Indonesian Journal of Electrical Engineering and Computer Science
Vol. 16, No. 3, December 2019, pp. 1600~1608
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i3.pp1600-1608

Overlapping issues and solutions in data visualization techniques

Nur Diana Izzati Husin, Nur Atiqah Sia Abdullah
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia

Article Info

Article history:

Received Mar 30, 2019
Revised Jul 22, 2019
Accepted Jul 30, 2019

Keywords:

Big data
Data visualization
Multidimensional
Network visualization
Overlapping

ABSTRACT

The tremendous growth of big data has caused the data visualization process becomes more complex and challenging, and yet, data is expected to be increased from time to time. With these massive and complex data, it is getting harder for the data analyst to interpret or read the data in order to gain new knowledge or information. Therefore, it is important to visualize these data using different techniques. However, there are many remaining issues in data visualization techniques. These issues make the data visualization a big challenge to the data analyst. The most common issue in data visualization techniques is the overlapping issue. This paper reviews the overlapping issues in multidimensional and network data visualization techniques. The existing solutions are also reviewed and discussed in term of advantages and disadvantages. This paper concludes the advantages of the overlapping issues and solutions, before discussing their drawbacks. This paper suggests the color-based approach, relocation, and reduction of data sets to solve the overlapping issues.

Corresponding Author:

Nur Atiqah Sia Abdullah,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia.
Email: atiqah@tmsk.uitm.edu.my

1. INTRODUCTION

The representation of data in a form of pictorial or graphical format is referred to as data visualization [1]. Data visualization is the concept of representing data using the use of pictures and it has been around for centuries [1]. Data visualization also can be described as making something visible [2] and helps the users to analyze difficult datasets by revealing a variety of information [3, 4]. Thus it saves time by making the process of knowledge acquisition much faster [2, 5]. Nevertheless, the visualization helps to grasp any difficult concept and identify hidden patterns in the data [2, 6].

However, before using any data visualization, there are things that need to be considered such as what are the main goals, the needs of visualization, and the audience. Besides that, the user also needs to take into consideration on the main big data challenges, which are the velocity, volume and variety. This is because the data generated faster that it can be managed and analyzed. The biggest problem in visualizing data is choosing the suitable technique.

There are many types of data visualization techniques available. For example, geometric, parallel coordinate, stick figure, icon based, hierarchical, graph based and pixel oriented techniques [1]. In order for the data visualization to be more accurate, the user needs to choose the right technique. There are several classification of visualization such as treemap, circle packing, sunburst, parallel coordinate, stream graph and circular network diagram [7].

Apart from all the advantages of data visualization, there are also some drawbacks. Data visualization can be a challenging task as there are many techniques that can be used. Different types of data

set use different type of techniques. Certain types of data may not be suitable for certain types of visualization technique. Therefore, it is crucial to choose the right technique to visualize a dataset. Although the right visualization technique has been chosen, there is also some other issues in data visualization, such as overlapping [8], relationship, interpretation and connections. This paper aims to find out the types of overlapping issues and their solutions in different data visualization techniques.

2. RESEARCH METHOD

This section explains the reviews of the overlapping issues and their solutions in multidimensional and network data visualization techniques.

2.1. Overlapping Solutions in Multidimensional Data Visualization

Overlapping issues in Euler and Venn diagrams are the most talked issue as it is among the oldest and most popular set visualizations [9-11]. In Euler diagram, any set exclusion, inclusion and intersection can be represented as there are no restrictions on how the curves overlap. Meanwhile, for Venn diagram, it is more restricted than the Euler diagram as it has to show all possible combination of curve overlaps. Thus, Venn diagram quickly becomes visually complex as more sets are depicted.

Variants of Euler diagrams for different purposes can be used to tackle the overlapping issues. For instance, where the Euler diagram cannot properly represented will be visualized by splitting or duplicating certain sets and subsets into disjoint parts, and connecting these parts using edges [12] as shown in Figure 1.

