Authors' preprint. Published in GI 2016.

Assessing the Readability of Stacked Graphs

Alice Thudt* University of Calgary
Jagoda Walny† University of Calgary
Charles Perin‡ University of Calgary
Fateme Rajabiyazdi§ University of Calgary
Lindsay MacDonald¶ University of Calgary
Riane Vardeleon‖ University of Calgary
Saul Greenberg** University of Calgary
Sheelagh Carpendale†† University of Calgary

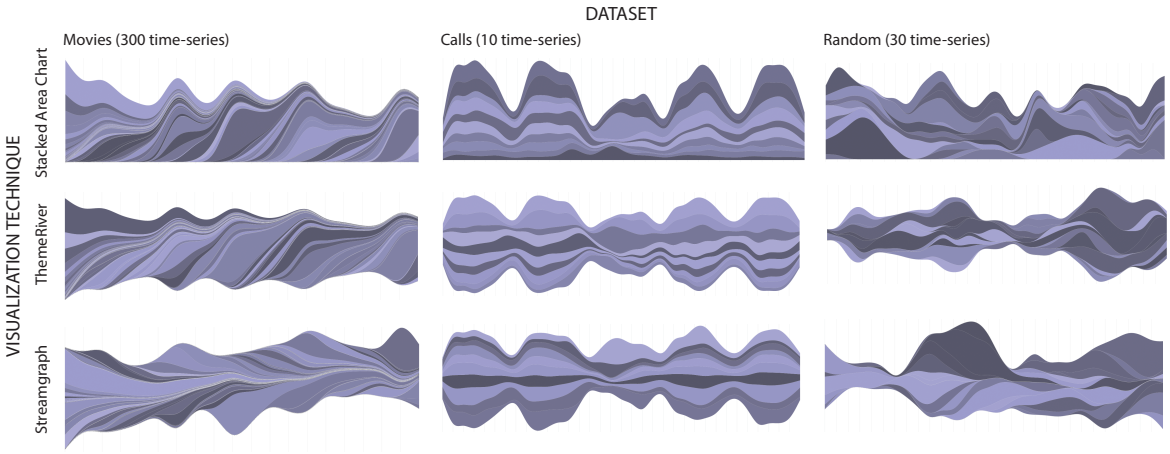

Figure 1: Using a simple stacked area chart, ThemeRiver, and streamgraph to visualize the Box Office Revenue (Movies) Dataset [13], the 311 Calls Dataset [27, 40], and a randomly generated dataset.

Abstract

Stacked graphs are a visualization technique popular in casual scenarios for representing multiple time-series. Variations of stacked graphs have been focused on reducing the distortion of individual streams because foundational perceptual studies suggest that variably curved slopes may make it difficult to accurately read and compare values. We contribute to this discussion by formally comparing the relative readability of basic stacked area charts, ThemeRivers, streamgraphs and our own interactive technique for straightening baselines of individual streams in a ThemeRiver. We used both real-world and randomly generated datasets and covered tasks at the elementary, intermediate and overall information levels. Results indicate that the decreased distortion of the newer techniques does appear to improve their readability, with streamgraphs performing best for value comparison tasks. We also found that when a variety of tasks is expected to be performed, using the interactive version of the themeriver leads to more correctness at the cost of being slower for value comparison tasks.

Keywords: Visualization, Streamgraphs, Stacked Area Charts, Interaction Technique, Readability

Index Terms: H.5.m [Information Interfaces and Presentation (e.g., HCI)]: Miscellaneous

*e-mail: alice.thudt@googlemail.com
†e-mail: jkwalny@ucalgary.ca
‡e-mail: charles.perin@ucalgary.ca
§e-mail: frajabiy@ucalgary.ca
¶e-mail: macdonla@ucalgary.ca
‖e-mail: vriane.vardeleon@gmail.com
**e-mail: saul.greenberg@ucalgary.ca
††e-mail: sheelagh@ucalgary.ca

1 Introduction

Stacked area charts and their variations are time-series visualizations that stack multiple time-series on top of each other. Stacking causes distortion to the shape of each individual time-series representation. It has been assumed that this affects readability due to perceptual experiments that have shown the human perceptual system to be less accurate at estimating and comparing values on curved slopes than on straight ones [12]. In response, the evolution of this stacked area technique has focused on reducing the distortion of individual streams. Despite these concerns about readability, stacked graphs are popular outside of a scientific context, in more casual scenarios [35], in which a key challenge is balancing the goal of readability—supporting accurate and efficient extraction of information—with making the visualization aesthetically appealing to evoke curiosity, draw people's attention to the visualization, or create a pleasurable experience for the viewer. Stacked graphs have been used to create attractive representations of data from personal music listening histories [8], the box office revenue of movies [10, 13] and social media content [17]. Given that stacked area charts are being promoted in spite of their limitations, understanding their relative readability (and what affects readability) is important.

We study a series of techniques that have been proposed to improve the balance of readability and aesthetic appeal in stacked graphs (see Figure 1). A basic stacked area chart (top row) stacks all time-series on a straight bottom baseline, causing maximal distortion to the time-series positioned at the top of the chart. ThemeRiver [23] organizes time-series symmetrically along a horizontal center axis, effectively reducing the outermost possible position of any stream by half (center row). Streamgraphs [10] further reduce the distortion, or "wiggle", in individual layers, resulting in an asymmetrical outer shape (bottom row). Byron and Wattenberg extracted anecdotal evidence of the issues and benefits of streamgraph readability [10]; however, no study has formally tested how the different techniques compare in terms of readability.

We contribute to this ongoing discussion by providing a formal investigation of the relative readability of stacked area charts, The-

meRiver, and streamgraphs for both real-world and randomly generated datasets (as seen in the columns of Figure 1). We also include an interactive ThemeRiver with baseline straightening to assess if simple interaction can help to mitigate readability problems.

Results indicate that the decreased distortion of the newer techniques does appear to improve their readability, in particular for value comparison tasks. We also found that when a variety of tasks is expected to be performed, using the interactive version of the themeriver leads to more correctness but is slower for value comparison tasks. Overall, we recommend using the last iteration of stacked area charts, streamgraphs, in static conditions; and to use either streamgraphs or an interactive themeriver in interactive conditions, depending on the tasks to performed.

