The Sampling Lens: Making Sense of Saturated Visualisations

Geoffrey Ellis
Computing Department,
Lancaster University
Lancaster, LA1 4WA, UK
g.ellis@comp.lancs.ac.uk

Enrico Bertini
Dipartimento di Informatica e Sistemistica
Universita di Roma "La Sapienza"
00198, Roma, Italy
bertini@dis.uniroma1. it

Alan Dix
Computing Department,
Lancaster University
Lancaster, LA1 4WA, UK
alan@hcibook.com

ABSTRACT
Information visualisation systems frequently have to deal with large amounts of data and this often leads to saturated areas in the display with considerable overplotting. This paper introduces the Sampling Lens, a novel tool that utilises random sampling to reduce the clutter within a moveable region, thus allowing the user to uncover any potentially interesting patterns and trends in the data while still being able to view the sample in context. We demonstrate the versatility of the tool by adding sampling lenses to scatter and parallel co-ordinate visualisations. We also consider some implementation issues and present initial user evaluation results.

Keywords
Sampling, random sampling, lens, clutter, density reduction, overplotting, information visualisation.

ACM Classification
H5.2 [**User Interfaces**]: Graphical user interfaces (GUI)
G.3 [**Probability And Statistics**]

INTRODUCTION
With all visualisation techniques, apart from space-filling approaches, there is the possibility that portions of the display will be saturated. By saturated, we mean that data points or lines are overplotted or the points are clustered as to be indistinct and in many cases, patterns will be hidden. Even when trends or patterns are discernable with overplotting, a quantitative assessment is difficult unless some technique, such as colouring, is used to indicate the amount of overplotting [6,11]. However, if colour and/or shape is already used to represent other attributes, clearly overplotting will result in a loss of data and the appearance of the visualisation may be unduly influenced by the order in which the points are plotted.

Our previous work proposed that random sampling could be used to make structure within saturated areas visible [4]. But this can make less dense areas become so sparsely plotted that the structures in those areas may disappear completely. Ideally we would like to see the overall picture, which includes both less and more highly dense regions. One technique to achieve this is adaptive sampling [2,5], which samples the denser areas more heavily, but the sampling rate is determined by maintaining a monotonic relationship between the density of data and density of plotted points.

In this paper we investigate an alternative technique, the Sampling Lens, which uses the lens metaphor that has become common in visualisation [3]. This allows a different level of sampling to be applied within the lens region, thus enabling an interactive examination of denser areas. This is like a focus + context method where we can see the broader context at a full density or with light sampling whilst also investigating the more densely plotted regions with greater sampling.

In the following section, we review some alternative methods for dealing with overplotting and density reduction. We then describe the Sampling Lens in more detail and look at several examples where the lens reveals structure in the data, which might otherwise be missed. Finally, we briefly discuss some implementation issues and present some initial user evaluation results.

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

BACKGROUND
Various techniques have been used to reduce clutter in a visualisation: some reduce the number of data points while others attempt to re-organise the points to remove overplotting.

General Density Reduction
Zooming into a saturated area of a display makes close points or lines appear separate (assuming points are not plotted larger) but does not solve the problem of overplotting. Zooming can also lead to a loss of context. Distortion techniques, such as Fisheye lens try to overcome this by enlarging an area of display at the expense of the rest, but these do not help with overplotting and in fact often cause a negative disorientating effect. Dynamic filtering has been employed to deal with cluttered visualisations by allowing the user to filter out uninteresting items but this presupposes that the user actually knows what is ‘uninteresting’. On the other hand, clustering techniques, as used by Artero et al [1] can improve an overview of saturated data but at the expense of detail. Other non-uniform density reduction methods have

been proposed, such as VIDA [12], which uses the principle of constant information density; however, this is more suited to cartographic data where semantic zooming is applicable.

Avoiding Overplotting

Space-filling algorithms such as TreeMaps avoid overplotting but suffer from the spatial aspect of scatter plots, although some are designed to provide correlation information [7]. Mobile Liquid 2D scatter space [10] uses a distance manipulation-based expansion lens to expose hidden points, but this does not deal with direct overplotting. Trutschl's [9] method of intelligently resolving point occlusion spreads out overlapping points to neighbouring pixels, though this only works for low density data. The Geospatial data viewer [8] instead distorts the underlying spatial area to accommodate all the data points, yet this only applies to maps where users can understand the spatial distortion.

THE SAMPLING LENS

Our Sampling Lens uses random sampling as a density reduction technique and offers the following controls:

- diameter – the user can drag the lens with the mouse around the display area and can change the diameter of the lens via a slider, thus allowing areas of different sizes to be explored.
- sampling rate – the user can uncover patterns at different density levels by either adjusting the sampling proportion manually via a slider or selecting the auto-sampling mode. In the latter case, the system dynamically adjusts the sampling rate by applying 'greater sampling' in saturated areas and 'light sampling' in sparse areas so low density patterns are not removed; hence maintaining a constant proportion of overlapping items.

