Multimed Tools Appl (2018) 77:11459–11487
https://doi.org/10.1007/s11042-017-5495-y

# Big network traffic data visualization

**Zichan Ruan**[1] · **Yuantian Miao**[1] · **Lei Pan**[1] ·
**Yang Xiang**[2] · **Jun Zhang**[3]

Received: 25 February 2017 / Revised: 9 November 2017 / Accepted: 5 December 2017 /
Published online: 19 April 2018

**Abstract** Visualization is an important tool for capturing the network activities. Effective visualization allows people to gain insights into the data information and discovery of communication patterns of network flows. Such information may be difficult for human to perceive its relationships due to its numeric nature such as time, packet size, inter-packet time, and many other statistical features. Many existing work fail to provide an effective visualization method for big network traffic data. This work proposes a novel and effective method for visualizing network traffic data with statistical features of high dimensions. We combine Principal Component Analysis (PCA) and Mutidimensional Scaling (MDS) to effectively reduce dimensionality and use colormap for enhance visual quality for human beings. We obtain high quality images on a real-world network traffic dataset named 'ISP'. Comparing with the popular t-SNE method, our visualization method is more flexible and scalable for plotting network traffic data which may require to preserve multi-dimensional

✉ Zichan Ruan
zichanr@deakin.edu.au

Yuantian Miao
myuanti@deakin.edu.au

Lei Pan
l.pan@deakin.edu.au

Yang Xiang
yxiang@swin.edu.au

Jun Zhang
junzhang@swin.edu.au

[1] School of Information Technology, Deakin University, Geelong, VIC 3220, Australia

[2] Digital Research, Innovation Capability Platform, Swinburne University of Technology,
John Street, Hawthorn, Victoria 3122, Australia

[3] School of Software and Electrical Engineering, Swinburne University of Technology,
John Street, Hawthorn, Victoria 3122, Australia

information and relationship. Our plots also demonstrate the capability of handling a large amount of data. Using our method, the readers will be able to visualize their network traffic data as an alternative method of t-SNE.

**Keywords** Visualization · Network traffic · Multidimensional data · MDS · PCA

## 1 Introduction

Network traffic has gained it popularity in the last decade, as increasing of volume, variety and velocity of the data. By in-depth analysis of the traffic, researchers are trying to address the the problems of security, abuse of network, the personal information security and so on [19, 43, 47]. Visual data analysis works as an impressed tool when analyzing large quantity of data, which suits the requirements of network traffic data [44, 46]. Patterns, data structures and discriminations could be perceived after visualization of even a highly complicated dataset. Thereby, visualization explicitly enlightens audience of the implicit properties and relationships buried in the data. An example provided in Shiravi's work [32] reveals the efficiency and importance of data visualization. Researchers and analysts can observe a month's worth of intrusion alerts within a single graph instead of scrolling down the figures and information in tables [32].

Three big traffic network properties were considered when designing the proposed visualization algorithm. Inherently, big traffic network dataset have some similar properties of big data [14, 18], a big traffic network dataset usually contains large volume of data, and the contents and features IP packets possesses could extremely different from each other. Except for the Volume and Variety properties, a big traffic network dataset contains partial encrypted entries [24] due to security concern, which prevent attackers from hacking and getting users' information, but set barriers for researchers and analysts when exploring network traffic meanwhile.

However, with the importance of visualization emphasized in analysis of traffic network, limitations of current visualization approaches and algorithms are identified [19, 34]. When the dataset to be plotted are multidimensional, heterogeneous and of large volume, some visualization methods can be powerless. For example, some visualization approaches could satisfy the requirement to deal with dimensionality, like parallel coordinate, are found limited in large scale dataset. Meanwhile, other visualization approaches like histograms, could perform well with quantity of data, will be useless in processing of data with various features. Also, according to Staheli et al. [34], novel cyber visualizations observed are limited in integrating with human factor and adaptability. Which means, the visualization algorithms to date are either too complex or too basic for intended users, and the lack of human factors in the design leads to the gap between the graphic presentation and human understandings.

The workflow of our visualization algorithm is shown in Fig. 1. We make the following two assumptions — 1) the dataset should be pre-labeled before visualization, and 2) there is the sufficient amount of time for visualization process to complete. Thus, we aim to design an effective visualization method for big network traffic with the consideration of human beings. The main contributions of this paper are listed below:

- We propose an novel visualization algorithm which combines human factors including color perception and vision limitation, provides esthetic visualization results with enlightenments from the big traffic network. With address in the three properties of big

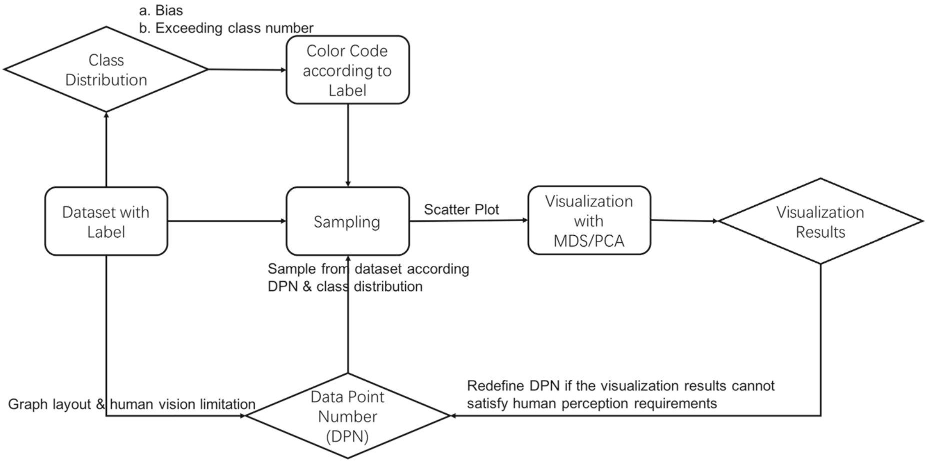

a. Bias
b. Exceeding class number
Class Distribution
Color Code according to Label
Dataset with Label
Sampling
Scatter Plot
Visualization with MDS/PCA
Visualization Results
Sample from dataset according DPN & class distribution
Graph layout & human vision limitation
Data Point Number (DPN)
Redefine DPN if the visualization results cannot satisfy human perception requirements

**Fig. 1** Flow chart of the new proposed algorithm

traffic network, our proposed visualization algorithm is able to process dataset with heterogeneous and multidimensional data, and provide satisfactory user experience. Which means, observers can understand graphs precisely and consistently with what graphs indicate.

- We obtain visualization results by using our proposed visualization algorithm. Compared with t-SNE [37], our algorithm gives a better data point distribution in the graph. Moreover, our algorithm take human factor into account, which make our results easier and accurate to understanding for intended users than graph results generated from t-SNE.
- We combine the mature techniques — PCA and MDS, which makes easy for others to adopt. As it is difficult for human being to observe information from high dimensional images, we utilized PCA and MDS to fit multidimensional dataset into 2-dimensional image. Adaptability of our algorithm in all big traffic network is gained by employing PCA and MDS.

The rest of the paper are structured as below: Section 2 lists related literature about multidimensional data visualization. Section 3 illustrates the methodology and visualization algorithm of this work with the emphasis on the PCA and MDS dimensionality reduction techniques. Section 4 describes the setup of our experiments. Section 5 provides the visualization results images and the comparisons with the results driven from an other visualization techniques. Section 6 concludes the paper.

## 2 Literature review

### 2.1 Multidimensional data visualization

There are multiple methods in visualization, and different methods aim at different aspects: Pie chart, line graph and histogram are three examples of fundamental visualization methods. However, those methods are effective when there are only two or less dimensions, and presentations for simple relations, like data distribution and the trend and changes one value will have when the other value varies [31]. Because of volume and variety properties of big traffic network, visualization methods that are effective for multidimensional data are required. Thus, this section will discuss different visualization methods used in multidimensional dataset.

### *2.1.1 Scatter plot*

A scatter plot displays values for two variables for a set of data as a collection of points. Which is a fundamental way of visualizing multidimensional data.

In 2006, Xiao, Gerth and Hanrahan [43] proposed a new system in multidimensional data visualization. To be specific, this system is for traffic data analysis from visualization aspect. Herein, there were some innovations discovered in the paper [43] analyzed from visualization according to the knowledge of traffic patterns. These inoovations are valued in visual tool designing and analyzing. In addition, the results of visual analysis are also decided by the sense making loop according to their statement [43]. Through choosing a pattern sample from visualization results, this work provide a knowledge process to find out new patterns. Based on the new patterns, the model will be constructed correspondingly. And with this model settled in the application, this novel system is cable to have more deep analysis and therefore discover more complex new patterns accordingly.

The novel system proposed by Xiao, Gerth and Hanrahan [43] set up multiple iterations for the loop of sense-making. The basic knowledge for next iteration is the discoveries that users stored. This paper [43] presents the usage of this kind of knowledge in terms of the improved selection in visual representations. Apart from this aspect, this knowledge base technique also enhanced the filtering and changing results in basic analysis. Herein, the visual representations is helpful in traffic classification. To be more specific, this work [43] shows the successful employment of this novel system in traffic classification and results in high performance in one data sample that the accuracy is up to 80%. Moreover, the network traffic analysis is cable in annotating and classifying network patterns due to the leveraging interactive settled in visualization process. And it also present the way that the discoveries mentioned above in improving the next analysis.

As for the detailed information about this system in paper [43], it can be demonstrated in two parts including the visualization analysis and declarative knowledge representation, where the latter is based on the knowledge concluded from the former step. This system focus on the balance of the association between visual analysis and knowledge. Specifically, the logical model which is used to demonstrate the traffic patterns can stand for the knowledge base. In the meanwhile, the visual analysis can create logical models after each interaction. Through applying the chosen data samples, a series of predicates can be named by this system according to the knowledge base that mentioned above, which are regarded as the candidates in more compound provision. The provision will be evaluated after each analysis iteration in forming a model. Additionally, as if the model is considered to be completed as expected, this model will be used in the whole data set according to this system.

Concerning the part two [43], the system is aiming at promoting the efficiency of visual analysis to get the best leverage of the knowledge base which is used to dominate the flows' attributes. It is helpful and insightful to use the visualization conclusions into the later analysis process. For instance, the color can be mapped according to the types that traffic flows obtained. The level of details can be changed in this part, which is meaningful since the classifying the traffic flows in low-level aspect is always suffered from overwhelming problem. That is, this system is targeting at reason the abstraction in higher level and the change of detail level can be derived by visualizing results. As for the filtering results enhanced, more flexible filtering in terms of sample's features is provided in this system with the usage of pattern labels. Additionally, this system can recognized traffic patterns according to the logical classes which is similar to the construction of pattern models in performing the filtering function.

Despite of the advantage of this system demonstrated as above, there are some disadvantages that hinder this system to gain more higher assessment. Specifically, Xiao, Gerth and Hanrahan [43] proved that this system's performance relies on the amount of traffic data to some extent. And the false positive and false negative values reveal that the misclassified traffic samples are near to the decision boundary which is difficult to analyze and classify into the correct type confidently. There are several interesting patterns discovered in Fig. 2, SSH attacks only uses the upper several port numbers, scan contains various of patterns.

The t-Distributed stochastic neighbor embedding (t-SNE) has gained its popularity in big data visualization. It was first proposed in [37], van der Maaten explored the t-SNE technique, which is an embedding technique used for visualization of the heterogeneous data with diverse dimension, using scatter plot.

