Clipped Graphs: A Compact Time-Series Encoding

Haihan Lin* Carolina Nobre† Amanda Bakian‡ Alexander Lex§

University of Utah

Abstract

The analysis of multiple time series and associated quantitative or categorical attributes is an important task. Air pollution data, for example, is captured along many dimensions. This data is valuable for studying the association between air quality and diseases risks. In these cases, researchers need to view a large amount of data for multiple cases simultaneously, limiting the space that is available for each time series. In this abstract, we introduce clipped graphs, a hybrid clipped area chart that uses redundant color coding for visualizing time series data with skewed distributions and relatively rare peaks. We designed clipped graphs for use in compact tabular layouts. We use binned color scales for the full data range but clip outliers above a pre-defined threshold. The clipped peaks can be revealed through interaction. We integrate clipped graphs into an existing multivariate data visualization system, which visualizes clinical genealogies and detailed data about individuals in a tabular layout.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Time series

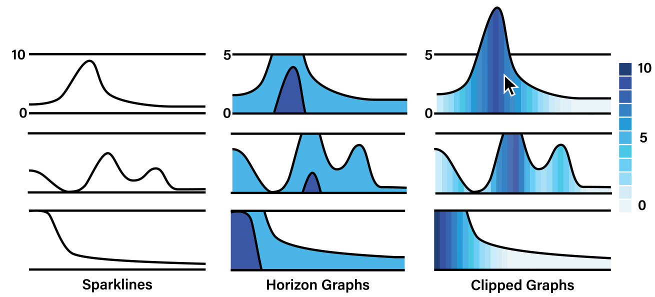

Figure 1: Comparison of three compact time series visualizations: sparklines, horizon graphs, and our clipped graphs. Clipped graphs redundantly encode data with the height of the line and color encoding. The line is clipped for outliers, to preserve details for lower values. Clipped peaks can be revealed on hover, as indicated in the top row.

1 Introduction

Air pollution can affect many aspects of our life. Exposure to air pollution has been associated not only with poor physical outcomes, but also an increased risk of mental illness. A recent study [1], for example, revealed relationships between short term acute air pollution exposure and suicide death risk. Short term air pollution data, as used in the study above, is collected as a time series of weather stations' recordings of air pollutants concentration during a period of time before and after a suicide incident.

Line and area charts are typical ways of visualizing time series. Given enough space, traditional line and area charts show the variation of time series data well. However, when analyzing many different time series for different items, and of different data types, the available space is limited. In this case, it is useful to embed a time series visualization in the cell of a table [5].

Air pollution data also have a property that visualization designers need to consider: while they have a tremendous range of possible values, most values tend to be within a narrow band close to the zero line, resulting in a skewed distribution. According to the US Air Quality Index (a compound index considering multiple pollutants), for example, values exceeding 50 are considered detrimental to health. The scale of the index, however, reaches up to a value of 500. These data characteristics pose challenges for visualization design: analysts need to understand variation on a small scale, as a change between, for example, 30 and 70 can have health effects, but also need to be aware of peaks.

In this abstract, we introduce clipped graphs, a new design for visualizing time series data with skewed distributions. We use a hybrid color-coding and positional encoding approach. Lower-scale variations are shown redundantly with color and area height, while values above a certain threshold are clipped and only encoded with color. On demand, the clipped region of the line chart can be revealed. We integrate clipped graphs into Lineage, our tool for visualizing clinical genealogies [4], to demonstrate the use of clipped graphs.

*e-mail: hhlin@cs.utah.edu
†e-mail: cnobre@sci.utah.edu
‡e-mail: amanda.bakian@hsc.utah.edu
§e-mail: alex@sci.utah.edu

2 Related Work

The most common approaches for visualizing time series data are line and area charts. To visualize a time series, the x-axis is used to encode time and the y-axis encodes the dependent variable. A curve or a polyline shows the values of the dependent variable over time.

