Paper Summary: “UpSet: Visualization of Intersecting Sets” – 998273

1.1 Objective

The aim of this essay is to summarize the paper “UpSet: Visualization of Intersecting Sets” which has been written by Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister and published in IEEE Transactions on Visualization and Computer Graphics in December 2014 (Lex et al., 2014). The authors of the paper have created a novel visualization technique with the respective software which is named “*UpSet*” to demonstrate quantitative analysis of sets, their intersections between the features and aggregations of them. This software uses two major different views in order to illustrate as more characteristics as possible by slicing the original dataset of sets to all possible pairs and combinations and also by using queries and task-driven aggregates that can be personalized properly for each use as well. The types of tasks which were examined, seem to be used in a wide interdisciplinary spectrum and their implementations have been done through a few datasets with very promising results and a decent implacability to the most set-related tasks.

1.2 Differences with the literature

Set-related tasks consist a wide spectrum of visualization tasks and there is an abundance of studies in literature, trying to approach that kind of datasets. Most usual approaches in set-related tasks seem to be the usage of Euler’s and Venn’s diagrams which tend to perform very well while in some studies (D’Hont et al., 2012); (Neale et al., 2014) tended to focus more on the visualization of each possible intersection, by increasing the amount of effort needed. Furthermore, element-centric techniques even though they perform quite well in some cases, face far-reaching limitations under some circumstances and especially when highly overlapping sets are analysed. *Upset*, on the other hand, focuses majorly on the relationships between each feature instead of each specific element.

Another technique that used for set visualization is a matrix-based set visualization approach like the case of *ConSet (*Kim et al., 2007). In this case, data is represented in a matrix which the user can filter sets and aggregate them through a dynamic control view where either sets vs elements or sets vs sets are presented. In contradiction to *ConSet, Upset* uses a matrix as well, but it has significant differences in functionality. Columns and rows represent the existing sets and intersections between them, respectively. Another approach that differs quite significantly with *UpSet,* is an aggregation-based technique named *Set O’ Grams* by Freiler et al., (2008). This type of view illustrates each set into bars by dividing it into different pieces and segments. Although the identification of intersecting sets and overlaps requires both interaction and effort, this visualization technique looks very similar to *UpSet. However, it has a* major difference which is the limitation to the number of sets while also not providing aggregations of the overlapping groups unlike *UpSet*. Lastly, *Radial Sets (*Alsallakh et al., 2013) are far more like *UpSet* while the former uses a more complex encoding in comparison with the latter.

1.3 Claim of the paper

In Upset, intersections tend to be the fundamental elements where the analysis is based on. The software has been developed to solve set visualization tasks such as set-related, element-related and attribute-related tasks as Alsallakh et al., (2013) describe them in their paper. The process of *UpSet* divide the whole dataset to each available intersection which authors named “*exclusive intersections*”. Afterwards, based on those exclusive intersections the whole analysis is constructed. Moreover, the user can produce aggregates which are far-reaching in order sets to be comparable, while simultaneously they clarify the rules on which the aggregations are produced. The two principal views, where the visualization is underpinned, is Set View and Element View. Additionally, the fundamental channel of encoding is position. In the Set View the relationships between features such as cardinality, intersections etc. are presented by using matrix-based techniques and sorting them properly. On the other hand, element view is based majorly on the aggregation and summary statistics that are produced and they are encoded with colour as well to highlight the represented features. Also, element queries can be defined at the element view by either selecting features from set view or applying filters to the Element view. Thus, as the authors mention they used this strategy of "*Divide and Conquer*" in order to answer research questions like "*Which is the biggest intersection of degree 3? or Which two-set intersection has the highest average attribute value?*" (Lex et al., 2014).