Figure 3 shows the t-SNE used in the visualization of MNIST dataset, which contains 6,000 grayscale images of handwritten digits. Scatter plot in Fig. 3 illustrated that t-SNE

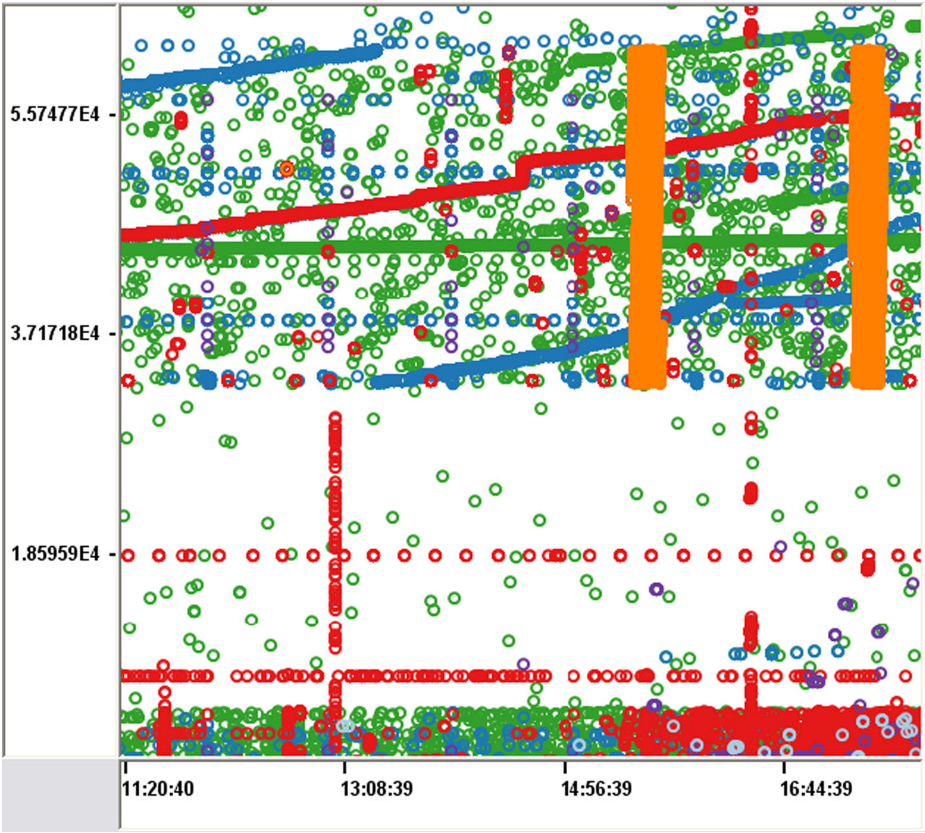


**Fig. 2** An event diagram of source-port for the following six traffic protocols: mail (green), DNS (blue), scan (red), SSH logins (purple), SSH attacks (orange), chat (indigo) [43]

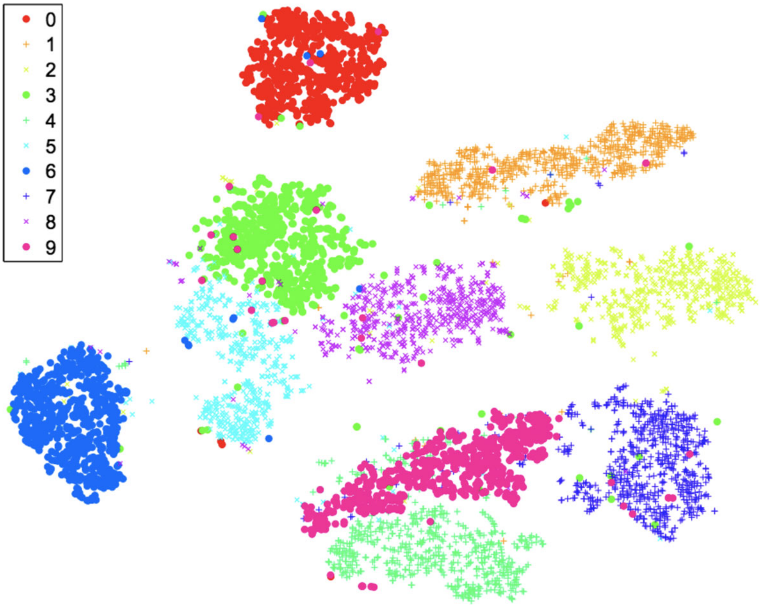


**Fig. 3** T-SNE visualization results plotted with 6,000 handwritten digits from the MNIST data set [37]

performs great in MNIST dataset. Ten classes by digits are clearly separated with acceptable number of wrong placed data points. T-SNE focuses on the local structure of data, because it processes dimensionality reduction according to data local structure. And it is obvious that t-SNE does a much great job of revealing the natural classes in the data.

### *2.1.2 Hyper graph*

In 2014, Glatz et al. [17] worked on communication logs visuallization which can be referenced by network traffic analysis tasks including network planning, performance monitoring and troubleshooting. Since the size of communication logs is always large and the number of network traffic is growing speedly, the visualization methods devised should be effective on large scale data set [17]. Targeting at this issue, a novel network traffic visualization scheme is proposed by this paper [17]. And the main ideas in their work obtain two points generally. The first is to pick up some traffic patterns to form a representative data set which is more concise than the original communication logs. And then visualize the new data set with frequent itemset mining (FIM). The second idea is to visaulize the processed data set into hypergraphs which can reveal the relationship attributes clearly. The experiments done by Glatz et al. [17] showed the advantages of this novel scheme comparing to other network traffic visualization methods in view of coordinating plots or graphs in parallel. Additionally, when visualization technique is applied in network monitoring, the speed of exploring large scale of data manually can be significantly improved for a good picture can express the information straightly and clearly than words demonstration. In detail, the changes are displayed obviously in visualization results which usually occurs in traffic patterns, troubleshooting problem, security incident detection and network growth plans. Further, FIM exploits the frequent patterns based on transactions.

As for the scheme in [17], hypergraphs are visualized based on the selected patterns. As Glatz et al. [17] stated, hypergraph is a high level graph of traditional visualization which displays the connection to multiple attributes. In contrast, traditional visual graph can only provide the relationship between each two attributes. A frequent itemset shows that plenty of traffic flows are classified in low accuracy which routed via an uplink provider with some port numbers. One of the main purposes in network measurement research is to find out the insight logic of communication attributes. And the scheme proposed in this paper can achieve this goal effectively. Moreover, the hypergraphs can be employed in revealing the relationship among multi-attributes which can be extracted from frequent itemsets. Last but not least, the computational cost spent by this scheme in their work is reasonable.

### *2.1.3 Force graph*

A force graph represents nodes as dots and links as line segments to show how a data set is connected. The fundamental idea of a force graph is a network, thus, relations among data points within a dataset are revealed.

Another paper published by Braun et al. [6] points out that traffic measurement visualization is one of the major target using flow inspector as well. And these measurement elements demonstrate the hosts' behaviors in network connection. Lots of web techniques developed for data analysis to gain the process in browser applications. Some advanced browser engines is capable in visualizing traffic data interacively in real-time. It is an efficient work in understanding the compound data. Flow-Inspector is proposed in this paper [6] as an advanced interactive open-source framework in network to visualize network traffic flow clearly. The outcomes of this Flow-Inspector obtains the process and storage of the large-scale network traffic flow. And the web application which is based upon JavaScript can show and control the traffic informaiton in real-time. A set of tools are provided in the operations suggested in this work [6] that make the scientific community to visualize novelly according to the advanced framework. There are verious components of visualization employed in their methods which can reveal the topologic understanding of the network flows. In additon, they also found that the implementations can be more flexible as if the developers can identify the flows individually according to a series of arbitrary keys. Therefore, the attributes of summplemental flows can be defined accordingly. It is unnecessary to modify the main implementational code to forming another new visualization results. Herein, nodes in node-flow pictures display the patterns which are regarded as the prepresentatives for the hosts or networks. And the lines between each connected nodes present the communication flows according to their statement [6]. Additionally, line color is set up based on the collected time [6]. Thus, the differences between communication patterns formed under various time collections can be observed in visual. Through customizing the time interval, the force picture layout is needed to be rendered and calculated in Flow-Inspector. And the flows information among this time slot means the connected knowledge between each nodes. Moreover, the multiple iterations is applied in visualization to gain a steady status. In the meanwhile, users can stop the interactions manually.

As shown in Fig. 4, weak geometric constraints are presented are the links between nodes with settled length. An iterative process of optimization is operated upon the image to find the best location of each data points.

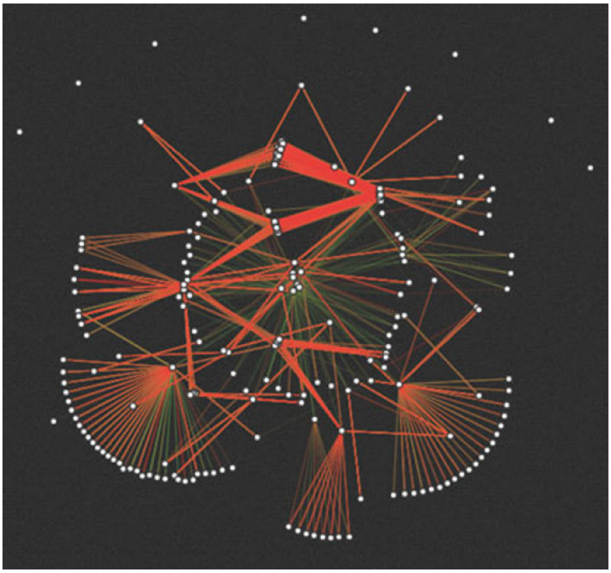


**Fig. 4** An example force graph produced by Flow-Inspector [6]

### *2.1.4 Parallel coordinate*

In 2015, an other iterative visualization method is used in forensic analysis explored by Promrit and Mingkhwan [30] to describe the communication in network. Specifically, this method analyzes the data from two aspects obtaining timeline and parallel coordinates. At the beginning of the forensic analysis, the authors reformed the events' timeline based on the knowledge extracted from the traffic logs. And the unexpected events can be tracked according to this analysis process description. Additionally, the parallel coordinates can display the information of both normal and anomaly event with various attributes.

However, it is worth to point out that the contribution of this work [30] in not the results shown successfully in timeline and parallel coordinate. Instead, the iterative feature of this visualization framework is the main point in demonstrating the abnormal traffic and traffic pattern for applications. The frequent item-set mining technique is also applied in singing out the traffic flows with most informative knowledge in classification. Thus, it is simple to classify these flows according to the shape and entropy the technique worked out. The advantages of this research comparing to other similar works include two parts — firstly, combining the frequent itemset mining and parallel coordinates methods in visualization allows the traffic flow's pattern to obtain lots of features, which led a deep understanding of the significant attributes of abnormal flows and application flows respectively; secondly, frequent itemset mining can work out a best entropy value in terms of the classification accuracy of application traffic. And this method can protect the privacy of user's information.

The above methods are integrated in a visual network communication tool by Promrit and Mingkhwan [30]. And their published paper shows some performance of this advanced tool. To be specific, the knowledge of abnormally traffic and normal one can be shown directly in a graph analyzed by its outcome graph. And the overall performance toward application traffic classification is up to 92% accuracy which can be regarded as a powerful classifier.

