Hierarchical Sankey Diagram: Design and Evaluation

William P. Porter, Conor P. Murphy, Dane R. Williams,
Brendan J. O'Handley, and Chaoli Wang

University of Notre Dame

Abstract. We present the hierarchical Sankey diagram that aims to augment the original Sankey diagram by enabling users to examine inflow links and levels of detail through four different variants. We provide the details of our design along with results on a student course performance dataset. Finally, the effectiveness of the four variants for the hierarchical Sankey diagram is evaluated via a user study.

Keywords: Hierarchical Sankey diagram · Inflow · Level-of-detail

1 Introduction

Sankey diagrams and their variants [1, 4, 8, 9, 11–15] have been an area of significant research in data visualization and utilized to study different applications [2, 3, 5–7, 10]. An important question researchers have studied along this topic is how to convey more information concerning data flow. Existing layouts attempt to improve the original Sankey diagram through various modifications. Riehmann et al. [8] presented several concepts for improving a static Sankey diagram with interactive features. Among these concepts is the ability to adjust grouping and level of detail on nodes, making it possible to drill down on a node to see how the flows pass through the subnodes contained in the hierarchy. Furthermore, their system enhances the Sankey diagram by introducing *flow tracing*, enabling users to select a node so that the contributing links are highlighted and moved to the foreground. Another interactive modification to Sankey diagrams was given by Kosara et al. [4], where a parallel sets layout is utilized and then improved by interactive queries. To facilitate hierarchical analysis, they grouped nodes into a single combined node. This design works similarly to the example given in [8], except that instead of breaking down a hierarchy into more specific nodes, the specific nodes can be combined into a larger node in the hierarchy. Sankey diagrams have also been modified to visualize data flows better without an interactive system. In one case, this is achieved by modifying the color of flows as exhibited in Lupton and Allwood [6]. For example, by adjusting the coloring of flows to correspond with the source node, they demonstrated how to convey information regarding the context of the data.

While the original Sankey diagram is helpful for quick summarization of prominent trends of data, there exist two main limitations. First, the original

Sankey diagram does not preserve the history of links. Thus, when viewing the outflow links from a node, it is impossible to identify which inflow links comprise that particular link and to what degree. Second, the original Sankey diagram cannot visualize the hierarchical structure within a particular node. Although the modifications to the Sankey diagram mentioned above improve the demonstration of data flow, there is still no solution that allows users to dynamically visualize the context of a data flow and change the specificity of nodes.

In this paper, we present the *hierarchical Sankey diagram*. Our hierarchical version of the Sankey diagram expands upon a standard Sankey diagram by addressing these limitations while preserving the core ability of the diagram to visualize data flow quickly. To make a Sankey diagram visually appealing and easy to understand, we usually avoid showing a large number of nodes when enabling level-of-detail exploration. Often, these nodes, which are essentially groupings of individual data points, serve as a general category for data. Within these categories may exist several subcategories, which may contain their distinct trends of data flow. We advocate an *in-place* approach by presenting two types of variants: *inflow* and *level-of-detail*. By "in-place", we mean modifying the original Sankey diagram rather than supplementing it with another separate view. These variants are built upon the fundamental idea of splitting nodes and merging them to enable more dimensions of comparison and generate more insight.

The contributions of our work are the following. First, our work offers an in-place solution to extending the capabilities of the Sankey diagram by splitting nodes dynamically. Second, unlike previous approaches that are practically limited to a single column [6], we present a new solution (i.e., vertical separation) to visualizing the inflow history across multiple columns. Third, previous works show that Sankey diagrams can either only depict change among groups over columns or break down the flow of a hierarchical relationship. By splitting nodes with our level-of-detail variants, we enable the Sankey diagram to communicate both dimensions: level-of-detail and change among groups over columns. Fourth, we conduct a user study to evaluate the effectiveness of the variants and assess user preference.

