Visualizing distributions of aggregated data sharing part-whole relationships

Hayden Coffey Ashwin Poduval

1 Introduction, Background and Identification of Tasks

Our goal involved trying to incorporate elements of the distributions of data which had been aggregated in some way while happening to share part-whole relationships. We tried to solve this problem by incorporating elements of visualizations used to represent distributions in those used for part-whole relationships, such as stacked bar charts and TreeMaps.

We felt that the nature of the objective of this theme lent itself to a high-level task in itself, because it set several constraints on the type of data. The first, obvious one - the data must share a part-whole relationship. Secondly, the task must involve aggregated data. It must be aggregated from elements that each share distinct values - they must be quantitative variables.

For our designs, we tried to visualize the distribution of time spent on different categories by the individual and the part-whole relationships of their means, since these mean values sum to 1440. This was done over the reduced data set provided in class. We chose this data set because it had a large, but not overwhelming number of categories. Ultimately, by using this data set, we had to rely on some aggregation to reduce the burden on the viewer.

We can think of at least one other problem using the ATUS data set which generalizes to the same high-level problem we try to solve. What if instead of looking at the distribution of values aggregated into the mean of a category (such as T01/Personal Care or T05/Work), we were to investigate the distribution of the means for time spent on different sub-categories which are aggregated by totaling to form a higher-level category? For example, the average time spent on T01 consists of many sub-categories like T010101/Sleeping, T010102/Sleeplessness, etc. We might want to know about the distribution of their means, since the individuality of these values is lost when they are aggregated by totaling into a higher-level category such as T01. In the following sections, we provide some explanation about 3 of our designs.

2 Proportional Area Distributions

Fig 1 was inspired by stacked bar charts. Stacked bar charts utilize one axis for comparison, i.e., for allowing multiple stacked bars to be plotted and compared against one another. Position is widely recognized amongst the most powerful methods of encoding information, as discussed in numerous works such as [2] and [5]. We accordingly chose to use the position along the x-axis to encode information about the distribution of parts in a part whole. We encode this information using both tick marks and violin plots to provide detailed information to viewers. Violin plots help users understand whether there is a very high density of values at certain points in the distribution like at zero, which can be hard to capture with tick marks. Tick marks can help the viewer understand why the violin plots have long tails while remaining unobtrusive.

While the positions of the axes could easily have been swapped, we chose to encode part-whole information along the y-axis to take advantage of viewers' familiarity with stacked bars, area, and stream charts. Distribution means are marked, and their values are available along the twinned x-axis above the chart. We also mark out the categories which correspond to each area on the right-hand side of the image. We initially used color to encode different categories. However, many color mappings implied some sense of ordering, which doesn't exist between the categories in the data set, and was therefore undesirable for us. Ironically, [1] identifies the lack of natural ordering to be a major issue with rainbow mappings, but we tested a few designs with rainbow colors for precisely that reason. However, some of the other issues mentioned in the paper contributed to it scaling poorly with the number of categories. We therefore abandoned encoding category type using colors, and directly displayed them using text. We use category codes here