Moreover, the prototype can be concluded in contributing its valued finding to visualization analysis.

Nevertheless, Meryem et al. [13] employed parallel coordinate in their work as well. As shown in Fig. 5, they proposed a visualization framework combining machine learning, sampling, feature selection and parallel coordinate to monitor real-time network traffic. Their visualization framework gives a pathway of paradigmatic processing of real-time traffic data. Which ensure that the analysis is up to date and mitigates the negative influences of delayed information.

### *2.1.5 Comparison of visualization methods in multidimensional data*

Among these visualization methods, scatter plot was employed in my algorithm. By adding spacioul elements into the presentation, scatter plots could give us more insights about the dataset, and appropriate clusters will be evidence of the classification as shown in [37]. Though, a force graph could give a similar outcome, the line segments are hard to accurately and completely define. Researchers and analysts could be misled once the line segments were wrongly defined [6]. As for the parallel coordinate, the limitation lies in the dimensions [41]. Feature characters within a traffic network dataset can be 20 or even more, in which case, a parallel coordinate is too complicated for human beings to discriminate clearly. In the aspect of a hyper graph, understanding a hyper graph for a human being is no easier than understanding a set of figures, not to mention the underlying insights it represents.

There are some fundamental visualization methods we put aside when comparison, like pie chart and histogram. They are important and useful, but their limitation under the circumstance of multidimensional dataset is obvious. Which is that they can only reveal one aspect of the dataset at a time.

Table 1 summarizes the mentioned works about the multidimensional data visualization from the aspects of visualization technique, flow extraction, human factors and datasets. The network traffic visualization could be applied into many areas in traffic analysis. However, there is no literature considering human factors when developing their visualization methods. As it is humans that will observe the visualization results, I believe it better that we take human vision and perception into account. Although, t-SNE gives a great outcome

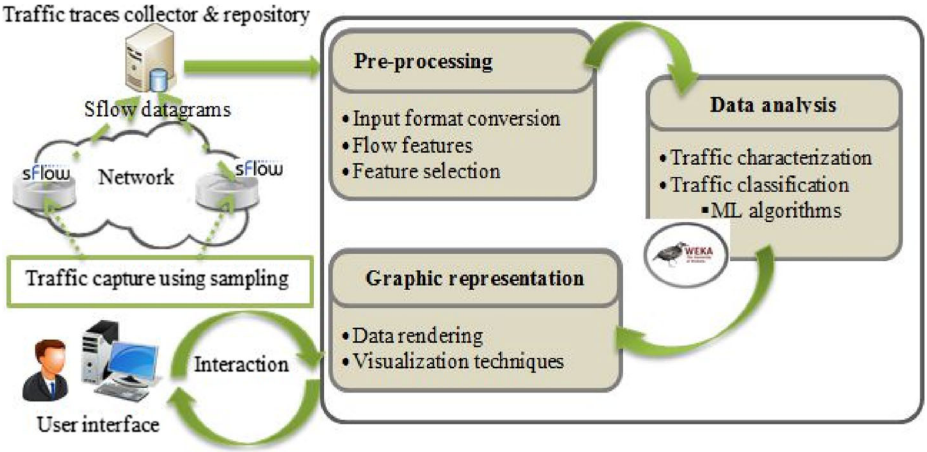


**Fig. 5** The visualization frame in Meryem's work [13]

**Table 1** Summarize of related research features

| | Visualization technique | Flow extraction | Human factors | Datasets |
|---|---|---|---|---|
| (Ling Xiao et al. 2006) [43] | Scatter Plot | Argus sensor | N/A | Network traffic |
| (Laurens van der Maaten and Geoffrey Hinton 2008) [37] | t-SNE based Scatter Plot | N/A | N/A | MINIST (natural data) |
| (Eduard Glatz et al. 2014) [17] | Hyper Graph | FIM | N/A | Simple flows |
| (Lothar Braun et al. 2014) [6] | Force Graph | IPFIX & NetFlow | N/A | One day of network traffic |
| (Nuttachot Promrit and Anirach Mingkhwan 2015) [30] | Parallel Coordinate | FIM | N/A | Simple flows, network traffic |
| (Meryem Elbaham 2016) [13] | Parallel Coordinate | NetMate | N/A | Collected from network |

at MINIST dataset, it is consisted of natural data. As a traffic network dataset could be extremely different from a natural dataset, it is worthy considering if we employ t-SNE in our work or not.

## 2.2 Color design

There is an established body of knowledge in color design and perceptual science, to better formulate the appropriate colormaps for the information visualization task, we took advantage of it and exploited some critical rules relative to our goal of visualization.

The crucial point when presenting clustering classes in data set is to utilize contrast colors and to distinguish those classes by colors [26]. We use hues to display the classes, however, there could be overlapping among classes even data points in the image, so the color selection is important. An example to draw the attention of contrast is given in Fig. 6 [25].

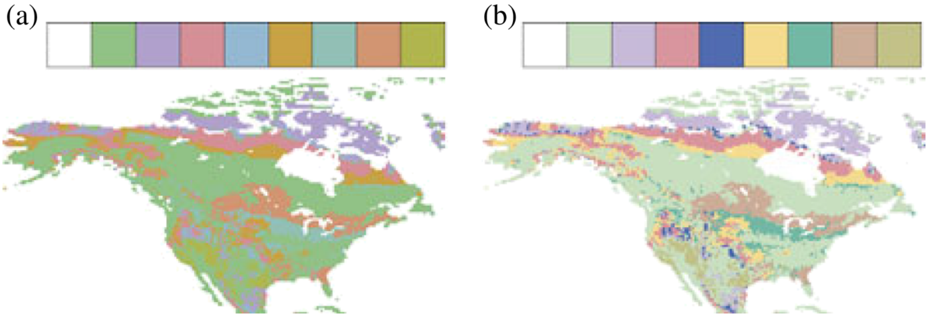


**Fig. 6** A categorical colomap, as shown in (**a**), results in a limited visibility of categorical differences on a map dataset. Lee et al. [25] propose a visibility optimization method that results in the colormap shown in (**b**), which provides much better contrasts in the data

An automated optimization method to improve categorical data perception was introduced in [25], to better display even the small spatial context in the map. The idea is to give much contrast colors to the small regions and the regions around them, thus to highlight those small regions. Lee et al. [25] define a point saliency measurement to compute the saliency of each point in the image by calculate the differen between the color of that point and the average color of the points around it in the CIELCH color space, which is a cylindrical representation of the CIELab space. Then, apply a class visibility metric to the data set to get the weighted sum of the point saliency of all pixels contained in current class. Finally, optimize class visibility of all classes to form a better visual effect and attention guidance.

### *2.2.1 Colormap selection*

The combination of three channels red, blue and green, the three primary colors, is used as RGB color space to display colors in digital devices with an additive fashion [35]. RGB color space provide three coordinates, which separately present the amount of red, green and blue channels and form the color displayed. For example, as defined in IEEE format [22], all color figures should be generated in RGB or CMYK color space. However, these device-oriented color spaces are less inconvenient to employ than cylindrical color spaces [48]. The connection between cylindrical color spaces and RGB or CMYK is weak, thus hue value and chroma (HVC) color space is proposed, which is a Munsell-like [28] discrete color space could be easily convert into RGB or CMYK color spaces employed by digital devices.

Rainbow colormap, widely adopted as the standard and default colormap by many visualization functions and applications, though providing pleasing presentation, was found important flaws which may lead to misinterpretation of the data. Rogowits et al. [20] stated that the by presenting data with rainbow colormap, the detailed data information could be ignored while the nonexistent boundaries were perceived and thus this false structure in the image lead to misinterpretation of data. Instead of the default rainbow colormap, Brewer [7] generated new groups of colormaps designed in HVC color space, which are perceptual and able to provide comprehensive understanding in view of data without misinterpretation. Harrower and Brewer implemented an online tool for further utilization of those colormaps [20]. And we employed the tool for our visualization task, the colormap that we choose will be displayed in the following Section.

Past several decades have witnessed multifarious efforts to integrate human color vision into visualization design to acquire information within data. Color design paly an important part in visualization in both esthetics aspects and data perception improvement part [48], as color could guide viewer attention to salient features [23] and therefore allow more critical insights into the data itself. The most significant part within the color design is the colomaps, which is the design of transfer functions intend for mapping scientific data information into RGB attributes [39]. Li et al. [26] also mentioned the importance of selection of color channels, they proposed a novel color image processing tool called quaternion hadamard transform, which can generate invisible watermarks against a wide range of attacks. Visualization is not only concerned with providing esthetic images, colormaps should be properly utilized to present the features of interest in the data to be visualized and help viewers gain an efficient and effective comprehension and insights of the data set [42]. Consequently, general color design rules are needed for visualization designers to visualize specific data. And proper colormaps is required for information visualization in our case.

# 3 Methodology and algorithm

## 3.1 Dimensionality reduction methods

The limitations of human vision and the requirements of big traffic network data prescribe the necessity of effective dimensionality reduction methods prior to visualization. But it is difficult to plot images beyond three dimensions on a flat surface for human beings to understand. Human perception and cognition is the key for human beings to understand their situations and surrounding objects. In fact, human beings can only directly 'see' two dimensions, and they 'feel' the third dimension though the eye movements and parallax [3, 16]. Thus, human perception limitation sets a boundary that we visualize dataset in a two-dimensional manner. However, big traffic network we are handling is almost always vast, heterogeneous, containing many different features as an accurate delivery of information. Which makes it important for our visualization algorithm to select dimension reduction methods that preserve significant information and details as much as possible.

### *3.1.1 Principal component analysis*

Principal Components Analysis (PCA) is one popular method of dimensionality reduction, and PCA implement dimensionality reduction by feature derivation, which enables critical feature identification and feature combination. By employing a matrix to the origin data matrix, transformations like moving and rotating will be applied to the whole data matrix, consequently, new features derived and new coordinate system of the graph produced [27].

The operation of PCA proposes a particular set of axes that enable the removal of the negligible coordinates or the coordinates of little impacts to reduce the dimension of graph. And the main question in processing principal components analysis is how to choose axes. A principal component means a direction that contains the largest variation [27].

The result will show diagonal on the covariance matrix, because all the variations are along the coordinate axes after the processing. In other word, each new derived variable is unrelated to every derived variable except itself. Some least important components have little effect on the data, those axes can be tailored and reducing the dimensionality of the graphic. An example about deciding the principal components from a two dimension image is given in Figure 7 [12].

Principal components analysis is wide spread in many applications for its idea to simplify. And this is also the reason why we implement it in our algorithm. However, PCA also gets drawbacks, by reducing dimensions, some information in the dataset will be discarded, which may cause the insufficient analysis results. Moreover, running PCA can be time consuming for datasets of high dimensions.

### *3.1.2 Multi-dimensional scale*

Multi-dimensional scale(short for MDS) is about to measure and map a set of multivariate data, in an ideal low dimensional space, using the points in the space presenting each data and the Euclidean Distances between each pair of points reflecting the dissimilarities between each pair of data entities to the greatest extend. That is to say, the multi-dimensional scale is a method measuring the dissimilarities arising from various situations, in our case, the multivariate data set is the network traffic dataset. The importance of graphical display of the data provided by MDS is that researchers could explore the structure of the data visually and discover the underlying regularities among those arrays of numbers [5].

