GroupSet: A Set-Based Technique to Explore Time-Varying Data

L. Liu & R. Vuillemot

LIRIS, École Centrale de Lyon

Abstract

We introduce GroupSet, a technique to facilitate the exploration of temporal charts using a set-based approach. GroupSet operates in a twofold way: first it classifies temporal data into categories (sets) for each time point, second it enables to explore such membership to categories (sets) over time. This approach enables to reveal temporal similarities of elements by categories (sets) memberships, usually hidden by overplot. We demonstrate the applicability of the technique to two case studies (traffic data and sport data) and report on usability feedback of an interactive prototype implementing the technique. Our code and datasets are published as an open-source project and we expect further research towards efficient set creation and temporal manipulation, which remain under-explored areas in the domain of set visualization and interaction.

CCS Concepts

• ***Human-centered computing*** → ***Interaction Design;***

1. Introduction

Making sense of time-varying data is frequent in many domains, as data are continuously recorded and analyzed. For instance, road operators monitor the traffic density of road segments every day in cities. In soccer field, analysts check teams' ranking weekly to understand their performance during a season. For those domains, a temporal chart—such as a line chart—usually is the goto representation as few assumptions can be made on the data attributes, volume and their distribution. However, such a chart usually generates overplot and hides patterns as datasets get increasingly large.

We introduce GroupSet, a visualization technique to explore changes within large temporal datasets using line charts. The technique relies on set-based tasks [AMA*14] to understand the relationships between lines, using partial memberships and change-related metrics. GroupSet workflow is summarized and operates as follows: 1) users manually categorize temporal elements (lines) based on their value at each time point, 2) the aggregation of elements generates set intersections listed by partial sets membership, and 3) the details of each set intersection are provided with partial membership distributions and other attributes (e. g., change, cardinality). Our contributions are as follows:

- a *partial membership visualization* (using pie charts) to make sense of temporal elements grouped by similarities;
- an *interactive visualization technique*, called GroupSet, implementing the partial memberships visualizations and interactions;
- two *case studies* demonstrating the applicability of the technique and preliminary usability feedback.

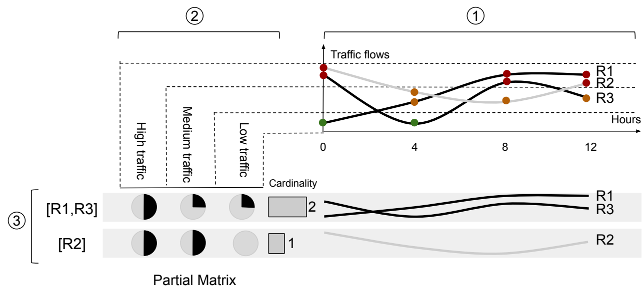

Figure 1: GroupSet *technique overview. Sets are created from time-varying data (line chart) in ①. The same sets are organized as columns to indicate partial membership as pie charts in ② and set intersections are listed as rows to represent groups of similar lines (along with their partial membership representation) in ③.*

2. Related Work

Our work primarily relates to large time-based values exploration, where an independent variable (time) is plotted along with a dependant value (quantity). Such a research area has already been widely investigated [AMST11]. Many line chart variations have been proposed such as slop charts [Tuf85] or temporal glyphs [FFM*13] to plot a handful of elements at once. Changing graphical properties such as using radial layouts ChronoLenses [ZCB11], introducing discrete density representation [MF18] or stacking [JE10] provide better comparison of values. Similarly searching by patterns [SC00], filtering values by boxing time intervals [HS04] or direct manipulation enable local patterns retrieval [VP15]. Self-organizing layouts such as StoryFlow [LWW*13], compos-

