Data Visualization (CSCI 627/490)

Filtering & Aggregation

Dr. David Koop
Composite Visualization Techniques

(a) Juxtaposed views. (b) Integrated views. (c) Superimposed views.

(d) Overloaded views. (e) Nested views.

Table 1: Classification of common composite visualization techniques using our design space.

- (a) Juxtaposed views: Easy correlation, limited space for client visualizations, compact representation.
- (b) Integrated views: Efficient approach to link each visualization, high visual design dependency.
- (c) Superimposed views: Any any juxtapose, none.
- (d) Overloaded views: Glyphs nested item-group, treemap node-link overload item-item, node-link matrix nested item-group.
- (e) Nested views: Scatterplot glyphs nested item-group, map text superimpose item-item, time line view area visualization juxtapose item-item, node-link matrix nested item-group.

Drawbacks:
- Very compact representation, easy correlation.
- Limited space for the client visualizations, clutter.

Applications:
- High rate what is a composite visualization and what is an "atomic" (or part) visualization.
- Limited space for the client visualizations, clutter.

Figure 12: Example of composing a scatterplot and bar graph using different methods.

ONCLUSION

We have proposed a novel framework for specifying, designing, and evaluating compositions of multiple visualizations in the same view. The benefit of the framework is not only to provide a way to unify a large collection of existing work where visual representations are combined, but also in evaluating their strengths and weaknesses.

However, the design patterns presented in this paper are all based on different conditions, such as the available view space, user perspectives, etc. It is also not always straightforward to separate what is a composite visualization and what is an "atomic" visualization. It is also not always straightforward to separate what is a composite visualization and what is an "atomic" visualization. It is also not always straightforward to separate what is a composite visualization and what is an "atomic" visualization.

REFERENCES

What is this technique?
What is this technique?

[NodeTrix, N. Henry et al., 2007]
## Multiple Views

<table>
<thead>
<tr>
<th>Encoding</th>
<th>Data</th>
<th>All</th>
<th>Subset</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>Redundant</td>
<td><img src="image1.png" alt="Graph" /></td>
<td>Overview/Detail</td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td>Different</td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
<td><img src="image5.png" alt="Graph" /></td>
<td>No Linkage</td>
</tr>
</tbody>
</table>

[Harder (ill. Maguire), 2014]
Brushing
Fig. 2: The Cerebral display of the TLR4 graph (V=91, E=124) with associated LPS and LPS+LL-37 time series. The small multiples show an overview of all 8 experimental conditions. The most noticeable differences between the LPS and the LPS+LL-37 condition occur at hour 4. By selecting the hour 4 conditions, the main window shows the computed difference between the two conditions.

Furthermore, the biologists’ assessment of what constitutes a good layout varies depending on the nature of the biomolecules involved. In the undirected portion of the graph, which comprises protein-protein interactions that propagate a signal from membrane to nucleus, they wish to see the network structure so that they can follow the signaling cascade. Thus for this section of the graph, it is important to minimize edge crossings, even if it places interacting nodes somewhat far apart.

In contrast, for the directed portion of the graph, representing the genes whose expression was altered in response to the signaling cascade, the biologists want to see the nodes grouped tightly by function, even at the expense of not being able to clearly see the interactions between them. Translating these desires into automated graph layout requires an algorithm that uses metadata associated with the nodes, in addition to the direct graph structure, for node placement. Positioning nodes according to biological meta-data defines a semantic substrate [34].

3.2 Small multiple views for multiple conditions

Cerebral uses small multiples [38] to simultaneously display multiple experimental datasets. Each small multiple contains a complete copy of the interaction graph with the same spatial layout, but with different coloring according to the experimental data it is displaying. Our design target was to handle from two to a few dozen gene expression conditions, and from 50 to 3000 nodes in the interaction graph.

One obvious alternative to multiple small views would be a single changeable or animated view, where the color coding changes over time rather than being distributed over space [33, 32]. Comparing something visible with memories of what was seen before is more difficult than comparing things simultaneously visible side by side [31]. Thus, the limitations of human memory make comparing the few dozen conditions of our design goal through animation quite difficult [40]. Although small multiples would not scale to hundreds of conditions, they handle the current usage of 8-10 easily and will certainly accommodate the projected usage of few dozen conditions.