2 Design

2.1 Inflow Variants

We design two variants for inflow links: vertical separation (i.e., splitting the original node based on the inflow links) and color distinction (i.e., coloring the outflow links based on the inflow links). Note that neither variant affects nodes with no inflow links. In the following discussion, let us consider a Sankey diagram with four nodes in each column and four outflow links from each node.

The first variant, vertical separation, replaces the original node with four copies corresponding to the inflow links. As shown in Figure 1, the four nodes still represent the same data category as the original node, and together, they

contain the same data points as the original one. However, these data points are split based on the source node. This variant has the advantage of quickly summarizing the distribution of outflow links based on inflow links (i.e., which node did data points go to next based on where the data points previously were) and comparing these distributions to the other inflow separations. However, if many nodes are split apart, the diagram may appear cluttered and lose the advantage of rapidly assessing information trends.

The second variant, color distinction, colors the outflow links based on the inflow links. This is similar to partitioning bundles of flows [6]. If a node contains four inflow links, each outflow link will show four partitioned color bands corresponding to the respective colors of the inflow links. The size (i.e., bandwidth) of the original outflow link will remain the same, and the sizes of color bands that comprise it are proportional to their contributions to the outflow link. Refer to Figure 2 for an example. We point out that color distinction does not scale to multiple columns as vertical separation does. Moreover, it does not work well for a thin outflow link associated with many inflow links.

2.2 Level-of-Detail Variants

We design two variants to show the level-of-detail node information: horizontal split and vertical split. Data points can be grouped into categories differently: numerical data can be categorized by the specific numbers as well as intervals of varying sizes, and categorical data can be grouped based on similarities between data points. Regardless of whether a dataset is numerical, categorical, or a combination of both, there are often subcategories generalized by nodes.

The first variant, horizontal split, adds a new column of the subcategories to the right side of the column where the original node belongs, as shown in Figure 3. While the original node and inflow links are preserved, the original node now has new outflow links to each subcategory. The advantage of this variant is that the original node is kept, and it is easier to see its breakdown. However, chaining together subcategories by applying this variant to more subcategories could significantly increase the diagram's horizontal space. Furthermore, the columns of the Sankey diagram are usually distinctive and naturally represent different groupings. Therefore, adding a new column for subcategories may confuse users.

The second variant, vertical split, means replacing the original node with new nodes of further specificity based on subcategories. These subcategorical nodes can be further split into their subcategories, replacing the node being broken down. Note that vertical split replaces a node with its child nodes showing the next level of detail, while vertical separation duplicates the same node based on the inflow links. Refer to Figure 4 for an example. The advantage of replacing the original node is that it allows users to identify each subcategories' inflow and outflow. Subcategories of a node may have their own trends, and this variant is useful for identifying them. However, it can be more difficult to quickly assess the size and trend of the original category as users would need to recombine the inflow and outflow links of these subcategories mentally. Adding marks to the

