EuroVis Workshop on Visual Analytics (2018)
C. Tominski and T. von Landesberger (Editors)

A set-based visual analytics approach to analyze retail data

Muhammad Adnan and Roy A. Ruddle

University of Leeds, UK

Abstract

This paper explores how a set-based visual analytics approach could be useful for analyzing customers' shopping behavior, and makes three main contributions. First, it describes the scale and characteristics of a real-world retail dataset from a major supermarket. Second, it presents a scalable visual analytics workflow to quickly identify patterns in shopping behavior. To assess the workflow, we conducted a case study that used data from four convenience stores and provides several insights about customers' shopping behavior. Third, from our experience with analyzing real-world retail data and comments made by our industry partner, we outline four research challenges for visual analytics to tackle large set intersection problems.

CCS Concepts

•***Human-centered computing*** → ***Visual analytics;*** •***Information systems*** → *Data mining;*

1. Introduction

"What products do our customers buy together" is a common question that retailers want to answer, because it provides them with insights about customers' shopping behavior. Such insights drive strategic changes in the layout of stores, products that are stocked and marketing campaigns.

To analyze customers shopping transactions, retailers use a variety of statistical modeling approaches that are validated by a retailer's domain knowledge and customer surveys. However, the models are not very accurate, and require a large amount of human effort to update as product lines change. It follows that retailers want to know how (or if) visual analytics methods can improve their understanding of customers' behavior, to make the models more accurate and easier to update.

The analysis of shopping transactions can be approached as a set analysis problem, where each transaction in a dataset is an element, each unique product is a set, and each unique combination of products (an itemset) is a set intersection. The present paper describes how a visual analytics approach that is based on set visualization techniques may be used to analyze shopping behavior, and makes three main contributions. First, we document the scale and characteristics of shopping transactions using anonymized data from a major supermarket. Second, we describe a workflow and visual analytic methods that can be used to quickly identify patterns in shopping behavior. Third, we outline research challenges for visual analytics to tackle large set intersection problems.

2. Related work

As would be expected, the details of how retailers model customers' shopping behavior are commercially confidential, but the general approach of our collaborator is as follows. Surveys show that the 'mission' (purpose) of most shopping transactions falls into one of a small number of categories (e.g., "eat now" or "food for a couple of days"). Our collaborator classifies missions by using a statistical model that is based on clustering and includes factors such the number and type of products, time of day and cost. The model is validated using customer surveys, which show that accuracy is improved by using a hybrid model (i.e., both positive and negative components, which classify a transaction as belonging to or not belonging to a mission, respectively). However, the model falls well short of 100% accuracy. Many misclassified missions make sense to human analysts, because they both know intuitively that the model is wrong and understand why the model got it wrong.

The remainder of this section briefly reviews computational and visualization techniques that could be applied to analyze missions. In the field of data mining, itemsets are computed by assigning all transactions that include items 'a' and 'b' to an itemset {a, b}, regardless of other items in these transactions [FVLV*17]. In set terminology, this approach is known as 'full set intersection' [AR17]. Since a dataset with 'm' unique items could generate as many as $2^m - 1$ itemsets, it is generally infeasible to compute all possible itemsets in a dataset using this approach. Therefore, some of the well-known itemset mining algorithms (e.g., Apriori [AS*94], Eclat [Zak00], and FP-Growth [HPYM04]) only compute frequent itemsets that meet a user-specified minimum support threshold. The support of an itemset {X} can be defined as the number of transactions containing X. However, choosing an appropriate support threshold is a non-trivial task for a user, as setting the support too high could exclude the interesting itemsets, while setting it too low could generate too many itemsets, or even make the computation infeasible. Furthermore, this itemset computation approach does not perform a proper partitioning of the transaction dataset because of potential overlap between itemsets. Consequently, comparing itemsets based on the attributes of their corresponding transactions (e.g., date and time of a transaction and store location) becomes a challenging task.

There are a number of approaches that have been proposed to visualize itemsets, computed using 'full set intersection' method (e.g., [Yan03], PowerSetViewer [MKN*05], FIsViz [LIC08], Fp-VAT [LC10], and [BSH13]). However, these approaches suffer from the same limitations that we highlighted above.

