.L. Huang and D. Hoang, “Visual analytics for intrusion detection in spam emails”, International Journal of Grid and Utility Computing, vol 4, no 2/3, pp 178-186, 2013

[23] A. Inselberg, “The plane with parallel coordinates”, The Visual Computer 1 (2) , pp. 69-91, 1985

[24] E. Wegman, “Hyperdimensional data analysis using parallel coordinates”, Journal of the American Statistical Association 85 (411), pp. 664-675, 1990

[25] J. H. Claessen, J. J. van Wijk, “Flexible linked axes for multivariate data visualization”, IEEE Transactions on Visualization and Computer Graphics 17 (12) , pp. 2310-2316, 2011

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

[27] M.-H. Huh, D. Y. Park, “Enhancing parallel coordinate plots”, Journal of the Korean Statistical Society 37 (2) , pp. 129 -133, 2008

[28] H. Zhou, X. Yuan, H. Qu, W. Cui, B. Chen, “Visual clustering in parallel coordinates”, Computer Graphics Forum, 27(3), pp. 1047-1054, 2008

Appendix

ISCX2012 dataset contains 20 dimensions shown as below.

When (T)	StartDateTime, StopDateTime
Where (P)	Source, SourcePort, TotalSourceBytes, TotalSourcePackets
How (X)	ProtocolName, Direction
Why (Y)	AppName, SourceTCPFlagsDescription, DestinationTCPFlagsDescription
What (Z)	Tag, SourcePayloadAsBase64, SourcePayloadAsUTF, DestinationPayloadAsBase64, DestinationPayloadAsUTF
Who (Q)	Destination, DestinationPort, TotalDestinationBytes, TotalDestinationPackets