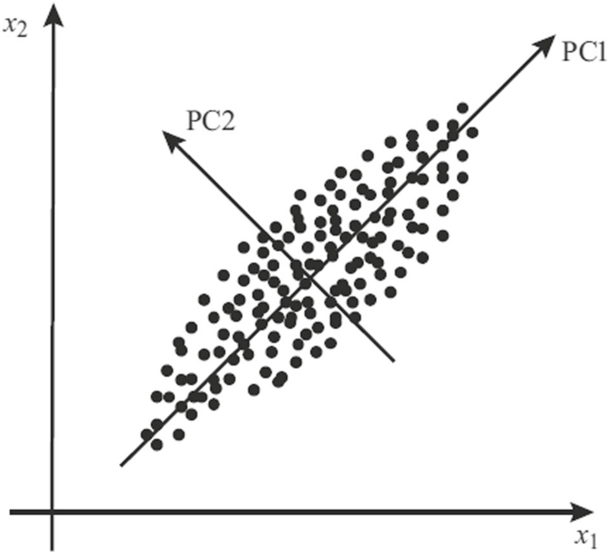


**Fig. 7** An example of PCA coordinates resetting for a two dimensional image [12]

As mentioned in the previous section, t-SNE performs great at MINIST dataset, which is a natural dataset. In order to choose second dimensionality reduction method from MDS and t-SNE, we ran a pilot experiment on plotting of randomly selected 10,000 DPN from 'ISP' dataset [40] with t-SNE and visualization result is shown in Fig. 8. And the result is

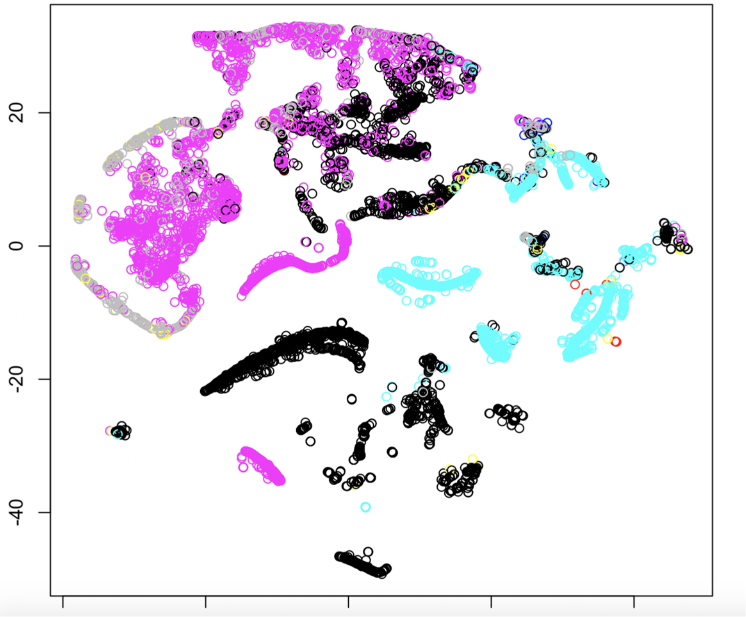


**Fig. 8** The plotting of randomly selected 10,000 DPN from original dataset 'ISP' with t-SNE

unsatisfactory, the reason may be that the ‘ISP’ dataset is a traffic network dataset with 21 statistical features. Additionally, t-SNE is a techniques based on converting the Euclidean Distances [37], we decide to choose MDS instead of t-SNE to better suit the visualization initiatives.

### 3.2 Our proposed visualization algorithm

Visualization is considered as a critical pathway for further insights in the data [34]. Proposed visualization algorithm combines human factors, scatter plot and effective dimensionality reduction methods when illustrating the multidimensional heterogeneous network traffic flow data. Human labor will take into place in evaluation of visualization results. The flow chart of the proposed visualization algorithm is shown in Fig. 1 at Section 1. Two assumptions should be noted: datasets are pre-labeled and time is limitless. Before a dataset is visualized, a preprocess is required.

In the first stage, class distribution of the dataset is calculated, and color codes according to the labels are generated. However, there could be 2 exceptional conditions in class distribution — one is severe bias problem, the other is exceeding class number. When there exists severe bias problem, the dataset will be divided into two parts, and visualization against these two parts will be processing separately. Majority part contains only classes whose entries occupy over 5% of the whole dataset, and the minority part is whatever classes left. And they will be visualized respectively according to the work flow. In which way, the detailed information of dataset could be revealed. The other exception is that too many classes in the dataset. Though humans have the potential to recognize over a million colors according to measurements of difference thresholds, it is difficult for us to distinguish them when two similar colors were placed together [4, 15, 38]. In most cases, human beings have a trichromatic vision. Trichromacy is the possessing of three independent channels for conveying color information, derived from the three different types of cone cells in the eye, and the three types of cone cell in the retina are preferentially sensitive to blue, green and red. Apart from the trichromacy, there are three properties of color percepts: hue, saturation, and brightness [4]. Thus, a generalized color classification are: Red, orange, yellow, green, cyan, blue and purple. All those characteristics combined, it is safe to say that we can use up to 14 colors for each visualization. When the number of classes a dataset contains exceed the threshold, the algorithm will combine classes that contains few entries into one for easier visualization.

Data point number (DPN) within an image will be determined at the second stage. Combining the human vision limitation, calculated from the Standard Snellen Chart and standard visual acuity with the image layout reality limited by the paper size, the initial DPN is defined. According to the data point distribution in the produced images, decide increase or reduce DPN within a certain limited range.

Sampling is the third procedure of the algorithm, followed by visualization with scatter plot and MDS/PCA. Then, expertise will determine if the visualization results are easy for human to understand or not. By adjusting the DPN again, we can get a final visualization results from proposed visualization algorithm.

### 3.3 A visualization example using the proposed algorithm

There is an example using BT and HTTP dataset, which contains 20 records from BT and 20 records from HTTP, as input to plot nice layout. Figure 9 shows the work flow of the small volume of multidimensional data in plotting. It is interesting to dig the underling

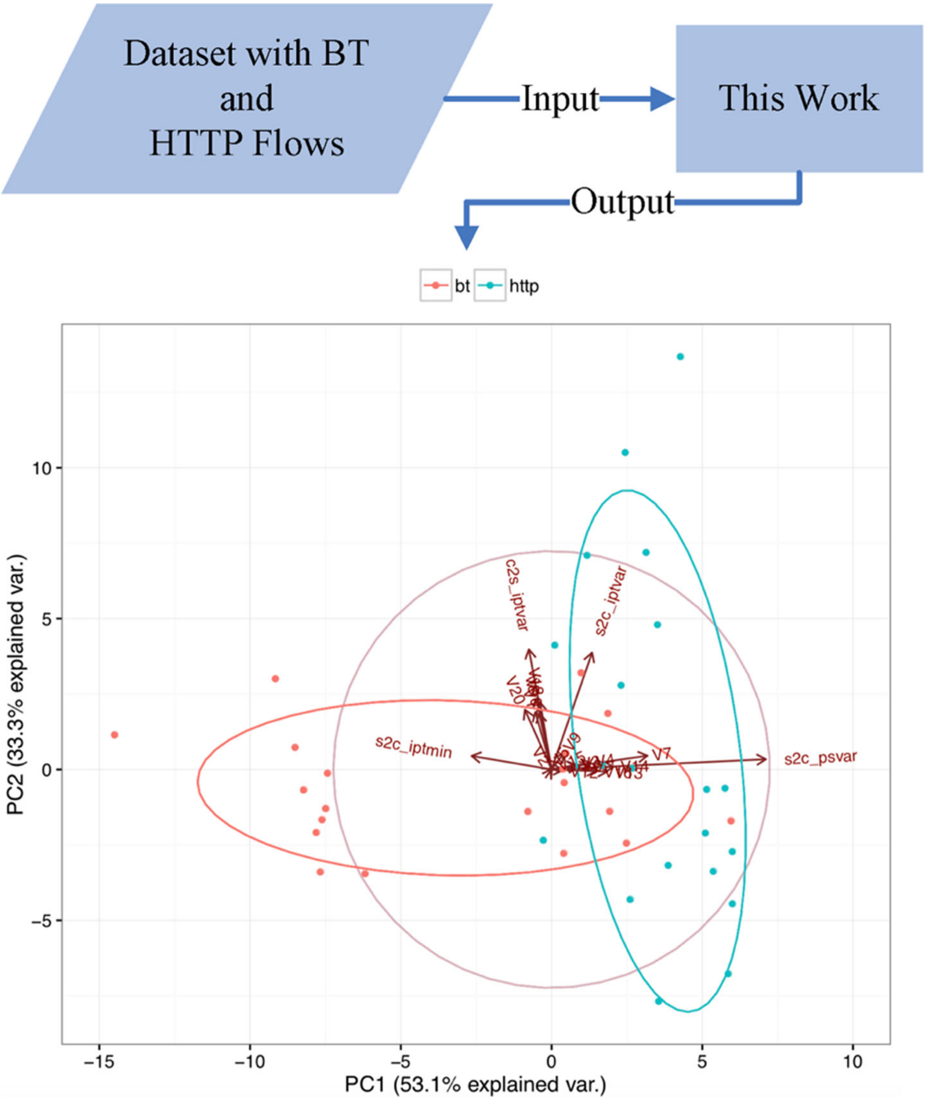


**Fig. 9** Our work flow when plotting the 40 records in BT and HTTP protocols

relationship between BT and HTTP protocols from the visualization view. As there is only two classes, color code is easy to generate. And only 40 entries are in the dataset, DPN is set to 40. The BT class is presented as red solid circle and the HTTP class is presented as green solid circle. From the output figure, it could be easily found that the two classes formed two clusters with partial overlapping as from the two confidence ellipses. The confidence ellipses were plotted according to the 95% spatial confidence intervals. The confidence level, 95%, indicates how frequently the data will be contained in the settled spatial intervals. Jerzy Neyman [29] is the first person to introduce the confidence interval in statistic, he stated that the given confidence level will be the proportion of those confidence intervals include the real value of the parameter. In this case, the frequent that the data points located in their corresponding confidence ellipses will be 95%. Thus, those ellipses could serve as a evidence of clustering in the produced images.

## 4 Experiments

### 4.1 Network traffic data set

The original data set we used in the experiment is the ‘ISP’ dataset [40]. There are 18 application protocols — BT, DNS, EBUDDY, EDONKEY, FTP, HTTP, IMAP, MSN, POP3, RSP, RTSP, SMB, SMTP, SSH, SSL2, SSL3, XMPP, and YAHOOMSG present in 3,380,993 entries with 21 features. As mentioned in the Section I, a big data properties is the presence of the encrypted protocols, in this case, SSH, SSL2 and SSL3. Which leads in difficulties when analyzing traffic flows in traditional ways like payload-based classification and also may cause privacy issues [45]. Thus, feature extraction and selection plays an significant role to avoid further using specific payload in IP packets. In [9–11], the importance of feature extraction and feature selection are addresses. Different scenario requires different feature set. For example, semantic features were extracted in [11] to capture movements and even emotions in untrimmed videos. In ‘ISP’ dataset, Wang et al. [40] extracted statistical features from ‘ISP’ without detailed information.

