# Visualization as an Analysis Tool:

## Presentation Supplement

This document is a supplement to the presentation "Visualization as an Analysis Tool" given by Phil Groce and Jeff Janies on January 9, 2008 as part of FloCon 2008. The intent of the presentation was to demonstrate how simple tools used together can provide significant insight into network behavior.

This supplement includes annotations to the presentation slides and a set of key points in the presentation, as well as supplementary points that were not included for the sake of brevity. Except where noted, all examples are drawn from real network flow data.

## Key Points

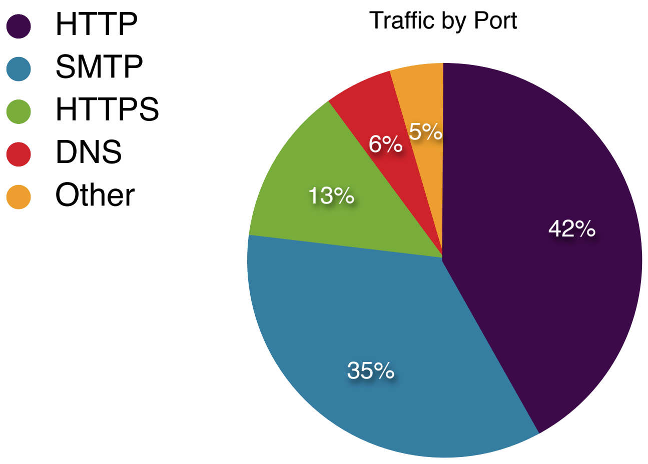
**Always tell the truth.** Many of the same techniques for making the eye aware of important distinctions in the data can also highlight unimportant distractions, or even create false impressions that the data do not support. Always ensure that the visualization presents an accurate picture of the data.

**Learn how to use your tools.** It's better to have a limited set of tools you know how to use than a whole bag of tools you don't understand. If nothing else, understand the limitations of the tools so you know when your question demands a new one.

**Facilitate direct comparison.** Use human perception to your advantage. Put comparable visualizations on the same page so people can see them in peripheral vision and switch back and forth quickly. Align visualizations along common axes. Make sure similar things look similar (e.g., by using the same scales).

**Combine complementary visualization techniques.** Use visualizations whose insights complement each other in ways that facilitate easy comparison between them.

**Tables are visualizations, too.** For small sets of data, a table may be the best way for people to consume the data. Compare the following (sample) data formatted as a pie chart and a table:



| | **Traffic** |
|---|---:|
| **Port** | **(% Volume)** |
| HTTP | 42 |
| SMTP | 35 |
| HTTPS | 13 |
| DNS | 6 |
| Other | 5 |

The table takes less space to communicate the same information, and it is easier to map the type of traffic to the value. In the pie chart, the reader must consult a key to map a color to a traffic type, then find the color on the chart. (Putting the type label directly on the pie slice causes problems with the smaller pieces; using callout-style labels makes the pie chart even larger.)

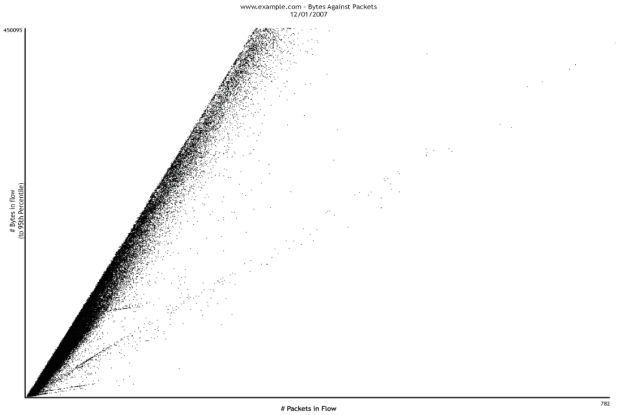
**Distinguish between useful and useless precision.** The right numbers in the right place are critical to understanding a visualization. Examples include maximum and minimum observed values; selected local maxima and minima; and relevant baselines such as the start and end values of the scale, medians or other useful "average" values, and specific "important" data points (e.g., a point representing an important host or a flow thought to be the source of a compromise).