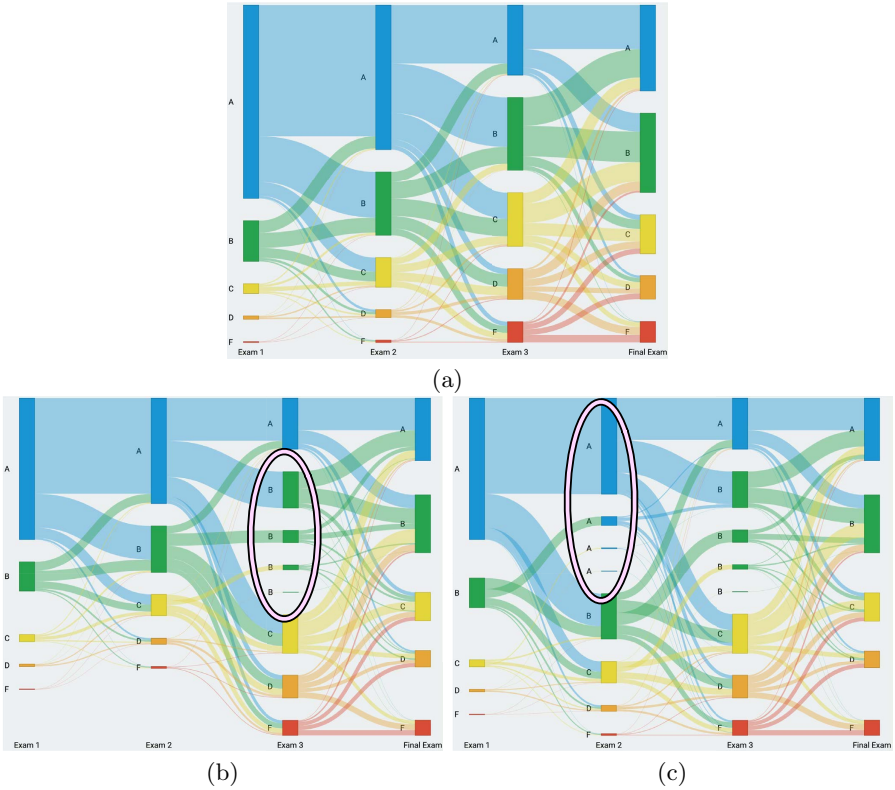


Fig. 1. Inflow: vertical separation. (a) shows the original Sankey diagram. (b) separating B on Exam 3 column. (c) continuing on (b) and separating A on Exam 2 column.

visualization indicating that certain nodes belong to the same initial node (e.g., dashed vertical lines connecting them or a bounding box around the split nodes in a group) would help. Still, it may lead to visual clutter with multiple such instances.

3 Results

3.1 Dataset and Web Application

The dataset was collected from student performance data of a course. The performance data include student grades in three exams and the final exam. The grades are quantitative (0 to 100), and we used standard groupings for letter grades. A, B, C, D, and F represent the groupings [90-100], [80-90), [70-80), [60-70), and [0-60), respectively. A '+' indicates a grade that the ones-digit is ≥ 7, and a '-' indicates a grade that the ones-digit is ≤ 3. B, C, and D all have both '+' and '-' while A only has '-' and F has neither. We created a web application by

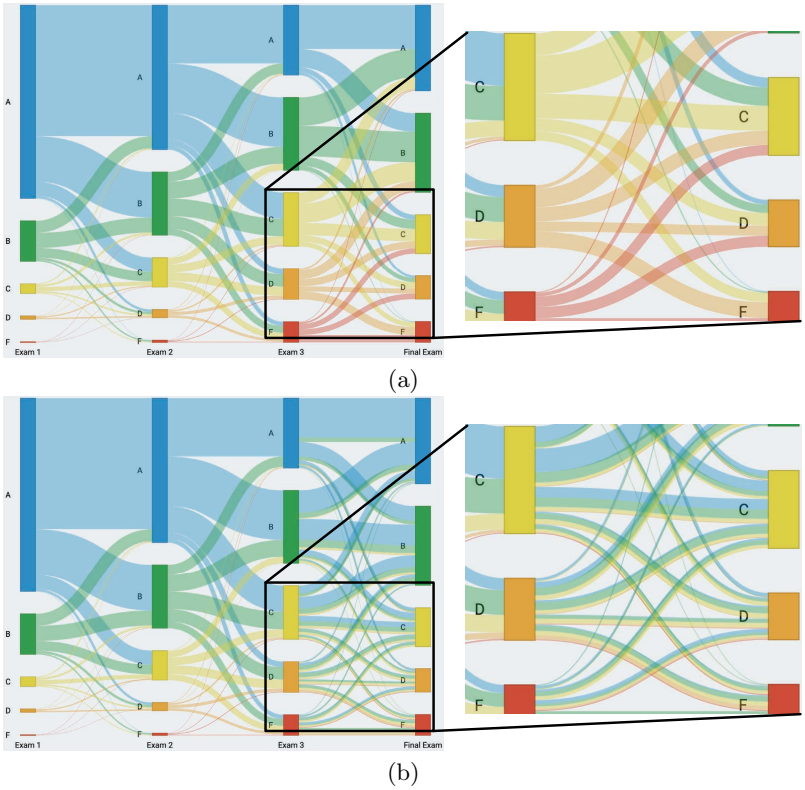


Fig. 2. Inflow: color distinction. (a) and (b) show the results before and after applying this variant to all nodes on Exam 3 column.