### 4.2 Color code

The color code is generated after the analysis of ‘ISP’ dataset class distribution as shown in Fig. 10. It is clear that only 5 classes possess entries occupying over 5% of all data, namely BT, HTTP, SMTP, SSH and SSL3. Thus, there will be two sets of visualization result from ‘ISP’ dataset. The majority part visualization is generated from 5 classes mentioned above,

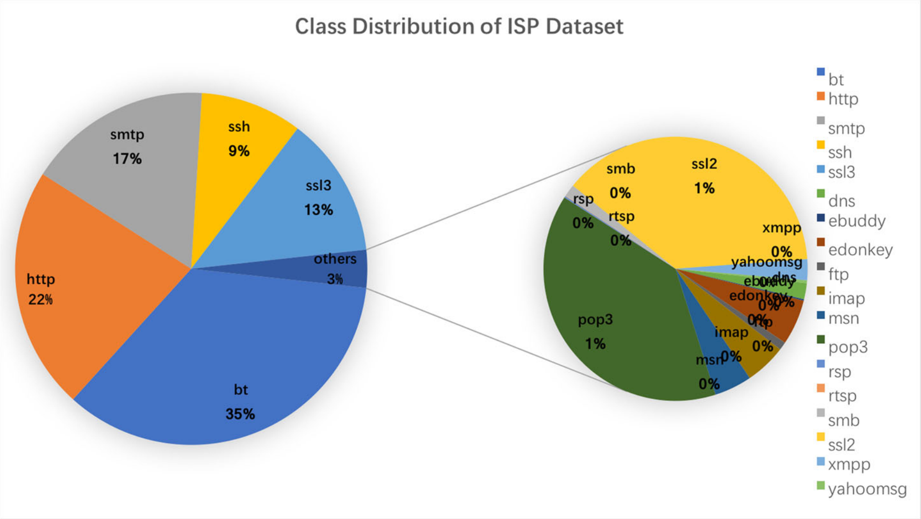


**Fig. 10** The class distribution of ‘ISP’ dataset

and the minority part visualization is generated from the rest 13 classes. Since class numbers in two visualization are 5 and 13, which could be presented and distinguished by color, the color codes for them is generated according to Brewer et al. [7]. Figure 11 shows two sets of color code accordingly, with class, color and HEX color code listed.

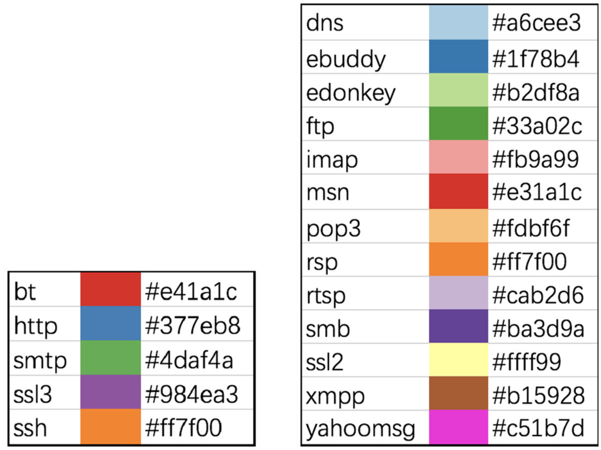
## 4.3 Two lists of data point number

Herman et al. [21] stated that it is easier to present less data element information in a largely scaled graph. The size of the graph to view is a key issue in graph visualization. It will make performance consumption even damage when too large elements are required within a graph to present. Even if there exists a large enough space to contain and display all elements, it is difficult for human to distinguish every elements in the layout due to human vision limitation. The quality of the graph can be low due to the large amount of the elements within it.

To address the problem mentioned above, data reduction or sampling is required for the next stage. However, before sampling, it is critical to define the number of data points in the graph and satisfy both the requirements of qualified graph and presenting sufficient information contained in the original data set.

### *4.3.1 Figure size reality*

In reality, there is boundaries of the graphs within thesis or papers, normally, those thesis and papers will be printed on a A4 paper, which defines the limitation of a graph layout. The specifications of a A4 paper is 210 millimeters width and 297 millimeters depth (8.27 × 11.69 inches) [1]. Usually, the publishers of the journals and magazines require certain layouts or formats of those publications, which may differ from publisher to publisher. Assume that the graph we make is of golden ratio, which is the best ratio in aspect of esthetics and for human to view. The graph size will be (3.5 × 2.2 inches) in column wide and (7.16 × 4.48 inches)in page wide.



| | | |
|---|---|---|
| bt | 🟥 | #e41a1c |
| http | 🟦 | #377eb8 |
| smtp | 🟩 | #4daf4a |
| ssl3 | 🟪 | #984ea3 |
| ssh | 🟧 | #ff7f00 |

(a) Color code for majority part visualisation.

| | | |
|---|---|---|
| dns | 🟦 | #a6cee3 |
| ebuddy | 🟦 | #1f78b4 |
| edonkey | 🟩 | #b2df8a |
| ftp | 🟩 | #33a02c |
| imap | 🟥 | #fb9a99 |
| msn | 🟥 | #e31a1c |
| pop3 | 🟧 | #fdbf6f |
| rsp | 🟧 | #ff7f00 |
| rtsp | 🟪 | #cab2d6 |
| smb | 🟪 | #ba3d9a |
| ssl2 | 🟨 | #ffff99 |
| xmpp | 🟫 | #b15928 |
| yahoomsg | 🟪 | #c51b7d |

(b) Color code for minority part visualisation.

**Fig. 11** Color code generated from ColorBrewer [8]

It is easy to conclude that better resolutions provided by higher DPI, a common laser printer will be equipped with resolution of 600 dots per inch. Dots is also known as pixels, which in digital imaging are small little dots that make up the images, also the basic unit or smallest element of programmable color in a digital image. From which, we could calculate the total pixels within the graph we could utilize: (2100 × 1320 dots) for one column width graph and (4296 × 2688 dots) for one page column width graph. We selected the filled circle to present each data point, and it is easy to conclude that even the smallest circle needs at least 3 × 3 pixels for printing.

### *4.3.2 Human vision limitation*

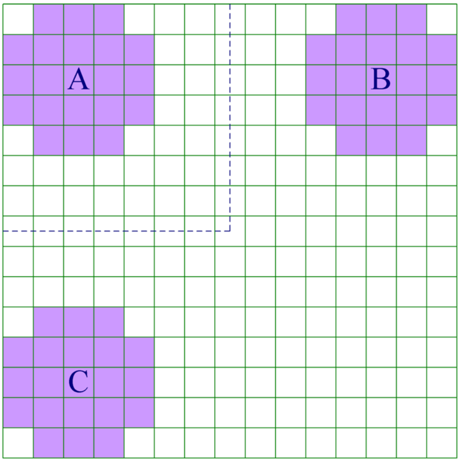
Human factor have strong impacts in visualization, as the way people interact and perceive with a visualization tool could vary dramatically and influence their understanding of the data presented [36]. Thus, it is importance to consider human factor when designing visualization. Human factors include many aspects, like emotion, intelligence, vision ability etc. Herein, we assume that readers are capable to understand the basic graph, emotional controlled and able to identify different colors. And we will concentrate on the human vision limitation, the visual acuity.

The Standard Snellen Chart is the most popular tool for visual acuity (VA) measurement, which is still the major method of VA assessment in clinical settings nowadays. Dutch ophthalmologist Herman Snellen [33] proposed the prototype of the Standard Snellen Chart in 1862, Snellen Chart. In [33], Herman also defined the ability to recognize one of his optotypes when it subtended 5 minutes of arc as *standard vision*. Some individuals with myopia problems may not be able to read the 'E' in 6/6 line, but they can read the 6/6 line with configured glasses. Thus it is reasonable to assume that the observers of graphics displayed in this thesis are of normal visual acuity with or without glasses.

VA of 6/6 is the ability to distinguish the direction of 'E' sized at 8.86 mm at distance of 6 meter. Which means person with VA of 6/6 could recognize the objects at least 1.772 mm length from 6 meter to the objects. From which, we could get the numbers of pixels contains in a graph: (392 × 246 pixels) for column width graph and (800 × 500 pixels) for page width graph. To allow better visualization from human perspective, the pixels a data point require for presentation will be at least 3 × 3 pixels. Which satisfy only the visible requirement, but not analyzable. Thus, we set a data point to 5 × 5 pixels, which allow insights within the data to be further researched. Moreover, the distance between two data points also needs to be considered when defining the number of data points within the image. As shown in Fig. 12, we can see point A, cause there are distances between A and B and A and C, which could be distinguished be human. The distance is defined as 5 pixels as the data points. As a conclusion, 10 x 10 pixels are needed to present each data points. Assume the data points were evenly distributed all over the image, the number of data points within an image will be 965 for column width image and 4,000 for page width image.

To simplify the visualization progress, we will only run the experiment upon page width image and utilize 4,000 data points when visualization. However, because the limitation of the space, some images were shrunk. The Fig. 13a and b shows that the data points in the images were distributed in funnel shape with most of the data were located in the tip and sides of the funny, the pixel occupancy arrived at around 20% of the whole graph layout. Thus the minimum of the data point number should be determined as 4000 × 20% = 800 according to my algorithm, and reduction of the data point number will be applied with a range of [800, 4000] in this case. If the figures shows clear boundaries among classes and clear clusters and the pixel occupancy was 20% as in Fig. 13a and b, the opposite condition

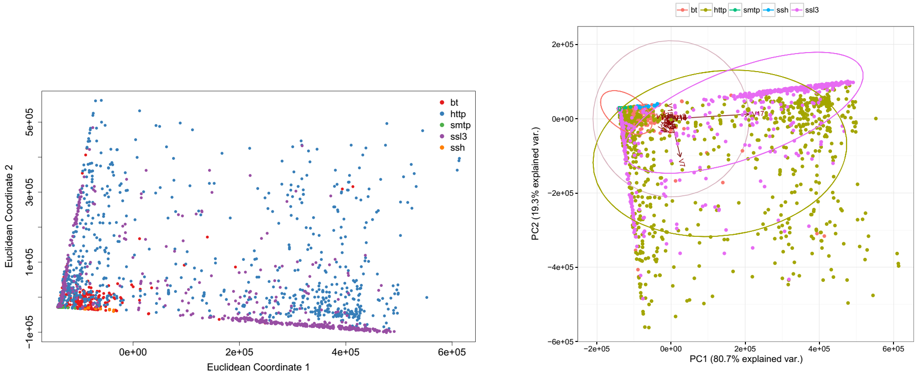
**Fig. 12** There should be distinguished distance between points, so that every points could be recognized from the image



of increasing the data point number happens and the maximum of the data point number should be determined as $4000 \div 20\% = 20000$ and the data point number should ranging from 4,000 to 20,000.

### *4.3.3 Determination of the data points in plotting*

After the range of data point number should be in figures had been defined as 800 to 4000 in the previous chapter, random and even sampling starts. And we tried to find a balance between the sufficient data information and clear and separate data clusters layout for esthetic visual experience and comprehension of the multidimensional data by comparison and analysis of the figure results plotted according to different sampling numbers from the dataset. Figure 13 shows disorderly distribution of the data points and we have to reduce to data points in visualization for better visual effect. Thus, 6 figures pictured with different random and even sampling numbers from the 5 selected classes are shown in Fig. 14.