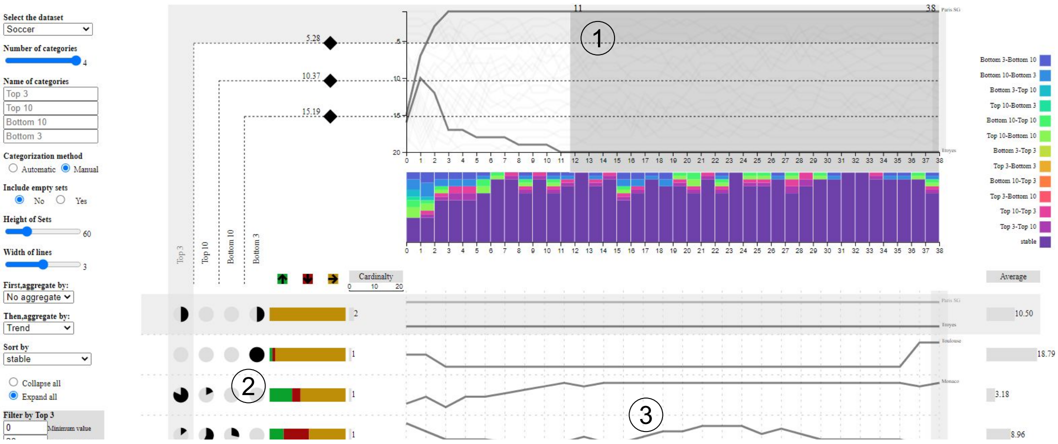

Figure 2: GROUPSET *to explore 20 soccer teams rank in France during a season (38 games). In ① teams are categorized by rank into 4 sets (*`Top 5`*,* `Top 10`*,* `Bottom 10`*, and* `Bottom 5`*). Users can then explore the teams based on their memberships to those sets over a season and identify temporal patterns by manipulating a combination matrix of set intersections ②. The final view provided by* GROUPSET *is a filtered line chart corresponding to each set intersection showing teams with similar patterns ③ along with other attributes (e. g., change).*

ite visualization [JE12] or Timecurves [BSH*16] provide further comparison by similarities intra or inter temporal events. However, those techniques do not address scalability issues. RankExplorer [SCL*12] and RankEvo [LXG*16] are scalable trends detection and interpretation using glyphs, but they rely on visual aggregation hiding individual patterns within each group. Our work is also based on aggregation, but it introduces an approach to avoid sacrificing the representation of individual details.

Our work is also closely related to categories exploration, where parallel coordinates charts [Ins85] are the flagship example. Building on this approach, ParallelSets [KBH06] displays groups of elements to reveal trends by cohorts. In [vLBA*12], temporal categories are presented using a flow chart that emphasizes changes across pairs of categories. But both approaches aggregate elements. Sets are a natural way to analyze categories [AMA*14] using set-based data models. TimeSet [NXWW16] shows groups changes over time using contours shapes. AggreSet [YEB16b] uses aggregation of elements to show their global membership to a category. PowerSet [AR17] is a scalable technique that enumerates all set intersections as a treemap and UpSet [LGS*14] uses a matrix approach. Our work is deeply connected to those scalable techniques such as [ATWB20], which conveys set dynamics over time. Finally, our work is also connected to interactive set creation [LV21] which remains an under-explored area in the visualization community [AMA*14].

3. The GROUPSET Technique

The core of the GROUPSET technique is a partial membership metric that captures temporal changes. Such changes are represented using pie charts where the black wedge encodes the % of membership (that we define it with *set membership degree*) to a category (that we now call *set*) which is one of the value intervals of the dependant quantity domain. If a temporal element is constant over time, and its value interval is divided into three categories, its membership is the following: [, ,] (black circle indicating 100% membership to a single set, gray circles 0% membership to the two others). If the membership is spread across multiple sets, the representation can be as follows e. g., [, ,]. Total membership across sets is always 100%. GROUPSET addresses the following set-based tasks [AMA*14]: **T1**: Create new sets. **T2**: Find elements with their set memberships and specific time intervals. **T3**: Find intersections with specific time intervals. **T4**: Analyze intersection relations with time intervals. **T5**: Analyze elements distribution with time intervals. **T6**: Filter time intervals.