implementing the four variants of the hierarchical Sankey diagram using D3.js. We have released the application at `https://www.nd.edu/~cwang11/hsd/`. The figures shown in the paper are screenshots of this web application. We added highlights in Figures 1, 3, and 4 and zoom-in views in Figure 2 to show the intended changes, which are easy to observe when interacting with the application.

3.2 Visualization Results

As depicted in Figure 1(b), the vertical separation variant to inflow links grants us further insight. Before splitting node B on Exam 3 (refer to Figure 1(a)), users cannot identify the relation between inflow and outflow links. After the split, B is broken down into four new nodes: one from each of the Exam 2 nodes (A, B, C, and D). Note that F is not shown as there is no inflow from it. The ability to relate the outflow and inflow links dramatically extends the capabilities of a Sankey diagram, which is premised on the relationship between nodes. This variant allows users to draw considerable more insight into the data quickly. By

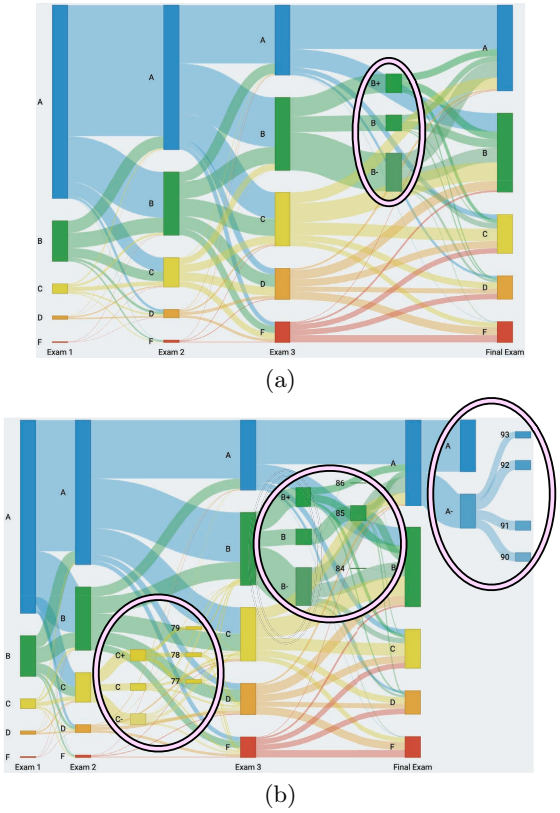


Fig. 3. Level-of-detail: horizontal split. (a) splitting B on Exam 3 column. (b) splitting C on Exam 2 column then further splitting C+, B on Exam 3 column then further splitting B, and A on Final Exam column then further splitting A-.

also splitting another node from another column (e.g., A on Exam 2), Figure 1(c) demonstrates how multiple nodes can be split to extend the capabilities of inflow history even further. Figure 2 depicts another variant to this idea using color distinction. Rather than create a new node, this variant simply colors the outflow link according to the contribution of the inflow links. While the result may appear more challenging to read, it has the advantage of not creating new nodes to avoid overcrowding. Consequently, as shown in Figure 2(b), where this hierarchy is applied to all nodes on the Exam 3 column, it is possible to quickly assess and compare inflow history across all nodes of a particular column.

As shown in Figure 3(a), the horizontal split for level-of-detail creates a new column to represent the breakdown of specificity. Here, we split the general category of B into their subcategories of B+, B, and B-. Figure 3(b) demonstrates this successive breakdown into a different level of detail. Notice how new columns are created for new levels of detail, and the inflow to each hierarchy is its parent.