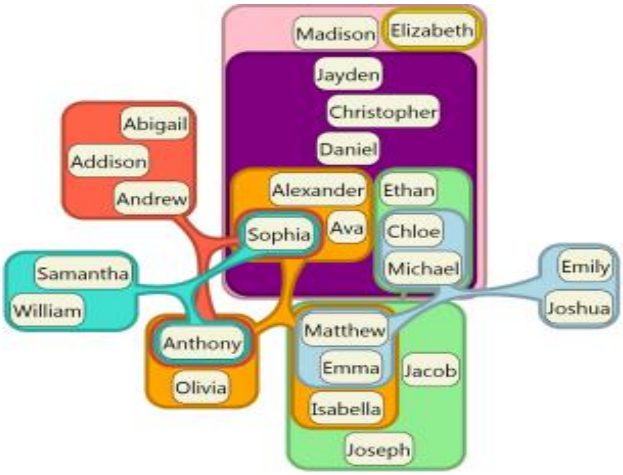

Figure 1. Disjoint Subsets and Edge Connection in Euler Diagram [9]

The overlapping issue is also found in bubble sets [9]. Normally, the bubble sets are assigned with semi-transparent colors to reveal their overlapping and to keep the context visualization visible [13]. This technique can only handle between four to twenty sets and still retains enough visibility of the context [14], [15]. However, if the datasets are more than that, the overlapping issue can be resolved by using texture splatting [14, 15], which it depicts the area of interests into a diagram. Splatting is applied to a skeleton constructed from the diagram elements according to their sizes and positions. Overlaps between multiple areas of interests are emphasized using subtractive color blending, which creates darker overlapping region. Then, texture and color are used to encode the area of interest as shown in Figure 2.

The line-based techniques are the most effective visualization type for users to analyze and explore data. It is also the most common way to represent any continuous data. Line-based techniques provide a visual patterns for slopes, curvature, crossing and further line patterns [16]. LineSets [9, 17] is proposed to overcome the overlapping issue in line-based techniques, as shown in Figure 3. It improves the readability of the complex set and hence to minimize the clutter by reducing the set regions by computing a line for each set that passes through its elements using travelling salesman heuristic that minimizes the line length. Although it is claimed to be better than bubble sets methods, the use of simple lines imposes an artificial ordering on the set elements.

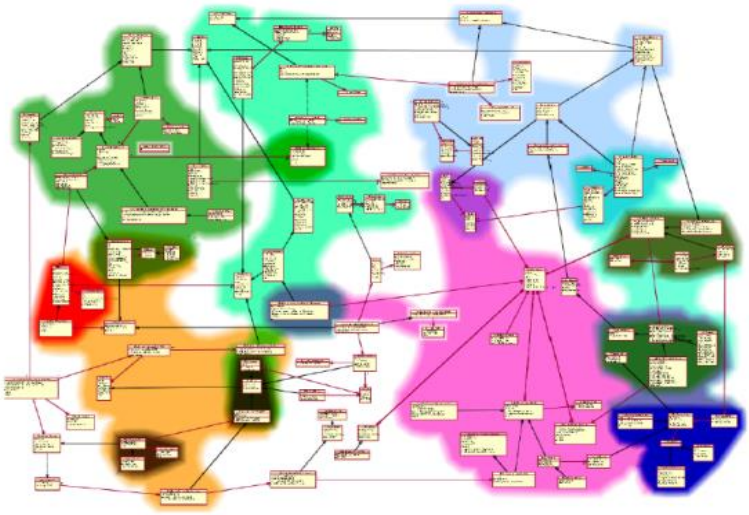

Figure 2. Texture Splatting in Bubble Sets [9]