A second alternative is to embed a glyph, such as a line graph or heat map, near or within the node itself [24, 32, 41]. While embedded glyphs provide good detail when zoomed in for a local view, they become indistinguishable when zoomed out for a global view of graphs larger than a few dozen nodes. The biologists often need to see such a view, as it more readily allows for the identification of interacting genes/proteins whose expression behaves similarly across several conditions. Thus, glyphs would not be appropriate in this domain.

Saraiya et al. [32] evaluated four approaches to integrating graph and time series data, comparing one versus two views and slider-controlled animation versus embedded glyphs. While they used 10 time series data points, in a good match for our problem domain, their graph contained only 50 nodes. They found many tradeoffs between task type, speed, and accuracy. Our design can be considered an attempt to combine the strengths of the four different interfaces they studied into a single interface for a problem where the tasks are complex, accuracy outweighs raw speed, and the graph is large.

3.3 Parallel coordinates and clustering for data-driven exploration

Cerebral’s main views focus on the interaction graph model of the biological system or process of interest. We also provide a data-driven exploration to help users discover interesting relationships within the data. Parallel coordinates are a useful tool for visualizing multivariate data, allowing users to see how variables change together across different conditions. Clustering algorithms can be applied to group similar data points together, providing an overview of the data distribution and highlighting clusters of interest.

Multiform & Small Multiples

[Barsky et al., 2008]
Partitioned Views

- Split dataset into groups and visualize each group
- Extremes: one item per group, one group for all items
- Can be a hierarchy
  - Order: which splits are more "related"?
  - Which attributes are used to split? usually categorical
I page. In Figure 2 there are 6 panels, 1 column, 6 rows, and 1 page. Later, we will show a Trellis display with more than one page. We refer to the rectangular array as the trellis because it is reminiscent of a garden trelliswork.

Each panel of a trellis display shows a subset of the values of panel variables; these values are formed by conditioning on the values of conditioning variables. In Figure 1 the panel variables are variety and yield, and the conditioning variables are site and year. On each panel, values of yield and variety are displayed for one combination of year.

[Becker et al., 1996]
Recursive Subdivision: HiVE System

[Slingsby et al., 2009]
Project Design

• Feedback:
  - Data Manipulation?
  - Questions lead, not technique!
  - Be creative! (interaction too) https://xeno.graphics

• Work on turning your visualization ideas into designs

• Turn in:
  - Two Design Sketches (like sheets 2-4 from 5 Sheet Design)
  - One Bad Design Sketch (like sheets 2-4: here, justify why bad)
  - Progress on Implementation

• Due Friday, Nov. 13
Assignment 5

• To be released soon
• Citi Bike NYC Data
  - Trips between neighborhoods
• Covers
  - Multiple Views
  - Filtering
  - Aggregation
  - Brushing
Overview: Reducing Items & Attributes

Filter

Items

Attributes

Aggregate

Items

Attributes

[Munzner (ill. Maguire), 2014]
Filtering

- Just don't show certain elements
- Item filtering: most common, eliminate marks for filtered items
- Attribute filtering:
  - attributes often mapped to different channels
  - if mapped to same channel, allows many attributes (e.g. parallel coordinates, star plots), can filter
- How to specify which elements?
  - Pre-defined rules
  - User selection
Filter vs. Query

- Queries start with an empty set of items and **add** items
- Filters start with all items and **remove** items
Restaurant locations are derived from the New York City Department of Health and Mental Hygiene database. Due to the limitations of the Health Department’s database, some restaurants could not be placed.

By JEREMY WHITE

Source: New York City Department of Health and Mental Hygiene

© 2013 The New York Times Company

The New York City Department of Health and Mental Hygiene performs unannounced sanitary inspections of every restaurant at least once per year. Violation points result in a letter grade, which can be explored in the map below, along with violation descriptions. The information on this map will be updated every two weeks. For menus and reviews by New York Times critics, visit our restaurants guide.