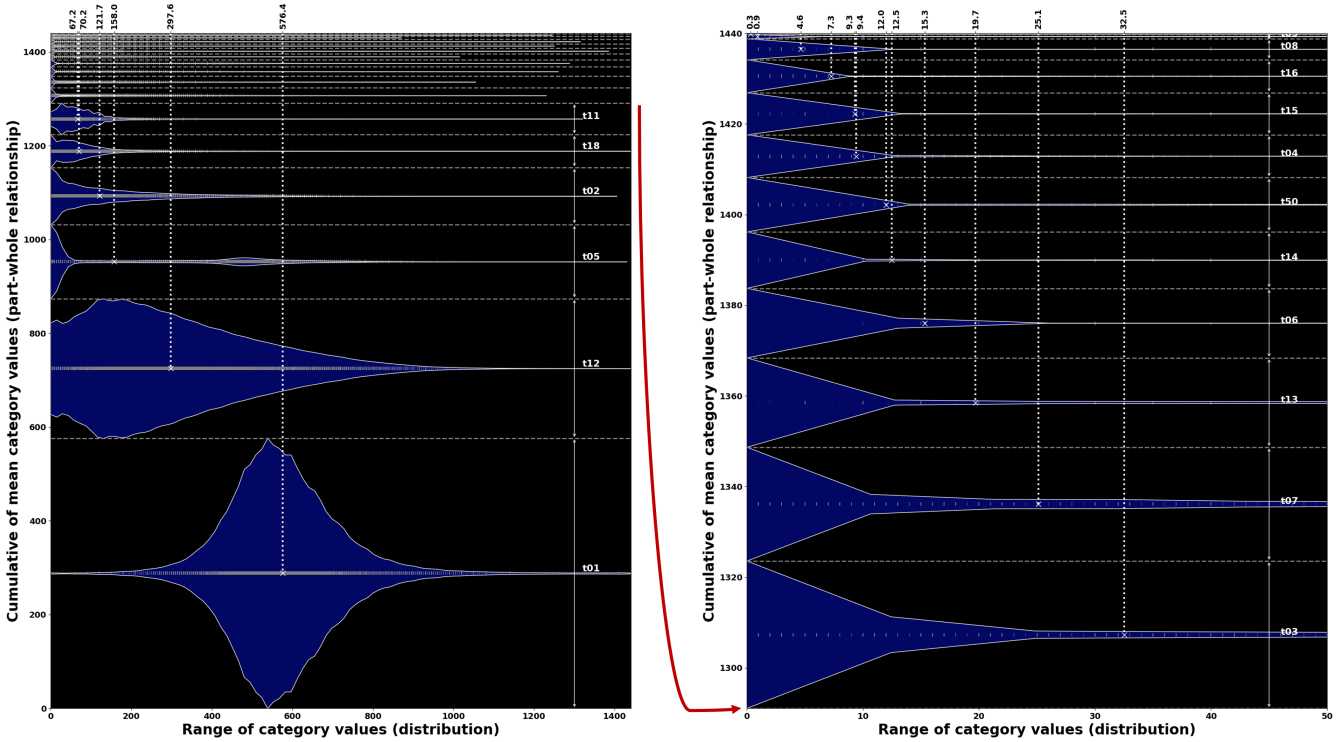

Figure 1: Distributions of ATUS categories with areas proportional to their means (which follow a part-whole relationship with total time available in a day.) The subplot on the left presents a complete picture using data for all categories, while the one on the right is a zoomed in view of the 12 smallest categories which are hard to distinguish in the complete plot.

because they are concise, but we realize that an unfamiliar audience would find them confusing - however abbreviated category names would easily fit, and besides, the position of the labels could be adjusted to avoid overlap with the distribution, so we felt this was a satisfactory solution. Also, many categories are clustered together (as seen in the top left of the left subplot) because their contribution to the whole is relatively low. We consequently made use of a second subplot to zoom into these categories and help the viewer gain as much information as possible.

Pros This design primarily focuses on two things - providing detailed information about the distribution of aggregated variables, and representing the part-whole relation. It does the first well, providing information comparable to a faceted boxplot or histogram. Since the details shown and choice of distribution is customizable, it can be modified to provide as much information (like that of medians and quartiles) as needed. It works better for part-whole studies than the faceted designs. It is also capable of supporting a fairly large number of categories (10-20 or so.)

Cons Like with stacked bars or stream charts, the part-whole information is fairly high-level. For example, in Fig. 1 we have one category with a mean of about 576 minutes and two categories with means less than 1 minute. While it is easy enough to understand the contribution of the first, it is hard to visualize the latter here. Additionally, if we had an even larger number of categories, this problem would likely be magnified. Also, performing comparisons is non-trivial. For example, if we wanted to generate such visualizations for data taken on each day of the week, comparing them could not be as easy as it is with something like stacked bars or even the faceted designs.

3 Node Link Design

Figure 2 displays our final node link design which took inspiration from [4]. We capture the part-whole relationship of the data with size encoding of the nodes, as well as the number of categories feeding into each node through the edge count. The red rings around each node encode the range of the standard deviation and give a sense of how ”fuzzy” a node is, while the blue rings detail 25, 50, 75, and 99% quantiles to give a sense of the distribution of the data.

