Dynamic Switching of Data Visualization Method for Increased Plotting Scalability

Angie Mikhail-Morozov
ABB Corporate Research
konzeptmeister@gmail.com

Mika P. Nieminen
Aalto University
mika.nieminen@aalto.fi

NordiCHI '14, Oct 26-30 2014, Helsinki, Finland
ACM 978-1-4503-2542-4/14/10.
http://dx.doi.org/10.1145/2639189.2670189

Abstract

In information visualization every method performs best in its own range of plotting density. When the available area for the visualization is unknown or when the data is supplied in packages of various sizes, the best suited method cannot be chosen in advance, and the wrong choice leads to the problems of under- or overplotting. This note suggests a solution to dynamically change the visualization method based on the sample size and plotting area. The concept is illustrated with an example application for multivariate data. Suggested future work includes applying the solution to other data types, defining the switching boundaries and fading functions. The proposed dynamic switching of methods increases the plotting scalability of interactive data visualizations and thus can improve the usability of information-intense graphical user interfaces.

Author Keywords

Data graphics; information visualization; data visualization; view transformations; visual structures; big data; overplotting; underplotting; HCI; GUI; user interface; plotting scalability; plotting density

ACM Classification Keywords

H.5.2 Graphical user interfaces (GUI)
H.1.2 Human information processing

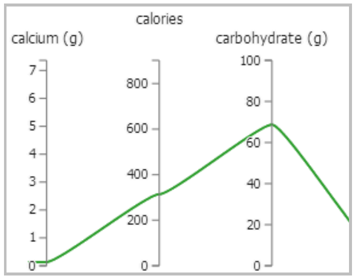

Figure 1. Underplotting example. The nutrition facts of sweet green and red peppers. A fragment from [1]

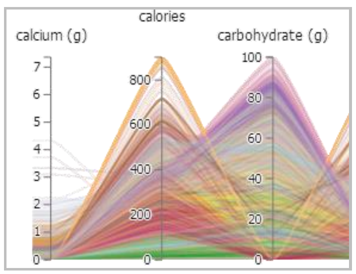

Figure 2. Overplotting example. The nutrition facts of 7637 products in 25 color-coded categories. A fragment from [1]

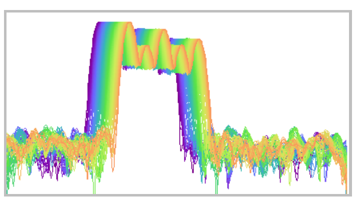

Figure 3. Using color to mitigate overplotting. 97 power spectrum estimations of Welch method with a sliding window. Own work

Introduction and definitions

The interactive data visualizations play an important role in complex data-intense human-computer interfaces. When designing a visual representation of information we as designers can make use of different *visualization methods* such as line charts, bar charts, scatterplots or bullet graphs. In each case the choice of method is individual and not formalized. With complex datasets the methods are often combined, remixed, adjusted or developed from the scratch. During the selection and adaptation of the method we consider how it performs in the given *plotting density*: some methods are good for just one sample per a given area, some work for tens or hundreds of samples, while some support any large datasets by showing a distribution or statistical summaries instead of individual points [4]. Outside of the comfortable range of the plotting density the methods often run into the problems of under- and overplotting. Underplotting creates areas of unnecessary low data density that look overly simple (fig.1). Overplotting is the opposite problem (fig.2), often encountered in Big Data applications: "Multiple objects are sharing the same space, positioned on top of one another. This makes it difficult or impossible to see the individual values, which can undermine analysis." [5] To mitigate these problems it is advised to choose the visualization method based on the sample size at hand and the available area. For example, to avoid underplotting the visual representation can be replaced by just numbers and to avoid overplotting the markers that show individual values can be replaced with their statistical distributions. It is relatively easy to pick the right method, when the plotting density is fixed. But when the data is supplied in packages of various sizes or the target physical area varies a lot, we cannot know in advance which method is best suited for the task.