2 Related Work

Many approaches exist for the visualization of multiple time-series, including line graphs, braided graphs [26], horizon graphs [20, 36], reduced line charts [41], and stacked graphs [10, 23]. Evaluating these different approaches in terms of readability is necessary to assess their efficiency (e. g., [25, 26, 34, 39]).

2.1 Evaluation of Time-Series Visualizations

Previous studies have evaluated graphical perception of multiple time-series visualizations. Horizon graphs [20, 36] overlay high values on lower values using a two-tone pseudo colouring technique, allowing for a vertically space-efficient time-series visualization for visualizing multiple time-series. Heer et al. [25] compared the readability of filled line charts and two variations of horizon graphs—mirroring or not mirroring the negative values. They measured speed and accuracy of discrimination and estimation tasks at various chart sizes and used a randomly generated dataset. They found that mirroring does not impair readability and that horizon graphs improve readability at smaller chart sizes.

Javed et al. compared four visualizations of multiple time-series: simple line graphs, braided graphs, small multiples, and horizon graphs [26]. They measured the correctness and completion time of three tasks (finding a global maximum, assessing global slope, and local point discrimination) for a synthetically generated dataset with 2,4, and 8 time-series. They found that superimposed line graph techniques work best for local tasks, and line graphs that create separate charts are more efficient for juxtaposed tasks.

Furthermore, Perin et al. [34] compared reduced line charts, horizon graphs and interactive horizon graphs. They measured binary correctness, error magnitude, and completion time of: finding the maximum value among several time-series for a given time point, discrimination of values among several time-series and time points, and finding a reference time-series. In a departure from the previously mentioned studies, they used a real-world financial dataset with 2, 8, and 32 time-series. They found that the interactive condition was most effective for datasets with large numbers of time-series.

As these studies illustrate, there is considerable interest in evaluating visualizations of multiple time-series. However, while the previous studies compare overlaid or small multiple-style visualizations, no studies have compared stacked graph visualizations.

2.2 Stacked Graphs of Multiple Time-Series

Stacked graphs (Figure 1, top row) are an approach to visualize multiple time-series by stacking filled shapes ('streams') that represent individual time-series on top of each other, on a straight baseline. At each point on the time axis, the height of each stream represents its value. The end result is an outer shape that is an aggregate view of all the time-series. This technique distorts the baseline of each individual stream (except for the bottom stream). The outer shape showing the value of the aggregated time-series is not distorted. The distortion has been the impetus for several incremental improvements to this technique [4, 10, 23], detailed below.

ThemeRiver [23] (Figure 1, middle row), stacks individual time-series around a central axis, resulting in a symmetrical outer shape. As the shapes are stacked both upwards and downwards from the axis, the outermost stream in a ThemeRiver is less distorted than the outermost stream in a stacked area chart. Havre et al. ran a small experiment comparing the readability of ThemeRiver with stacked bar charts and found ThemeRiver to be useful for identifying an overview of the changes, but less useful for identifying minor trends. Participants also expressed interest in interacting with the visualization, particularly to reorder the time-series vertically. One weakness of ThemeRiver is that it disproportionately emphasizes streams that happen to be arranged in the middle of the river [1].

Streamgraphs [10] (Figure 1, bottom row) sort individual streams in a way that smooths the distortion of each stream by reducing their 'wiggle-factor'. This results in an asymmetric outer shape. The authors claim that the reduced distortion improves readability over ThemeRiver. This line of reasoning based on foundational perceptual studies [12] is commonly accepted in the visualization community. Heer et al. deliberately excluded stacked graphs from their study of horizon graphs, due to their lack of support for negative value display, and based on [12]. A blog post [29] analyzed several examples of casual streamgraphs published on the web and concluded that static or printed streamgraphs are difficult to read due to their uncommon shapes, and that interaction is a way to mitigate this problem.

This strategy has also been investigated by Baur et al. who introduced interaction to stacked graphs with the aim of mitigating their stated perceptual issues [4]. They developed a hierarchical ThemeRiver for touch-interactive devices with interactive stream reordering. The approach of adding interaction to mitigate downsides of visual representations is not recent [16] and has been proven to improve the efficiency of some time-series visualizations [34].

Despite the concerns regarding their readability, stacked graphs have aesthetic value, leading to widespread use on the web, e. g., the Ebb and Flow of Box Office Sales[13], World Cup Twitter streamgraph [22], the NameVoyager [44], and ThemeRiver [23]. *Artifacts of the Presence Era* [43] is an installation in a museum that samples video recordings of the space around it and displays them as stacked sedimentary layers. The NameVoyager [44] is a popular web visualization that represents baby names' popularity over time. ColourVis [31] is another aesthetically appealing visualization that maps to a stacked line graph the proportions of colours used in sets of images over time. ColourVis can be viewed in numerous configurations, including with a baseline at an arbitrary position.

Stacked graphs have become widespread due to their aesthetic appeal, and increments of the original technique have been proposed to overcome their supposed limitations. Despite this, no formal studies have compared the readability of stacked graphs and their variations. In this paper, we derive evaluation criteria for stacked time-series visualizations and assess the readability of stacked graphs.

3 Readability

We define *readability* as the extent to which a visualization supports graphical perception—"the visual decoding of information encoded on graphs" [12]. Readability of visualizations has been of fundamental importance in the InfoVis community, beginning with the perceptual classification of visual variables [6]. Bertin classifies visual variables such as location, color, size and orientation into different "levels of organization". At these levels of organization he distinguishes "selective perception" (i.e. determining the category of a visual mark), "ordered perception" (i.e. comparing the orderings of two categories) and "quantitative perception" (i.e. numerically defining the difference between two visual marks).