Upset was applied to two datasets to identify whether it is considerably applicable or not. The first dataset had to do with trade products; with a primary goal to find product similarities and anomalies. For this task to be solved, the field expert used the software for the analysis. Afterwards, he commented that *UpSet* was highly useful to the purpose of the task. Furthermore, he mentioned that he is going to use the software again while expanding the task with extra data. In the second case it was used for a biology task. The field expert in this case said that not only was it extremely useful to her task, but also, she might have not thought about some things if she had not used the software. Overall, software tends to give analysts a better opportunity besides Venn's and Euler's diagrams which seemed to be the predominant choice in the most set-related tasks.

Finally, *UpSet* presents decent results to the tasks that it can be applied for. Nevertheless, there are 3 out of 26 tasks which Alsallakh et al., (2013) described in their paper that *UpSet* cannot serve. A future development might be that users could create their own sets. Another feature which also could be added is the option to analyse and compare similarities between the features which is the primary plan of authors for the future. A final improvement could be in relation to the scalability of the software. Due to a bottleneck was observed when more than 40 to 50 sets were put into analysis. Also, there was interest for the ability to analyse more than 100 sets. Ultimately, a dimension reduction technique could be a potential complement to the handle of the data.

References

B. Alsallakh, W. Aigner, S. Miksch, and H. Hauser, (2013). Radial sets: Interactive visual analysis of large overlapping sets. *IEEE Transactions on Visualization and Computer Graphics* (InfoVis '13), 19(12):2496–2505.

A. D'Hont, F. Denoeud, J.-M. Aury, F.-C.Baurens, F.Carreel, O.Garsmeur, B.Noel, S.Bocs, G.Droc, M.Rouard, C.DaSilva, K.Jabbari, C.Cardi, J.Poulain, M.Souquet, K.Labadie, C.Jourda, J.Lengellé, M. Rodier-Goud, A. Alberti, M. Bernard, M. Correa, S. Ayyampalayam, M. R. Mckain, J. Leebens-Mack, D. Burgess, M. Freeling, D. MbéguiéA-Mbéguié, M. Chabannes, T. Wicker, O. Panaud, J. Barbosa, E. Hribova, P. Heslop-Harrison, R. Habas, R. Rivallan, P. Francois, C. Poiron, A. Kilian, D. Burthia, C. Jenny, F. Bakry, S. Brown, V. Guignon, G. Kema, M. Dita, C. Waalwijk, S. Joseph, A. Dievart, O. Jaillon, J. Leclercq, X. Argout, E. Lyons, A. Almeida, M. Jeridi, J. Dolezel, N. Roux, A.-M. Risterucci, J. Weissenbach, M. Ruiz, J.-C. Glaszmann, F. Quétier, N. Yahiaoui, and P. Wincker, (2012). The banana (musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, 488(7410):213–217

W. Freiler, K. Matkovic, and H. Hauser, (2008). Interactive visual analysis of set typed data. *IEEE Transactions on Visualization and Computer Graphics* (InfoVis '08), 14(6):1340–1347.

B. Kim, B. Lee, and J. Seo., (2007). Visualizing set concordance with permutation matrices and fan diagrams. *Interacting with Computers*, 19(5-6):630-643.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). *UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics, 20(12), 1983–1992.* doi:10.1109/tvcg.2014.2346248

D. B. Neale, J. L. Wegrzyn, K. A. Stevens, A. V. Zimin, D Puiu, M. W. Crepeau, C. Cardeno, M. Koriabine, A. E. Holtz-Morris, J. D. Liechty,P.J.Martínez-García,H.A.Vasquez-Gross,B.Y.Lin,J.J.Zieve, W. M. Dougherty, S. Fuentes-Soriano, L.-S. Wu, D. Gilbert, G. Marçais, M. Roberts, C. Holt, M. Yandell, J. M. Davis, K. E. Smith, J. F. Dean, W. W. Lorenz, R. W. Whetten, R. Sederoff, N. Wheeler, P. E. McGuire, D. Main, C. A. Loopstra, K. Mockaitis, P. J. deJong, J. A. Yorke, S. L. Salzberg, and C. H. Langley, (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, 15(3):R59.