We chose a node link design as it gives us a few different visual elements to work with:

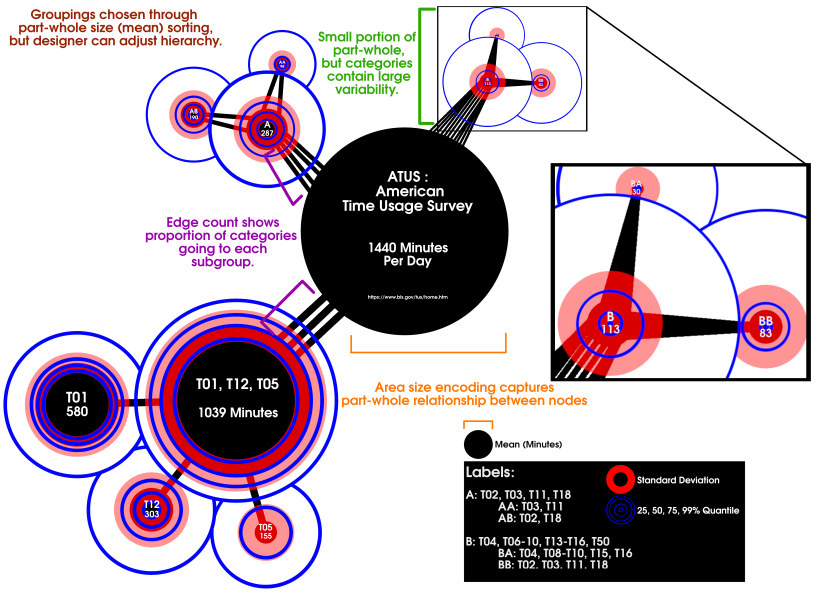

Figure 2: Node link representation of ATUS categories incorporating standard deviation and quantile information.

- First, we can control the size of the nodes and the number of connecting edges to capture part-whole information such as total sub-grouping size and number of unique categorical elements in said grouping.
- Next, a node-link layout allows for us to organize the data into hierarchies in a modular fashion and treat any node as the ”root” of its own unique diagram. A potential interactive use case would be to allow for the viewer to select a node to generate a ”zoomed-in” view of that particular branch of the diagram as we see with the annotated *B* branch. Each branch can stand on its own for a part-whole evaluation if the viewer is not interested in the other categories.
- Lastly, we utilize aliasing for larger subgroups that include many categories that comprise a small portion of the overall part-whole data.

Pros If the goal is for the viewer to quickly gather a general sense of part-whole information and data variability, the size encoding and hierarchical layout work well for establishing a size order. The viewer can quickly build a mental map of the data groupings layout by looking at the child nodes of the root node as their size puts a maximum constraint on their own children. Also, this hierarchical design aids scalability by providing the designer a way to aggregate data while preserving the design layout by merging nodes up into their parent nodes as discussed above.

The quantile and standard deviation encodings aid the viewer in quickly identifying highly variable data and enhance the visualization over a standard node-link graph. In a more traditional design, branch *B* could easily be disregarded due to its size, but with the added distribution information, we can see that this data is noisy and has a large range of values.

As compared to a faceted design, this approach allows for the viewer to capture both part-whole and distribution data within a single focal point of the visualization instead of scanning between multiple figures.

Cons One of the largest issues with this design would fall into the ”Algorithm” stage of Munzner’s nested hierarchy [5] in that due to the largely custom design, this figure was entirely hand drawn in GIMP. As a result, applying the design to a new data set would be largely time consuming in its current state. This can be an issue with scalability as well, as more categories means more nodes, and more time spent drawing. However, node-link graphics libraries do exist and the quantile and standard deviation information is drawn as circles of varying sizes, so creating a programmed implementation that handles drawing the node-link layout and applying visual annotations on top of the nodes is not outside the realm of possibility.