With tools such as Tableau [Tab18], it is straightforward to generate bar charts and histograms from sales data, so that analysts may visualize shopping behavior in terms of transaction length (i.e., number of items in a transaction) or commonly bought products. However, such tools are poorly suited for visualizing itemsets.

Lastly, several set visualization systems have been proposed to explore relationships between sets (items) and their intersections (itemsets). There are two main types of these systems. The first is set visualization systems that do not visualize all set intersections in a dataset, but provide reduced information about them (e.g., SEEM [GSG*14], Set'o'gram [FMH08], Radial Sets [AAMH13], and AggreSet [YEB16]). While these systems are scalable to tens of sets, they do not provide an overview of all set intersections.

The second type of set visualization systems are designed to visualize all set intersections (e.g., InfoCrystal [Spo95], Mosaic plots [HK81], Parallel Sets [KBH06], UpSet [LGS*14] and PowerSet [AR17]). To address the limitations of 'full set intersection' approach, these systems usually use 'exclusive set intersections', which only assigns transactions that exclusively include item 'a' and 'b' to an itemset {a, b} [AR17]. In contrast to potentially producing $2^m - 1$ itemsets, this approach limits the maximum number of generated itemsets to the total number of transactions in a dataset. Moreover, the one-to-one mapping of a transaction to an itemset allows easy comparison of itemsets, based on the attributes of their corresponding transactions. However, even the state-of-the-art systems in this category can only effectively visualize tens of sets and a few thousand set intersections.

3. Scale and characteristics of shopping data

The dataset that was used in the present research contained a total of 366,072 transactions with 1,198,650 products from four local 'convenience' stores. Convenience stores stock a range of everyday products and are designed for the convenience of customers who did not have time to do a full shop in a supermarket. The transactions were provided by our collaborator as a representative sample. Our collaborator uses a four-level hierarchy to classify products, and the present research used the second finest level, which they term 'sub-categories' (e.g., bread or lottery draws).

Although there were only 418 unique sub-categories in the dataset, there were 140,986 unique itemsets in the transactions. That low ratio of itemsets to transactions (1:2.5 when products were defined at the sub-category level; there were similar ratios for the other three hierarchy levels) is one of the characteristics that makes it so difficult to compute patterns from and visualize retail data. For example, frequent itemset mining algorithms (e.g., Apriori [AS*94] and FP-Growth [HPYM04]) could have only been used if aggressive minimum support thresholds were specified, because this dataset could potentially generate as many as $2^{418} - 1$ different itemsets. Furthermore, none of the existing set visualization tools can effectively handle hundreds of sets, and hundreds of thousands of set intersections [AMA*16].

4. Visual analytics of retail data

This section is divided into two parts. First, we describe a workflow that we developed to identify patterns in shopping behavior, and the computational and visualization methods that were adopted in each stage of the workflow. Then we describe the application of that workflow and methods to the convenience store dataset.

4.1. Analysis workflow

The proposed workflow is comprised of three main steps.

1. Data cleaning: A shopping transaction may include items that were not bought, but instead, refunded by a customer. Such items are usually identified by a binary flag in the dataset. Further, there may be items that are supplementary to the actual transaction (e.g., plastic bags, which in the UK must be paid for). The analyst needs to exclude these items from the dataset.

2. Itemset computation: Following in the footsteps of existing set visualization systems that are designed to visualize all set intersections (e.g., UpSet [LGS*14] and PowerSet [AR17]), we compute all the itemsets in a transaction dataset using the 'exclusive set intersections' approach.

3. Iterative analysis: The final step in this workflow is the iterative analysis of transaction data, which combines computational data processing and interactive visualizations (Figure 8 in supplementary material provides an example of this iterative analysis). Collectively, the iterative steps need to take a number of perspectives. Some are high-level, for example, analyzing set cardinality (the number of times each sub-category was bought) or the degree of set intersections (the number of sub-categories in an itemset). Other perspectives take account of the composition of set intersections, to investigate patterns within intersections of a given degree (e.g., itemsets of length two), or patterns that involve related intersections (e.g., supersets).