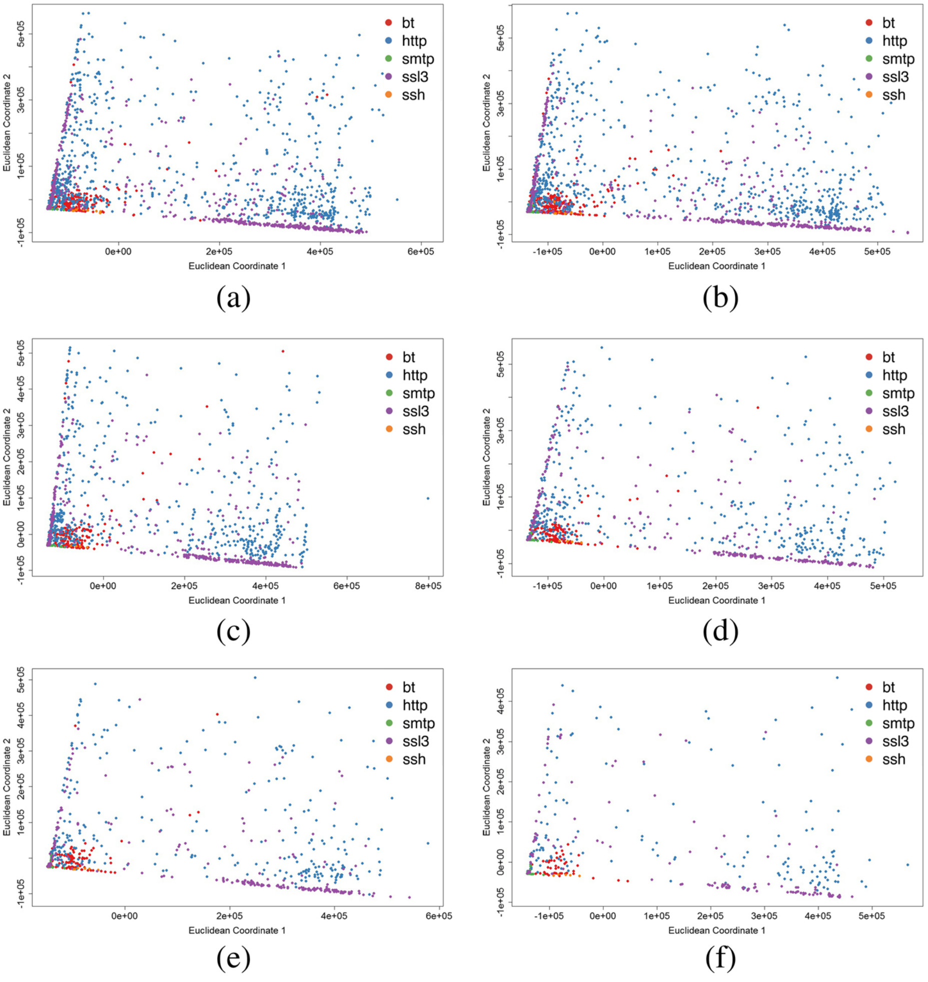
(a) Random Sampling 4000 Data Records in Network Traffic Visualization with MDS

(b) Random Sampling 4000 Data Records in Network Traffic Visualization with PCA

**Fig. 13** Images produced by 4000 random sampling data records

After times iteration of reduction and comparison the final data number was set to 800 when visualization for better visual effect with simultaneous comprehensive data information contained.

The random sampling of data number 1440 and 800 are processed from the original dataset, Fig. 14e and f were plotted with MDS technique accordingly. The data points in the Fig. 14e were separately located at the layout, though the data points in the funnel tip still keep overlapping. The main progress compared with the Fig. 14d were made on the left edge of the funnel, the data points in this position were separate rather than overlapping between HTTP and SSL3 classes as the previous graph. And Fig. 14f outweights it from the perspective of the clear distribution of the data points at the funnel tip. Much more clear boundaries and points were shown separately on the figure, without loss of the precision of data information, which implies that the clusters and distributions of each classes are remain similar as the 4000-sampled Fig. 14a. To perform visible data points in the graphs with main



(a)

(b)

(c)

(d)

(e)

(f)

**Fig. 14** The comparison of 6 figures random sampling from the whole data set with MDS. Separately, 4000 (**a**), 3360 (**b**), 2720 (**c**), 2080 (**d**), 1440 (**e**), 800 (**f**) data points sampled randomly and evenly from 5 classes

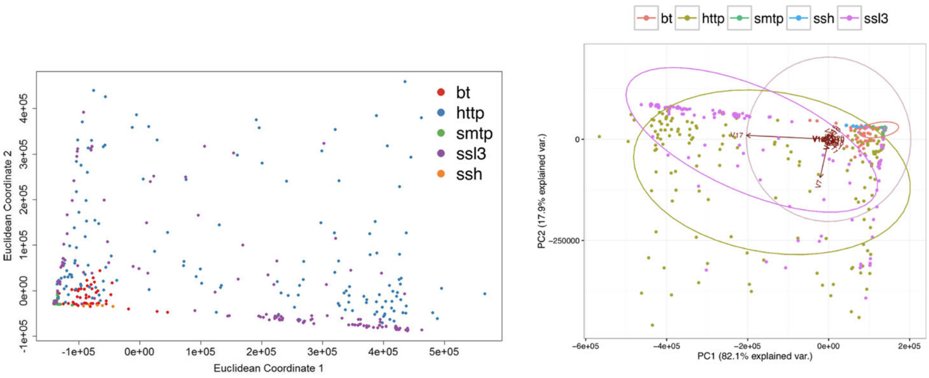
data information contained, and to forms high compactness clusters, the sampling number is set to 800 from the whole dataset and 160 from each class.

# 5 Results and evaluation

## 5.1 Random sampling

The 160 data points of each 5 classes mentioned above were randomly selected and form a new data set contains total of 800 data points for picturing. Visualization process employed two techniques, multi-dimensional scaling and principal components analysis separately, producing 2 figures as in Fig. 15. When utilizing PCA, only first two the most important components were set as axes for pictures to be in 2D manner and another set of colormap was employed with also high contrast and pleasing presentation among classes presented.

As shown in Fig. 15a, the data points were funnel-shaped distributed in the layout with lots of data points were gathered and overlapped on the funnel tip (bottom left corner). It is obvious that the clusters of flows BT, SMTP and SSH were much focused while clusters of HTTP and SSL3 flows were randomly distributed in the graph with no clear boundaries to classify those 5 classes. As for the PCA Fig. 15b, the first principal component presented 82.1% of the information while the second principal component presented 17.9% of the information, which means these two components contains almost all the variation. The arrowhead lines (vectors) present 21 features of data selected to visualization with their directions and lengths together consisting the two new coordinates. Among those 21 features, the 7th and the 17th features contain most significance information to present BT, HTTP, SMTP, SSH and SSL3 classes. It is worth mentioning that, there is a clear boundary as the data presenting, which is the inherent nature of the network traffic flows when collecting them like the locations and time. The result is not satisfactory as the data points were overlapping and it is difficult for people to distinguish 5 classes neither with the MDS or PCA graph.



(a) Random Sampling 800 Data Records in Network Traffic Visualization with MDS

(b) Random Sampling 800 Data Records in Network Traffic Visualization with PCA

**Fig. 15** Images produced by 800 random sampling data records

A different method of sampling is needed for visualization in next stage to improve the graph quality, along with clear boundaries among those class clusters.

## 5.2 Centralized sampling

To address the problem arose from random sampling, the gathering and overlapping of the data points in the graph, a new sampling method is proposed. The MDS technique was by exploiting Euclidean Distance to present the dissimilarities between each pair of different data tuples. In other word, there exists distance, no matter how small it is, between any pair of data points in the graph. Enlarge the graph layout to spread data points in the graph is a theoretical feasible way to fully demonstrate information contained in data points and to discover potential underlying relationship among data records. However, the graph layout is limited in reality, also a large graph is hard to be comprehended [21].

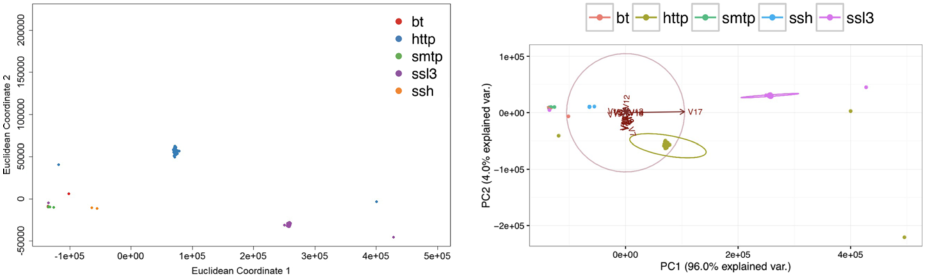
Consequently, we to find the closet data records in each flows, thus to reduce the variance of the distances between pairs of data points. Normally, the median indicates a cluster or a data set best than other statistical characteristics like minimum, maximum and mean. And median remains stable, even the cluster or the data set is changing. As a consequence, we selected 160 data records that are closest to the median point of the 5 classes to form a new 800-recorded data set for visualization.

Median Point of each class is determined by the medians of all 21 features. After collecting 21 medians $m^i$, $i = 1, ..., 21$ derived from the 21 features, 5 fake data records is created as Median Points $MP$ for each class:

$$MP_p = \{m_p^1, m_p^2, ..., m_p^{21}\} \tag{1}$$

where $p \in \{BT, HTTP, SMTP, SSH, SSL3\}$. And the $\{MP\}$ contains 5 fake Median Points of the 5 protocols. And the distance between each data points with the median point is decided by Euclidean Distance.

As shown in Fig. 16, compared with Fig. 15, dramatic improvement were made. Severe imbrication still occurs between BT and SMTP flows, though the clusters of HTTP and SSL3 were clearly defined with few outliers. In the PCA image, 96% of the variation was explained by the first principal component and 4% of the variation explained by the second principal component. To make the data more separate within each class, log transformation was employed.



(a) Centralized Sampling 800 Data Records in Network Traffic Visualization with MDS

(b) Centralized Sampling 800 Data Records in Network Traffic Visualization with PCA

**Fig. 16** Images produced by 800 centralized sampling data records

## 5.3 Log transform

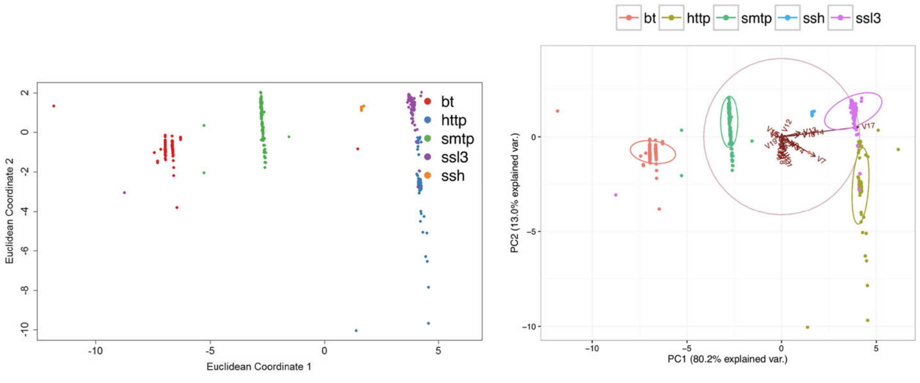
Applying log transform to the data set is another way to spread gather data points and form a clear and pleasing view of the graph. It is obvious from Fig. 17, that all the data points was spreading within their clusters especially classes BT and SMTP. However, the SSH traffic flow data is implicit to distinguish, though it has same point number as other 4 classes, the points we can see are little. As for the PCA graph, the first principal component contains 80.2% of variation and the second principal component consist 13% of variation, which may cause incomplete expression of data information. And all 4 traffic flows, BT, HTTP, SMTP and SSL3 were successfully clustered with clear boundaries and only slight overlapping between classes SSL3 and HTTP.

