Icicle Parallel Coordinates Plots for visualizing hierarchical data

Hugh Garner

Newcastle University

High dimensional hierarchical datasets are challenging to explore, particularly those collected for many different samples (for example individual network traffic sources). Datasets of data center traffic metrics that contain an explicit hierarchical structure, with each sample including values for centre, cluster, rack, and workload may require analysis to understand the dynamics of traffic flow, aiding discovery of potential bottlenecks or bugs. Visualization methods can augment early stage exploratory analysis and provide the user with an overview of data features, anomalies and patterns. However, early stage visualizations typically focus on aggregate measures or use drill-down mechanisms, limiting initial visual comprehension throughout the hierarchy. Following Schneiderman's information seeking mantra, ‘overview first’, ‘zoom and filter’, ‘details-on-demand’, a more complete overview including aggregate measures and individual sample dynamics would provide for a more complete picture of datasets, leading to a greater likelihood of serendipitous discovery, aid aggregation and analysis choices and reduce the risk of information loss.

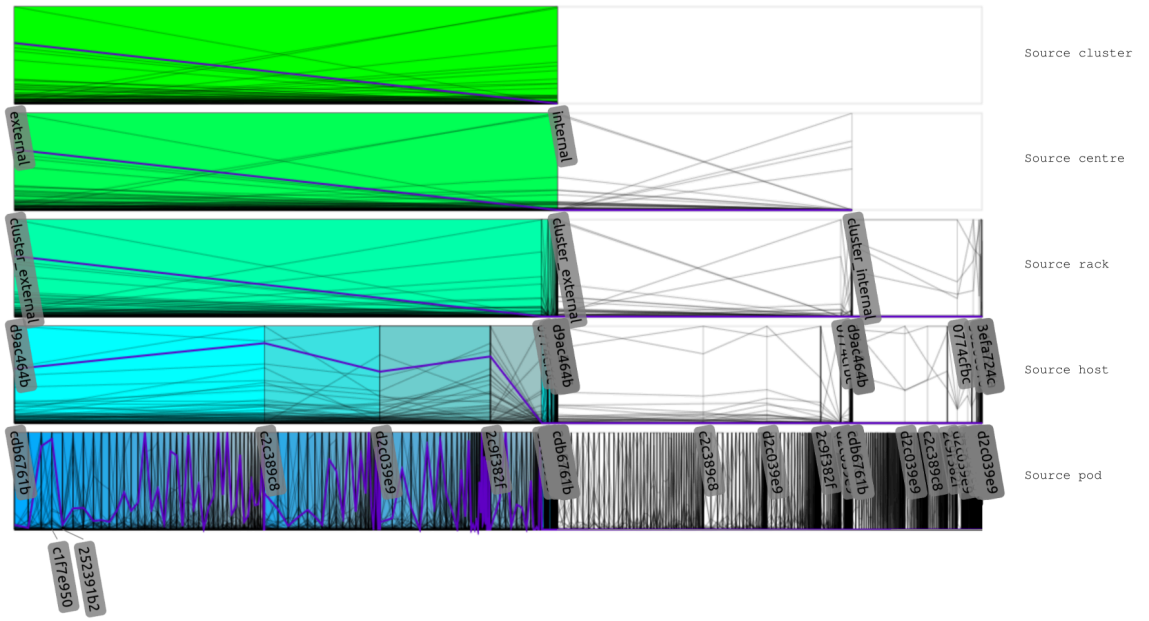
Icicle plots are commonly used for visualizing hierarchical data, either for single samples or aggregate measures. The width of any individual element is proportional to its value, with child elements represented in the next vertical level, similarly scaled. However, the icicle plot either only enables visualization of a single sample (for example, traffic flow destination from a single rack), or aggregates multiple samples (for example, aggregated traffic flow destination from multiple racks).

Parallel coordinates plots are well suited representing relationships in multivariate datasets, with single parallel axes for each variable, linked by a polyline for each sample. Common issues around axis ordering and the limitations of screen space provide obstacles in viewing very high-dimensional datasets. Frequently, axes are filtered to select the most relevant or interesting measures, requiring extensive domain and problem knowledge to ensure appropriate relevancy criteria.

To address the challenge of enabling a viewer to see the individual sample measures for each aggregated category, we combine the icicle plot with a PCP - the Icicle Parallel Coordinates Plot (IPCP). An icicle plot represents aggregate values, and the edges of the icicle blocks are used as PCP axes, with the PCP enabling representation of individual sample measures (for example traffic from a single source).

Using the icicle width for x-separation of PCP axes presents a solution for the axis ordering problem, with higher levels of the icicle measure more clearly represented with a large separation, while the lower levels are reduced in prominence.

Originally developed for visualizing microbiome data, the IPCP has now been applied to explore data centre network traffic, suggesting potential relevance across a wider field (for example OpenTelemetry data). While the visualization can be easily adapted for many datasets, understanding useful interactions, features and evaluation of the potential for novel insights requires individual domain and problem expertise. In particular, I hope to elicit feedback on the existing application to data centre traffic datasets, and gain insight into the challenges across the systems research community that may be addressed with the IPCP or similar overview exploratory visualization methods.



Example IPCP for datacentre traffic source by destination pod.