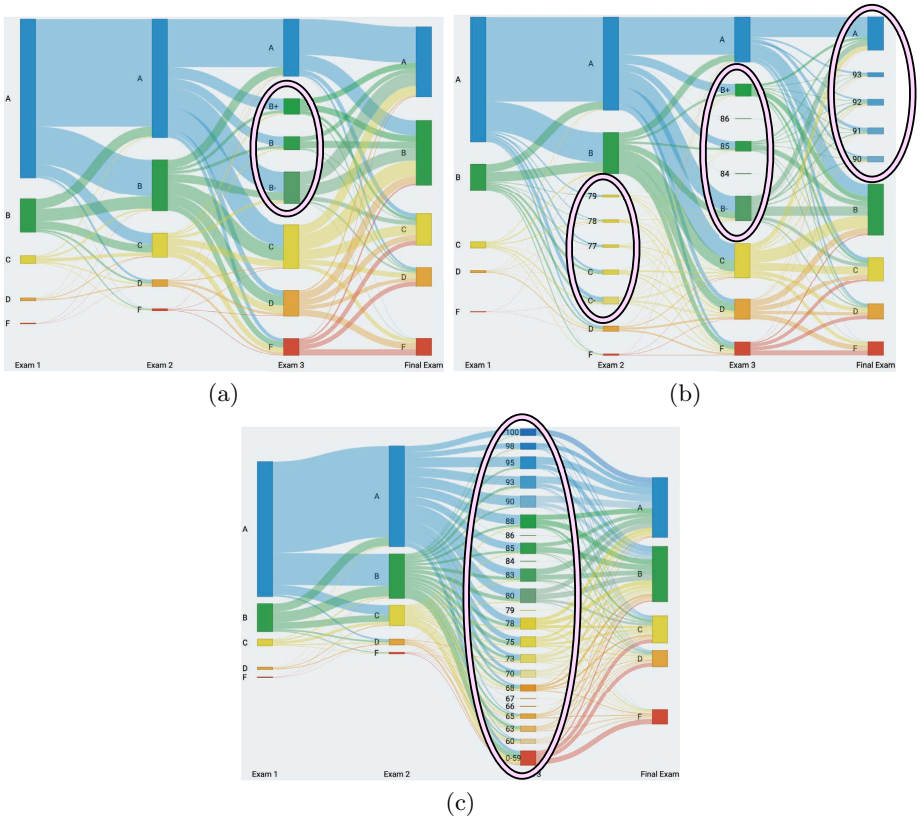
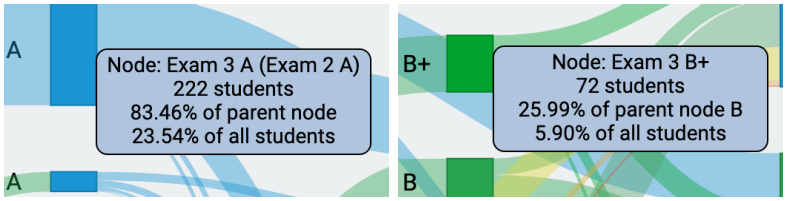


Fig. 4. Level-of-detail: vertical split. (a) and (b) show the same breakdowns as in Figure 3 (a) and (b), respectively. (c) applying this variant to all nodes on Exam 3 column to the finest level of detail.

In Figure 4, (a) and (b) show the same breakdowns using vertical split rather than horizontal split. As depicted, the original node is replaced by the new categories. However, the Sankey diagram may begin to lose its readability as the number of nodes increases due to constant expansions of the level-of-detail hierarchy of nodes. Such a result is shown in Figure 4(c).