We define *plotting scalability* as the extent to which a data visualization avoids under- and overplotting at any given number of data values per physical area (plotting density). In this paper we propose a concept to improve the plotting scalability by dynamically switching visualization method as the plotting density changes.

Known techniques that improve scalability

Besides changing the visualization method, other techniques are known to improve the plotting scalability. The ones reviewed here are concerned with scaling up. Scaling down can be achieved by doing the opposite.

Reducing the prominence of data markers

Stephen Few advices to reduce the prominence of each visual marker by decreasing the size of the markers, removing their fill-color, changing their shape to a more simplistic, or making them transparent [5]. On fig.2 the 7637 markers are 86.4% transparent.

Reducing the number of data values

These techniques are based on rejection of data values: statistically resample the dataset (pick only some values), filter the data or aggregate the points into clusters. Scalability through clustering was previously applied to maps in [7] and to graphs in [2].

Separating the markers by adding a variable

Data markers can be artificially separated into visual layers by adding a variable to change their aesthetics. In fig.3 overplotting is avoided by changing the color for the data markers. Each of the 97 power spectrum estimations has its own color as the sliding window moves across the dataset. We can see how the frequencies shift with time, running from violet to orange.

See the interactive demo: http://angieskazka.github.io/plotting_scalability/vectors/

Dynamic switching of visualization methods

Since it is advised to choose the visualization method based on the plotting density, we propose a concept to create the sequences of different methods that would automatically make such a choice and replace one another as the plotting density changes.

Below we illustrate how the concept can be applied to multivariate data with a fixed plotting area and gradually increasing sample size. The chosen methods are: Bullet graph, Parallel coordinates, Dash plot and Box plot. These are not necessarily optimal or preferable, other methods could be used as well. We took a modified retro-car mileage dataset [8]. Each car-sample is a vector in 4 dimensions. With just one sample (fig.4) the dimensions are easily expressed in a set of bullet graphs [3] and two or three samples can be compared side by side (fig.5). When scaling up to 7 samples only the tips of the bars are shown, connected by the lines (fig.6), turning the visualization into a classic parallel coordinates view [6]. The graph can no longer fit labels for each sample, but they can be claimed on hovering. With over 100 samples the lines for each are removed and the graph turns into a set of four dash-plots (fig.7).

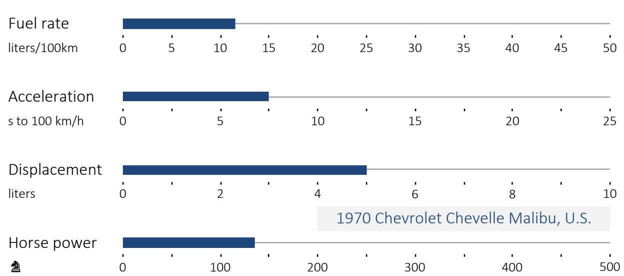

Figure 4. An illustration of visualization method switching idea for multivariate data. Sample size: 1