Step 1: Sets are manually created by defining intervals over the Y-axis of the line chart (Fig. 1, ①). By default, 3 sets are included, but users can add or remove sets, customize their value intervals, and name them. The interval customization uses direct manipulation of the Y-axis, so the intervals are not necessarily uniform.

Step 2: Aggregation exploration is achieved using the set permutation matrix and aggregate all elements in a similar way to the UpSet matrix [LGS*14]. Each column is a set and each row encodes a single type of intersection, where the pie chart shows the intersection's partial membership inspired by AggrSet [YEB16a]. It aggregates the elements if they have the same distribution among sets. Such as shown in Fig.1, $R1$ and $R3$ both have $2/4$ time points in `High traffic`, $1/4$ time points in `Medium traffic` and $1/4$ time points in `Low traffic` respectively. Moreover, GROUPSET also allows users sorting and filtering the aggregated groups to change the vertical resolution. Finally, a horizontal bar chart encodes the intersection's cardinality (i. e., number of elements). Each set intersection embeds a stacked chart encoding a change across two sets (Fig.2 ②), which present the average change values of all elements in each set intersection, where green indicates ups, red downs, and orange stable slopes. The stack height (horizontal) shows the most frequent changes for each category of change.

Step 3: Element exploration is enabled as a filtered line chart for each row (Fig.2, ③). It provides a less cluttered representation of an individual inspection of trends. Brushing the global line chart (Fig.2, ①) selects time range for the intersections visualizations, which helps analysts focus on specific time intervals. Fol-

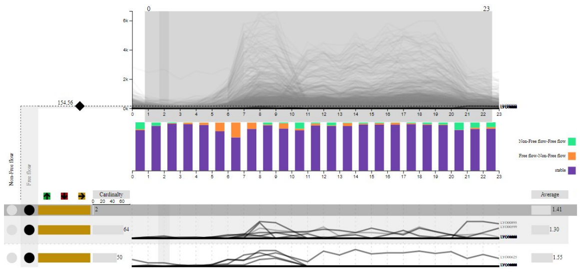

Figure 3: *Road segments that are always not busy. An aggregated group with two subgroups contains 64 and 50 road segments that only belong to* `Free flow` *during 24 hours.*

lowing with element view, it is the distribution view of other elements/groups attributes.

4. Prototype and Illustrative Case Studies

We now present 2 scenarios to illustrate the technique using representative tasks and datasets and the implementation of GROUPSET as web application `https://llqsee.github.io/groupset/` using JavaScript and D3.js.

4.1. Traffic Density Analysis

With the development of a variety of sensors to collect the data in cities, traffic operators are able to efficiently monitor roads [CGW15] and eventually detect anomalies or unexpected traffic variations on an hourly basis. We loaded a traffic dataset from a public open data repository in Lyon, FR, for a single day of observation (Fig. 3). The datasets include 1334 road segments during a day (24h), monitored by sensors using inductive loops. The X-Axis represents 24 hours and the Y-Axis represents the traffic densities collected by the sensors. Generally, the traffic densities are low from 0:00 AM to 6:00 AM, and then start rising and reaching the morning peak around 8:00 AM. After the afternoon, it descends again to a lower traffic density. As [CGW15] mentioned, one of the challenges in visualizing the traffic data is that it has too many data items to track. A general line chart cannot handle too many road segments without overplot. Thus, for the traffic dataset, we aim at achieving the followings:

Detecting roads with low traffic. Finding the roads that are *not busy* (free flow) is critical work to help traffic planners optimize traffic load. These traffic flows can be re-planned from the roads where the traffic congestion always happens. Thus, assisting traffic operators find the roads that are always not busy (during 24h) is able to improve the traffic conditions. As shown in Fig. 3, we focus on the entire period (24h) and create two sets (`Non-free flow` and `Free flow`) (**T1**), with a low density threshold to identify the free flows. By using the filter panel, we keep only the `Free flow` roads which are characterized with a membership to this set only (**T3**). The result is shown in Fig. 3, and two groups exist (as we aggregated by change pattern): one contains 64 road segments and another one contains 50 road segments (**T2**). These road segments mostly have less traffic flow. Further investigations will be needed for their relevance to bear more traffic.