Cleveland and McGill experimentally investigated graphical perception of visual encodings [12]. They asked participants to estimate the ratio between two marks and measured their accuracy after displaying the graph for 2.5s. This study resulted in a refined perceptual

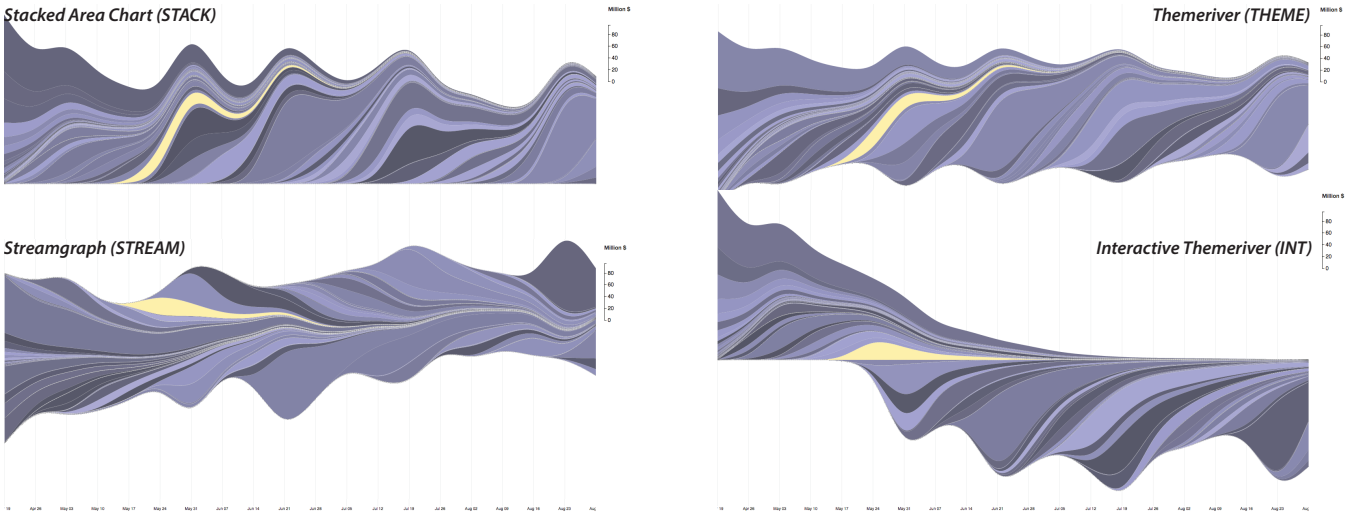

Figure 2: The four evaluated techniques: stacked area chart (STACK), ThemeRiver (THEME), streamgraph (STREAM), and ThemeRiver with interactive baseline straightening (INT), all displaying the Movies dataset.

classification of Bertin's visual variables [6], and was later confirmed using Amazon's Mechanical Turk crowdsourcing [24]. These results have been used to justify design choices in visualization, such as the low-distortion design of streamgraphs based on the perception of curved slopes and slope ratios [10]. The issue of slope ratio comparisons is further addressed for line charts [39], however, this has not been studied in the context of stacked graphs directly. Our experiment studies these issues with the goal of better understanding the effectiveness of the different stacked graph variations.

The information extractable from a visualization consists of more than individual data points. Bertin suggests that three information levels should be readable from an information graphic [5]:

- *Elementary level:* Extraction of individual values. For multiple time-series, reading the value of one time series at one point. This is an integral part of values comparison; it is necessary to read the individual values to be able to compare them.
- *Intermediate level:* Comparisons and trends in subsets of characteristics. We classify the reading and comparison of time-series trends (e.g. identifying growth or peaks) as intermediate level tasks. These differ from elementary tasks as reading one or two values is not enough to identify a trend in a time-series.
- *Overall information level:* Global values and trends. This involves tasks that require reading values of the combination of streams. An example would be comparing the aggregated values of all time-series at multiple time points, or reading trends at the global level (such as finding peaks or growth of all time-series combined).

We take all three of these information levels into account when choosing tasks to measure the readability of stacked graphs.

4 Experiment

Our experiment is an empirical contribution to the ongoing discussion on the readability of stacked graphs. In particular, we investigate the impact of static and interactive straight baselines, symmetry, and wiggle on the extraction and comparison of individual and aggregated values and on the readability of trends.

4.1 Techniques

We compared the readability of four stacked graph techniques for visualizing time-series data, illustrated in Figure 2: a basic stacked area chart (STACK), a ThemeRiver (THEME), a streamgraph (STREAM), and our own interactive technique, a ThemeRiver with interactive baseline straightening (INT). INT initially shows multiple time-series using the ThemeRiver technique. By clicking on one individual stream, the baseline of that stream is straightened (see Figure 2 (INT)). Clicking on the bottom layer turns the streamgraph into a stacked area chart. Using the ThemeRiver technique initially causes less distortion in the rest of the graph when an arbitrary stream's baseline is straightened than if a streamgraph were used initially.

All of these visualizations display both individual time-series and the aggregation of multiple time-series (by stacking them on top of each other). This supports reading and comparing values within time-series and across multiple streams, as well as global comparisons within the overall stream. It also allows for reading both local and global trends. We kept the qualities of all four techniques as constant as possible by using the same style and amount of curvature, and the same colour scheme. The color scheme we used is similar to the one used in the original streamgraphs paper [10] and scales up well to large numbers of streams. We added simple highlighting interaction, common for these types of graphs, that changes the colour of a stream to light purple on hover and to light yellow on selection.

4.2 Datasets

The types of time-series datasets that can be displayed using the discussed stacked graphs can vary greatly in size and data distribution, and thus impact the performance of each technique. For validity, and to vary both size and data distribution, we used two real-world datasets that stacked graphs have been applied to, the Box Office Revenue Dataset (Movies) [13] and the 311 Calls Dataset (Calls) [27, 40], as well as a randomly generated dataset (Random). Using randomly generated datasets is common in such perceptual studies (e. g., [21, 25, 26, 30]).

The Movies dataset is a subset of the data used by the New York Times in their streamgraph visualization entitled "The Ebb and Flow of Box Office Sales" [13]. The data that we used was collected over the course of 20 weeks from 4/20/2012 until 8/31/2012 and contains 300 time-series with 20 time points, each time-series representing the revenue of one movie in U.S. dollars. Time-series have non-zero values over an average of 7.27 time points (weeks). This results in short streams that are similarly shaped, due to the similar development of revenue for movies.