## 5.4 Protocols contains minority of the original dataset

Another set of images Fig. 18 are plotted by the data points sampling from the rest 13 classes: DNS, EBUDDY, EDONKEY, FTP, IMAP, MSN, POP3, RSP, RTSP, SMB, SSL2, XMPP and YAHOOMSG. 60 data points are randomly sampled from each of the classes and forms a new dataset with 780 data records, because we want to address the imbalance problem in the origin dataset and the smallest class, RTSP, contains only 69 data records. It is easy to discover that those data points distributed as a funnel, too. As same as the previous finding, the data points are mainly located at the funnel tip and the boundaries. From the Fig. 18b, we could see that the clusters derived from the spatial confidence ellipses are parallel or vertical to each other according to the first two principal components. This is an interesting distribution of those clusters, some relative information may be ignored when people utilizing those protocols. However, due to the time limitation, we will leave it to the future work.

## 5.5 Results compared with t-SNE based visualization

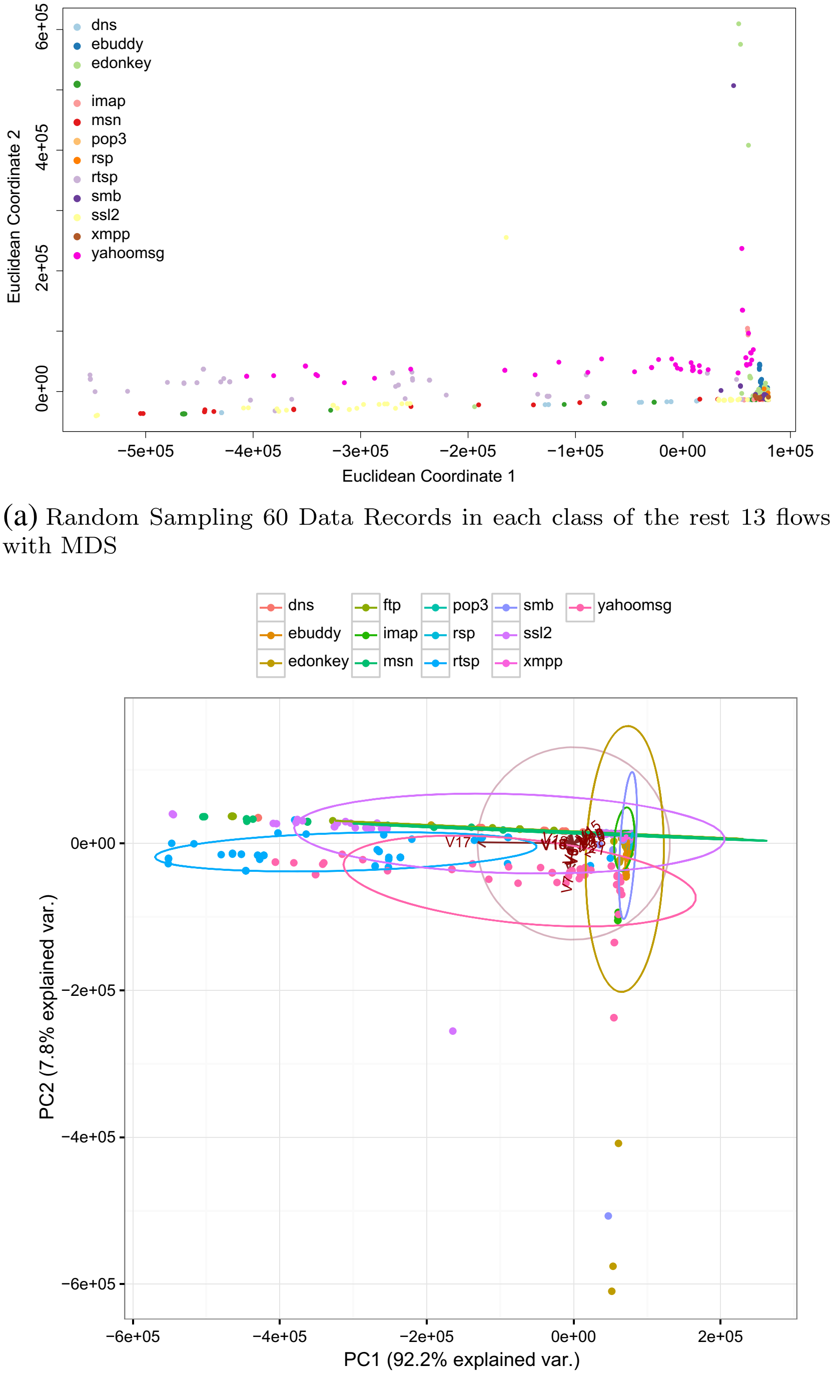
To compare our approach and t-SNE, we selected randomly 160 datapoints from each of BT, HTTP, SMTP, SSH and SSL3 class and built up a 800 dataset for t-SNE based visualization



(a) Centralized Sampling 800 Data Records in Network Traffic Visualization with MDS after Log Transformation

(b) Centralized Sampling 800 Data Records in Network Traffic Visualization with PCA after Log Transformation

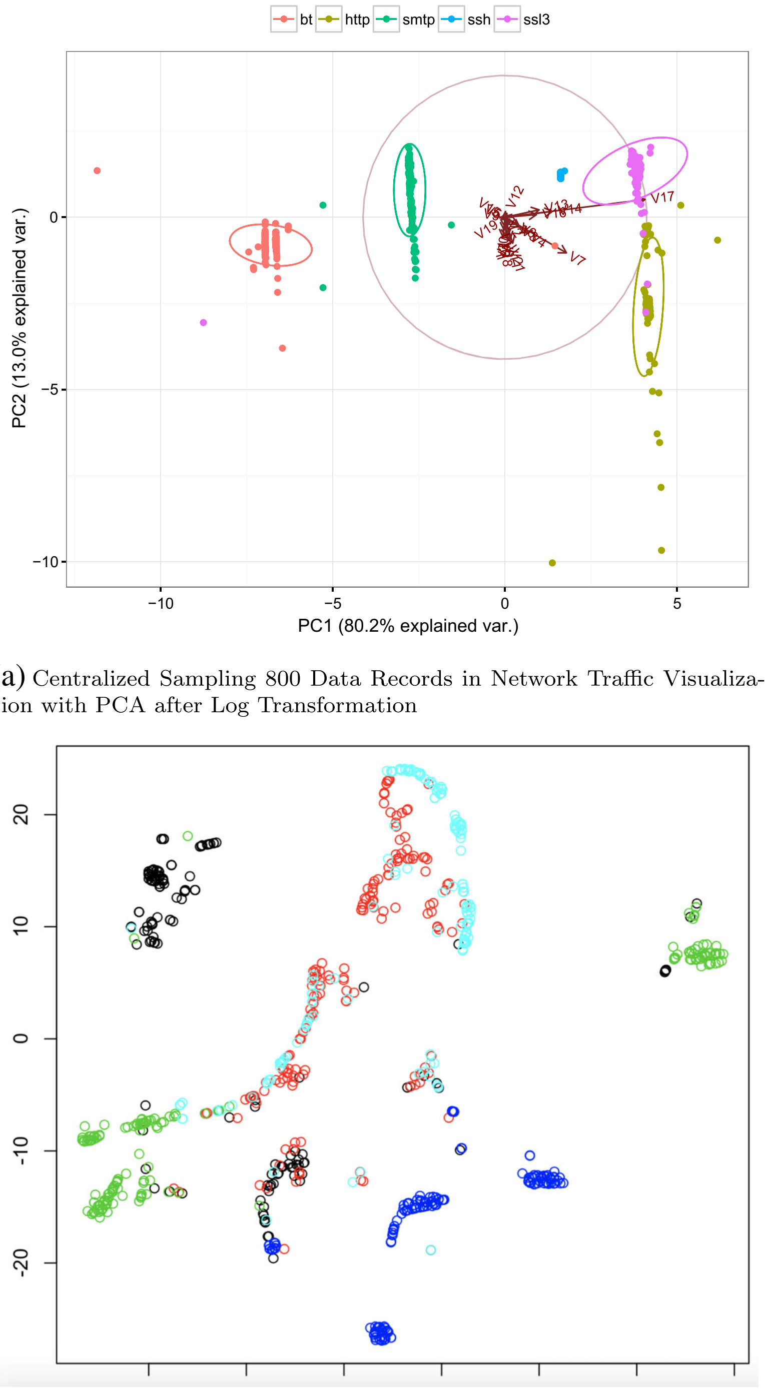
**Fig. 17** Images produced by 800 centralized sampling data records with log transformation



(a) Random Sampling 60 Data Records in each class of the rest 13 flows with MDS

(b) Random Sampling 60 Data Records in each class of the rest 13 flows with PCA

**Fig. 18** Images produced according to the 60 random sampling data records from each of 13 classes



(a) Centralized Sampling 800 Data Records in Network Traffic Visualization with PCA after Log Transformation

(b) The Random Sampling 800 DPN Plotted by t-SNE

**Fig. 19** The output image from our algorithm compared with t-SNE based visualization of randomly and evenly selected 800 DPN from majority part of ‘ISP’

as in Fig. 19b. And the final results generated from proposed visualization algorithm is shown in Fig. 19a. It is obvious that the data point distribution in Fig. 19a is better than that in Fig. 19b. There exists overlapping among all 5 classes and two separation clusters from a same class. Which is an evidence that proposed visualization algorithm performs better in 'ISP' dataset than t-SNE based visualization. And by using the default plot function introduced in R with t-SNE, the color code Fig. 19b applied emphasized on the blue and black colors, which can leads in misunderstandings or neglects over potential information.

## 5.6 Discussion

Visualization results generated from proposed algorithm is satisfactory. According to the comparison between the plotting of randomly selected 10,000 DPN from 'ISP' with MDS in Fig. 20 and final visualization results in Fig. 19a, a remarkable improvement can be easily observed. The overall look of 'ISP' dataset is a mess, even expertise could not perceive any useful information from the image. After proposed algorithm, the visualization result is readable, and the distribution of data point is perfect with a pattern.

The dimensionality reduction methods utilized in this work is PCA and MDS. PCA is a popular methods with wide spread influence and has proven sound effects in effective multi-dimensionality reduction. Although t-SNE performed great in natural dataset MINIST, it didn't give a satisfactory performance in 'ISP' dataset, which indicated a negative effect of t-SNE on big traffic network. Hence, we employed MDS, the initiate in using Euclidean Distance, in the proposed algorithm. And the combination of the two methods made our algorithm more adaptive for big traffic network dataset.