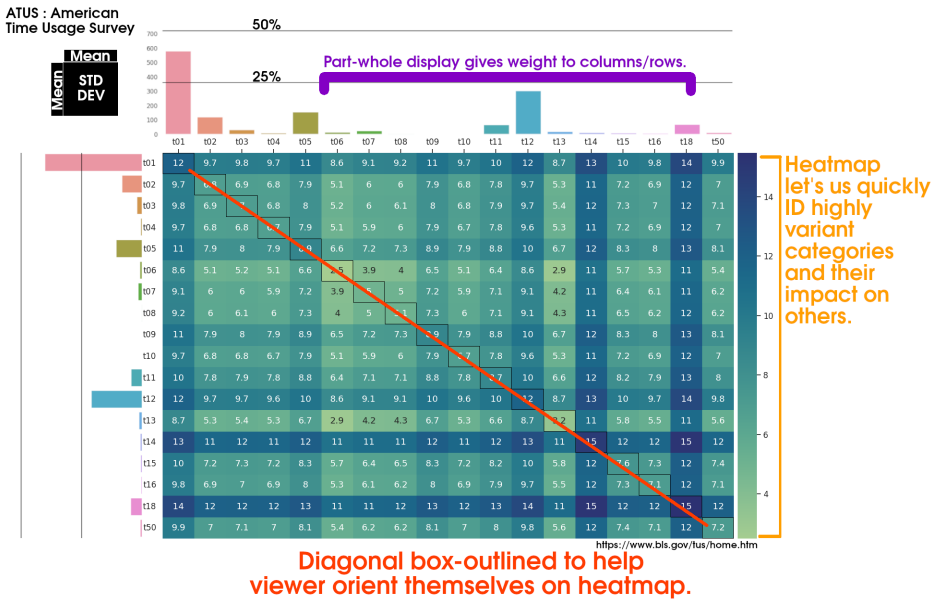

Figure 3: Heatmap and bar chart hybrid design. Bar chart provides visual weight and part-whole information for columns/rows in standard deviation table.

If the goal is to provide the viewer with fine details regarding the information, then the lack of ticks makes doing so challenging as the viewer must ”eyeball” and estimate what the number the sizes of various markings represent.

4 Heatmap / Bar Chart Design

Figure 3’s design combines a heatmap with a bar chart to capture standard deviation and part-whole information. The heatmap is the central feature of the figure, displaying average standard deviation for different combinations of part-whole data, while the bar chart provides visual weight to the rows and columns with its part-whole display.

One of the challenges with trying to capture distribution information alongside part-whole we see a rapid growth in dimensionality if we want to compare variability across all possible category combinations. A heatmap [3] provides us with a way to capture information pertaining to these combinations which scales reasonably well up to the number of categories tested in our reduced data set (18). The bar chart displaying part-whole information for each category is redundantly placed on both the columns and rows to add visual weight to them and provide the viewer with part-whole context. With this design the viewer is able to quickly select a row or column, obtain an understanding of what portion of the data it makes up, and trace it along the table to see how its variability interacts with the variability of other categories.

Pros As compared to Figure 2, this design provides the viewer with higher information granularity as cells are numbered and the bar charts have tick lines demarcating the 25 and 50% values. With the color encoding in the heatmap, the viewer can quickly identify which categories have the highest variability (i.e. t14/t18), and the bar chart provides a visual filter the viewer can use (i.e. ignore rows/columns that are tiny). Additionally, unlike Figure 2, none of the categories are grouped in this layout, so no category can ”hide” the information of another. This design is similar to a multi-faceted layout, but emphasizes the connection between the two primary charts by using the bar charts as a direct annotation to the heatmap. This provides less visual separation between the two and allows the viewer’s focus to travel in a straight line from part-whole to distribution data for a particular category.

Cons Scalability is one issue with this design, as we are looking at N^2 elements for N categories. However, the design does not preclude itself from being able to utilize aggregation as a work around for this at the cost of information granularity.

REFERENCES

[1] D. Borland and R. M. Taylor II. Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications*, 27(2):14–17, 2007.

[2] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[3] G. Grinstein, M. Trutschl, and U. Cvek. High-dimensional visualizations. In *Proceedings of the Visual Data Mining Workshop, KDD*, volume 2, page 120, 2001.

[4] L. A. Lucas. Infographic: The 1,234 satellites orbiting earth, Dec 2014.

[5] T. Munzner. *Visualization analysis and design*. CRC press, 2014.