4 Evaluation

We conducted an uncontrolled user study to evaluate the effectiveness of each of the variants in communicating inflow or level-of-detail information to users and gauge which variant users prefer to use. We did not include the original Sankey diagram in the study because it only supports the examination of inflow. We recruited students from the Department of Computer Science and Engineering at our university who responded to a department-wide email soliciting paid



(a) (b)

Fig. 5. Tooltips when hovering over nodes in the hierarchical Sankey diagram.

volunteers. A total of 19 participants completed the study, and each was compensated $20 for a session that lasted less than one hour. The majority of the participants are juniors and seniors (nine juniors and seven seniors). In terms of demographic breakdown, four are Asian, 11 are Caucasian, one is Hispanic, and three are other ethnicities. We hosted our implementation as a web application and distributed the URL to participants on the day they chose to complete the study. Participants completed the study using their personal computers. One of the student co-authors of this work was available via a Zoom link at all times to answer questions from the participants or help troubleshoot the application.

4.1 Tutorial

As the first step of the study, we provided a tutorial that introduces participants to the variants and helps them understand how to use the variants to gain a particular type of insight from the dataset. In addition, the tutorial prepares participants for the format of questions that will be asked in the next phase. The tutorial is divided into four sections, one for each variant. Participants were instructed to use the application to select a variant and then interact with the diagram for that variant. For example, in the Inflow: Vertical Separation section, participants separated node Exam 3 A and were told how one of the new nodes, Exam 3 A (Exam 2 A), represents students that received an A on Exam 2 and an A on Exam 3. The instructions also describe the corresponding tooltip as shown in Figure 5(a) that appears when the Exam 3 A (Exam 2 A) node is hovered. The tooltip includes title, number of students corresponding to the node, percentage of the parent node (Exam 3 A) the previous number represents, and percentage of all scores from Exam 3. In the Level-of-Detail: Horizontal Split section, participants clicked on node Exam 3 B and took note of the three resultant nodes: Exam 3 B+, Exam 3 B, and Exam 3 B-. The instructions describe the corresponding tooltip as shown in Figure 5(b) that appears when the Exam 3 B+ node is hovered. The tooltip includes title, number of students, percentage of the source node (all B grades), and percentage of all grades for Exam 3. Participants completed the tutorial by following a Google Form and checking a checkbox after finishing each section. The purpose is to dissuade them from skimming through the tutorial without thoroughly reading it. More than two-thirds of the participants agreed

that the tutorial was sufficient to learn how to use and interpret the hierarchical Sankey diagram.

4.2 Survey

After participants completed the tutorial, they moved on to a new Google Form with the survey including seven sections of questions. The four initial sections, one for each variant, come with multiple-choice and short-answer questions. The questions were designed to test the understanding of the variant and how effective it is at conveying either inflow or level-of-detail to the participants. We randomized the order of these four sections for participants to mitigate the possible accumulation of learning effects in answering questions. For the inflow variants, participants were asked to identify the most common score from Exam 1 received by students who also received the specified scores on the following two exams and state the number of students who received these scores on the three exams. This question assesses the ability of the variants to allow a user to compare the outflows of a given node, separated by the inflow to that node, to a specified target node. They were also asked to answer the number of students who received a given grade on Exam 3 with specified grades on Exam 1 and Exam 2. This question assesses the ability of the variants to convey to a user how a path can be created between three nodes using inflow information. For the level-of-detail variants, participants were asked to identify the number of students who received a specific letter grade (A- or B-) on Exam 3 and state the percentage of the overall letter grade those students represent. These questions assess the ability of the variants to convey the size of a node's subcategories.

The fifth section of the survey asks two questions that can be answered with an inflow variant and two questions that can be answered with a level-of-detail variant. In this section, there is no guidance on which variant to use in answering a question. Participants were asked to give the answer to the question and also state which variant was used to come up with the answer. In this way, we can examine whether users understand the differences between the variants and which one is optimal for a given task. The sixth section of questions asks participants to describe the usage of a Sankey diagram to assess their baseline knowledge of the diagram. We also asked various questions that require comparing and contrasting the inflow and level-of-detail variants against each other, stating the advantages and disadvantages of each variant and stating the preferred choice between each pair of variants. These questions were asked to assess further the participant's understanding and preference of the variants. The seventh section asks participants to rate on a scale of 1-5 how strongly they feel that (1) the variants, in general, provide additional use beyond the standard Sankey diagram and (2) the pair of inflow and pair of level-of-detail variants, are effective in understanding the prior grades of students that comprise nodes and the more specific grades that can be revealed by breaking down nodes. These questions were asked to understand the participant's judgment of the effectiveness of our in-place approaches to the hierarchical Sankey diagram.