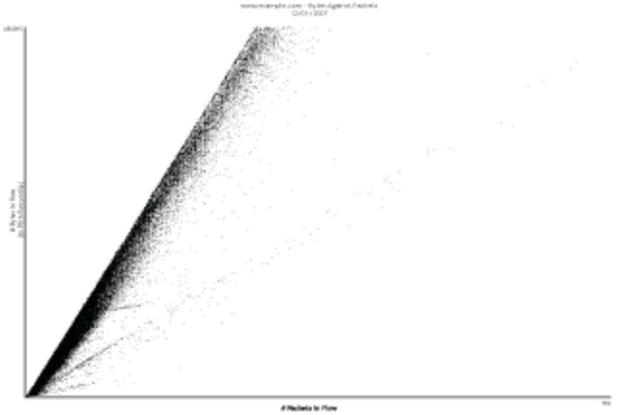
In general, if the exact numbers don't provide an important perspective on the whole visualization, it's probably a distraction. If all the numbers are important, the best visualization may be a table.

**Consider sampling.** For constant-magnitude data, a visualization over a sample of the data may tell the story as convincingly as a visualization over the full set of data. If the generation of the visualization takes significant time, sampling may improve performance.

**Choose your display media wisely; don't underestimate paper.** The primary display media available to most analysts are paper and computer monitors. In our experience, analysts overwhelmingly use monitors over paper. Both have advantages, however. Computer screens allow types of interactivity that paper cannot, and digital copies of visualizations can be easily sent electronically. Visualizations on paper can be read without special equipment, and annotated with only a pencil. Moreover, paper resolution ranges from 300 to 1000 dpi. Screen resolution typically ranges from 72-100 dpi.

For comparison, here is the same scatterplot rendered at 300 dpi and 100 dpi:





**Use the appropriate number of dimensions.** Paper and screens are naturally two-dimensional; color, perspective and motion can provide some additional dimensionality, but will never be as effective as length and width on a flat screen or piece of paper. For example, decorating a scatterplot axis with a histogram of the data along that axis communicates an additional dimension of data density often plotted to lesser effect with color or an isomorphic rendering of a three-dimensional plot.

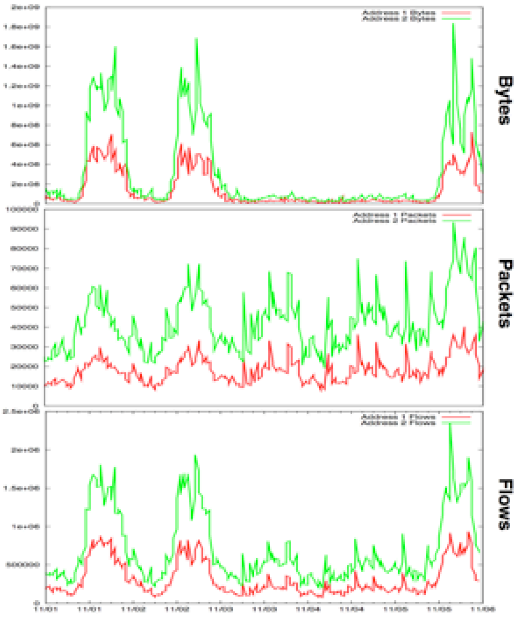
**Make everything explicit.** When generating visualizations for personal consumption, the most important thing is finding an insightful view on the data. When passing visualizations to others (including yourself in the future), annotate the visualization with everything required to understand where the data came from and what processing has been done on it. (E.g., the command used to extract the data, how the data may have been trimmed for readability, what other transformations have been done on the data.) In particular, if your scale doesn't start at zero for linear scale plots or one for log scale plots, make a special point of noting it.

# Basic Visualization Types

## Time series

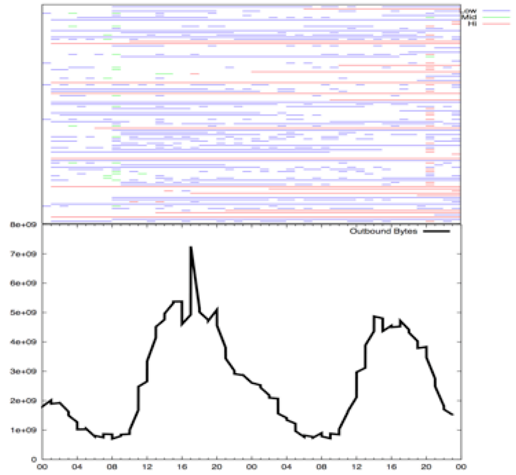
Relating data to time is so intuitive and useful that it risks being used to the exclusion of other visualization tools. It makes sense, then, to optimize time series visualizations as much as possible.

