Data Visualization (DSC 530/CIS 602-02)

Filtering and Aggregation

Dr. David Koop
Design Space of Composite Visualization

- Composite visualization views (CVVs)
  - Includes Coordinated multiple views (CMV)
  - + More!
- Design Patterns:
  - Juxtaposition: side-by-side
  - Superimposition: layers
  - Overloading: vis meshed with another
  - Nesting: vis inside a vis (recursive vis)
  - Integration: "merge" views + links

[W. Javed and N. Elmqvist, 2012]
Juxtaposition

[ComVis, K. Matkovic et al., 2008]
Integration

(later known as a Sankey Diagram)

[Napoleon's March to Moscow, C. J. Minard, 1869]
Integration

"best statistical graphic ever"

[Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

Napoleon's March to Moscow, C. J. Minard, 1869

(later known as a Sankey Diagram)

[Napoleon's March to Moscow, C. J. Minard, 1869]
Superimposition

[M. Bostock, http://bl.ocks.org/mbostock/3884955]
Fig. 13. A software system and its associated call graph (caller = green, callee = red). (a) and (b) show the system with bundling strength $\beta = 0.85$ using a balloon layout (node labels disabled) and a radial layout, respectively. Bundling reduces visual clutter, making it easier to perceive the actual connections than when compared to the non-bundled versions (figures 2a and 11a). Bundled visualizations also show relations between sparsely connected systems more clearly (encircled regions); these are almost completely obscured in the non-bundled versions. The encircled regions highlight identical parts of the system for (a), (b), and figure 15.

Fig. 14. Using the bundling strength $\beta$ to provide a trade-off between low-level and high-level views of the adjacency relations. The value of $\beta$ increases from left-to-right; low values mainly provide low-level, node-to-node connectivity information, whereas high values provide high-level information as well by implicit visualization of adjacency edges between parent nodes that are the result of explicit adjacency edges between their respective child nodes.

Changing the bundling strength $\beta$ and by switching between different tree layouts. The participants from academia were our fellow researchers, PhD students and MSc students from the Computer Science department of the Technische Universiteit Eindhoven. They all had experience with either software development, software visualization, or information visualization in general. Participants from industry were representatives of the Software Improvement Group (SIG) in Amsterdam, which delivers insight in the structure and technical quality of software portfolios, and representatives of FEI Company Eindhoven, which produces software to operate with FEI's range of electron microscopes.

The majority of the participants regarded the technique as useful for quickly gaining insight in the adjacency relations present in hierarchically organized systems. In general, the visualizations were also regarded as being aesthetically pleasing. SIG and FEI Company Eindhoven are currently supporting further development by providing us with additional data sets and feedback regarding the resulting visualizations.

More specifically, most of the participants particularly valued the fact that relations between items at low levels of the hierarchy were automatically lifted to implicit relations between items at higher levels by means of bundles. This quickly gave them an impression of the high-level connectivity information while still being able to inspect the low-level relations that were responsible for the bundles by interactively manipulating the bundling strength.

This is illustrated in figure 14, which shows visualizations using different values for the bundling strength $\beta$. Low values result in visualizations that mainly provide low-level, node-to-node connectivity information. High values result in visualizations that provide high-level information as well by implicit visualization of adjacency edges between parent nodes that are the result of explicit adjacency edges between their respective child nodes.

Another aspect that was commented on was how the bundles gave an impression of the hierarchical organization of the data as well, thereby strengthening the visualization of the hierarchy. More specifically, a thick bundle shows the presence of two elements at a fairly
Superimposition

Superimposed views overlay two or more visual spaces on top of each other (Figures 6 and 7). The resulting visualization becomes the visual combination of the component visualizations, often using transparency to enable seeing all views. Superimposed views are generally used to highlight spatial relations in the component visualizations. In other words, the spatial linking present in these views is one-to-one, i.e., all the overlay visualizations share the same underlying visual space. Line graph visualizations with several data series, where more than one graph is superimposed in a single chart (e.g., [19]), is a very commonly used example of this design pattern. The spatial linking in the superimposed views allows for easy comparison across different datasets because the user does not have to split their attention between different parts of the visual space. Furthermore, the fact that visualizations are stacked means that they can each use the full available space in the view. However, because the composition simply adds the component visualizations together, the visual clutter may become significant, and it is also likely to cause conflicts arising from one visualization occluding another.