Table 1. Participant performance for the four variants of the hierarchical Sankey diagram, corresponding to the first four sections of the survey.

type	variant	# questions	aggregate score	percentage
inflow	vertical separation	3	37/57	64.91%
inflow	color distinction	3	46/57	80.70%
level-of-detail	horizontal split	2	37/38	96.37%
level-of-detail	vertical split	2	37/38	96.37%

4.3 User Study Results

Table 1 reports the aggregate scores of participants for the first four sections of the survey, which evaluates how effective each variant is in helping the participants answer the questions correctly. Participants were more accurate in their responses when using color distinction to answer questions in the first and second sections of the survey, which measure the performance of the inflow variants. Only 11 students were able to correctly answer "*Of the students who scored a B on Exam 2 and an A on Exam 1, how many students received an A on Exam 3?*" using vertical separation, whereas 17 students correctly answered "*Of the students who scored an A on Exam 2 and a B on Exam 1, how many students received a B on Exam 3?*" using color distinction. Both variants saw 18 correct responses to the questions "*How did most of the students perform on Exam 1 who scored a C on Exam 2 and a C on Exam 3?*" (vertical separation) and "*How did most of the students perform on Exam 1 who scored a B on Exam 2 and a B on Exam 3?*" (color distinction). However, only seven participants correctly identified how many students the group contained using vertical separation, compared to 13 correct answers using color distinction.

Performance was relatively similar when using the level of detail variants. Using horizontal split, all participants correctly answered "*How many students received an A- on Exam 3?*" and 17 correctly answered "*What percentage of all students who scored an A on Exam 3 does this make up?*". Using vertical split, all participants correctly answered "*How many students received a B- on Exam 3?*" and 16 correctly answered "*What percentage of all students who scored a B on Exam 3 does this make up?*". The concept of hierarchical subcategories that make up the level-of-detail variants is perhaps very straightforward to users, explaining the high level of correctness in participant responses to these questions. On the other hand, the inflow variants attempt to solve a problem that potentially requires more thought from users, even with an effective visual tool. In this case, we can see a more apparent separation in understanding between the two inflow variants for the questions asked, suggesting that color distinction is more effective at creating a continuous path between nodes that a user can interpret.

We found that participants preferred color distinction over vertical separation when asked to choose a variant to answer an inflow-related question about the data. Still, the results were reversed when asked directly to select a preference between the two. Of the two questions that require a participant to choose an inflow variant without guidance, the question, "*How many students received an*

A on Exam 1, a B on Exam 2, and a C on Exam 3?" was answered by ten participants using the color distinction variant and eight participants using vertical separation (one student chose horizontal split). The other question of this type, "*Of the students who scored a C on Exam 3 and a B on Final Exam, how did most of these students perform on Exam 2?*" was answered by 11 participants using color distinction and eight participants using vertical separation. However, when asked to select their preference between the two variants, 14 participants picked vertical separation, and only five preferred color distinction. A similar outcome occurred in the comparison of preference between level-of-detail variants. "*How many students scored 78/100 on Exam 3?*" was answered by ten participants using horizontal split and nine using vertical split, and "*How many students received a B+ on Exam 3 and an A on Final Exam?*" was answered by ten participants using horizontal split and seven using vertical split (two participants used vertical separation). When asked directly which of the two variants they preferred, 11 participants selected vertical split, while eight picked horizontal split.