Example: NYC Health Dept. Restaurant Ratings

Gracie’s Cafe
Grade pending
Violation points 27
Click for details

Gracie’s Cafe
Grade pending
Violation points 27
Click for details

Chicken Indian Pizza
Improper chemicals 14+ points

FIND A RESTAURANT | FIND A LOCATION | FILTER

Search All NYTimes.com
FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE

J. White, New York Times
Dynamic Filters

• Interaction need not be with the visualization itself
• Users interact with widgets that control which items are shown
  - Sliders, Combo boxes, Text Fields
• Often tied to attribute values
• Examples:
  - All restaurants with an "A" Grade
  - All pizza places
  - All pizza places with an "A" Grade
Scented Widgets

For each task, subjects with one of three scenting methods were instructed to complete three tasks.

We gave them an introductory tutorial to the system, and then presented them with a total of 1096 visits and 172 comments on views. To test this hypothesis, we compared these visits with one of the conditions. We removed the starting overview from consideration, because it was broken down by gender, from 1850 to 2000. The current task hypothesis was that scented widgets would help us make unique discoveries using scented widgets, and would express a preference for scented widgets over traditional widgets. The study employed a 3 (Task) x 3 (Scent) between conditions design. Task and scent pairings and presentation order were counterbalanced, and so we scaled them logarithmically for display in the scented widgets.

Next, we analyzed the data to check if scented widgets help us make unique discoveries. Our hypotheses were that scented widgets would increase the likelihood that users would visit views that were visited in the seed data than users in the no scent condition. We found that the correlations are not very strong. We believe that the semantics of the tasks also affect the underlying activity measure used to seed the scented widgets. Using prior visits to a view, we used [Willett et al., 2007] to find that performance would improve over subsequent trials, regardless of conditions; make unique discoveries. Our hypotheses were that scented widgets would express a preference for scented widgets over traditional widgets. The study employed a 3 (Task) x 3 (Scent) between conditions design. Task and scent pairings and presentation order were counterbalanced. Task and scent pairings and presentation order were counterbalanced, and so we scaled them logarithmically for display in the scented widgets.

The results are shown in Figure 7, which exhibits a power law distribution, and so we scaled them logarithmically for display in the scented widgets.

[Willett et al., 2007]

D. Koop, CSCI 627/490, Fall 2020
Scented Widgets

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hue</td>
<td>Varies the hue of the widget (or of a visualization embedded in it)</td>
<td>Option A</td>
</tr>
<tr>
<td>Saturation</td>
<td>Varies the saturation of the widget (or of a visualization embedded in it)</td>
<td>Option A</td>
</tr>
<tr>
<td>Opacity</td>
<td>Varies the saturation of the widget (or of a visualization embedded in it)</td>
<td>Option A</td>
</tr>
<tr>
<td>Text</td>
<td>Inserts one or more small text figures into the widget</td>
<td>(2) Option A</td>
</tr>
<tr>
<td>Icon</td>
<td>Inserts one or more small icons into the widget</td>
<td>Option A</td>
</tr>
<tr>
<td>Bar Chart</td>
<td>Inserts one or more small bar chart visualizations into the widget</td>
<td>Option A</td>
</tr>
<tr>
<td>Line Chart</td>
<td>Inserts one or more small line chart visualizations into the widget</td>
<td>Option A</td>
</tr>
</tbody>
</table>

[Willett et al., 2007]
Star Plots (aka Radar Charts)

- Aberfeldy: Malt, Fruity, Nutty, Spicy, Honey, Smoky
- Aberlour: Floral, Body, Sweetness
- AnCnoc: Floral, Body, Sweetness
- Ardbeg: Floral, Body, Sweetness
- Ardmored: Malt, Fruity, Nutty, Spicy, Honey, Smoky
- ArranIsleOf: Floral, Body, Sweetness
- Auchentoshan: Floral, Body, Sweetness
- Auchroisk: Floral, Body, Sweetness

K. Schaul
Star Plot / Radar Chart

• Use:
  - Compare variables
  - Similarities/differences of items
  - Locate outliers