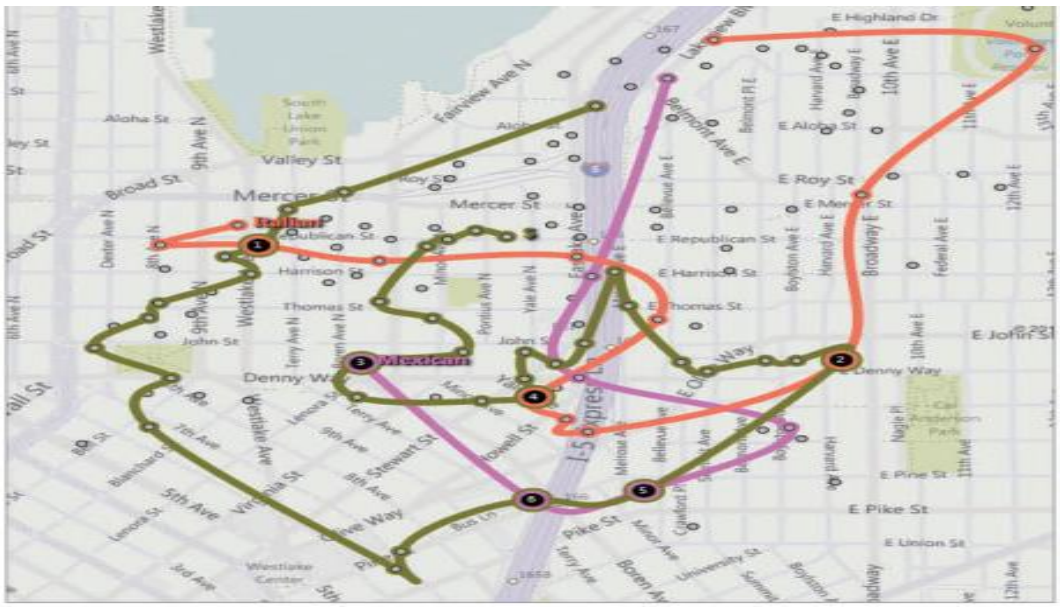

Figure 3. Example of LineSet [17]

Kelp diagram [9, 11, 18] is used to avoid overlaps in line-based methods that are based on spanning graph. This method incorporates classical graph drawing that consists of bubble and sticks or a tree spanner over the member points in a set [11]. It connects the elements in a set using a graph structure instead of a simple line and surrounding each element with a circle clipped to its Voronoi cell. Figure 4 shows the nested style that draws links over each other, with thinner links on the top to ensure their visibility.

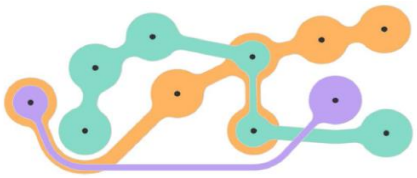

Figure 4. Nested Style of Kelp Diagram [11]

Meanwhile, the striped style (see Figure 5) uses alternating stripes for areas that contain elements of multiple sets. Another method that is similar to the Kelp diagram method is the KelpFusion [11]. KelpFusion incorporates the use of a proximity graph that is called shorted-path graph. With the use of shortest-path graph, KelpFusion can fill faces when many points are spatially closes to each other. By using this method, it can visualize corresponding boundary efficiently and enabling interactive manipulation of the visualization. There also exists overlapping problems in glyph-based technique as glyphs can be used to simply overlays set memberships. However, overlap can occur if the membership is too many. Therefore, colored pie-like glyphs were used in order to visualize the fuzzy membership of overlapping [9, 19].

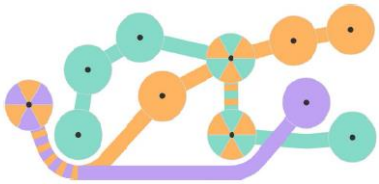

Figure 5. Striped Style of Kelp Diagram [11]

Another way to overcome overlapping in glyph based technique is by using scatterplot [20]. Scatterplot is a technique that consists of two axes and glyphs that are being used to represent the data points. The primary feature of scatterplot is that it represents every data in the view individually. This makes it an excellent technique at highlighting clusters, outliers and trends. Moreover, by decreasing the amount of glyphs and engaging opacity can manage the overlap thus achieving high data intensity.

2.2. Overlapping Solutions in Network Visualization

Overlapping in network visualization is the major issue of the visualization that happens in everyday life when involves the visual clutter of nodes and edges [19, 21, 22]. The issue of overlapping communities had been an attention and there has been many algorithms or techniques that have been developed [19, 23]. In large networks, normally it is a challenge to read node-link diagram due to the overlap, overdraw and clutter [24]. Overlapping of nodes and links may cause occlusion and ambiguity in the graph representation thus reducing the potential usefulness of the visualization. Some static techniques can be used to overcome overlapping issue in network visualization [21].