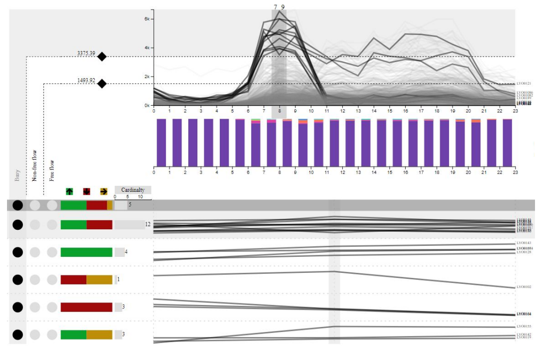

Figure 4: *The traffic flow patterns during morning peak. An aggregated group with 5 subgroups reaches the peak time in the morning, but some do not reach the peak in the evening.*

Analyzing the traffic flow patterns with a specific time interval. Managing the traffic during peak time is very important for a traffic manager, usually during rush hours around 8:00 AM and 6:00 PM. To analyze this peak pattern of traffic density, we first brush the time interval to the period which we are interested in (7:00 AM to 9:00 AM) **(T6)**. We then add `Busy` as shown in Fig. 4 to capture the busiest roads at peak time (**T1**). We observe there is an aggregated group (containing 5 subgroups) only belonging to `Busy` (**T4**). A first insight is that the largest group has roads with a clear increase and decrease change pattern (green and red bars), which means these roads are not stable being either rising up or falling down. A second insight is that a certain number of roads peak in the morning, but they do not have peak time in the evening, as shown in the global line chart (**T5**). These two patterns were difficult to identify in the full dataset, and the roads that match them can further be explored in the tool using the attribute panel of GROUPSET (e. g., road segments average length, etc.)

4.2. Soccer Rankings

Soccer data analysis is becoming increasingly important as a wealth of data is now available to data scientists. In particular, games outcomes are indicators of teams performance that can be aggregated and represented over a season. The standard representation usually is a ranking of usually around 20 teams (for European leagues) over time 19×2 games, i. e., 38 games. We created 3 sets {`TOP5`, `NORMAL`, `BOTTOM3`} which rank corresponds to the teams qualified for European competitions, the non-qualified teams, and teams relegated to the inferior division (**T1**). The `NORMAL` group usually is not the aim of the top-performing teams, but it is for less-performing teams (which usually never get to the `TOP5` group).

Discarding early ranking variability. An essential property of permutations in soccer championships is that at the beginning, teams may change rank with a higher probability than later in the championship (when there are high points differences). So it is essential to discard the first days as shown in Fig. 5 by selecting the time interval from e. g., day 11 to the end as an analyzed period **(T6)**. Once the beginning period is discarded, we sort the soccer teams by a trend to find the most stable soccer teams. We notice that three sets are being more stable than other sets, which contain four teams (Fig. 5): Paris, Troyes, Toulouse, and Monaco (**T2**).

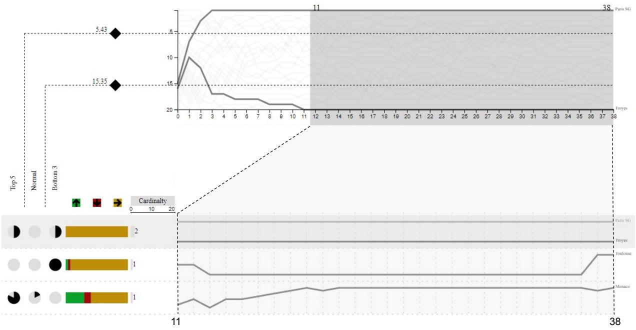

Figure 5: *Discarded time interval in soccer data using the brush feature on the line chart (to keep time points between day 11 and day 38). Groups are then ranked by teams with less changes of sets (resulting in both the best and worst teams).*

The most important thing is that these 4 teams are either stable at the top or bottom levels. Especially for Paris and Troyes, they did not change their rank from 11 to 38, and they are ranked first and last (**T5**).