5.1 Mapgets

Mapgets [38] is a geographic visualization system that allows users to interactively perform map editing and querying of geographical datasets. The maps generated using Mapgets are built on an underlying presentation stack that superimposes multiple dataset layers on top of each other. The users can dynamically select the dataset to use for each layer and the total number of layers to compose. Different layers in the presentation stack allow users to independently interact with each of the associated visualization and control the layer attributes. The technique also allows the users to reorder layers in the presentation stack to achieve the desirable map result. Figure 6 shows an example of a European map generated in Mapgets. The presentation stack associated with this map consists of three layers: the bottom layer visualizes rivers, the center layer is used to depict the country borders, and the topmost layer is used to display the country labels.

5.2 GeoSpace

GeoSpace [22] allows users to interactively explore complex visual spaces using superimposed views. It permits progressively overlaying different datasets, based on the user queries, in a single view. Beyond allowing users to explore datasets through dynamic queries, GeoSpace also supports pan and zoom operations for navigation. Figure 7 shows GeoSpace system being used for exploring crime around the Cambridge, MA area. The figure shows a 2D view of the visualization, where red dots that are spatially coupled to the underlying layer show the reported crime cases in the region.

Overloading

Overloading points into the region bounded by two axes in the parallel coordinate plot. [Links on Treemaps, J.-D. Fekete et al., 2003]
Nesting

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

<table>
<thead>
<tr>
<th>Encoding</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
</tr>
<tr>
<td>Same</td>
<td>Redundant</td>
</tr>
<tr>
<td>Different</td>
<td>Multiform</td>
</tr>
</tbody>
</table>

[Munzner (ill. Maguire), 2014]
Multiform

[Improvise, Weaver, 2004]
Overview-Detail View
Small Multiples

- Same encoding, but different data in each view (e.g. SPLOM)

[http://bl.ocks.org/mbostock/4063663]
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
Example: Grouped Bar Chart

[Example: Grouped Bar Chart]

[Example: Grouped Bar Chart]

[Example: Grouped Bar Chart]

[Example: Grouped Bar Chart]

[Example: Grouped Bar Chart]

[Example: Grouped Bar Chart]

[Example: Grouped Bar Chart]

[M. Bostock, http://bl.ocks.org/mbostock/3887051]
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

[Slingsby et al., 2009]
Assignment 3

- Soccer data
  - Draw two choropleth maps
  - Draw a teammate graph using force-directed layout
  - Country lookup: Create an associative array, d3.map or Map
Exam 2

- Monday, April 3
- Similar Format to Exam 1: Multiple Choice + Free Response
- Cumulative but emphasis on the topics covered since Exam 1
  - Trees
  - Geospatial Data
  - Color & Colormaps
  - Interaction
  - Multiple Views
- Reading: Chs. 8.1-8.3, 10, 6, 11, 12, 13
- Be prepared for a similar D3 question to the last exam
- More information to be made available on the course web page
Reducing Complexity

- Too many items or attributes lead to visual clutter
- Interaction and Multiple Views can help, but often lose the ability to start understanding an entire dataset at first glance
- **Reduction** techniques show less data to reduce complexity
- Can reduce items or attributes (both are elements)
- **Filtering**: eliminate elements from the current view
  - "out of sight, out of mind"
- **Aggregation**: replace elements with a new element that represents the replaced elements
  - summarization is often challenging to design
- Another method is **focus+context**: show details in the context of an overview
Overview

Reducing Items and Attributes

- **Filter**
  - Items
    - [chart showing filter process]
  - Attributes
    - [chart showing filter process]

- **Aggregate**
  - Items
    - [chart showing aggregate process]
  - Attributes
    - [chart showing aggregate process]

Reduce

- **Filter**
  - [chart showing reduce process]

- **Aggregate**
  - [chart showing reduce process]

- **Embed**
  - [chart showing reduce process]

[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
Example: NYC Health Dept. Restaurant Ratings

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

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

[Related Article »]

For menus and reviews by New York Times critics, visit our restaurants guide.

[Jeremy White, New York Times]
Dynamic Queries

- Interaction need not be with the visualization itself
- Users interact with **widgets** that control which items are shown
  - Slides, 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, we presented users with one of three scenting conditions. We created three vectors for each view, each indicating the number of visits to that view. The conditions consisted of no scent, social navigation scent cues, and a combination of both. The task hypotheses were that social navigation cues would increase the number of comments made on a view, and that the combination of social navigation and scenting would lead to even more unique discoveries.