5 Conclusions and Future Work

We have presented the design and evaluation of the hierarchical Sankey diagram, which includes inflow support via vertical separation or color distinction and level-of-detail support via horizontal split or vertical split. The evaluation results show that color distinction is more effective than vertical separation for inflow-related questions, although vertical separation was preferred in use. We note that these results may not indicate the efficacy of both variants as color distinction is limited to a single column, but vertical separation can scale to multiple columns. Thus, the survey could only assess questions related to a single column for comparison. Furthermore, horizontal split and vertical split perform similarly for level-of-detail-related questions, and there is no clear preference over horizontal split and vertical split. We believe that the general in-place approach presented for augmenting the standard Sankey diagram can be helpful in many cases. Therefore, besides making the current code open source, we will generalize our implementation as a library to benefit others.

Acknowledgments

This research was supported in part by the U.S. National Science Foundation through grants IIS-1455886, DUE-1833129, IIS-1955395, IIS-2101696, and OAC-2104158. The authors would like to thank the anonymous reviewers for their helpful comments.

References

1. Burch, M., Timmermans, N.: Sankeye: A visualization technique for AOI transitions. In: Proceedings of ACM Symposium on Eye Tracking Research and Applications (Short Papers). pp. 48:1–48:5 (2020)

2. Chou, J.K., Wang, Y., Ma, K.L.: Privacy preserving event sequence data visualization using a Sankey diagram-like representation. In: Proceedings of ACM SIGGRAPH Asia Symposium on Visualization. pp. 1:1–1:8 (2016)
3. Huang, C.W., Lu, R., Iqbal, U., et al.: A richly interactive exploratory data analysis and visualization tool using electronic medical records. BMC Medical Informatics and Decision Making **15**, 92:1–92:14 (2015)
4. 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), 558–568 (2006)
5. Lehrman, B.: Visualizing water infrastructure with Sankey maps: a case study of mapping the Los Angeles aqueduct, California. Journal of Maps **14**(1), 52–64 (2018)
6. Lupton, R.C., Allwood, J.M.: Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. Resources, Conservation and Recycling **124**, 141–151 (2017)
7. Müller, G., Sugiyama, H., Stocker, S., Schmidt, R.: Reducing energy consumption in pharmaceutical production processes: Framework and case study. Journal of Pharmaceutical Innovation **9**, 212–226 (2014)
8. Riehmann, P., Hanfler, M., Froehlich, B.: Interactive Sankey diagrams. In: Proceedings of IEEE Symposium on Information Visualization. pp. 233–240 (2005)
9. Sansen, J., Lalanne, F., Auber, D., Bourqui, R.: Adjasankey: Visualization of huge hierarchical weighted and directed graphs. In: Proceedings of International Conference on Information Visualisation. pp. 211–216 (2015)
10. Verma, J., Luo, H., Hu, J., Zhang, P.: DrugPathSeeker: Interactive UI for exploring drug-ADR relation via pathways. In: Proceedings of IEEE Pacific Visualization Symposium. pp. 260–264 (2017)
11. Vosough, Z., Hogräfer, M., Royer, L.A., Groh, R., Schulz, H.J.: Parallel Hierarchies: A visualization for cross-tabulating hierarchical categories. Computers & Graphics **76**, 1–17 (2018)
12. Vosough, Z., Kammer, D., Keck, M., Groh, R.: Mirroring Sankey diagrams for visual comparison tasks. In: Proceedings of International Conference on Information Visualization Theory and Applications. pp. 349–355 (2018)
13. Xia, M., Velumani, R., Wang, Y., Qu, H., Ma, X.: QLens: Visual analytics of MUltistep problem-solving behaviors for improving question design. IEEE Transactions on Visualization and Computer Graphics **27**(2), 870–880 (2021)
14. Zarate, D.C., Bodic, P.L., Dwyer, T., Gange, G., Stuckey, P.: Optimal Sankey diagrams via integer programming. In: Proceedings of IEEE Pacific Visualization Symposium. pp. 135–139 (2018)
15. Zhou, K., Wu, W., Zhao, J., Li, M., Qian, Z., Chen, Y.: Click or not: Different mouseover effects may affect clicking-through rate while browsing interactive information visualization. Journal of Visualization **23**(1), 157–170 (2020)