Plot Balalaika:
Simple Chart Designs for Long-Tail Distributed Data

Mark M. Shovman
Eyeway Vision
Tel Aviv, Israel
mark.shovman@eyeway-vision.com

Ran Wolff
Yahoo Research
Haifa, Israel
ran.wolff@yahoo-inc.com

***Abstract*—Current approaches to summarising large arrays of data for presentation and communication mostly comprise reporting means with, e.g., bar-charts. These methods are well-suited for unimodal, ideally normally- or near-normally distributed data, but are misleading for long-tail distributions that comprise most of the Big Data. We propose a succinct visualisation format, parallel in simplicity to bar-charts, that is suitable for communicating the gist of long-tail distributions, and show its efficiency empirically.**

***Keywords*-data visualization; human computer interaction;**

I. Introduction

The aim of explanatory (as contrasted to exploratory) visualisation is to efficiently communicate evidence, in order to enable fast, yet accurate, judgement, sense- and decision-making. Often that means summarising large data sets in a succinct form – typically, by presenting the central tendency. The central tendency is usually approximated with arithmetic mean, with other metrics, such as median, tri-mean or mid-hinge used more rarely. The central tendency is used in a variety of use-cases: to compare several datasets in a bar- or pie-chart; to see changes in data over time on a time-line; *etc.*

When a more complete, yet still succinct, description is required, a measure of spread is added to a central tendency, such as the standard deviation, standard error, or an inter-quartile range. The spread is then visualised as error-bars, confidence intervals, or used in compound visualisations such as a box-and-whiskers plot [1].

More elaborate visualisations show distributions as they are – as histograms, PDF, CDF or CCDF plots, and the relatively recent violin plots [2]. These, however, are difficult to communicate to laymen, and, perhaps more importantly, hold too much information for quick analysis – even by experts.

A major underlying assumption for using a central tendency is that the dataset does actually have one. In other words, it is assumed that data is drawn from a distribution that has a single mode, and tapers out around that mode – in short, that it is more-or-less bell-curve-shaped, ideally – normally distributed. It then follows that: that the majority of the measurements do not fall far from the central tendency; that the central tendency is the most common value; that the range of the data is symmetrical around it, and probably does not spread too far from it *etc.*. It does not matter which measure of a central tendency is used because in a normal distribution they are all equal. Even if the data does not exactly fit a normal distribution, all these assumptions hold to a large extent.

Recently, however, more and more real data sets are analysed that do not satisfy the assumption of normality at all. Most of the data regarding social networks, natural language, online behaviour – most of the so-called Big Data – are not normally distributed. Instead, these data have highly skewed distribution that are truncated on one side and taper to a long heavy tail to the other side. The underlying theoretical distributions for many of these are either truncated power-law [3] or log-normal[1] [4].

As an example, we can compare a normally distributed variable: human height, and a power-law distributed variable: town size. The factoid that an average height of a US male is 180cm is useful, because it represents a central tendency of a normally-distributed data. In contrast, the factoid that an average population of a US town is 8.2k is useless, because it describes a power-law distributed dataset that has no central tendency.

Big Data are increasingly used for sense- and decision-making, often by summarisation and visualisation of the arithmetic mean. A common suggestion to use median or tri-mean instead of the mean [5] does not address the heart of the matter – that these data can not be summarised by any measure of central tendency, simply because they do not have one. Thus, while the value of arithmetic mean or tri-mean *etc.* can be calculated, it would not have much meaning.

In this study, we aim instead to develop a visualisation format, similar in simplicity to common bar-charts, that would be specifically suited for long-tail distributed data. Our requirements are:

- ease of calculation and rendering, especially for large

[1] It is often non-trivial to distinguish between power-law and log-normally distributed data, especially from a noisy and/or low-resolution sample.