Our proposed visualization algorithm takes human factors into account when processing. Little literature gives attention on human factors, despite its importance. Humans will be the final stage of a visualization process as evaluation. And the limitation on human vision

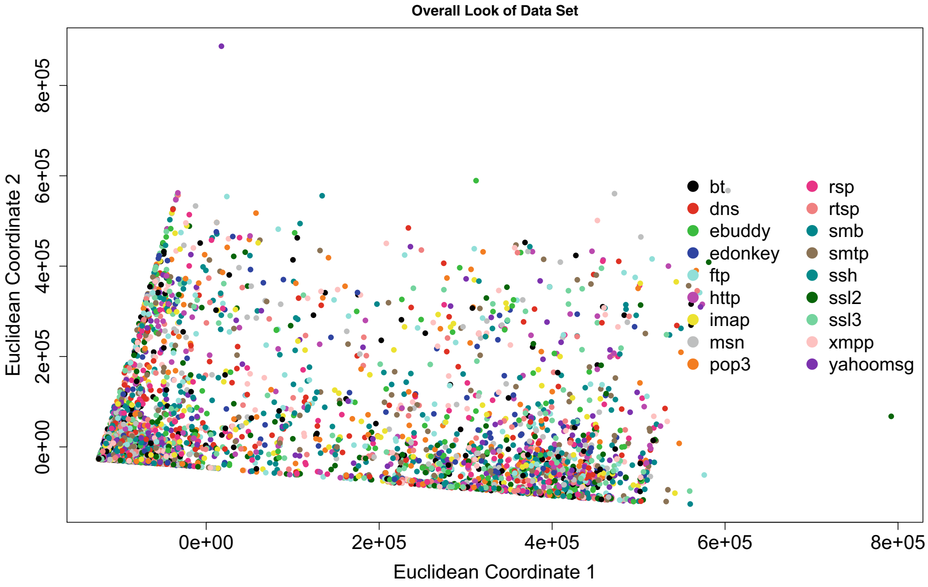


**Fig. 20** The plotting of randomly selected 10,000 DPN from original dataset 'ISP' with MDS

and perception should be considered when visualizing for accurate information delivery. Our algorithm set DPN and color code according to human vision limitation and color perception. Which is a pathway for later researchers who interested in visualization.

The time consumption is significant. And there needs some improvement upon this issue. It took averagely 30 minutes to perform 800 DPN visualization with proposed algorithm, which is an obvious limitation. Our visualization algorithm employs MDS, which is a technique plotting the data points by pairwise Euclidean Distance, thus, running MDS is inherently time-consuming as data points increasing. Also, running time can vastly grow when PCA takes into place as dimensions of origin dataset increasing.

Some visualization results may be not complete and comprehensive enough. Human factors are fully satisfied at the cost of accurate delivery of information. Take 'ISP' dataset for example, two sets of visualization results are generated for easier observation, more esthetics data points layout in the images, and distinguishable color code for classes. Which leads to a lack of evidence in relationships among the majority part and the minority part.

Limitations of the proposed algorithm also lies in human factors. The evaluation of visualization results relies on human labor and knowledge of expertise of this field. However, colors can miss-lead to false understanding of information. And limited usage of colors because of human color perception is a barrier of visualization. Human vision limitation also preserves observation of tiny points.

## 6 Conclusions and future work

Visualization is a bridge connected humans and data, especially, multidimensional, heterogeneous data, like network traffic. It helps people to understand underlying information of dataset, and enlightens researchers and expertise of cyber security by defining anomaly traffic flows. This paper propose a novel visualization algorithm combining human factors in plotting big network traffic. Color code and DPN will be defined before the visualization according to class distribution, and human vision and perception. PCA and MDS are two dimensionality reduction methods employed in proposed algorithm. The final results generated from our algorithm are better compared with t-SNE based visualization, in human-oriented information delivery and readability.

We also indentify several limitations for future improvement. Firstly, running time of proposed algorithm is large. This is an inherent flaw due to calculation of pairwised Euclidean Distance employed in MDS. A in-depth study of MDS in aspect of algorithm can help this situation. On top of that, we need to handle with the problems raised as a consequence of human factors. Human vision limits the datapoint size in the image, while human color perception limits the color code we can use. One solution is utilizing image reduction and segmentation[2] to partition a large image into several smaller ones. In addition, we can employ deep learning algorithms to inspect images and assess their readability for human beings.

## References

1. a4papersize.org: a4 paper size. https://www.a4papersize.org/. Accessed July 4, 2016
2. Al-Ayyoub M, AlZu'bi S, Jararweh Y, Shehab MA, Gupta BB (2018) Accelerating 3d medical volume segmentation using GPUs. Multimedia Tools and Applications 77(4):4939–4958
3. Ashby FG (2014) Multidimensional models of perception and cognition. Psychology Press, Hove

4. Boothe RG (2001) Perception of the visual environment. Springer Science & Business Media, Berlin
5. Borg I, Groenen PJ (2005) Modern multidimensional scaling: theory and applications, Springer Science & Business Media, Berlin
6. Braun L, Volke M, Schlamp J, von Bodisco A, Carle G (2014) Flow-inspector: a framework for visualizing network flow data using current web technologies. Computing 96(1):15–26
7. Brewer CA (1999) Color use guidelines for data representation. In: Proceedings of the section on statistical graphics, american statistical association, pp 55–60
8. Brewer C, Harrower M, The Pennsylvania State University (2013) Colorbrewer 2.0. Accessed July 4, 2016. http://colorbrewer2.org/#type=qualitative&scheme=Set1&n=5
9. Chang X, Ma Z, Lin M, Yang Y, Hauptmann AG (2017) Feature interaction augmented sparse learning for fast kinect motion detection. IEEE Trans Image Process 26(8):3911–3920
10. Chang X, Ma Z, Yang Y, Zeng Z, Hauptmann AG (2017) Bi-level semantic representation analysis for multimedia event detection. IEEE Transactions on Cybernetics 47(5):1180–1197
11. Chang X, Yu YL, Yang Y, Xing EP (2017) Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans Pattern Anal Mach Intell 39(8):1617–1632
12. Dzemyda G, Kurasova O, Zilinskas J (2013) Multidimensional data visualization. Methods and Applications Series: Springer Optimization and its Applications 75:122
13. Elbaham M, Nguyen KK, Cheriet M (2016) A traffic visualization framework for monitoring large-scale inter-datacenter network. In: 12th international conference on network and service management (CNSM), 2016. IEEE, pp 277–281
14. Erl T, Khattak W, Buhler P (2016) Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, Englewood Cliffs
15. Feng Z, Yuan W, Fu C, Lei J, Song M (2018) Finding intrinsic color themes in images with human visual perception. Neurocomputing 273:395–402
16. Fisher DF, Monty RA, Senders JW (2017) Eye movements: cognition and visual perception, vol 8. Routledge, Evanston
17. Glatz E, Mavromatidis S, Ager B, Dimitropoulos X (2014) Visualizing big network traffic data using frequent pattern mining and hypergraphs. Computing 96(1):27–38
18. Gupta B, Agrawal DP, Yamaguchi S (2016) Handbook of research on modern cryptographic solutions for computer and cyber security. IGI Global, Hershey
19. Harrison L, Lu A (2012) The future of security visualization: lessons from network visualization. IEEE Netw 26(6):6–11
20. Harrower M, Brewer CA (2011) Colorbrewer.org: an online tool for selecting colour schemes for maps. The map reader: Theories of mapping practice and cartographic representation, pp. 261–268
21. Herman I, Melançon G, Marshall MS (2000) Graph visualization and navigation in information visualization: a survey. IEEE Trans Vis Comput Graph 6(1):24–43
22. IEEE.org (2017) Ieee publications and standards. Accessed July 4, 2016. https://www.ieee.org/publications_standards/index.html
23. Kim Y, Varshney A (2006) Saliency-guided enhancement for volume visualization. IEEE Trans Vis Comput Graph 12(5):925–932
24. Kumano Y, Ata S, Nakamura N, Nakahira Y, Oka I (2014) Towards real-time processing for application identification of encrypted traffic. In: International conference on computing, networking and communications (ICNC), 2014. IEEE, pp 136–140
25. Lee S, Sips M, Seidel HP (2013) Perceptually driven visibility optimization for categorical data visualization. IEEE Trans Vis Comput Graph 19(10):1746–1757
26. Li J, Yu C, Gupta BB, Ren X (2017) Color image watermarking scheme based on quaternion Hadamard transform and Schur decomposition. Multimedia Tools and Applications pp. 1–17
27. Marsland S (2015) Machine learning: an algorithmic perspective. CRC Press, Boca Raton
28. Munsell AH (1915) Atlas of the Munsell color system. Wadsworth, Howland & Company, Incorporated, Printers, Boston
29. Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Philos Trans R Soc Lond A Math Phys Sci 236(767):333–380
30. Promrit N, Mingkhwan A (2015) Traffic flow classification and visualization for network forensic analysis. In: 2015 IEEE 29th international conference on advanced information networking and applications. IEEE, pp 358–364
31. Robbins NB (2012) Creating more effective graphs. Wiley, New York
32. Shiravi H, Shiravi A, Ghorbani AA (2012) A survey of visualization systems for network security. IEEE Trans Vis Comput Graph 18(8):1313–1329
33. Snellen H (1873) Probebuchstaben zur bestimmung der sehschärfe (Vol. 1). H. Peters

34. Staheli D, Yu T, Crouser RJ, Damodaran S, Nam K, O'Gwynn D, McKenna S, Harrison L (2014) Visualization evaluation for cyber security: trends and future directions. In: Proceedings of the eleventh workshop on visualization for cyber security. ACM, pp 49–56
35. Stone M (2016) A field guide to digital color, CRC Press, Boca Raton
36. Tory M, Moller T (2004) Human factors in visualization research. IEEE Trans Vis Comput Graph 10(1):72–84
37. van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(Nov):2579–2605
38. Varela FJ, Thompson E, Rosch E (2017) The embodied mind: cognitive science and human experience, MIT Press, Cambridge
39. Wang L, Giesen J, McDonnell KT, Zolliker P, Mueller K (2008) Color design for illustrative visualization. IEEE Trans Vis Comput Graph 14(6):1739–1754
40. Wang Y, Xiang Y, Zhang J, Zhou W, Wei G, Yang LT (2014) Internet traffic classification using constrained clustering. IEEE Trans Parallel Distrib Syst 25(11):2932–2943
41. Ward MO (1994) Xmdvtool: integrating multiple methods for visualizing multivariate data. In: Proceedings of the conference on visualization'94. IEEE Computer Society Press, pp 326–333
42. Ware C (2012) Information visualization: perception for design, Elsevier, Amsterdam
43. Xiao L, Gerth J, Hanrahan P (2006) Enhancing visual analysis of network traffic using a knowledge representation. In: 2006 IEEE symposium on visual analytics science and technology. IEEE, pp 107–114
44. Zhang J, Chen C, Xiang Y, Zhou W, Xiang Y (2013) Internet traffic classification by aggregating correlated naive bayes predictions. IEEE Trans Inf Forensics Secur 8(1):5–15
45. Zhang J, Xiang Y, Wang Y, Zhou W, Xiang Y, Guan Y (2013) Network traffic classification using correlation information. IEEE Trans Parallel Distrib Syst 24(1):104–117
46. Zhang J, Chen X, Xiang Y, Zhou W, Wu J (2015) Robust network traffic classification. IEEE/ACM Trans Networking (TON) 23(4):1257–1270
47. Zhang Z, Sun R, Zhao C, Wang J, Chang CK, Gupta BB (2017) Cyvod: a novel trinity multimedia social network scheme. Multimedia Tools and Applications 76(18):18,513–18,529
48. Zhou L, Hansen CD (2016) A survey of colormaps in visualization. IEEE transactions on visualization and computer graphics 22(8):2051–2069

**Zichan Ruan** recieved her Bachelor of IT and Bachelor of IT (Honors) degrees at Deakin University, Melbourne, Australia.