The first method is by reducing the number of items [21]. However, the downside of using this method is that by reducing the number of items, it causes the loss of information or relevant links and nodes. The second approach is by using color-based technique [21]. This technique can be used in different ways in order to overcome the issue. By using the color-based technique, map the orientation of links to the color, which it reduces the ambiguity between crossing links but this may be a problem when the links have small crossing angle as they may have similar color. The other technique that can also be used is by relocating the node and links. Figure 6 shows an example of color-based links.

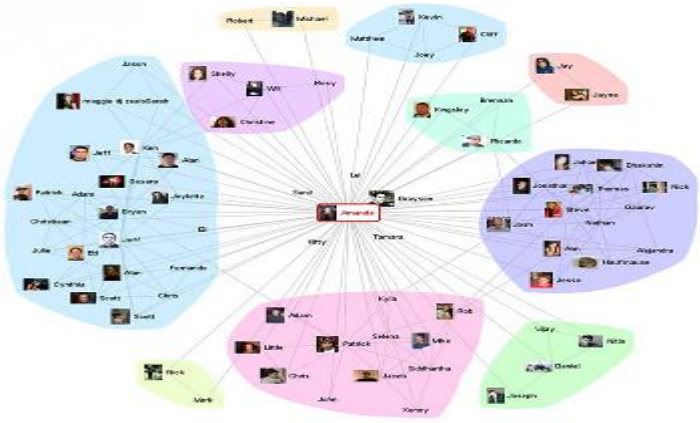

Figure 6. Color-based link [21]

Meanwhile, layered visualization technique overcomes the fuzzy overlapping communities [19]. This technique uses different aggregation levels that are described by a level of interest function. The function will aggregate the nodes of a particular degree of fuzziness, which is being described by the threshold θ. The example on the technique is shown in Figure 7 with the sequence of graph showing the fuzzy overlapping community.

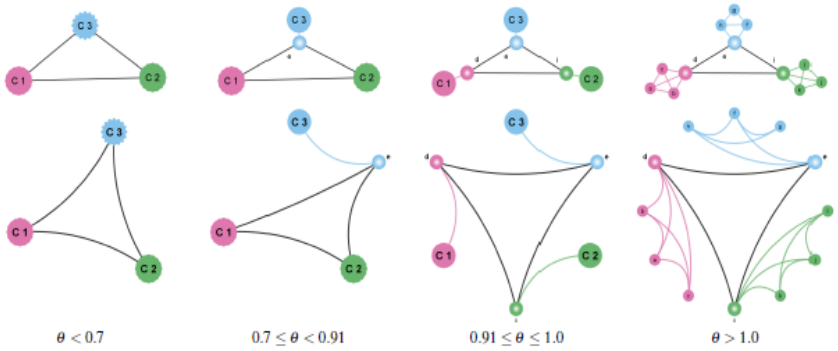

Figure 7. Fuzzy overlapping community [19]

Overlapping issue also occurs when a set of point that is in a fixed positioning that needs to be visualized. For an overlapping point set to be visualized effectively, it needs to have certain criteria and one of it is that the data points need to be unambiguous and it should also need to represent the geometrical layout of the points as close as possible [25]. Overlapping in geographical techniques is also a common issue. A solution by solving the movement of data in a geographical visualization by formulating a new model of circos figure is proposed [26]. This figure is used to show the interchange patterns as a junction nodes and optimizing the assignment of color to the respective connections within and between the junction nodes. However, the first circos design has a certain issue that is visual confusion and visual cluttering as shown in Figure 8.