Although the 'exclusive set intersections' approach guarantees that the number of itemsets is always ≤ the number of transactions, they could still range from a few hundred to millions. Therefore, our workflow involves a suite of summary (a-d) and detailed (e-g) visualizations that, together, support a diverse set of tasks.

a) **Workflow overview:** Shows the proportion of transactions and certain products that have been accounted for during each iterative step of the analysis.
b) **Product frequency histogram:** Shows the distribution of the frequency of products.
c) **Itemset length histogram:** Allows analyst to see most common itemset lengths.
d) **Itemset frequency histogram:** Shows the distribution of the frequency of itemsets.

e) **Product frequency bar chart:** Allows analyst to see the number of times individual products were bought.
f) **Itemset frequency bar chart:** Shows the frequency of individual itemsets.
g) **Itemset heatmap/matrix plot:** Allows analyst to see the frequency, length, and composition of each itemset.

The illustrations of the above visualizations are provided in the supplementary material. It is important to note that histograms (b-d) scale to any volume of data because their performance does not deteriorate with increasing numbers of observations. Bar charts (e and f) are less scalable than histograms, but their scalability can be improved by grouping low frequency items into 'other'. Lastly, the heatmap (g) at most can only show a few thousand itemsets (as is the case with gene expression heatmaps), and is more effective with many fewer itemsets (say, a maximum of 50).

4.2. Convenience store case study

We applied the above workflow to explore the transaction data from four convenience stores. The analysis was performed at the granularity of sub-categories (not products). The rationale for this was that sub-categories provide sufficient abstraction of items while losing relatively little information about a transaction, for example, a single sub-category of milk encompasses several varieties of milk (e.g., 2 pint semi-skimmed milk and 2 pint whole milk). The dataset has 366,072 transactions and 418 unique sub-categories.

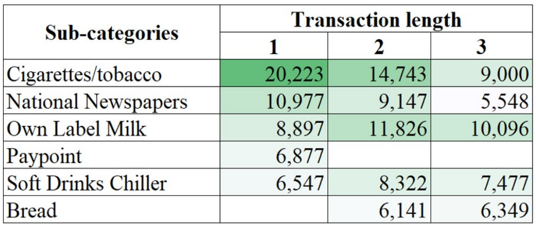
1. Data cleaning: To begin with, 582 refunded items were excluded from the dataset. This resulted in the removal of 258 transactions. Next, we removed the sub-category of 'System Test', which was bought 28,372 times. 'System Test' includes four items, three of which are related to shopping bags. The forth item is 'Tobacco Think 25' (bought only four times), which is recorded when a customer fails to prove his/her age when trying to buy a restricted product. The removal of 'System Test' resulted in the removal of 58 transactions, which only contained this sub-category. Figure 7 in supplementary material provides an illustration of the data cleaning process.

2. Itemset computation: Data cleaning was followed by the computation of all itemsets in the remaining 365,756 transactions, which contained 1,169,696 items. This generated 140,986 itemsets of length 1-38, 41, 45, and 55.

The itemset computation only took 1.1 seconds on a desktop PC with a 3.6 GHz Intel i7 processor and 16 GB of RAM. This meant that it would have been possible to perform additional computations (e.g., of itemsets at the finest level of the product hierarchy) on demand, as the iterations progressed.

3. Iterative analysis: The first iterative step performed a high-level analysis of the convenience store data and showed that, as expected, the distribution was highly skewed. A few sub-categories were bought much more often than the others, and 323,631 (88.5%) transactions contained five or fewer sub-categories. Those transaction accounted for 98,918 (70.2%) itemsets.

The second iterative step focused on the most common lengths of transactions (1, 2 or 3) and, within them, the sub-categories that were bought most frequently (see Figure 1). This highlighted that, as the number of sub-categories in a transaction increases, the most frequently bought sub-categories change. For example, 'Cigarettes/tobacco' becomes relatively less common as transaction length increases, and although 'Paypoint' is the fourth most common sub-category in transactions of length one, it drops to 11th and 16th place in transactions of length two and three, respectively.