A compact version of line charts are sparklines or reduced line charts [5] (shown in Figure 1), which are essentially down-scaled line charts that omit axis information. Typically, sparklines are placed in floating texts to provide context about an entity that is being referred to, and a variety of placements have been studied [2]. Due to the space constraint in a tabular layout, the range of data that can be represented with a sparkline is limited, which is aggravated by strongly skewed distributions.

Horizon graphs [3] alleviate the space problem by segmenting a line chart and overlaying the segments, as shown in Figure 1. The segments of the line chart are put into different color-coded layers and then vertically translated to overlap. A peak in the data will hence appear to "wrap around" the chart and come back in from the bottom. Though it can be difficult to read at first, study has shown that with equal or fewer than four bands, horizon graphs outperform sparklines for certain tasks [3].

3 Clipped Graphs Design

We propose a new technique for visualizing compact time series: clipped graphs. Our design is inspired by horizon graphs, but instead of wrapping high values around, we simply clip them and only reveal the peaks on mouse-hover, as shown in Figure 1. In addition, we use a redundant color-encoding on a not-clipped scale to visualize areas with extreme values. This hybrid encoding leverages perceptually efficient positional encoding through the lines for low values, while high values are indicated through the color coding, and details can be revealed on demand. Clipped graphs are explicitly designed for

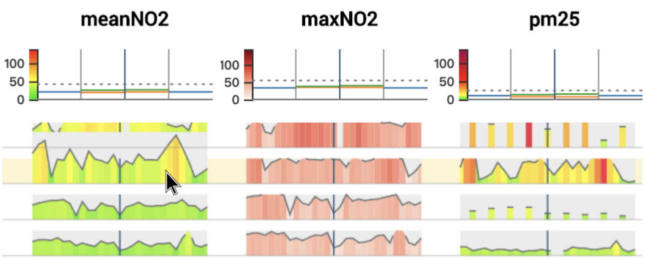

Figure 2: Clipped graphs implementation in a tabular layout visualizing air quality data. The cell in the second row of the first column is drawn to full scale as the user's mouse hovers on the row. The otherwise clipped peaks overlap with the row above. These dataset shows time series immediately before and after an event, which is indicated by a vertical line in the center of each horizon graph. The different columns use different color scales. Columns 1 and 3 use an EPA color scale, while Column 2 uses a single hue color scale. Only data presented in the data set is drawn; missing values are left empty, as is evident in the rightmost column. The header contains color legends, in addition to an indication of the clipping threshold (dashed line) and summary values.

data that follows a skewed distribution. We automatically determine the clipping point based on the distribution of the data. By default, we set it such that 90% of all data points are below the clipping threshold, but this value can be adjusted to suit the dataset and tasks.

Figure 2 shows our implementation of clipped graphs in a tabular layout using an air pollution dataset. In the header of each column, a color map visualizes the color scale used in the column. On a shared axis, we visualize properties of the dataset. A horizontal dashed line marks the clipping threshold of the attribute. The header also contains an overview of the data in the column, visualizing average values before or after an event, which is indicated by a center line. We also experimented with an overlaid view of all time series in the header, yet we found that this approach lead to excessive clutter, as our data is too variable for this approach.

In order to support tasks related to air quality data, we use three color schemes for different types of time series data. For air quality data, we use the U.S. Environmental Protection Agency (EPA) standardized color used for their Air Quality Index (AQI). While this color-scale has well known perceptual deficiencies, for example it is not suitable for users with red-green color blindness, the strong domain convention convinced us to nevertheless employ it. These weaknesses are partially mitigated by the fact that we use color only redundantly with positional encoding. The EPA scale is defined for discrete bins, however, we decided to use a continuous version that can communicate more nuance. We use CIELAB color space for color interpolation to ensure it is perceptually accurate. In addition to the EPA scale, we use a blue-red diverging color scale for temperature, following conventional practice, and a single-hue red scale for other data types, such as NO2 concentration.