Since the data points within the lens are sampled randomly to determine which ones are to be displayed, this can generate some important sampling issues [5]. To ensure display continuity, the points that are removed as the sample size is reduced, need to reappear in reverse order if the user increases the sample size again. A continually changing set of points would be very distracting. A similar technique is applied when the user is moving the lens around the visualisation so that the same data points appear when the user returns the lens to the same area of the display.

Furthermore, random sampling can give rise to artefacts in the visualisation that are a feature of the sample rather than being inherent in the underlying data. It is therefore deemed important to allow the user to question "is this a real pattern?". Consequently, we have provided an additional user control:

- 'reality check' – the user can click on this button to view a completely new sample within the lens, thus 'real' patterns will persist whilst sampling induced artifacts will disappear.

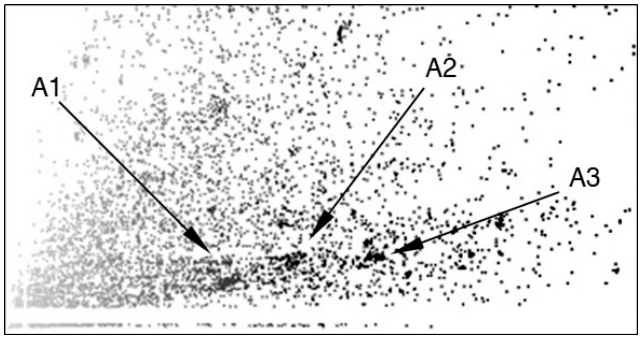
In addition to revealing previously hidden detail in a visualisation, the lens can also provide a visual indication of the proportion of different attribute values within it. For example, if the points in a scatter plot are coloured to represent a categorical attribute, say gender in a census dataset, then applying the lens to remove overplotting should result in a good approximation of the proportion of male and female in the region covered by the lens.

EXAMPLES

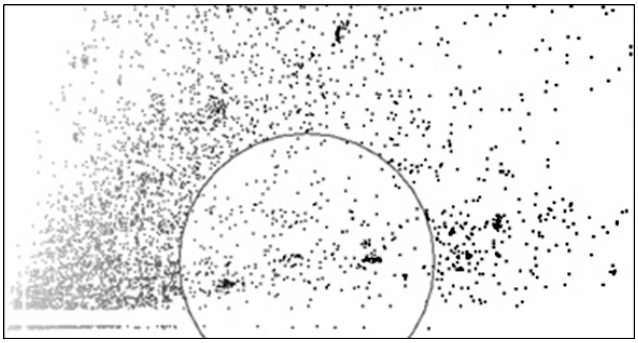
We will now demonstrate the general applicability of the sampling lens through some examples.

Clusters of Different Overall Densities

The sampling lens can be used to investigate saturated areas to compare clusters that look similar. Because of overplotting, it is possible that some aggregations with different density appear the same, even when they are not.



(a) without lens



(b) lens over clusters

Figure 1. Parcel data scatter plot example

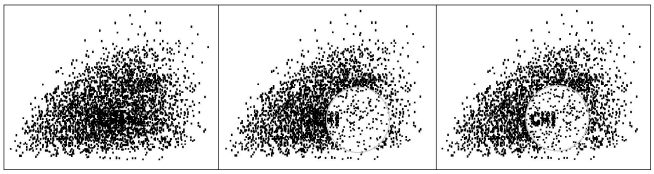
Figure 1 represents a visualisation of mail parcel data from the German post-office plotted according to weight and volume. Some overall structure of the data is clear: most of the points lie within a 60 degree arc and there are more small light parcels than heavier large ones. However, if we look at the dense areas (clusters A1, A2, A3 in figure 1a) it becomes harder to distinguish details. Are these areas equally dense or are there differences between the clusters? It is hard to tell because the areas are saturated.

But by applying the Sampling Lens over the area of interest (figure 1b), we can now see that A2 has disappeared. This implies that clusters A1 and A3 must be of higher density,

hence more significant, representing the most common combination of goods and boxes.

Clusters With Hidden Pattern

Figure 2 has been generated from a synthetic data set as another example to clarify the basic idea behind the Sampling Lens. No trends are apparent from the visualisation in figure 2a. But when we move the lens across from right to left (figure 2b and 2c), the hidden pattern becomes evident.