Sub-categories	Transaction length		
	1	2	3
Cigarettes/tobacco	20,223	14,743	9,000
National Newspapers	10,977	9,147	5,548
Own Label Milk	8,897	11,826	10,096
Paypoint	6,877		
Soft Drinks Chiller	6,547	8,322	7,477
Bread		6,141	6,349

Figure 1: *The five most frequently bought sub-categories in transactions of length 1-3. Length is the number of different sub-categories in a transaction.*

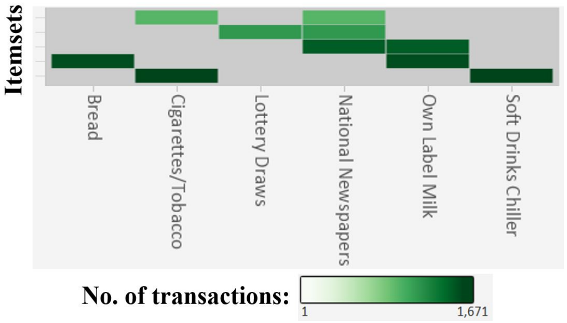
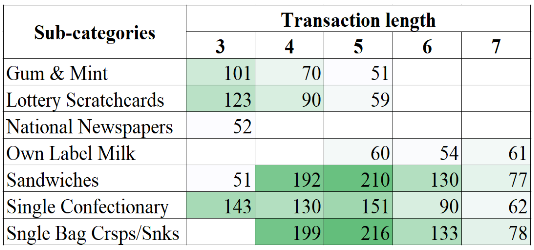


Figure 2: *The five most frequent itemsets of length 2.*

The third iterative step explored the components of common itemsets (the sub-categories in itemsets). For this, we created two heatmaps to view the composition of the five most frequent length two and three itemsets, and the findings were counterintuitive. For the length two itemsets, the sub-categories that were bought most often (Cigarettes/tobacco and Own Label Milk; see Figure 1) were typically bought with other sub-categories, not together (see Figure 2). For length three itemsets, we noticed that none of the three most frequent itemsets contained either of the two most frequent sub-categories for this transaction size (Own Label Milk and Cigarettes/tobacco; see Figure 1). Furthermore, three (i.e., Sandwiches, Sngle Bag Crsps/Snks, and Front of Store Juice) out of the four sub-categories in the two most frequent length three itemsets were not even listed for this length in Figure 1.

The fourth iterative step investigated supersets - longer itemsets that contained other frequent ones. As an example we use the most frequent length two itemset {Cigarettes/tobacco, Soft Drinks Chiller}, which was bought 1,671 times. Computation showed that this itemset was bought 3,837 times with other sub-categories, and only seven sub-categories appeared more than 50 times in a transaction with this itemset (see Figure 3).

Each of the above iterative steps took place from a particular perspective (see Section 4.1) and involved a number of computa-



Sub-categories	Transaction length				
	3	4	5	6	7
Gum & Mint	101	70	51		
Lottery Scratchcards	123	90	59		
National Newspapers	52				
Own Label Milk			60	54	61
Sandwiches	51	192	210	130	77
Single Confectionary	143	130	151	90	62
Sngle Bag Crsps/Snks		199	216	133	78

Figure 3: *Sub-categories that were bought more than 50 times in longer length transactions with the most frequent length 2 itemset.*

tions and visualizations (see Figure 8 in supplementary material). As is the norm in visual analytics, the analyst used the results of one iterative step to make choices for the next.

5. Research agenda

The analysis of customers' shopping behavior is essentially a large-scale set analysis problem. In this section, we outline research challenges that need to be addressed if visual analytics is to be effective for tackling such problems. The four challenges are drawn from our experience of analyzing real-world data (the convenience store case study) and comments made by our industry partner.

1. Scalability of set visualization techniques: As a recent review of set visualizations shows [AMA*16], few current techniques scale to even 100 sets. By contrast, our case study involved 417 sets (after data cleaning), that number can be 100 times larger (e.g., product-level data for supermarkets), and the number of set intersections is an order of magnitude larger still.