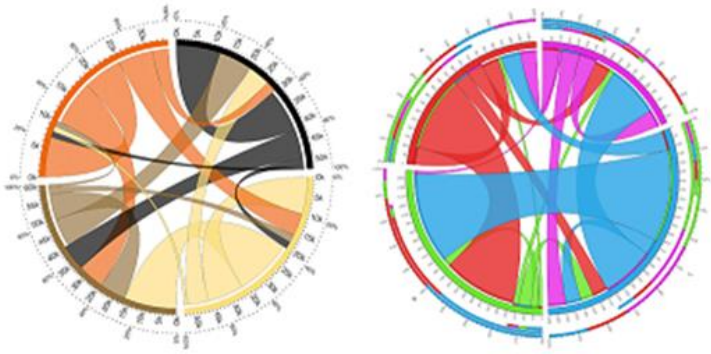

Figure 8. Example of circos figures (Zeng et al., 2013)

Therefore, they proposed second technique [26] to overcome the overlapping of arc element as shown in Figure 8. Every arc element in the visualization is positioned so that the arc is point towards its link direction. If there is any neighboring arc elements that are too close to each other and if does, make them repel from each other. The process is repeated iteratively until every pair has minimum gap of 10 degree from each other. The example is shown is Figure 9.

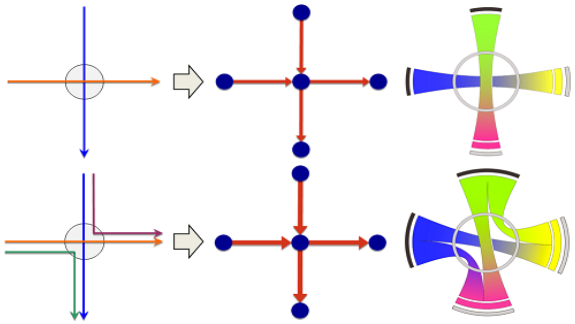

Figure 9. Comparison of existing visualization and circos diagram [26]

Some visualizing techniques overcome the problem of overlapping by limiting the number of sets and overlaps that can be visualized once. Some other visualizing techniques avoid the overlapping problems by explicitly and convey more abstract information about the set system instead. The reason behind its complexity is the exponential growth of possible overlaps according to the number of sets.

3. RESULTS AND ANALYSIS

Overlapping in data visualization techniques are mostly being solved by reducing the data set. This issue occurs because of the extensive amount of data that being visualized at one time. Most techniques cannot handle too many data set or points hence this will produce the overlapping in the visualization. Although there have been many solutions for every overlapping problem in various data visualization techniques, there still have some drawbacks for these solutions. The solutions are basically categorized into three main approaches, which are using color-based approach, relocation, and lastly reduction of data sets.

For instance, Bubble sets [9], Kelp diagram [9, 11, 18], graph-based [21] and LineSets [9, 17] are using color-based approaches to overcome the overlapping issue. The overlapping issue in Bubble set can be solved by assigning semi-transparent color; however it can only handle data sets between 4 and 20 in order to remain its readability. If the data sets are more than 20, then it can be resolved by using splatting approach. Meanwhile for Kelp diagram, it uses color to represent its nested and stripes kelp diagram. The advantage of using Kelp diagram is because the consistency and easy to interpret. However the routing algorithm used by Kelp diagram is too slow for interactive use. Other than that, a graph-based also uses color-based approaches to overcome the overlapping. By using the color-based method, it reduces the ambiguity between crossing links but it might have similar color if the crossing angle is small. LineSets uses different color to differentiate the relationships. The perk of using LineSets is better than using bubble set method. Nevertheless, this method occupies more area when anode contains many datasets.

Secondly, overlapping issues can be solved by using relocation approach. As for relocation approach, it is adopted by several data visualization solutions such as the Euler and Venn diagram [12], ScatterPlot [20], graph-based [19], network based [19] and circos figure [26]. In Euler and Venn diagram the sets is split or duplicated into disjoint parts. This method preserves the continuity of the set regions but the hyperedges contain no elements hence the mutual crossing show no shared elements between the sets. Next technique that uses relocation approach is Scatterplot. This technique is excellent for highlighting clusters, outlier and trends. However, this technique not all similarity measures defines a distance function thus limiting the applicability of a 2D projections.

