

The Need for Information Loss Metrics in Visualization

Aritra Dasgupta*

Robert Kosara†

1 INTRODUCTION

Information visualization lacks a sound framework for verification and validation of the established techniques. While we have seen a steadily increasing amount of interest in the real-world about the applicability of different visualization techniques, there is a dearth of metrics that would help in creating a baseline that can be used for comparison among them. The subjectivity of the perceptual space adds complexity to the problem because evaluation of a visualization technique depends on users' comprehension. However if we find the intrinsic properties of a visualization system that guides the user comprehension, irrespective of subjective parameters, it will pave the way for quantitative verification and validation and establishing the ground truth.

1.1 Deficit of trust in visualization

In comparison with exploratory data analysis techniques in the field of data mining, one drawback of most information visualization techniques is that the user is not necessarily able to trust what he sees on screen [1]. This may sound contradictory because the very goal of visualization is to augment the trust of the user with visual aid and thereby move the analysis forward in an intelligent manner. This factor is often ignored in the current visualization pipeline. While the issue of large data analysis is handled through dimension reduction techniques, what is still an area of open research is how to quantify what is showed on screen, so that the user is not burdened with visual information overload.

1.2 State-of-the-art

We find several instances in the literature where researchers have devised qualitative metrics to estimate the quality of the rendered image [4] or the data abstraction [6]. These are important to maximize the perceptual benefits from the visualization and implicitly deal with information loss. However, there is a stronger motivation for quantitative metrics that describe how the visual structures relate to the underlying information space. In a nutshell, we should consider visual representation not just as an end product of visualization but as the guiding factor for the exploratory analysis of the user.

1.3 Problem of information loss

One of the reasons for the deficit of trust, is that the WYSIWYG paradigm does not often hold true, because although the user sees information on screen, he does not know how much of the data-space is being represented and how much is not shown. Because we are dealing with limited pixel-space this becomes a non-trivial issue. Most visualization techniques, thus can be conceptualized as an optimization process which balances the two constraints: the fidelity of the data space and the clarity of the visualization space. One of the ways to address this problem is to quantify the information content of the visualization. But that is highly subjective and dependent on the users' perspective, therefore hard to model [5]. A more feasible solution to this problem is to study the problem of

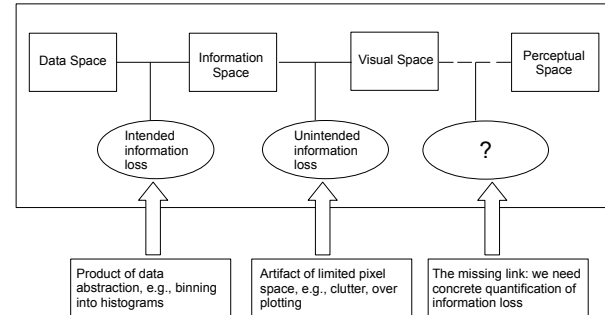


Figure 1: Different types of information loss at different stages of the visualization pipeline. the solid lines represent a strong coupling in terms of existing research, while a dotted line signifies weak coupling and need for more research

information loss and provide a quantitative analysis as part of the pipeline.

2 QUANTIFYING INFORMATION LOSS AS A MEANS TO THE ENDS

Mapping of billions of data points onto a limited screen space entails a loss in information and that is an underlying assumption in visualization, whether explicitly mentioned or not. In this section we study the variants of information loss and possible applications of controlled information loss. Information loss can be of two types: *intended* and *unintended* [8]. In Figure 1 we illustrate where they fit in the visualization pipeline.

2.1 Intended information loss

This is encountered mainly in the data-space when large data is abstracted to a summarized level so that the aggregated representation is used for visualization. There have been efforts to deal with intended information loss, i.e. measure the quality of abstraction at the data level through measures like data abstraction quality [6] and augment the visualization with that abstracted data, but more convincing metrics need to be found.

2.2 Unintended information loss

This occurs in the screen-space as a result of limited screen space and/or human perception. Many visualization techniques use panning and zooming type interaction techniques to enable the user to overcome perception related information loss. Also there have been efforts to judge the image quality [4]. However we need a concrete judgment of information loss related to different kinds of visualization tasks to be convinced about what data to show to the user without losing out on important information, as well as creating clear visual representations of that data. The inclusion of visual representation as part of the analytic loop is critically important.

While intended and unintended information loss have been implicitly addressed in the literature, no concrete quantification mechanism has been proposed so far. As shown in Figure 1 the link between the visual space and perceptual space needs significant inves-