Our case study exploited histograms, bar charts and heatmaps to visualize at three levels of detail the cardinality, degree and composition of sets and their intersections. However, this was only effective when a user customized the visualizations (e.g., to change the number of histogram bins so that particular cardinalities were separated). Even though each customization only required a few mouse clicks, the cumulative interaction cost was considerable. Therefore, research is needed to determine types of perceptual discontinuity (e.g., variable-width bins, gaps, and axis breaks) that should be incorporated into the visualizations, and methods for auto-detecting those types from set data so that the cost of user interaction is reduced [Lam08, HEFR17].

Research is also needed to investigate how we may combine different set visualization techniques (e.g., node-link vs. matrix; [AMA*16]) to provide a step-change in our ability to visualize large-scale set data. Again, the solution needs to encompass a range of tasks (cardinality vs. degree vs. composition; sets vs. intersections). Starting points are provided by tools such as UpSet [LGS*14] and PowerSet [AR17], although it should be noted that they have only been evaluated with modest-scale data.

2. Identify relationships between set intersections / itemsets: To identify the strength of an intersection, users need to see how it relates to other intersections. PowerSet [AR17] does not support this due to its label-based approach to represent the composition of set intersections. On the other hand, the matrix-based approach in UpSet [LGS*14] and our heatmap allows users to see the related occurrences of an intersection across the dataset, but these only scale to a few thousand set intersections (at most).

An open research challenge is to develop visual analytics techniques for identifying relationships in much larger scale datasets (e.g., hundreds of thousands of set intersections, as exemplified by the case study). This requires research into measuring the similarity of relationships, drawing on methods from itemset mining [FVLV*17] and graph theory (e.g., clustering coefficients and modularity [BGLL08]), and techniques for optimizing the methods that are chosen [vLFR17].

3. Explain set intersections: To understand shopping behavior, users need to provide explanations that tally with customer survey data and domain knowledge. In our experience, the explanations often involve correlating set intersections with the values of other fields in the data (e.g., date and time of a transaction, store type, and store location), and this may be subdivided into: (1) using domain knowledge to choose fields or group values, (2) calculating correlations of the fields with selected set intersections, and (3) ranking the fields to help decide on a robust explanation.

Tools such as UpSet [LGS*14] and PowerSet [AR17] offer good support for (1) and (2). However (3) has not been explored by existing set visualization tools. The use of feature selection methods (e.g., Pearson's correlation coefficient and mutual information [GE03]) could potentially prove useful in this regard.

4. Capture an analysis story: As shown in our case study, the vast scale of data requires users to take an iterative analysis approach. It is challenging to keep track of every step (including dead ends) in those iterative steps, but that is essential if a user is to be able to: (1) validate any assumptions, (2) review and refine the analysis in consultation with colleagues, and (3) gain sufficient confidence for a company to be willing to act on the analysis findings. Although analysis provenance and workflow have been the subject of previous research (e.g., [BCC*05]), capturing an analysis story has not been a focus of existing set visualization tools. Therefore, it is an open research problem in this domain.

6. Conclusion

In this work, we ask how a set-based visual analytics approach may be used to model customers' shopping behavior. We begin by documenting the scale and characteristics of a real-world retail dataset from a major supermarket. We then present a visual analytics workflow that could be used to analyze this data in a scalable manner. The workflow was assessed by conducting a case study with real-world shopping transaction data from four convenience stores. Finally, informed by our experience with the case study and feedback provided by our industry partner, we present a research agenda for visual analytics to tackle large set intersection problems.

Acknowledgments

This research was funded by the Engineering and Physical Sciences Research Council (EP/N013980/1). We gratefully acknowledge the data and feedback provided by our project partner. The license agreement limits the dataset's use to that funded project.