• Considerations:
  - Order of axes
  - Too many axes cause problems
Attribute Filtering on Star Plots

(a) (b)

(c) (d)

[Yang et al., 2003]
Attribute Filtering

• How to choose which attributes should be filtered?
  - User selection?
  - Statistics: similarity measures, attributes with low variance are not as interesting when comparing items

• Can be combined with item filtering
Aggregation
Aggregation

- Usually involves derived attributes
- Examples: mean, median, mode, min, max, count, sum
- Remember expressiveness principle: still want to avoid implying trends or similarities based on aggregation

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>10.0</td>
<td></td>
<td>y</td>
<td>8.04</td>
<td>10.0</td>
<td></td>
<td>9.14</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>8.0</td>
<td></td>
<td>y</td>
<td>6.95</td>
<td>8.0</td>
<td></td>
<td>8.14</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13.0</td>
<td></td>
<td>7.58</td>
<td></td>
<td>13.0</td>
<td></td>
<td>8.74</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9.0</td>
<td></td>
<td>8.81</td>
<td></td>
<td>9.0</td>
<td></td>
<td>8.77</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>11.0</td>
<td></td>
<td>8.33</td>
<td></td>
<td>11.0</td>
<td></td>
<td>9.26</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>14.0</td>
<td></td>
<td>9.96</td>
<td></td>
<td>14.0</td>
<td></td>
<td>8.10</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6.0</td>
<td></td>
<td>7.24</td>
<td></td>
<td>6.0</td>
<td></td>
<td>6.13</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td></td>
<td>4.26</td>
<td></td>
<td>4.0</td>
<td></td>
<td>3.10</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>12.0</td>
<td></td>
<td>10.84</td>
<td></td>
<td>12.0</td>
<td></td>
<td>9.13</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7.0</td>
<td></td>
<td>4.82</td>
<td></td>
<td>7.0</td>
<td></td>
<td>7.26</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.0</td>
<td></td>
<td>5.68</td>
<td></td>
<td>5.0</td>
<td></td>
<td>4.74</td>
<td></td>
</tr>
</tbody>
</table>
Aggregation

- Usually involves **derived** attributes
- Examples: mean, median, mode, min, max, count, sum
- Remember expressiveness principle: still want to avoid implying trends or similarities based on aggregation

### Data

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>10.0</td>
<td>y</td>
<td>9.04</td>
<td></td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
</tr>
<tr>
<td></td>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
</tr>
<tr>
<td></td>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
</tr>
<tr>
<td></td>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
</tr>
<tr>
<td></td>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
</tr>
<tr>
<td></td>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>19.0</td>
<td>12.50</td>
</tr>
<tr>
<td></td>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
</tr>
<tr>
<td></td>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
</tr>
</tbody>
</table>

| Mean of x | 9 |
| Variance of x | 11 |
| Mean of y | 7.50 |
| Variance of y | 4.122 |
| Correlation | 0.816 |
Anscombe's Quartet

[F. J. Anscombe]
Aggregation: Histograms

- Very similar to bar charts
- Often shown without space between (continuity)
- Choice of number of bins
  - Important!
  - Viewers may infer different trends based on the layout

[Munzner (ill. Maguire), 2014]
Aggregation: Histograms

Observed ranks of posts by subreddit

["The reddit Front Page is Not a Meritocracy", T. W. Schneider]
Common Distributions
Binning Scatterplots

- At some point, cannot see density
- Blobs on top of blobs
- 2D Histogram is a histogram in 2D encoded using color instead of height
- Each region is aggregated
Binning

- Hexagonal bins are more circular
- Distance to the edge is not as variable
- More efficient aggregation around the center of the bin
Spatial Aggregation

In cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results.
Modifiable Areal Unit Problem

- How you draw boundaries impacts the type of aggregation you get
- Similar to bins in histograms
- Gerrymandering

50 people

- 60% blue, 40% red
- 3 blue districts, 2 red districts
- 5 blue districts, 0 red districts
- 2 blue districts, 3 red districts

[Wonkblog, Washington Post, Adapted from S. Nass]
Drawing Different Maps: Compactness