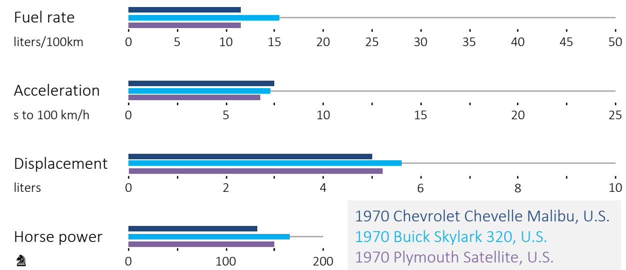

Figure 5. Sample size: 3

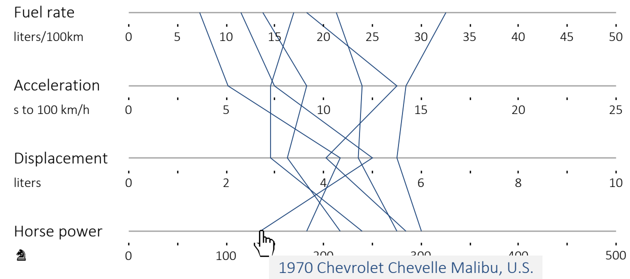

Figure 6. Sample size: 7

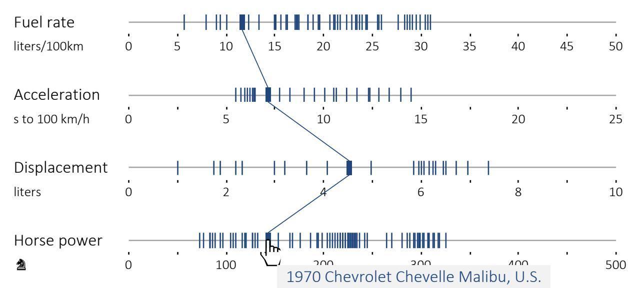

Figure 7. Sample size: 100

Finally, when approaching 400 samples, the data is aggregated into statistical box-and-whisker plot (fig.8). Individual markers are now invisible, but hovering still brings details on demand that help to identify outliers.

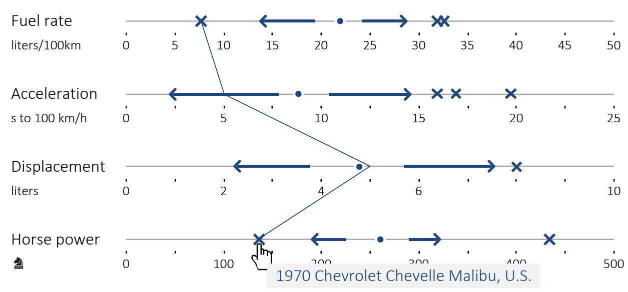

Figure 8. Sample size: 400

Future work

One direction for the future work is to apply the concept to other types of data, such as time series, parts of the whole and more, search illustrative situations and datasets, and validate the applications with appropriate usability testing. Another direction for the future work is to define the transition boundaries: when should one visualization method replace another. The threshold should be proportional to the sample size and inversely proportional to the area taken by the visualization. The limits can be found empirically based on user testing (measuring the performance of information retrieval by real users) or analytically by formulating the overlapping criteria. In order to make the switching easily understandable, the visualization methods in a sequence should be visually adjusted and unified to create a smooth transition between the methods, for example, by using the cross-fading function yet to be parameterized in the future studies.

Conclusion

To our knowledge, based on literature and interviews with visualization researchers and practitioners, the proposed dynamic switching of data visualization methods is not known to be applied at the moment. We hope that this technique will find its place among the practices of information graphics and spawn a new class of methods with increased plotting scalability.

References

[1] Chang, K. Nutrition Explorer. bl.ocks.org 2012 [Online]. Available: http://bl.ocks.org/syntagmatic/raw/3150059/

[2] Dunne. C. and Shneiderman. B. Motif Simplification: Improving Network Visualization Readability with Fan, Connector, and Clique Glyphs. In Proc. CHI 2013, ACM Press (2013), 3247-3256

[3] Few, S. Bullet Graph Design Specification. Perceptual Edge, 2006. [Online]. Available: http://www.perceptualedge.com/articles/misc/Bullet_Graph_Design_Spec.pdf

[4] Few, S. Now You See It: Simple Visualization Techniques for Quantitative Analysis. Analytics Press, Oakland, CA, USA, 2009

[5] Few, S. Solutions to the Problem of Over-Plotting in Graphs. Perceptual Edge, 2008. [Online]. Available: http://www.perceptualedge.com/articles/visual_business_intelligence/over-plotting_in_graphs.pdf

[6] Kosara, R. Parallel Coordinates. Eager Eyes 2010 [Online]. Available: http://eagereyes.org/techniques/parallel-coordinates

[7] Mahe, L. and Broadfoot, C. Too many markers! Google Developers Blog, 2010. [Online]. Available: developers.google.com/maps/articles/toomanymarkers

[8] MPG dataset. UCI Machine learning repository, 1993. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Auto+MPG