References

[AAMH13] ALSALLAKH B., AIGNER W., MIKSCH S., HAUSER H.: Radial sets: Interactive visual analysis of large overlapping sets. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2496–2505. 2

[AMA*16] ALSALLAKH B., MICALLEF L., AIGNER W., HAUSER H., MIKSCH S., RODGERS P.: The state-of-the-art of set visualization. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 234–260. 2, 4

[AR17] ALSALLAKH B., REN L.: Powerset: A comprehensive visualization of set intersections. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 361–370. 1, 2, 4

[AS*94] AGRAWAL R., SRIKANT R., ET AL.: Fast algorithms for mining association rules. In *Proceedings of the Twentieth International Conference on Very Large Data Dases (VLDB)* (1994), vol. 1215, pp. 487–499. 1, 2

[BCC*05] BAVOIL L., CALLAHAN S. P., CROSSNO P. J., FREIRE J., SCHEIDEGGER C. E., SILVA C. T., VO H. T.: Vistrails: Enabling interactive multiple-view visualizations. In *IEEE Visualization (VIS)* (2005), IEEE, pp. 135–142. 4

[BGLL08] BLONDEL V. D., GUILLAUME J.-L., LAMBIOTTE R., LEFEBVRE E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 2008*, 10 (2008), P10008. 4

[BSH13] BOTHOREL G., SERRURIER M., HURTER C.: Visualization of frequent itemsets with nested circular layout and bundling algorithm. In *International Symposium on Visual Computing* (2013), Springer, pp. 396–405. 2

[FMH08] FREILER W., MATKOVIC K., HAUSER H.: Interactive visual analysis of set-typed data. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008). 2

[FVLV*17] FOURNIER-VIGER P., LIN J. C.-W., VO B., CHI T. T., ZHANG J., LE H. B.: A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 7*, 4 (2017). 1, 4

[GE03] GUYON I., ELISSEEFF A.: An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, Mar (2003), 1157–1182. 4

[GSG*14] GOVE R., SAXE J., GOLD S., LONG A., BERGAMO G.: Seem: a scalable visualization for comparing multiple large sets of attributes for malware analysis. In *Proceedings of the Eleventh Workshop on Visualization for Cyber Security* (2014), ACM, pp. 72–79. 2

[HEFR17] HARRISON D. G., EFFORD N. D., FISHER Q. J., RUDDLE R. A.: Petminer–a visual analysis tool for petrophysical properties of core sample data. *IEEE Transactions on Visualization and Computer Graphics* (2017). 4

[HK81] HARTIGAN J. A., KLEINER B.: Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (1981), Springer, pp. 268–273. 2

[HPYM04] HAN J., PEI J., YIN Y., MAO R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery 8*, 1 (2004), 53–87. 1, 2

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics 12*, 4 (2006), 558–568. 2

[Lam08] LAM H.: A framework of interaction costs in information visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008). 4

[LC10] LEUNG C. K.-S., CARMICHAEL C. L.: Fpvat: a visual analytic tool for supporting frequent pattern mining. *ACM SIGKDD Explorations Newsletter 11*, 2 (2010), 39–48. 2

[LGS*14] LEX A., GEHLENBORG N., STROBELT H., VUILLEMOT R., PFISTER H.: Upset: visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1983–1992. 2, 4

[LIC08] LEUNG C. K.-S., IRANI P. P., CARMICHAEL C. L.: Fisviz: a frequent itemset visualizer. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2008), Springer, pp. 644–652. 2

[MKN*05] MUNZNER T., KONG Q., NG R. T., LEE J., KLAWE J., RADULOVIC D., LEUNG C. K.: Visual mining of power sets with large alphabets. *Department of Computer Science, The University of British Columbia* (2005). 2

[Spo95] SPOERRI A.: *InfoCrystal, a visual tool for information retrieval*. PhD thesis, Massachusetts Institute of Technology, 1995. 2

[Tab18] TABLEAU SOFTWARE: Tableau, 2018. URL: https://www.tableau.com/. 2

[vLFR17] VON LANDESBERGER T., FELLNER D. W., RUDDLE R. A.: Visualization system requirements for data processing pipeline design and optimization. *IEEE Transactions on Visualization and Computer Graphics 23*, 8 (2017), 2028–2041. 4

[Yan03] YANG L.: Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates. In *International Conference on Computational Science and Its Applications* (2003), Springer, pp. 21–30. 2

[YEB16] YALCIN M. A., ELMQVIST N., BEDERSON B. B.: Aggreset: Rich and scalable set exploration using visualizations of element aggregations. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 688–697. 2

[Zak00] ZAKI M. J.: Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering 12*, 3 (2000), 372–390. 1