The Calls dataset is a subset of complaint calls made over the New York 311 line available at the NYC OpenData website [40]. This dataset was originally used in a streamgraph for Wired Magazine [27]. We use a subset of this data extracted by Vallandingham [42]. The data was collected during Hurricane Sandy over the

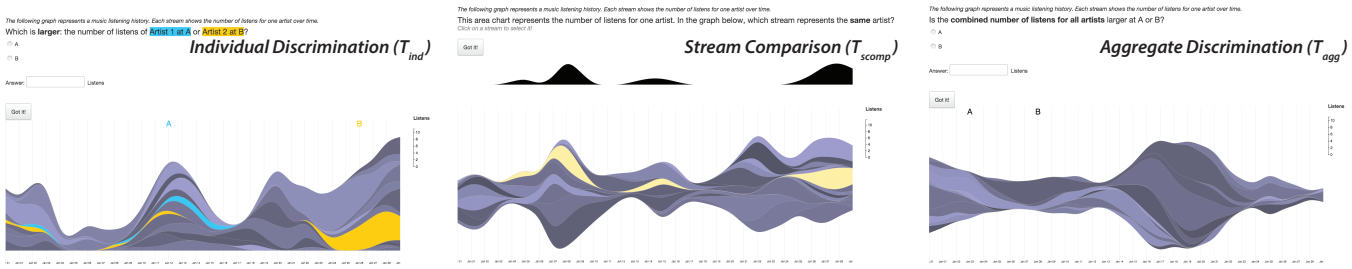

Figure 3: The three evaluated tasks: individual discrimination (T_{ind}) with a stacked area chart (left); stream comparison (T_{scomp}) with a streamgraph (center); and aggregate discrimination (T_{agg}) with a ThemeRiver (right).

course of 35 days from Oct 14th, 2012 to Nov 17th, 2012. The dataset contains 10 time-series over 35 time points, each time-series representing the number of calls on one topic. Each time-series has non-zero values over the entire period shown in the graphs. All streams show a similar weekly pattern.

We used the d3 [7] implementation of Byron's test data generator [9, 10] to create a series of random time-series datasets. The datasets all contained 30 time-series over 30 time points with varying temporal patterns. Following Byron's example, we referred to this dataset as a listening history in the study, where each time-series represents the numbers of listens for one artist during 30 days.

4.3 Tasks

We designed the tasks for this experiment based on 1) Bertin's three levels of information that should be readable from a visualization [5], and 2) frequently tested tasks for time-series, and related to Andrienko and Andrienko's taxonomy for time-series [3].

Individual discrimination (T_{ind}): *"Which is larger: the [value of time-series x] at A or [time-series y] at B?"* Participants were asked to identify the larger of two individual stream values in the graph at two given time points A and B (see Figure 3 (left)). This task can be associated with the elementary information level. In Andrienko and Andrienko's taxonomy, T_{ind} is an elementary task (about individual data elements) that requires direct comparison. This is also a standard benchmark task for evaluating time-series visualizations (e. g., [11, 19, 25, 37, 38]).

The discrimination points A and B as well as the corresponding streams were highlighted in bright yellow and blue in both the graph and the question. This choice ensured that we measured the time to assess the values only – not the time to find the right streams.

Stream comparison (T_{scomp}): *"The following area chart represents [time-series data]. In the graph below, which stream represents the same [time-series]?"* Participants were shown a stream with a straight baseline and asked to find its equivalent in the displayed graph (see Figure 3 (center)). This intermediate level task makes it possible to assess how people perceive trends. In Andrienko and Andrienko's taxonomy, T_{scomp} is a synoptic task (about a set of values) that requires direct comparison. Such a task has previously been used for evaluating time-series visualizations (e. g., [11, 34]).

Aggregate discrimination (T_{agg}): *"Is the combined [value of time-series] larger at A or at B?"* Participants were asked to compare the aggregate value of all time-series at two given time points A and B (see Figure 3 (right)). This overall information level task requires participants to make global comparisons. In Andrienko and Andrienko's taxonomy, T_{agg} is an elementary task (about individual values) that requires direct comparison. Similarly to T_{ind}, this task has been used extensively to assess the performances of time-series visualizations (e. g., [11, 19, 25, 26, 34, 37, 38]). It has also been found to be easier than T_{ind} [25, 34].

4.4 Hypotheses

We expected the following effects of visualization technique on tasks:

- **H1** For T_{agg}, we expect answers to be more correct in STACK and INT than THEME and least correct in STREAM.
- **H2** For both T_{ind} and T_{scomp}, we expect answer to be more correct from INT over STREAM, over THEME, to STACK.
- **H3** Overall, we expect INT to be slower than all three other techniques for all tasks.

We formulated **H1** because using STACK with its global baseline, T_{agg} simply consists of comparing the height of the aggregated chart from the baseline at two points. INT can be turned into a STACK thus should result in similar correctness. T_{agg} is difficult to perform using THEME because the technique does not provide a global baseline, and even harder using the asymmetric STREAM.

We formulated **H2** because both T_{ind} and T_{scomp} are more difficult to perform with distorted streams. INT makes it possible to limit the distortion by setting an appropriate baseline, STACK has more distortion for individual streams, THEME reduces distortion slightly, and STREAM reduces distortion significantly.

We formulated **H3** because INT requires interacting with the visualization to change the baseline. Such interaction costs have been observed in a similar evaluation [34].

4.5 Procedure and Apparatus

To test our hypotheses we decided to run an experiment in a controlled lab setting where we could ensure that the perceptual conditions are the same for all participants. This was necessary as we observed during a pilot tudy that participants make use of their hands and other objects to measure parts of the visualization on the screen instead of relying solely on their perceptual capacities. To avoid this bias we refrained from running an experiment with a larger number of participants on an online platform.

Participants were first asked to fill out a short questionnaire to determine demographic information, web usage, and their previous experience with visualizations. After that, the experimenter explained the conditions to the participants in the order determined by the Latin square design. Then, the experimenter instructed the participants to sit in front of the screen, a 30-inch monitor with 2560 x 1600 resolution, with their back on the backrest of the chair, and to only touch the mouse and keyboard. This constraint was necessary to make the results comparable and to prevent people from using their hands as a measurement aid rather than relying primarily on their perceptual capacities. Participants were then asked to follow the experiment in the browser application.

The experiment was broken into four parts, one for each visualization technique. Each part consisted of the three tasks (T_{agg}, T_{ind}, T_{scomp}), and each of these tasks was performed on the three datasets (Movies, Calls, and Random). For each visualization technique $\times$ task, participants first performed a training round using another random dataset to familiarize themselves with the task using the current technique. After participants completed all tasks for one technique, the experimenter asked them to comment on their experience.