Relocation is also used in graph-based data visualization techniques however this relocation approach may cause the context of the background map to be lost. Other mentioned solution that uses relocation is the network-based layered technique. This technique takes into account the fuzziness of the nodes memberships but this may cause the shared nodes to be far away from the communities that they belong. The last data visualization that uses relocation approach is the circos figure. The first version of the circos figure is good at examining mutual relationship among genomes but this version gives cluttered and confused visual. Meanwhile the second version of the circos figure able to present clearly the difference in the interchange patterns but this method depends on the time resolution chosen thus affecting the size of the interchanged data. Table 1 show summary of overlapping issues and solutions in multidimensional visualization and Table 2 show the summary of overlapping issues and solutions in network visualization.

Table 1. Summary of Overlapping Issues and Solutions in Multidimensional Visualization

Technique	Overlapping Issue	Solution	Advantage	Disadvantage
Euler and Venn Diagram	No restriction on curve overlap	Splitting or duplicating set and subsets into disjoint parts	Parts are connected with hyperedges	Hyperedges contain no elements; No shared elements between the datasets
Bubble Sets	Overlap when more datasets	Assigning semi-transparent color	Use splatting if more than 20 datasets	Only handle datasets between 4 and 20
Line-based	Overlaps in slope, curvature, crossing and line patterns	LineSets	Better than the bubble set method	Use more area when a node is contained in many sets
Line-based	Overlapping of lines	Kelp diagram connects elements use graph structure; Nested and stripes Kelp diagram	Consistent and easy to interpret	Routing algorithm is too slow for interactive use
Glyph-based	Too many overlaying the set membership	Colored pie-like glyph	Reduce the ambiguity	Similar color might cause visual confusion
Glyph-based	Too many overlaying the set membership	Scatterplot has two axes and glyph to represent the points	Excellent for highlighting clusters, outliers and trends	Not all set similarity measures a distance function and it limits the applicability of 2D projections

Lastly, reducing data sets or point is another common approach to overcome the overlapping issue. However; this method will cause loss of information as data with meaning may be removed during the process. The reducing data sets approach is used in graph visualization. Besides, overlapping in graph visualization can also be solved by relocating the nodes and links, or applying color-based techniques. The reduction approach is also applied to the Bubble set technique as the technique can only handle certain amount of data sets.

Table 2. Summary of Overlapping Issues and Solutions in Network Visualization

Technique	Overlapping Issue	Solution	Advantage	Disadvantage
Graph	Occlusion and ambiguity	Reduce the number of items; Color-based technique; Relocating the node and links	Increase readability; Reduce the ambiguity between crossing links;	Loss of information; Small crossing angle have similar color; Risk of losing background map context
Layered	Overlapping of communities	Using different aggregation level	Consider fuzziness of the nodes memberships	Shared nodes are positioned far from the communities
Geographical	Overlapping point set in a fixed position; Overlapping of arc element	Circos figure v1 has interchange patterns; Circos figure v2 arc element is positioned	Examine mutual relationship among genomes; Present difference in the interchange patterns emerged	Visual clutter and visual confusion; Scalability depends on the time resolution and the size of the interchanged data

4. CONCLUSION

Data visualization has been used for centuries and it is an emerging field as it is being used by many areas. With the use of data visualization, the user can understand any kind of data easily with the help of patterns. With the overlapping issues that happened in many data visualization techniques, this paper provides better understanding on the overlapping issues and suggested solutions in the previous studies. Many solutions have been developed to solve the overlapping issue in multidimensional and network type of data visualization. This paper reviewed these solutions and elaborated on the advantages and disadvantages for these solutions. Most of the solutions use data set reduction, color-based, and relocation approaches to overcome the overlapping issues.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA in supporting this paper.