A common problem in time series data is relating multiple independent series (e.g., volume measurements of bytes, packets and flows) by time to give a clear picture of an event. Often, the scales of each series diverge too widely to plot on the same scale without losing detail. Using different scales for each series is an option if the designer can communicate this decision clearly to the reader. Another (often simpler) solution is to plot the series independently and align them on a shared scale, as in the figure to the right.



When comparing very large numbers of series, the above approach breaks down, unless the data is very tightly coupled (e.g., EKG or seismic data). The existence plot trades measurement resolution for scale by defining value ranges and plotting each series as a single line, colored by the range in which the value for that time resides.

To mitigate the loss of resolution, it is often useful to pair an existence plot with a traditional time series that relates to all the series in the existence plot.

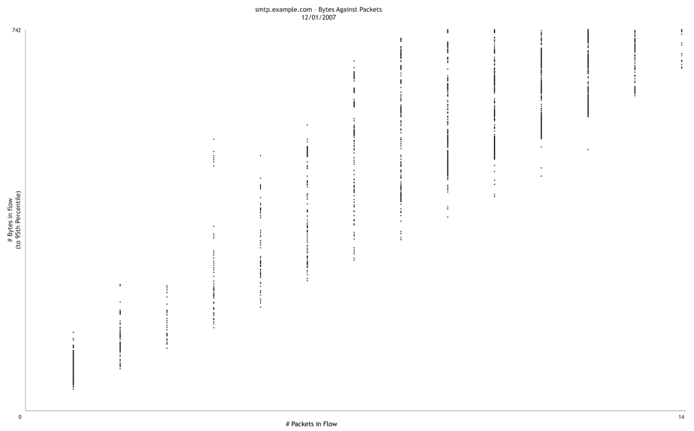


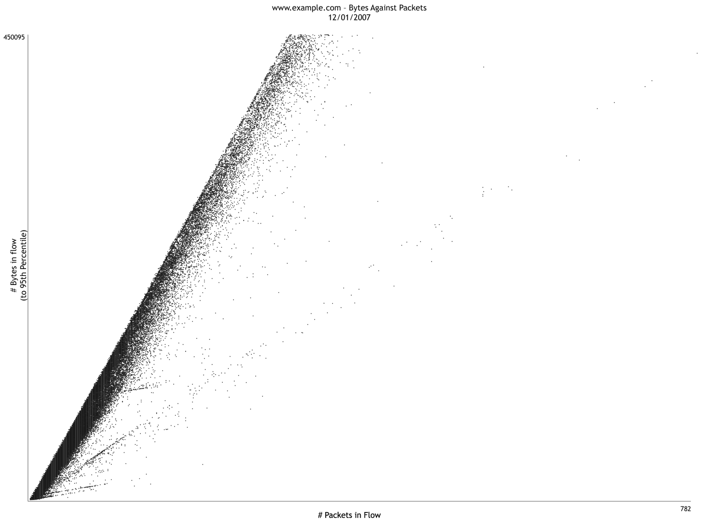
## Visualizing Relationships

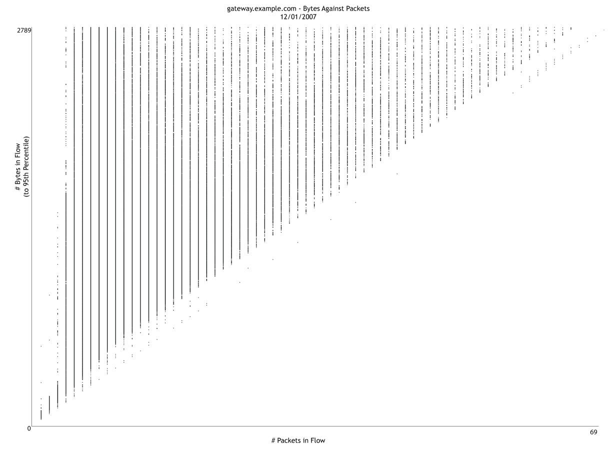
Time series plots are a specific instance of relating one dimension with another (time). Scatterplots are generalizations of this approach. By plotting points in space, the eye can perceive relationships between these dimensions as lines, shapes or other patterns.

