Set Visualization Challenges for Big Data

Luana Micallef

Helsinki Institute for Information Technology, Aalto University, Finland
luana.micallef@hiit.fi

Abstract. This talk will provide a brief overview of the state-of-the-art of set visualization, followed by an in-depth discussion of challenges and open questions when dealing with real-world set-typed data.

Keywords: Sets, visualization, big data.

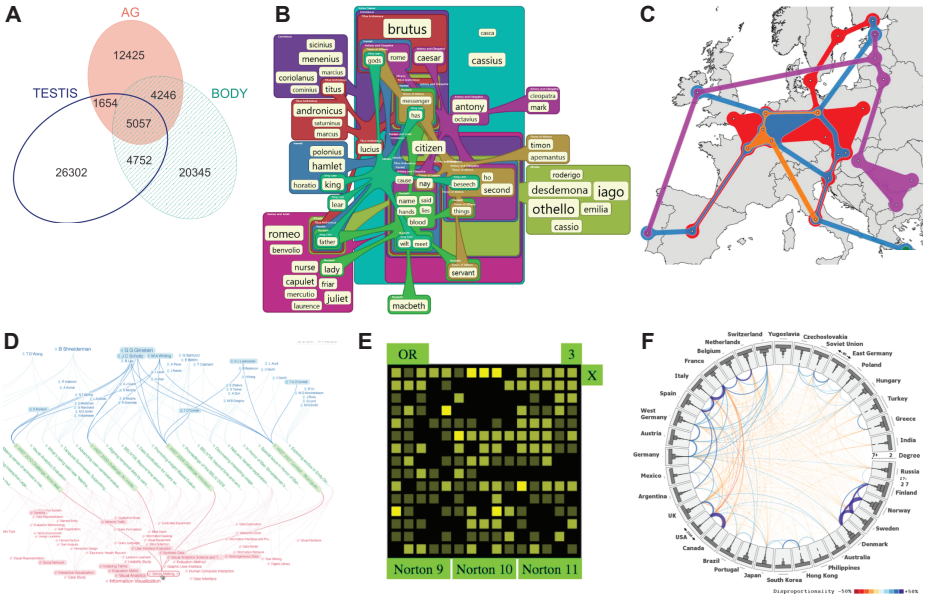

Fig. 1. Different set visualization techniques depicting real-world data. (A) eulerAPE's [8] *area-proportional 3-Venn diagram* showing genomic variations of three tissue types. (B) ComED's [10] *Euler diagram variant* visualizing commonly used words in Shakespeare's plays. (C) KelpFusion's [7] *overlay* visualization showing cities that are members of different EU communities like the Eurozone. (D) PivotPaths's [5] *node-link* diagram depicting connections between publications, authors and keywords. (E) OnSet's [12] *matrix-based* set visualization showing similarities between blood samples of different whale sharks. (F) Radial Set's [1] *aggregate-based* set visualization depicting relationships between IMDb movies produced in different countries.

Data is often organized into groups or sets to provide analysts with an overview of shared properties and help them identify patterns and relationships between the data items and the sets. For instance, links between social communities are analysed to predict and disrupt crimes, while relationships between groups of genes are studied to find cures to illnesses. Set visualization can help in the analysis of such set-typed data. However, due to advances in data collection technology, real-world data is getting bigger and more complex, imposing further challenges and a greater demand for scalable set visualizations that are optimized for the user's data analysis tasks.

Alsallakh et al. [3, 2] categorized set visualization techniques into seven categories: (A) Euler and Venn diagrams; (B) Euler diagrams variants; (C) overlays; (D) node-link diagrams; (E) matrix-based diagrams; (F) aggregate-based diagrams; (G) scatter plots and other. Figure 1 illustrates examples of the techniques in categories A-F, all of which visualize real-world set-typed data.