REFERENCES

[1] J. J. Hilda, et al., "A review on the development of big data analytics and effective data visualization techniques in the context of massive and multidimensional data," *Indian Journal of Science and Technology*, vol. 9, pp. 1-13, 2016.
[2] J. Liu, et al., "A Survey of Scholarly Data Visualization," *IEEE Access*, vol. 4, pp. 1-1, 2018.
[3] N. A. S. Abdullah, et al., "Budget Visual: Malaysia Budget Visualization," *Soft Computing in Data Science, 2017. SCDS 2017. Communications in Computer and Information Science*, Springer, Singapore, vol. 788, 2017.
[4] N. A. S. Abdullah, et al., "Review on sentiment analysis approaches for social media data," *Journal of Engineering and Applied Sciences*, vol. 12, pp. 462-467, 2017.
[5] Q. Liu, et al., "A novel data visualization approach and scheme for supporting heterogeneous data," *IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference, 2017. ITNEC 2017*, pp. 1259-1263, 2017.
[6] E. Liu, et al., "Analysis Disease Progression Using Data Visualization," *2017 IEEE International Conference on Internet of Things (IThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 877-882, 2017.
[7] S. M. Ali, et al., "Big data visualization: Tools and challenges," *2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 656-660, 2016.
[8] L. J. Baptiste, et al., "Rainbow boxes: A technique for visualizing overlapping sets and an application to the comparison of drugs properties," *International Conference on Information Visualisation*, pp. 253-260, 2016.
[9] B. Alsallakh, et al., "Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges," *Eurographics Conference on Visualization (EuroVis)– State of The Art Reports*, pp. 1-21, 2014.
[10] A. Lex, et al., "UpSet: Visualization of intersecting sets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, pp. 1983-1992, 2014.
[11] W. Meulemans, et al., "KelpFusion: A hybrid set visualization technique," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, pp. 1846-1858, 2013.
[12] N. H. Riche, et al., "Untangling Euler diagrams," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 1090-1099, 2010.
[13] C. Collins, et al., "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, pp. 1009-1016, 2009.
[14] H. Byelas, et al., "Visualization of areas of interest in component-based system architectures," *32nd Euromicro Conference on Software Engineering and Advanced Applications, SEAA*, pp. 160-167, 2006.
[15] H. Byelas and A. Telea, "Visualizing metrics on areas of interest in software architecture diagrams," *IEEE Pacific Visualization Symposium, PacificVis 2009*, pp. 33-40, 2009.
[16] K. Nazemi, "Adaptive Semantics Visualization," 2014.
[17] B. Alper, et al., "Design Study of LineSets," *Novel Set Visualization Technique*, vol. 17, pp. 2259-2267, 2011.
[18] K. Dinkla, et al., "Kelp Diagrams: Point Set Membership Visualization," *Computer Graphic Forum*, vol. 31, pp. 875-884, 2012.
[19] C. Vehlow, et al., "Fuzzy overlapping communities in networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, pp. 2486-2495, 2013.
[20] R. Sadana and J. Stasko, "Designing and implementing an interactive scatterplot visualization for a tablet computer," *2014 International Working Conference on Advanced Visual Interfaces - AVI '14*, pp. 265-272, 2014.
[21] A. Debiasi, "Study of Visual Clutter in Geographic Node-Link Diagrams," 2016.
[22] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Journal Phys. Rev. E*, vol. 68, 2003.
[23] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75-174, 2010.
[24] S. Elzen, et al., "Multivariate Network Exploration and Presentation: From Detail to Overview via Selections and Aggregations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, pp. 2310-2319, 2014.
[25] J. Vihrovs, et al., "An Inverse Distance-Based Potential Field Function for Overlapping Point Set Visualization," *2014 International Conference on Information Visualization Theory and Application, IVAPP*, pp. 29-38, 2014.
[26] W. Zeng, et al., "Visualizing interchange patterns in massive movement data," *Computer Graphics Forum*, vol. 32 (3 PART3), pp. 271-280, 2013.

BIOGRAPHIES OF AUTHORS

Nur Diana Izzati Husin is a postgraduate student, who is currently taking Master of Computer Science in Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia. Her master thesis is on Dengue Hospot Visualization using Relocation and Color-based Approach. This paper is a part of her unpublished thesis in order to review the overlapping issues and solutions before choosing the appropriate techniques to solve the visualization problem in Dengue dataset.

Nur Atiqah Sia Abdullah is an Associate Professor in Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia. Her PhD is in Software Effort Estimation Metric. Her current research interests include data visualization, sentiment analysis, and software engineering. She has a number of postgraduate students and consistently contributes academic writing to conferences and journals. She has several national research grants that lead to significant publications. She is also active in participating in various Invention, Innovation and Design competitions.