4 Implementation

We demonstrate the use of clipped graphs with the air pollution data, retrieved for 14 days before and after a completed suicide. It is calculated using all local air monitoring stations, weighted by inverse distance of the individual's residence. To enable the domain experts to find individuals exposed to high air pollution within a family structure, we added clipped graphs to show the time series in the attribute table component of Lineage [4], as shown in Figure 3. Users can select an attribute, including the time series, from the drop-down menu on the navigation bar, and add the attribute to the current display. A vertical line in the middle of the cell indicates the day of suicide. We support sorting on time series attributes based on the average values before suicide. The column header consists of four small columns, in addition to the aforementioned legend. The first and fourth columns are the global average of the attribute value before and after suicide across all cases in the dataset, whereas the two columns in the middle are the averages of each family selected in the current view. This helps analysts judge whether a family in general is more exposed to air pollution than expected.

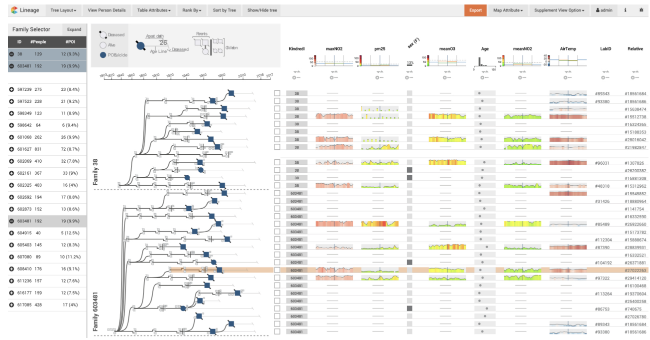

Figure 3: Overview of Lineage with integrated Clipped Graphs showing air pollution time series data. Two families with a history of suicide are shown. The family tree on the left reveals the family structure. Nodes in blue identify individuals that have committed suicide. Each node corresponds to a row in the table on the right. The table displays selected attributes, in this case various time-series and several numerical and categorical attributes.

5 Conclusion and Future Work

In this abstract, we have introduced clipped graphs, a novel time series visualization technique for use in compact layouts, such as in table cells. Clipped graphs are optimized for displaying data following a skewed distributions. We demonstrate the utility of clipped graphs by integrating them in an existing clinical genealogy visualization tool. Our technique can be applied not only to air pollution data, but also many other fields for time series data with large range and occasional peaks.

While we demonstrate that clipped graphs are useful in our context, we have not shown that they perform better than alternatives, such as linear spark lines, logarithmic spark lines, or horizon graphs. In the future, we would like to conduct a formal study on how clipped graphs perform compared to these alternatives under different conditions.

References

[1] A. V. Bakian, R. S. Huber, H. Coon, D. Gray, P. Wilson, W. M. McMahon, and P. F. Renshaw. Acute Air Pollution Exposure and Risk of Suicide Completion. *American Journal of Epidemiology*, 181(5):295–303, 2015. doi: 10.1093/aje/kwu341

[2] P. Goffin, W. Willett, J. Fekete, and P. Isenberg. Exploring the Placement and Design of Word-Scale Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2291–2300, Dec. 2014. doi: 10.1109/TVCG.2014.2346435

[3] J. Heer, N. Kong, and M. Agrawala. Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*, pp. 1303–1312. ACM, 2009. doi: 10.1145/1518701.1518897

[4] C. Nobre, N. Gehlenborg, H. Coon, and A. Lex. Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs. *Transaction on Visualization and Computer Graphics*, 25(3):1543 – 1558, 2019. doi: 10.1109/TVCG.2018.2811488

[5] S. C. Radakovitz, A. M. Buerman, A. Garg, M. J. Androski, M. K. Becker, and B. S. Ruble. Sparklines in the grid, Nov. 2009.