Euler and Venn diagrams are widely used to reason about sets and their relationships. For instance, Euler diagrams are used to teach set theory to schoolchildren, and to reason about biomedical data (e.g., Figure 1A). Their closed curves form clearly bounded *common* regions [9] with a preattentive popout effect that is stronger than the Gestalt powerful laws of proximity and similarity [6].

However, Euler diagrams are not scalable and are unable to depict large data collections with numerous sets and set relations [11]. This led to the development of diverse set visualization techniques, such as Figure 1B-F, where for instance relationships are depicted as the edges of a graph (Figure 1D) or the cell values of a matrix (Figure 1E). In some cases, the sets and their relations have to be shown on a pre-defined visualization where set elements have a fixed pre-defined position; for such cases, overlay set visualization techniques, like Figure 1C, have been devised. Visualizing aggregate information about the sets, such as their cardinality, is often helpful when reasoning about sets. For two or three sets, it is possible to have an area-proportional Euler diagram like Figure 1A, but for more sets, other techniques like Figure 1F would have to be used. No current technique is considered appropriate in handling more than around 100 sets [3], despite that most real-world data is set-typed, large and often multi-dimensional, particularly in areas like biosciences, security and social networking.

There are various set visualization challenges and open questions which need to be investigated further when dealing with big data:

- Faster drawing algorithms that are tailored to the user's data analysis tasks and needs are required.
- Established information visualization and human-computer interaction methodologies, such as focus+context techniques [4] or Shneiderman's [13] information-seeking mantra of overview first, zoom and filter, then details-on-demand, should be adopted for set visualization.
- Cognitive and perception theories should also be taken into account, so set visualizations exploit and mitigate the capabilities and limitations of the human information processing system.

- Data mining and machine learning techniques could facilitate the selection and visualization of important aspects, patterns and trends in set-typed data.
- Evaluation of the effectiveness of set visualization techniques for big data is also important and difficult due to the different features and characteristics of the visualization system and the data, and the data analysis tasks the user wants to accomplish.

We discuss such challenges and open questions in this talk, together with a brief overview of the state-of-the-art of set visualization.

References

1. Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and Helwig Hauser. Radial Sets: Interactive Visual Analysis of Large Overlapping Sets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2013.
2. Bilal Alsallakh, Luana Micallef, Aigner Wolfgang, Hauser Helwig, Silvia Miksch, and Rodgers Peter. Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges. *Proceedings of the 16th Annual Eurographics Conference on Visualization (EuroVis), State of the Art Reports (STARs)*, page 121, 2014.
3. Alsallakh Bilal, Micallef Luana, Aigner Wolfgang, Hauser Helwig, Miksch Silvia, and Rodgers Peter. The State-of-the-Art of Set Visualization. *Computer Graphics Forum*, 35(1):234260, 2015.
4. Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.
5. Dörk, Marian and Riche, Nathalie Henry and Ramos, Gonzalo and Dumais, Susan. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, 2012.
6. Kurt Koffka. *Principles of Gestalt Psychology*. Harcourt Brace, New York, NY, USA, 1935.
7. Wouter Meulemans, N Henry, Riche, Bettina Speckmann, Basak Alper, and Tim Dwyer. KelpFusion: a Hybrid Set Visualization Technique. *IEEE Transactions on Visualization and Computer Graphics*, 19(11):18461858, 2013.
8. Luana Micallef and Peter Rodgers. eulerAPE: Drawing area-proportional 3-Venn diagrams using ellipses. *PloS one*, 9(7):e101717, 2014.
9. Stephen E Palmer. Common region: A new principle of perceptual grouping. *Cognitive Psychology*, 24(3):436447, 1992.
10. Nathalie Henry Riche and Tim Dwyer. Untangling Euler Diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):10901099, 2010.
11. Peter Rodgers. A Survey of Euler Diagrams. *Journal of Visual Languages and Computing, Special Issue on Visualization and Reasoning using Euler Diagrams*, 25(1), 2014.
12. Ramik Sadana, Timothy Major, Alistair Dove, and John Stasko. OnSet: a visualization technique for large-scale binary set data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):19932002, 2014.
13. Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.