(a) (b) (c)

Figure 2 Revealing hidden pattern

It is worth noting that the idea of using a synthetic dataset is not new. In fact, our example resembles the famous Pollen dataset in which the word EUREKA is hidden and which is frequently used as a benchmark for visualisation tools.

Density Inspection on Parallel Coordinates.

Although parallel coordinates are effective in showing correlations among multiple variables, they do suffer strongly from saturation. In particular, trends that are typically conveyed by the direction of lines are difficult to perceive, as shown in the dense far left and right areas of figure 3a.

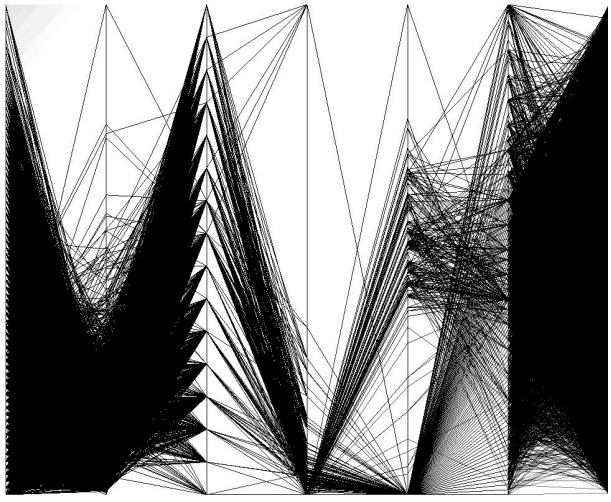


Figure 3a. Parallel coordinate example without lens

The Sampling Lens however allows these dense areas to be explored quickly to see if there are any interesting trends (figure 3b).

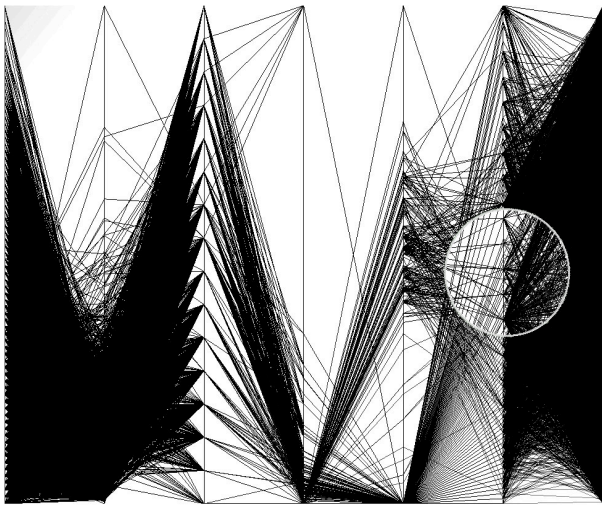


Figure 3b. Parallel coordinate example with lens

IMPLEMENTATION

The Sampling Lens has been implemented on top of the InfoVis Toolkit[1]. This has allowed us to focus on the sampling aspects of the lens and is acting as a layer on top of existing implementations of various visualisations. The toolkit employs a standard pipeline architecture (see fig 4a) with raw data being transformed into an abstract visualisation data-structure (shape descriptions, x-y coordinates etc.), which is then rendered onto a bitmap and hence appears on the user's display.

We have attempted to interfere as little as possible into the heart of this pipeline because this is where different types of visualisation are 'plugged in'. By keeping alterations to the beginning and end, we aim to make the Sampling Lens operate generically over a range of visualisations.

We have used two types of adaptations, which work better for different kinds of lens and visualisation. These adaptations would also work for other lenses, such as those with different semantic filtering criteria.

The first adaptation takes the lens coordinates and uses the abstract visualisation structure to determine which data points are rendered within the lens (fig 4b). This is then used to alter the filtering criteria (by decreasing sampling rate within the lens) and hence change the visualisation. This method is particularly good for point visualisations where there is an unambiguous meaning for which data items are within the lens. It also allows non-binary alterations such as having a graded sampling within a penumbra of the lens.

The other adaptation is to have two pipelines from the raw data with different sampling levels (fig 4c). These are used to create two complete bitmap images and the lens simply draws a portion of one image over the other. This demands that the two visualisation pipes generate images that precisely overlay, but otherwise the two can be manipulated

[1] http://ivtk.sourceforge.net/

independently. For parallel coordinates and other line-based visualisations this can be useful when one wants to sample the visualisation precisely within the lens, rather than sampling over the whole visualisation based on the lines crossing the lens.