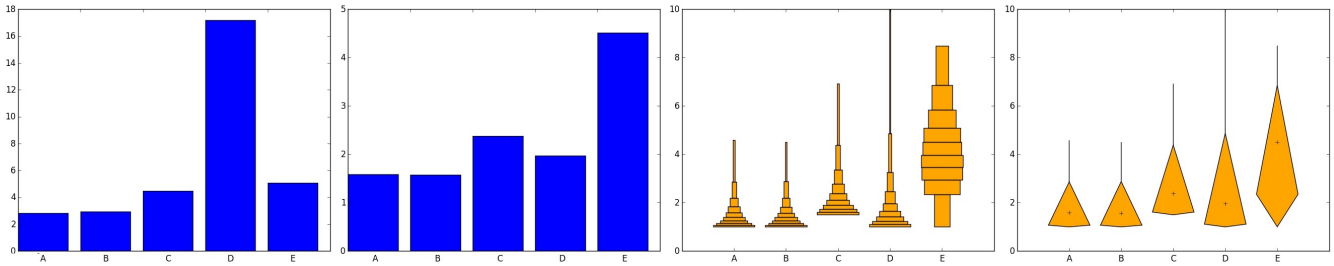

Figure 1. Different summary visualisations of the same five datasets (10k items each). Visualisations (left to right): mean; median; stackchart; and balalaika plot. Datasets (left to right): A and B: two samples from a power-law distribution with offset 1 and slope 2.5; C: offset 1.5; D: slope 2; E: log-normal distribution with mean 1.5 and standard deviation 0.5.

datasets;

- ease of comprehension in common use-cases such as outlier detection, comparison, and trend estimation;
- suitability to a large family of general long-tail distributions, and tolerance to noisy data.

Notably, comprehensive representation of the dataset is not a requirement; instead, we aim for summarising just the gist of it – just as a bar-chart of means represents the gist of normally-distributed datasets.

II. Choice of Metric

The first step of designing a visualisation is to choose a suitable succinct numerical representation of a dataset.

Just as normal distributions can be fully described by two numbers - mean and standard deviation, power-law distributions can be described by slope and offset of the linear probability density function on a log-log scale (the log-log linearity of the PDF is the definition of a power-law distribution).

The offset is simply the lower bound of power-law behaviour: the beginning of the long tail. The meaning of slope is difficult to communicate to non-statisticians; however, its reciprocal, ranging from 0 to 1, has a clear meaning of 'tail-heaviness' – the closer it is to 1 the more values would be found in the long tail. Commonly found power-law distributions have a slope of 2 to 3, and therefore a tail-heaviness of 0.3 to 0.5.

There are, however, problems with these metrics that are not as easy to address. First, the fit is complex to calculate, especially for high offsets; and second, the slope is only really relevant to pure power-law distributions and is misleading for long-tail distributions that are not linear on a log-log scale. That makes summarising a long-tail distribution by its 'tail-heaviness' useful only in the limited situations of *a priori* known power-law distribution.

Likewise, log-normal distributions can be fully characterised by a geometric mean, also know as log-average: $e^{\frac{1}{n}\Sigma_i^n lnx_i}$, that represents the central tendency of the underlying normal distribution. While it is easy to calculate, it is only really meaningful for purely, or nearly so, log-normal distributions.

One relatively widely-known metric of long-tail distribution is the Gini index, or the index of inequality. It is general, not assuming any specific distribution, but it is very computationally heavy to calculate for large samples. It is also, essentially, a secondary statistic, based on percentages, that disregards the actual values of the data.

With all this in mind, we have decided to use deciles as our metric. Deciles, especially the median (5^{th} decile) and the Pareto point (8^{th} decile) are already widely used in reporting long-tail distributions. They are easy to compute, and do not assume any specific distribution shape.

III. Visualisation Design

The design process of a new visualisation is by its nature less objective and harder to formalise than a choice of an appropriate metric. In this section we will attempt to outline and in some measure justify the design choices made, given the multitude of options.

Stackplot (fig. 1C) was our fist attempt at representing long-tail data using deciles. We followed the idea of a violin plot as a starting point, aiming to show a rough outline of data distribution. A stackplot is a stack of rectangles, with each rectangle representing a decile of the data. Thus the height and positioning of i^{th} rectangle is defined by deciles i and $i+1$; the width is scaled so that all rectangles in a stack are of equal area, with the widest ones in every stack being of the same width. Importantly, the last decile, the longest tail, is not shown: if it were, its height would obscure the rest of the plot without contributing any meaningful information except for the position of the largest datum in the long tail - which is spurious.

While useful, and much easier to calculate and render than a violin plot, a stackplot still has too much detail compared to, for instance, bar-chart. Therefore, our second design, the balalaika chart (fig. 1D), focused on the deciles that best describe the long-tail distribution: 0^{th} (minimal value), 1^{st} (widest point), 5^{th} (median, shown as a plus sign in the middle), and lastly 8^{th} and 9^{th} (the tip of the body and the tail). The exact width calculations were also discarded in favour of simplicity.