We gave subjects an introductory tutorial to the system, and then instructed them to make at least seven observations that provided evidence either for or against the current task hypothesis. At least two of the observations had to be unique findings on views not yet visited. Subjects were asked to note their observations by leaving new comments on the corresponding views. The study employed a 3 (Task) x 3 (Scent) between conditions factorial design. Task and scent pairings and presentation order were counterbalanced across 154 views.

The study recruited 22 participants, either graduate or undergraduate students, and were instructed to complete three tasks. Participant ages ranged from 19 to 32. We gave them an introductory tutorial to the system, and then asked the participants to complete three tasks. We removed the starting overview from consideration, representing the number of visits to each view in each scenting condition. We then compared these visits to the visitation vector for the nation's labor force, with a sharp drop after Prohibition. Scented Widgets are used in the dynamic query widgets to show visitation patterns and likely had an effect on these correlations.

Next, we analyzed the data to check if scented widgets help us make unique discoveries. Our hypotheses were that scented widgets would increase the number of comments made on a view, and that the combination of social navigation and scenting would lead to even more unique discoveries. We found that performance would improve over subsequent trials, regardless of conditions. We noted that the correlations are not very strong. We believe that the semantics of the tasks also affect underlying activity measure used to seed the scented widgets. Using the seed data, we scaled them to exhibit a power law distribution, and so we scaled them similarly to freely explore the data.

For the task hypothesis, we created three vectors, each representing the number of visits to each view in each scenting condition. We then compared these vectors to the visitation vector for the nation's labor force, with a sharp drop after Prohibition. Scented Widgets are used in the dynamic query widgets to show visitation patterns and likely had an effect on these correlations.

Willett et al., 2007
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><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Saturation</td>
<td>Varies the saturation of the widget (or of a visualization embedded in it)</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Opacity</td>
<td>Varies the saturation of the widget (or of a visualization embedded in it)</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Text</td>
<td>Inserts one or more small text figures into the widget</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Icon</td>
<td>Inserts one or more small icons into the widget</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Bar Chart</td>
<td>Inserts one or more small bar chart visualizations into the widget</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Line Chart</td>
<td>Inserts one or more small line chart visualizations into the widget</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
</tbody>
</table>

[Willett et al., 2007]
Star Plots

Aberfeldy
- Malty
- Fruity
- Floral
- Body
- Sweetness
- Smoky
- Honey
- Nutty
- Spicy

Aberlour

AnCnoc

Ardmore
- Malty
- Fruity
- Floral
- Body
- Sweetness
- Smoky
- Honey
- Nutty
- Spicy

ArranIsleOf

Auchentoshan

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

• 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>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</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

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</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]
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
Hexagonal Binning

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

[Bachthaler & Weiskopf, 2008]
Aggregation: Continuous Scatterplot

[Bachthaler & Weiskopf, 2008]
Spatial Aggregation

In cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results.
Spatial Aggregation

spatial aggregation
modifiable areal unit problem
in cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results

[Penn State, GEOG 486]
Spatial Aggregation

[Penn State, GEOG 486]
Modifiable Areal Unit Problem

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

[Wonkblog, Washington Post, Adapted from S. Nass]
Boxplots

- Show **distribution**
- Single value (e.g. mean, max, min, quartiles) doesn't convey everything
- Created by John Tukey who grew up in New Bedford!
- Show **spread** and **skew** of data
- Best for **unimodal** data
- Variations like vase plot for multimodal data
- Aggregation here involves many different marks
Boxplot Example

(a) Overall Activity  (b) Structural Activity  (c) Parameter Activity  (d) Layout Activity

Percentage (%)  0  10  20  30  40  50  60  70  80  90  100

Struct.  Param.  Layout  Task 1  Task 2  Task 3  Task 4  Task 5  Task 6

Task 1  Task 2  Task 3  Task 4  Task 5  Task 6

[L. Lins et al., 2008]
Hierarchical Parallel Coordinates

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. (See Color Plates).

[Fua et al., 1999]
D3 Multiple Views and Interaction

• http://codepen.io/dakoop/pen/jBQrYp