For each task, the question alone was displayed at first. After participants read and understood the question, the visualization was

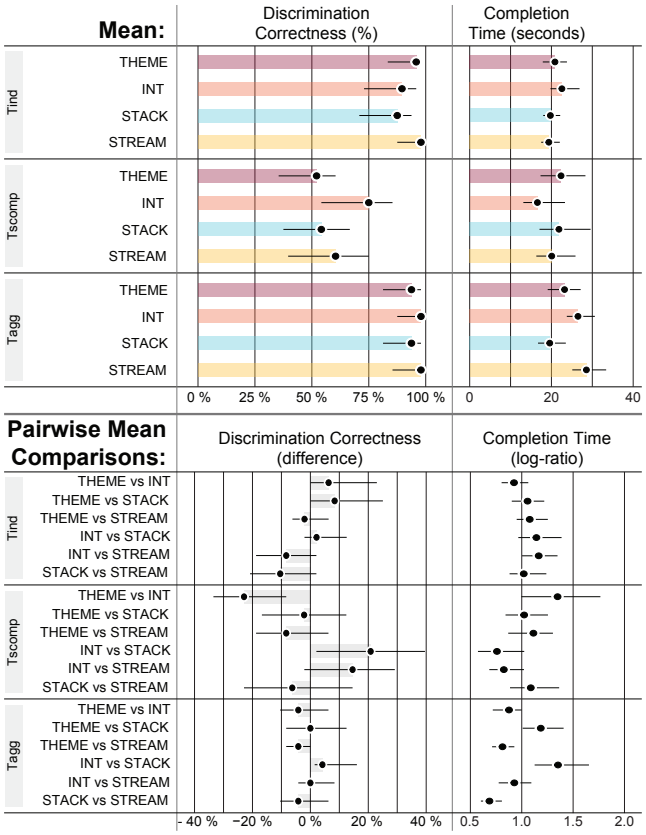

Figure 4: Mean and pairwise comparisons correctness and completion time by task and by technique. Error bars show 95% bootstrapped confidence intervals. Pairwise comparisons for A vs B are $A - B$ for correctness and $log(A)/log(B)$ for completion time.

displayed and the start time was logged. For both T_{agg} and T_{ind}, participants were asked to identify the larger value of two points (A or B) in the graph and select their answer from a radio button list. For T_{scomp}, participants selected the time-series in the graph by clicking on it. For all tasks, the given answer could be changed until the submit button was clicked, logging the end time.

After the experiment, we asked participants to score the visualization techniques based on their aesthetic preference and perceived legibility on [0–10] continuous scales.

4.6 Participants

We recruited 16 participants (9 male, 5 female, 2 declined to answer) aged 18–65 years old with various occupations (12 students) in a variety of fields (in consideration of the casual context of popular stacked graph visualizations). All participants frequently used computers, but had heterogeneous knowledge of visualization. We recruited these participants through posters and mailing lists at a university. They received monetary compensation of $20. One participant had previously seen STREAM, three had seen THEME, seven had seen STACK and one was very familiar with STACK.

4.7 Experiment Design

Our study used a within-subjects design, with conditions arranged using a balanced 4x4 Latin square [32] in order to mitigate learning effects. The independent variables were visualization technique (4: STACK, THEME, STREAM, INT) × task (3: T_{agg}, T_{ind}, T_{scomp}) × dataset (3: Movies, Calls, Random), or 36 trials per participant. With 16 participants, this produced a total of 576 trials.

Our dependent variables were: *correctness* (the ratio of correct answers compared to all answers) and task completion *time*.

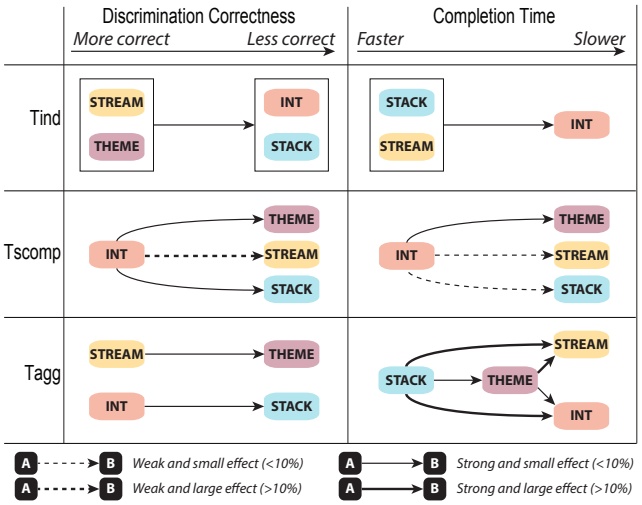

Figure 5: Summary of differences between techniques for discrimination correctness and completion time by task. In each cell, the horizontal position of each technique icon encodes its performance, ordered from best (left) to worst (right). Two icons with the same horizontal position mean that the two techniques perform similarly.

4.8 Performance Results

We base our analyses on estimation, i.e., effect sizes with bootstrapped [28] confidence intervals [15]. This approach, recommended by the APA [2], is an alternative to NHST (null hypothesis significance testing), whose limits are growing concerns in various research fields. Hundreds of articles criticize the indiscriminate use of NHST and various scientific disciplines are more and more recommending banning the use of NHST (for a summary see [18]).

Using confidence intervals, black points in pairwise comparison figures indicate the best estimate while intervals indicate all plausible values, with point estimates being about 7 times more likely than interval endpoints [14]. When performing pairwise comparisons, the measures are computed for each participant. For correctness, if the confidence interval graphical representation does not cross the 0% vertical line then there is a 95% chance of difference between the techniques (identical to $p < .05$). The same is true for time, with the 1.0 vertical line because we applied a log-transform to measures of time and thus compute ratios. We interpret the results visually as follows: we call an effect *small* if it is likely to be smaller than a 10% difference between two techniques, and *large* if it is likely to be larger than a 10% difference. We qualify these effects as being *weak* if there may be an effect, but the confidence interval is wide or crosses the 0/1 vertical line, and *strong* if there is confidently an effect, with the ratio between the part of the confidence interval which is on the opposite side of the vertical line and the total length of the confidence interval being small or null.