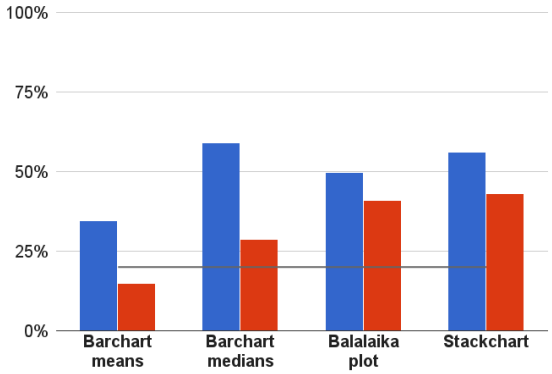

Figure 2. Accuracy of outlier detection in various chart types. Left/blue bars: offset outlier; Right/red bars: slope outlier; chance level at 20%.

IV. Experiment

To validate the design, we have run a short experiment using Amazon Mechanical Turk (MTurk) as the platform. Four visualisation formats were compared in terms of outlier detection accuracy: whether they allowed to reliably detect a sample drawn from a similar, but different distribution. Two visualisation formats were well-established: barcharts of either mean or median; two were new: the stackchart and the balalaika plot. Only pure power-law distributions were used for the sake of simplicity; experiments using other log-normal and real-life distributions are also planned.

A. Method

1) Stimuli: For each stimulus, four samples of 10k values were randomly generated. Three were drawn from a baseline power-law distribution, and one from either baseline (in 20% of the cases) or a differing one.

The baseline distribution started at 1 and had a slope of $1/0.4$. The outlier distribution differed in either slope: $1/(0.4 \pm 0.005)$ or start: 1 ± 0.1. These values were determined in a pilot study to be discernible but not trivially so. A rigorous study of a JND in either parameter is planned, but was out of scope of this study.

A combination of four outlier types (two slope and two start), and five outlier target positionings (A, B, C, D or none) produced 20 datasets. Each was visualised using four chart types, resulting in 80 stimuli (see fig. 1 for sample stimuli).

2) Procedure: Participants could rate any number of stimuli, but could only rate each stimulus once. Before seeing the stimulus, they were shown the following instruction: "You will see a barchart, or something similar to a barchart, with four items labelled A, B, C and D. Your task is to say which one of the items is different from the rest. If there are none or more than one, select the 'None' option. Should take 1 second." After that the stimulus image was presented with the above options as radio buttons. Response time was recorded but not restricted.

B. Results

Each stimulus was seen by 30 MTurk workers. The accuracy results (Figure 2) show near-chance performance for barcharts of means. Barcharts of medians perform well for offset outliers, but are near-chance for slope outliers. In contrast, both novel designs perform similarly well for either type of target outlier. Response time varied from 0.7 seconds to more than two hours, and was not analysed further.

V. Discussion

While definitely not optimal, both our designs outperform commonly used barcharts of either means or medians. They, especially the balalaika plot, are easy to calculate, render, and, most importantly, comprehend.

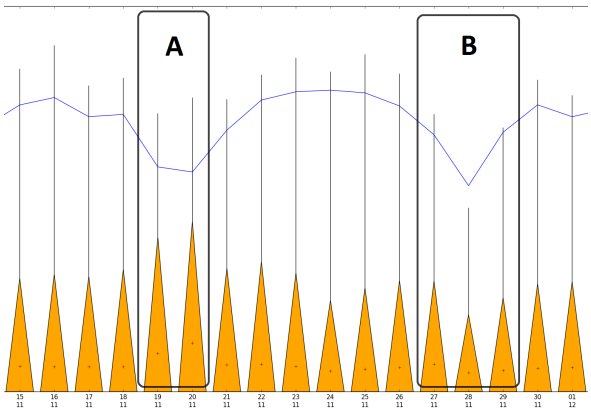

Figure 3. Balalaika plot of a real long-tail distributed data. Blue line at mean.

For example, fig. 3 shows a real-life sample of noisy, long-tail distributed data: daily page view times in one of Yahoo properties. The property attracts millions of daily views; of these, view times of 100k pages per day were randomly sampled for 10 days and visualised using both balalaika chart and arithmetic mean. It is clear from the balalaika plot that the two dips in mean, highlighted as A and B, are in fact due to different phenomena: A is a decrease in long tail extreme outliers, while B is a decrease in body.

To conclude, we argue that, while better design and more studies are of course necessary, these visualisation formats can already be used in real-life analysis of long-tail distributions.

Acknowledgement

This research was done at Yahoo Labs Haifa.

References

[1] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.

[2] J. L. Hintze and R. D. Nelson, “Violin plots: a box plot-density trace synergism,” *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.

[3] A. Clauset, C. R. Shalizi, and M. E. Newman, “Power-law distributions in empirical data,” *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[4] J. Aitchison and J. A. Brown, “The lognormal distribution with special reference to its uses in economics,” 1957.

[5] J. W. Tukey, *Exploratory data analysis*. Reading, Ma, 1977, vol. 231.