These three machines serve three very different network roles. (Mail server, web server and gateway, respectively.) This is reflected in the different (but consistent) ratios between bytes per flow and packets per flow in their traffic. This, in turn, is visible in the very distinctive "shapes" their traffic takes when these dimensions are plotted against each other.

In this plot, multiple points that share exactly the same values show up as a single point. The next section addresses this deficiency.



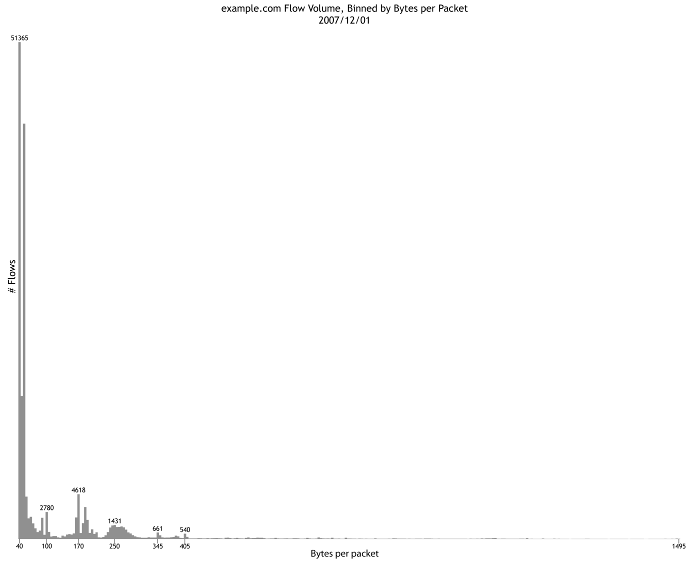




### Visualizing Distributions

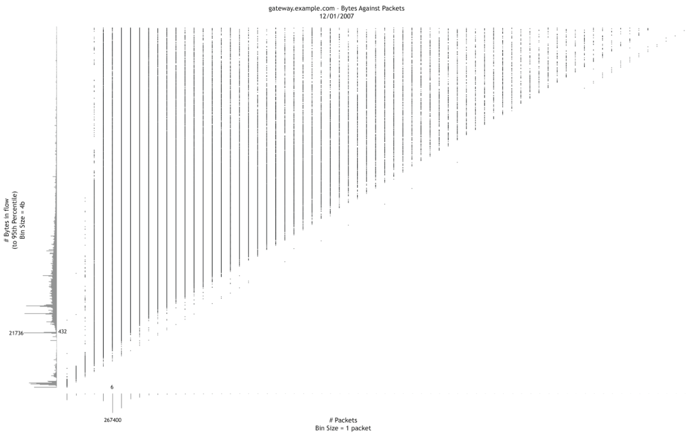
There are many ways to visually analyze the distribution of sets of single values—this document focuses on histograms, but box plots, whisker plots and violin plots; CDF (cumulative distribution function) plots are all in common use.

The distribution of bytes per packet for flows associated with a given host. Most connections fall within a few narrow ranges of values. As with the scatterplots above, the "shape" of the histogram is characteristic of the host behavior.



Because of the physical dimensions of a histogram, and the fact that it operates on a single (data) dimension, it complements scatterplots well to indicate density of data.

In the example at right, the distribution of values gives no indication that a disproportionate number of flows visualized have low packet and byte counts. (267,400 flows have 6 packets; 21,700 flows contain only 432 bytes.)



## Additional Reading

The R Project. *Introduction to R.* Chapter 13: Graphics. http://cran.r-project.org/doc/manuals/R-intro.html#Graphics

Tufte, E. R. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 1983.

Tufte, E. R. *Envisioning Information.* Cheshire, CT: Graphics Press, 1990.

Tufte, E. R. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press, 1997.

Tufte, E. R. *Beautiful Evidence.* Cheshire, CT: Graphics Press, 2006.

Wilkinson, L., et al. *The Grammar of Graphics*. New York: Springer-Verlag, 1999.