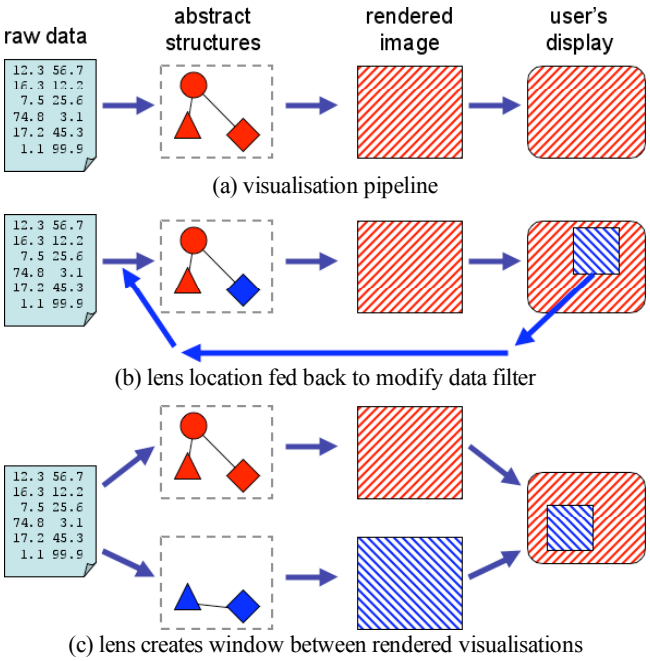


Figure 4. Implementation architectures for sampling lens

EVALUATION

Initial user studies have shown that in general, the concept of sampling and reality check is not as instinctive as we believed. Some users struggled to understand how sampling works and how reality check helps to discover 'fake' patterns. However, once users were guided through a set of examples, they found the Sampling Lens to be valuable and intuitive and started using it appropriately.

Although the display continuity feature (as discussed previously) was not pointed out to the users, some did in fact ask if sampled data items are returned in reverse order when one comes back to already visited sampling rates, a feature they found to be desirable.

The use of the Lens was perceived to be more valuable on dense parallel coordinates than scatter plots. Users discovered interesting trends in the saturated areas of parallel coordinates more easily, most probably due to patterns being more intrinsic in such representations.

Users quickly found that they could move the Lens around the screen easily and set the controls without much effort. Also, the lens metaphor was naturally accepted.

CONCLUSION AND FUTURE WORK

In this paper we have seen how the Sampling Lens allows the interactive exploration of saturated areas of visualised data. We have seen this applied to both point plot and parallel coordinates visualisation. Whilst there is only one obvious form of sampling for point data, this is more complex for parallel coordinate data and even more so for relationship data such as graphs. We intend to investigate appropriate sampling techniques for these richer forms of visualisation as well as refine and evaluate more formally the existing Sampling Lens. Furthermore, we plan to add more functions to the lens, such as display aggregated statistics.

REFERENCES

1. Artero, A O, Ferreira de Oliveira M C, and Levkowitz H. Uncovering Clusters in Crowded Parallel Coordinates Visualizations. Proc. Symposium on Information Visualization 2004, 131-136.
2. Bertini, E and Santucci, G. By chance is not enough: preserving relative density through non uniform sampling. Proc.IV'04, IEEE, 622-629
3. Bier, E A., Stone, M C., Pier, K., Buxton, W., De Rose, T D. Toolglass and magic lenses: the see-through interface. Proc. Computer graphics and interactive techniques 1993, 73-80
4. Dix, A and Ellis, G P. By chance: enhancing interaction with large data sets through statistical sampling. Proc. AVI'02, ACM Press, 167-176
5. Ellis, G P and Dix, A. Density control through random sampling : an architectural perspective. Proc IV'02, IEEE, 82-90
6. Fekete, J and Plaisant, C. Interactive Information Visualization of a Million Items. Proc. InfoVis'02, IEEE, 117
7. Keim, D A., Hao, M C., Dayal, U., Hsu, M. Pixel Bar Charts: A Visualization Technique for Very Large Multi-Attribute Data Sets. Information Visualization Journal, Palgrave, Vol. 1, No. 2, 2002
8. Keim, D A., Panse, C., Schneidewind, J., Sips, M. Geo-Spatial Data Viewer: From Familiar Land-covering to Arbitrary Distorted Geo-Spatial Quadtree Maps, WSCG 2004
9. Trutschl, M., Grinstein, G., Cvek, U. Intelligently Resolving Point Occlusion. Proc. Symposium on Information Visualization 2003, 131-136
10. Waldeck, C. and Balfanz, D. Mobile Liquid 2D Scatter Space. Proc. IV'04, IEEE, 494-498
11. Wilkinson, L., Rubin, M., Rope, D. and Norton A. nViZn: An Algebra-Based Visualization System. International Symposium on Smart Graphics 2001
12. Woodruff, A., Landay, J., Stonebraker, M. Constant density visualizations of non-uniform distributions of data. Proc. UIST 98. ACM Press, 19-2