Congressional districts drawn to be compact while trying to respect county borders

How often we'd expect a party to win each of the nation’s 435 seats over the long term — not specifically the 2018 midterms — based on historical patterns since 2006

[Bycoffe et al., 538]
## Drawing Different Maps

<table>
<thead>
<tr>
<th>MAP</th>
<th><em>USUALLY DEM. DISTRICTS</em></th>
<th><em>HIGHLY COMPETITIVE</em></th>
<th><em>USUALLY REPUBLICAN</em></th>
<th>EXPECTED SEAT SPLIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic gerrymander</td>
<td>263</td>
<td>27</td>
<td>145</td>
<td>250.6 184.4</td>
</tr>
<tr>
<td>Proportionally partisan</td>
<td>174</td>
<td>82</td>
<td>179</td>
<td>214.0 221.0</td>
</tr>
<tr>
<td>Majority minority</td>
<td>169</td>
<td>82</td>
<td>184</td>
<td>209.8 225.2</td>
</tr>
<tr>
<td>Highly competitive</td>
<td>94</td>
<td>242</td>
<td>99</td>
<td>209.4 225.6</td>
</tr>
<tr>
<td>Compact (borders)</td>
<td>155</td>
<td>99</td>
<td>181</td>
<td>203.9 231.1</td>
</tr>
<tr>
<td>Compact (algorithmic)</td>
<td>151</td>
<td>104</td>
<td>180</td>
<td>202.8 232.2</td>
</tr>
<tr>
<td>Current</td>
<td>168</td>
<td>72</td>
<td>195</td>
<td>200.6 234.4</td>
</tr>
<tr>
<td>Republican gerrymander</td>
<td>139</td>
<td>21</td>
<td>275</td>
<td>171.3 263.7</td>
</tr>
</tbody>
</table>

[A. Bycoffe et al., 538]
Boxplots

• Show **distribution**
• Single value (e.g. mean, max, min, quartiles) doesn't convey everything
• Created by John Tukey
• Show **spread** and **skew** of data
• Best for **unimodal** data
• Variations like vase plot for multimodal data
• Aggregation here involves many different marks
Aggregation: Boxplots

[Washington Post, 2015]
Four Distributions, Same Boxplot...
Hierarchical Parallel Coordinates

[See Color Plates.]

Figure 4: This image sequence shows a Fatal Accident data set of 230,000 data elements at different levels of detail. The first image shows a cut across the root node. The last image shows the cut chaining all the leaf nodes. The rest of the images show intermediate cuts at varying levels of detail.

[See Color Plates.]

Figure 6: Left image shows Iris data set without proximity-based coloring. Right image shows Iris data set with proximity-based coloring revealing three distinct clusters depicted by concentrations of blue, green and pink lines.

Hierarchical Parallel Coordinates

[Fua et al., 1999]
K-Means

Run

[K. Coop, 2014]
K-Means Issues

Shape

Number of Clusters

[D. Robinson, 2015]
Dimensionality Reduction

- Attribute Aggregation: Use fewer attributes (dimensions) to represent items
- Combine attributes in a way that is more instructive than examining each individual attribute
- Example: Understanding the language in a collection of books
  - Count the occurrence of each non-common word in each book
  - Huge set of features (attributes), want to represent each with an aggregate feature (e.g. high use of "cowboy", lower use of "city") that allows clustering (e.g. "western")
  - Don't want to have to manually determine such rules
- Techniques: Principle Component Analysis, Multidimensional Scaling family of techniques
Principle Component Analysis (PCA)

original data space

component space

PC 1

PC 2

Gene 3

Gene 2

Gene 1

PC 1

PC 2

M. Scholz, CC-BY-SA 2.0
Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize.

First, consider a dataset in only two dimensions, like (height, weight). This dataset can be plotted as points in a plane. But if we want to tease out variation, PCA finds a new coordinate system in which every point has a new (x,y) value. The axes don’t actually mean anything physical; they’re combinations of height and weight called “principal components” that are chosen to give one axes lots of variation.

Drag the points around in the following visualization to see PC coordinate system adjusts.