Identifying best and worst performing teams. The next step is to answer basic performance questions such as which teams are the best or worst, which can immediately be seen on the championship's last day. However, we may dig a little bit into the analysis to check if this has been the case for the whole season. The rank by set membership degree shows teams that do not change group, so they have always performed very well or very poorly (Fig. 6). We aggregate the soccer teams by *Degree* and then sort them by *Degree-Bottom 3* so as to get the soccer teams with the highest membership to the `Bottom 3` on top (and the ones with partial membership to this group below) (**T5**). The team Troyes is the only one always being in the `Bottom 3` from 11 to 38, even though its rank raised at the very beginning of the season (but got discarded with our initial temporal selection).

5. Feedback and Perspectives

We conducted a preliminary study to validate the usability of GROUPSET with 4 researchers that frequently use data visualization tools, to validate the tool's usability and detect any major design issues. We presented them with the tool loaded with the Soccer dataset as it is the easiest one to understand as all researchers were familiar with soccer. We then demonstrated the standard workflow from categorization to detailed view exploration. We then asked them to use the tool and follow a think-aloud protocol to capture their thoughts and understand their intentions. We first asked them to reproduce the demonstrated workflow and then they conducted an open-ended exploration with any dataset of their choice regarding the tasks. All participants found the flow of the tool relevant to explore such a dataset and logical from the exploratory chart to the detailed view. The main remark we collected was preserving the sequence of events we will discuss as a tool's limit. They also noticed some performance issues we will discuss in the next section.

Performance and scalability. The first feedback we collected during the usability study was the performance, especially when

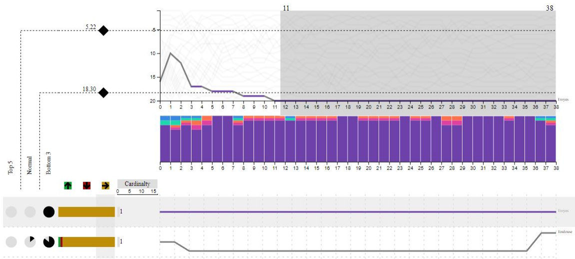

Figure 6: *Understanding the less performing team. They are the ones that had the worst rank for most of the season, except for Toulouse whose rank increased during the last games.*

many elements and sets were presented. For some application domains, it may be needed to include more than 3 sets either because it is semantically relevant or because a finer grain of analysis of changes is needed. Currently, GROUPSET supports up to 5 sets which potentially generates an important number of intersections. However, as we only focus on analyzing a subset of the data (due to sorting and aggregation), we argue there is no need to calculate and visualize all the set intersections.

Applicability to other datasets and beyond line charts. Our tool is generic to support any temporal dataset (e. g., University rankings) without any change in the design. Regarding the type of chart it supports, it is applicable beyond line charts to all charts with a single dependent variable (time points) plotted on the Y-axis to visualize the distribution of elements values in each time point among different sets (`Normal`, `Top 5`, `Bottom 5`, etc), such as histograms, density plots, or bar charts. 2D temporal charts, such as scatterplots or geo-map, will require some significant change in the design but GROUPSET could be used as a marginal plot to filter and group such charts over time.