We compare the results for each technique by task. Figure 4 shows mean and pairwise comparison correctness and completion time by task and technique. Figure 5 summarizes these comparisons.

For T_{ind}, in terms of correctness, we found that both THEME and STREAM performed better than both INT and STACK. In terms of time, INT was slower than both STACK and STREAM.

For T_{scomp}, we found that INT performed best overall. In terms of correctness, INT had strong, small advantages over THEME and STACK and a weak, small advantage over STREAM. Similarly in terms of time, INT had a strong, small advantage over THEME and small, weak advantages over STREAM and STACK.

For T_{agg}, in terms of correctness, STREAM had a strong, small advantage over THEME and INT had a strong, small advantage over STACK. In terms of time STACK performed best, followed by THEME

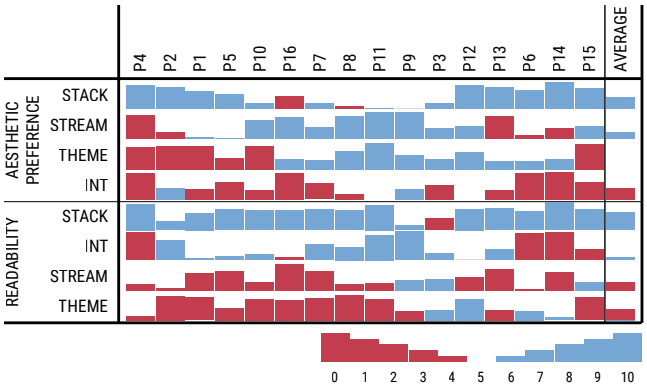

Figure 6: Participants' preferences for technique aesthetics and readability on 1–10 Likert scales.

then both STREAM and INT. STACK had a strong, large advantage over both STREAM and INT; and a strong, small advantage over THEME. THEME had a strong, large advantage over STREAM and a strong, small advantage over INT.

4.9 Questionnaire Results

Participants' preferences for technique aesthetics and readability are presented in Figure 6, created using Bertifier [33].

Overall, participants found STACK to be both the most aesthetically pleasing and most readable. They found STREAM to be the second most aesthetically pleasing visualization, but did not find it to be readable. THEME obtained mitigated results in terms of aesthetic preference, with some extreme divergent judgments. Finally, participants found INT to be the least aesthetically pleasing visualization, but the majority of participants found INT to be readable.

5 Discussion

We discuss our findings by summarizing the performance results as well as the results obtained from the questionnaires and providing possible explanations as to why these results occurred.

5.1 Discussion of Results

For T_{ind} and T_{scomp} we originally expected that the technique performance would be, from best to worst: INT, STREAM, THEME, and STACK, in terms of correctness (**H2**). This hypothesis was primarily based on the degree of individual stream baseline distortion in each technique. For T_{ind} we found that both THEME and STREAM performed better than both INT and STACK. The low performance of STACK is unsurprising, since this technique distorts individual streams the most. However, INT surprisingly performed worse than both THEME and STREAM while we expected the opposite. The poorer performance of INT may have been influenced by the number of small, similar streams: because the technique only allows for the straightening of one baseline at a time, participants might have been comparing differences from memory. Here, the advantage of the straight baseline is outweighed by the disadvantage of relying on memory. Also, THEME performed as well as its incremental redesign, STREAM. This suggests that when comparing values at specific times in T_{ind} reducing the baseline distortion of individual streams does not provide a substantial benefit in terms of accuracy.

For T_{scomp}, we found that INT performed best overall. This is generally expected, as the interaction technique is appropriate for this kind of task: one can quickly click through each stream to visually compare it to the given stream, whereas the other techniques require one to mentally "undistort" each stream. However, we did not find any other difference – in particular, STACK did not perform worse than THEME and STREAM.

For T_{agg}, we originally expected both STACK and INT to perform better than THEME, then STREAM, in terms of correctness (**H1**). This hypothesis was based on the presumed advantages of having a global baseline. We found that STREAM and INT performed better than THEME and STACK. Again, STACK performed worse than expected while STREAM performed better than expected. This runs counter to our original expectations, suggesting that STREAM wiggle reduction improves the graphical perception for this task. However, the differences in this task were very small with close to 100% correctness in all four techniques. Therefore, although we found a significant difference between STREAM and THEME and between INT and STACK these differences might still have happened by chance.

We expected INT to be slowest for all tasks (**H3**), due to the time it takes to interact with the visualization. This is confirmed for T_{ind}, and to a lower extent for T_{agg}. However, INT was the fastest technique for T_{scomp}. It seems that for this task, the time spent to interact and change the baseline is lower than the time to visually browse and compare all streams with the three other techniques. We observed very different strategies in the non-interactive conditions. While some people guessed the answer (which resulted in short answer times), others tried to read the top value and the bottom value of the combined stream and mentally calculated the difference. In the INT condition people could choose to straighten the bottom baseline so that they only had to read the top value and could save the time for mental calculation. The slightly lower efficiency of INT for value comparison tasks (T_{ind} and T_{agg}) could be explained by the extra time needed for deciding on a strategy as well as for the interaction of straightening the baseline. For example, one participant indicated a change of strategy partway through the session: *"There was a learning curve. In the middle of the session I realized that in some [tasks] I can turn the entire graph into a stacked area graph. That makes it much easier for comparing the total amounts"* - P13.

Participants' aesthetic preferences and perceived readability of visualizations were not indicative of either our results or the underlying graphical perception results. While we generally confirm that stacked area charts were outperformed by the other techniques, participants overwhelmingly preferred the aesthetic qualities of stacked area charts and estimated stacked area charts to be easier to read – though participants did not perform well with this technique during the experiment and even expressed frustration. For example, P14 said that *"[Stream comparison in the stacked area chart] was really hard. [...] If I wasn't asked to I wouldn't bother doing it"* and P16 that *"[The stacked area chart is] not so easy for looking at individual streams, especially the ones up higher as they are bumped up a lot by the ones below."*.

Participants found INT to be least aesthetically pleasing while this technique has both high perceived readability and high measured readability. On the other hand, participants found STACK to be the most aesthetically pleasing and the most readable visualization while this technique has the lowest measured readability. It appears that there can be a tradeoff between aesthetics and efficiency of a technique. However, participants found STREAM to be aesthetically pleasing and STREAM performed best for two out of three tasks, suggesting that STREAM has both aesthetics and efficiency value.