*e-mail: adasgupt@uncc.edu

†e-mail:rkosara@uncc.edu

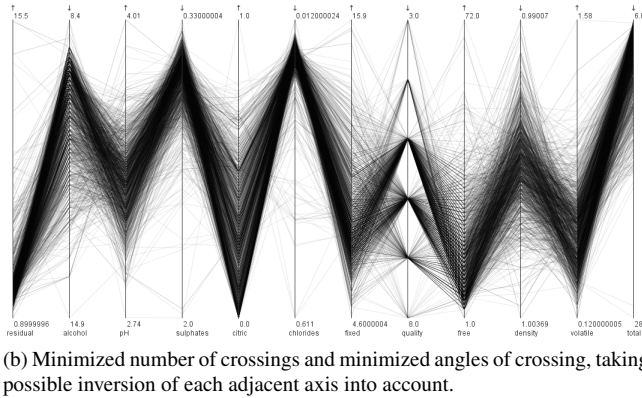
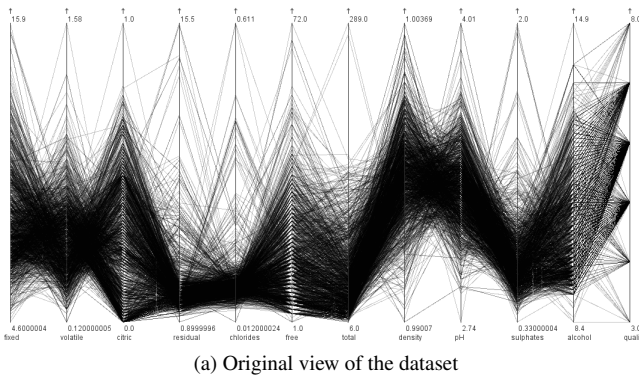


Figure 2: Optimization of parallel coordinates display using different metrics for the wine dataset.

tigation. A quantification of information loss would not only help the user understand what he is seeing, but how much he is seeing.

2.3 Uses of information loss

Controlling information loss through proper quantification can have some interesting applications. One of them is visualizing sensitive data, which involves protecting the privacy of the data. Analysis of privacy-protected data has received much attention in the field of data mining, but visualization has seen little or no work in this area. While visualization techniques continue to be widely accepted, one of the key challenges would be handling sensitive data. If we can quantify what factors contribute to information loss in each visualization technique, we can intentionally hide data, while at the same time we will know at the most how much we can hide without disturbing the fidelity of the data.

3 SPECIFIC CASES

In this section we demonstrate the merits of quantification of information loss by analyzing some existing visualization techniques:

3.1 Parallel Coordinates

Parallel coordinates [3] developed by Inselberg et al. has hitherto seen many extensions and modifications of the original technique. The problem in parallel coordinates can be summarized as visual information overload. The poly-lines spreading across multiple adjacent axes make it difficult for the user to find salient patterns. While the structure of the poly-lines convey useful information about the data-space, they lead to occlusion. Problems of adjacency and clutter have been well-studied in the literature. However the missing link between the visual and perceptual space still exists in this case. In Parnostics [2] we have tried to address this by coming with a

set of screen-space metrics that quantify the visual structures and understand information loss. Figure 2a shows the default layout of axes in parallel coordinates. Figure 2b shows the optimized layout based on the metrics like line crossings and angles of crossing. This helps in reducing the search space and guiding the users' exploration. With different quantifications, the user can more readily see clusters and correlations between axes, eliminate the causes for information loss and configure the visualization according to his need.

3.2 Other multi-dimensional visualization techniques

Similar argument can be extended to other multidimensional visualization techniques like IN-SPIRE [7] which is a text visualization tool. Its goal is to transform the high-dimensional text data into a reduced spatial representation by preserving the semantic relationships between documents. The underlying technique works by converting the unstructured text data to some mathematical vector-based representation after which clustering algorithms are applied. Like in every problem related to high-dimensional data, the challenge is to balance the loss of information with effective information retrieval. In this case too, the trade-off between necessary and sufficient information loss is fuzzy. The user does not know, quantitatively, how varying the screen resolution would affect the structure of the clusters and how his analysis, in turn, will be impacted by that. We believe, if there is a concrete quantification of the dependency between resolution and information loss, this will help more accurate and believable information retrieval.

4 CONCLUSION

We have outlined the motivation for quantifying information loss as part of the visualization pipeline and have briefly described the potential benefits by analyzing a couple of visualization techniques. We envision information loss metrics as serving a two-fold purpose: a) acting as a stepping stone towards understanding where the equilibrium lies between losing some information and maximizing insight from the represented data-space and b) forming a core part of the intrinsic properties of all visualization systems which will help us designing future techniques.

REFERENCES

- [1] E. Bertini and D. Lalanne. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proceedings Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, pages 12–20. ACM Press, 2009.
- [2] A. Dasgupta and R. Kosara. Parnostics: Screen-space Metrics for Parallel Coordinates. In *IEEE Conference on Information Visualization*, 2010, in press.
- [3] A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *IEEE Visualization*, pages 361–378. IEEE CS Press, 1990.
- [4] J. Johansson and M. Cooper. A screen space quality method for data abstraction. *Comput. Graph. Forum*, 27(3):1039–1046, 2008.
- [5] H. Purchase, N. Andrienko, T. Jankun-Kelly, and M. Ward. Theoretical foundations of information visualization. *Lecture Notes In Computer Science*, 4950, 2008.
- [6] E. A. Rundensteiner, M. O. Ward, Z. Xie, Q. Cui, C. V. Wad, D. Yang, and S. Huang. Xmdvtoolq: quality-aware interactive data exploration. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, pages 1109–1112, New York, NY, USA, 2007. ACM.
- [7] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *INFOVIS*, pages 51–58, 1995.
- [8] C. Ziemkiewicz and R. Kosara. Embedding information visualization within visual representation. In *Advances in Information and Intelligent Systems*, pages 307–326. 2009.