Limits and perspectives. GROUPSET currently does not capture intra-group variations and temporal changes in the pie chart: thus some temporal elements may be included in the same set intersection, despite having different patterns (e. g., one is increasing, the other decreasing). This issue could be addressed by adding a second level of aggregation to the intersections using a global trend indicator to group elements by either globally increasing or decreasing value. But this may generate additional intersections and slow performances. Another perspective of our work is to investigate alternative designs to the pie charts to encode intersections and memberships. While circles and pie charts are already used in [LGS*14, YEB16b] for such encoding, there currently is no formal evidence they are the best-suited representation. We plan to implement and formally evaluate alternatives using other statistical charts (e. g., bar charts, box plots). Finally, the traffic density case only helps us analyze how traffic flows change over time. However, spatial information should be taken into account. In the next step, we will consider connecting GROUPSET with map to analyze the spatio-temporal information.

6. Acknowledgements

We thank Alexander Lex and the SICAL LIRIS team for their valuable feedback on an early version of the technique.

References

[AMA*14] ALSALLAKH B., MICALLEF L., AIGNER W., HAUSER H., MIKSCH S., RODGERS P.: *Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges*. The Eurographics Association, 2014. URL: https://diglib.eg.org:443/handle/10.2312/eurovisstar.20141170.001-021, doi:http://dx.doi.org/10.2312/eurovisstar.20141170. 1, 2

[AMST11] AIGNER W., MIKSCH S., SCHUMANN H., TOMINSKI C.: *Visualization of Time-Oriented Data*. Human-Computer Interaction Series. Springer, London, 2011. URL: http://link.springer.com/10.1007/978-0-85729-079-3, doi:10.1007/978-0-85729-079-3. 1

[AR17] ALSALLAKH B., REN L.: PowerSet: A Comprehensive Visualization of Set Intersections. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 361–370. doi:10.1109/TVCG.2016.2598496. 2

[ATWB20] AGARWAL S., TKACHEV G., WERMELINGER M., BECK F.: Visualizing Sets and Changes in Membership Using Layered Set Intersection Graphs. In *VMV: Vision, Modeling, and Visualization* (2020), vol. VMV2020, The Eurographics Association, pp. 69–78. URL: http://oro.open.ac.uk/72010/, doi:10/8/069-078.pdf. 2

[BSH*16] BACH B., SHI C., HEULOT N., MADHYASTHA T., GRABOWSKI T., DRAGICEVIC P.: Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 559–568. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi:10.1109/TVCG.2015.2467851. 2

[CGW15] CHEN W., GUO F., WANG F.-Y.: A Survey of Traffic Data Visualization. *IEEE Transactions on Intelligent Transportation Systems 16*, 6 (Dec. 2015), 2970–2984. Conference Name: IEEE Transactions on Intelligent Transportation Systems. doi:10.1109/TITS.2015.2436897. 3

[FFM*13] FUCHS J., FISCHER F., MANSMANN F., BERTINI E., ISENBERG P.: Evaluation of Alternative Glyph Designs for Time Series Data in a Small Multiple Setting. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)* (New York, United States, Apr. 2013), ACM, pp. 3237–3246. URL: https://hal.inria.fr/hal-00781504, doi:10.1145/2470654.2466443. 1

[HS04] HOCHHEISER H., SHNEIDERMAN B.: Dynamic Query Tools for Time Series Data Sets: Timebox Widgets for Interactive Exploration. *Information Visualization 3*, 1 (Mar. 2004), 1–18. URL: http://journals.sagepub.com/doi/10.1057/palgrave.ivs.9500061, doi:10.1057/palgrave.ivs.9500061. 1

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer 1*, 2 (Aug. 1985), 69–91. URL: http://link.springer.com/article/10.1007/BF01898350, doi:10.1007/BF01898350. 2

[JE10] JAVED W., ELMQVIST N.: Stack zooming for multi-focus interaction in time-series data visualization. In *2010 IEEE Pacific Visualization Symposium (PacificVis)* (Mar. 2010), pp. 33–40. ISSN: 2165-8773. doi:10.1109/PACIFICVIS.2010.5429613. 1