5.2 Implications

Overall, the differences between techniques are usually small. The incremental improvements in stacked graph design, from STACK, to THEME to STREAM, and finally the addition of interactivity, has been largely justified using Cleveland and McGill's fundamental graphical perception studies [12]. These studies would have predicted that, overall, STREAM would perform better than THEME, which would perform better than STACK. Our results confirm these predictions, suggesting that this evolution indeed leads to small perceptual improvements for performing standard tasks on multiple time series.

Predictions Based on Theoretical Models: In the Information Visualization community, theoretical perceptual models are often used to predict the relative readability of visualizations. In the case

of stacked graphs Cleveland and McGill's perceptual model [12] was used to argue the advantage of streamgraphs over ThemeRiver and stacked area charts [10]. According to our study these predictions can work, however they do not shed light on the extent to which the techniques differ. Therefore empirical investigations can still be beneficial. In the case of stacked graphs our study suggests that while the predictions are generally correct, the effect sizes are small.

Empirically Based Suggestions for Using Stacked Graphs: Although the perceptual difficulties of stacked graphs are well known in the Information Visualization community and this form of visual representation has been critiqued a lot on the basis of previous perceptual studies [12], stacked graphs are still widely used on the web and in casual scenarios [10, 13, 17]. Our empirical investigation contributes to this ongoing discussion by offering recommendations on when to use which technique. Based on our empirical results, we make the following recommendations:

- *R1* STREAM performs best for value comparison tasks (T_{ind} and T_{agg}). Therefore, if only value comparison tasks are to be performed, we recommend using STREAM. Overall, in a static condition, STREAM appears to be the best choice, as STREAM leads to better results than both STACK and THEME.
- *R2* INT resulted in more correct answers for both T_{scomp} and T_{agg}, at the cost of being slower for T_{agg}. Therefore, if a variety of tasks are to be performed, we recommend using INT especially if people are expected to compare streams instead of comparing values at specific times. However, in a context where aesthetics are important, INT should be avoided.
- *R3* There is no performance advantage to using STACK in any condition, and we recommend *against* using this technique. However, STACK is subjectively interpreted to be both the most readable and the most aesthetically pleasing technique.

Interaction for Solving Perceptual Difficulties: While adding interactivity to a technique helped in some cases, it appeared to interfere in other cases where it required participants to rely on their memory of a perceived value for a comparison. This is interesting from an HCI perspective as it suggests that interaction has to be carefully designed to provide perceptual benefits. We recommend that visualization designers consider the tradeoff between supplying the interaction and the increase in memory load for perceptual tasks.

5.3 Limitations and Future Work

Our study is the first to quantitatively assess and compare the graphical perception of stacked area charts, ThemeRivers, streamgraphs and ThemeRiver with interactive baseline straightening. Although our findings suggest that iterations on stacked charts and interactivity lead to better graphical perception, like any controlled experiment, the results of our study are valid under the conditions of the study.

Interaction Techniques for Stacked Area Charts: Our interactive technique seemed promising for improving readability in some tasks. Future research could evaluate the impact of using a wider range of interactive techniques such as straightening more than one stream or reordering streams, as suggested by Baur et al [4].

Datasets: We picked two real-world datasets and a randomly generated datasets in order to vary the dataset properties widely. However, our selection could not be exhaustive and representative of all possible datasets. Replicating the study with other datasets would certainly lead to slightly different results.

Tasks: Our task selection was also not exhaustive. In particular we chose not to use value retrieval tasks. Value estimation tasks are also important, but as our experiment already included a large number of factors, we chose to use comparison tasks, since they would also be impacted by value estimation performance. As is the nature of controlled experiments, some of the choices necessary for this study would not normally be reflected in a real-world application of these four techniques. Although we chose our tasks to cover all three information levels, they do not cover the full range of tasks and combinations of tasks that one might attempt in a real use setting.

Colour Scheme: The color scheme we used may also have had an effect on our results. A follow-up study assessing the effect of color scheme would be an interesting complement to our findings.

Impact of Interaction Techniques: The most surprising of our results is that using INT led to lower accuracy than both STREAM and THEME for the T_{ind} task. Given that INT is an enhancement of THEME, we expected that people would perform better with the enhanced version than with the basic version, as this is usually the case [34] and as this is commonly accepted. Instead, our results suggest that adding interactivity to a static visualization technique can be detrimental. In our case, we explain the lower performance using INT due to the fact that in the interactive condition, participants may have felt that they *had* to use the interactive capabilities of the technique, and used the interactive baseline even when it did not help (for T_{ind}). Indeed, for performing T_{ind}, participants usually changed the baseline to read accurately the height of the stream x at A. Then, participants changed the baseline to read accurately the height of the stream y at B. By doing so, participants had to memorize the perceived height of x at A and compare it to the height of y at B. On the other hand, in a static condition, participants compared x and y at A and B at the same time, without having to store one value in memory. This last point raises an important question regarding interaction. Although interactive capabilities are usually designed to improve performance, the effects can be negative. Better understanding the interplay between interactive capabilities, perception, and memory, appears to be a direction worth pursuing.

6 Conclusions

We have assessed and compared the readability of stacked area charts, ThemeRivers, streamgraphs, and our own interactive ThemeRiver technique with baseline straightening for tasks covering the elementary, intermediate, and overall levels of readability for two real-world datasets and one randomly generated dataset.

This study is the first to measure the readability of stacked area charts and their incremental variations, whose design has been justified based largely on fundamental graphical perception studies [12]. Our results show that in general the expectations from graphical perception studies hold, but that the performance of each technique is highly dependent on the task. Therefor, to be able to apply knowledge from general perceptual models to predict the readability of visualizations, we have to carefully consider the task to perform.

Our study contributes empirically grounded recommendations for the use of stacked graphs. Indeed, using STREAM leads to better performance than the two other static visualization techniques for both individual and aggregated value comparison tasks. However, for stream comparisons, the INT led to better results, both in terms of correctness and completion time. Within the context of our experiment design, we recommend using STREAM for static representation of stacked time series – which reach their limits for stream comparisons. We recommend avoiding using STACK if efficiency is a criteria; but if the purpose is to create an aesthetically pleasing visualization, then stacked area charts should be considered.