[JE12] JAVED W., ELMQVIST N.: Exploring the design space of composite visualization. In *2012 IEEE Pacific Visualization Symposium* (Feb. 2012), pp. 1–8. ISSN: 2165-8773. doi:10.1109/PacificVis.2012.6183556. 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 (July 2006), 558–568. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi:10.1109/TVCG.2006.76. 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 (Dec. 2014), 1983–1992. doi:10.1109/TVCG.2014.2346248. 2, 4

[LV21] LIU L., VUILLEMOT R.: Categorizing Quantities using an Interactive Fuzzy Membership Function:. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (2021). URL: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010270801950202. 2

[LWW*13] LIU S., WU Y., WEI E., LIU M., LIU Y.: StoryFlow: Tracking the Evolution of Stories. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2436–2445. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi:10.1109/TVCG.2013.196. 1

[LXG*16] LEI H., XIA J., GUO F., ZOU Y., CHEN W., LIU Z.: Visual exploration of latent ranking evolutions in time series. *Journal of Visualization 19*, 4 (Nov. 2016), 783–795. URL: https://doi.org/10.1007/s12650-016-0349-7, doi:10.1007/s12650-016-0349-7. 2

[MF18] MORITZ D., FISHER D.: Visualizing a Million Time Series with the Density Line Chart. *arXiv:1808.06019 [cs]* (Sept. 2018). arXiv: 1808.06019. URL: http://arxiv.org/abs/1808.06019. 1

[NXWW16] NGUYEN P. H., XU K., WALKER R., WONG B. W.: TimeSets: Timeline visualization with set relations. *Information Visualization 15*, 3 (July 2016), 253–269. URL: https://doi.org/10.1177/1473871615605347, doi:10.1177/1473871615605347. 2

[SC00] SILVA S., CATARCI T.: Visualization of linear time-oriented data: a survey. In *Proceedings of the First International Conference on Web Information Systems Engineering* (June 2000), vol. 1, pp. 310–319 vol.1. doi:10.1109/WISE.2000.882407. 1

[SCL*12] SHI C., CUI W., LIU S., XU P., CHEN W., QU H.: RankExplorer: Visualization of Ranking Changes in Large Time Series Data. *Visualization and Computer Graphics, IEEE Transactions on 18* (Dec. 2012), 2669–2678. doi:10.1109/TVCG.2012.253. 2

[Tuf85] TUFTE E. R.: THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION. *The Journal for Healthcare Quality (JHQ) 7*, 3 (July 1985), 15. URL: https://journals.lww.com/jhqonline/Citation/1985/07000/THE_VISUAL_DISPLAY_OF_QUANTITATIVE_INFORMATION.12.aspx. 1

[vLBA*12] VON LANDESBERGER T., BREMM S., ANDRIENKO N., ANDRIENKO G., TEKUŠOVÁ M.: Visual analytics methods for categoric spatio-temporal data. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2012), pp. 183–192. doi:10.1109/VAST.2012.6400553. 2

[VP15] VUILLEMOT R., PERIN C.: Investigating the Direct Manipulation of Ranking Tables for Time Navigation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI '15, ACM, pp. 2703–2706. URL: http://doi.acm.org/10.1145/2702123.2702237, doi:10.1145/2702123.2702237. 1

[YEB16a] 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 (Jan. 2016), 688–697. URL: http://ieeexplore.ieee.org/document/7194854/, doi:10.1109/TVCG.2015.2467051. 2

[YEB16b] YALÇIN 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 (Jan. 2016), 688–697. doi:10.1109/TVCG.2015.2467051. 2, 4

[ZCB11] ZHAO J., CHEVALIER F., BALAKRISHNAN R.: KronoMiner: using multi-foci navigation for the visual exploration of time-series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, May 2011), CHI '11, Association for Computing Machinery, pp. 1737–1746. URL: https://doi.org/10.1145/1978942.1979195, doi:10.1145/1978942.1979195. 1