We discussed the introduction of interaction as a means to mitigate perceptual difficulties based on our results. Although interaction can help people perform some tasks more accurately and sometimes more quickly, if additional memory load is introduced, then the use of interaction can be detrimental.

The findings of our experiment can inform visualization designers when deciding which visual representation to choose. In a more general sense we discussed the use of theoretical models to predict readability of visualizations as well as the introduction of interaction to solve perceptual problems in visualization.

ACKNOWLEDGEMENTS

We would like to thank our participants as well as our Sponsors Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Innovates Technology Futures (AITF) and SMART Technologies for making this research possible.

REFERENCES

[1] W. Aigner, S. Miksch, W. Muller, H. Schumann, and C. Tominski. Visual Methods for Analyzing Time-Oriented Data. *IEEE Trans Vis Comput Graph*, 14(1):47–60, 2008.

[2] American Psychological Association. *The Publication manual of the American psychological association*. Washington, DC, 6th edition, 2013.

[3] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[4] D. Baur, B. Lee, and S. Carpendale. Touchwave: Kinetic multitouch manipulation for hierarchical stacked graphs. In *Proc. ITS'12*, pages 255–264, New York, NY, USA, 2012. ACM.

[5] J. Bertin. *Graphics and Graphic Information Processing*. Walter de Gruyter & Co, 1981.

[6] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983.

[7] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE TVCG*, 17(12):2301–2309, Dec. 2011.

[8] L. Byron. Listening History. `http://www.leebyron.com/what/lastfm/`. Last accessed March 3, 2015.

[9] L. Byron. Streamgraph generator. `https://github.com/leebyron/streamgraph_generator`.

[10] L. Byron and M. Wattenberg. Stacked graphs–geometry & aesthetics. *IEEE TVCG*, 14(6):1245–52, 2008.

[11] M. Cho, B. Kim, H.-J. Bae, and J. Seo. Stroscope: Multi-scale visualization of irregularly measured time-series data. *IEEE TVCG*, 20(5):808–821, May 2014.

[12] 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, Sept. 1984.

[13] A. Cox and L. Byron. The Ebb and Flow of Box Office Sales, Feb. 2008.

[14] G. Cumming. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Multivariate applications series. Routledge, 2012.

[15] G. Cumming and S. Finch. Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2):170, 2005.

[16] A. Dix and G. Ellis. Starting simple: Adding value to static visualisation through simple interaction. In *Proc. AVI '98*, pages 124–134. ACM, 1998.

[17] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE TVCG*, 16(6):1129–38, Jan. 2010.

[18] P. Dragicevic. Fair statistical communication in HCI. In J. Robertson and M. Kaptein, editors, *Modern Statistical Methods for HCI*. Springer, 2016. In press.

[19] P. Federico, S. Hoffmann, A. Rind, W. Aigner, and S. Miksch. Qualizon graphs: Space-efficient time-series visualization with qualitative abstractions. In *Proc. AVI '14*, pages 273–280. ACM, 2014.

[20] S. Few. Time on the horizon. *Visual Business Intelligence Newsletter*, Jun/Jul 2008.

[21] M. Ghoniem, J.-D. Fekete, and P. Castagliola. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135, 2005.

[22] M. Graves. The 2010 World Cup: a Global Conversation . `https://blog.twitter.com/2010/2010-world-cup-global-conversation`.

[23] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: visualizing theme changes over time. *IEEE Symposium on Information Visualization, INFOVIS' 00*, 2000.

[24] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proc. CHI'10*, pages 203–212. ACM, 2010.

[25] 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 *Proc. CHI '09*, pages 1303–1312, 2009.

[26] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE TVCG '10*, 16(6), 2010.

[27] S. Johnson. What a hundred million calls to 311 reveal about new york. *Wired Magazine*, January 11 2010.

[28] K. N. Kirby and D. Gerlanc. BootES: An r package for bootstrap confidence intervals on effect sizes. *Behavior research methods*, 45(4):905–927, 2013.

[29] A. Kirk. Making Sense of Streamgraphs. `http://www.visualisingdata.com/index.php/2010/08/making-sense-of-streamgraphs/`.

[30] S. Lewandowsky and I. Spence. Discriminating strata in scatterplots. *Journal of the American Statistical Association*, 84(407):682–688, 1989.

[31] S. Lynch, J. Haber, and S. Carpendale. Special Section on CANS: ColourVis: Exploring colour in digital images. *Computers & Graphics*, 36(6):696–707, Oct. 2012.

[32] D. W. Martin. *Doing Psychology Experiments, 7th Edition*. Thomson/Wadsworth, 2007.

[33] C. Perin, P. Dragicevic, and J.-D. Fekete. Revisiting bertin matrices: New interactions for crafting tabular visualizations. *IEEE TVCG*, 20(12):2082–2091, Dec 2014.

[34] C. Perin, F. Vernier, and J.-D. Fekete. Interactive horizon graphs: Improving the compact visualization of multiple time series. In *Proc. CHI '13*, pages 3217–3226. ACM, 2013.

[35] Z. Pousman, J. Stasko, and M. Mateas. Casual information visualization: depictions of data in everyday life. *IEEE TVCG*, 13(6):1145–52, Jan. 2007.

[36] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-tone pseudo coloring: compact visualization for one-dimensional data. *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 173–180, 2005.

[37] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 82(398):454–465, 1987.

[38] I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77, 1991.

[39] J. Talbot, J. Gerth, and P. Hanrahan. An empirical model of slope ratio comparisons. *IEEE TVCG*, 18(12):2613–2620, 2012.

[40] The New York Times. NYC Open Data. `https://nycopendata.socrata.com/`. Last accessed March 3, 2015.

[41] E. R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[42] J. Vallandingham. How to animate transitions between multiple charts. `http://www.flowingdata.com`.

[43] F. B. Viégas, E. Perry, E. Howe, and J. Donath. Artifacts of the presence era: Using information visualization to create an evocative souvenir. In *IEEE Symposium on Information Visualization, INFOVIS'04*, pages 105–111. IEEE, 2004.

[44] M. Wattenberg. Baby names, visualization, and social data analysis. In *IEEE Symposium on Information Visualization, INFOVIS'05*, pages 1–7. IEEE, 2005.