Lecture 11: High Dimensionality

Information Visualization
CPSC 533C, Fall 2009

Tamara Munzner

UBC Computer Science

Wed, 21 October 2009

Readings Covered

Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, Vol. 85, No. 411. (Sep., 1990), pp. 664-675.

Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner, IEEE Visualization '99.

Glimmer: Multilevel MDS on the GPU. Stephen Ingram, Tamara Munzner and Marc Olano. IEEE TVCG, 15(2):249-261, Mar/Apr 2009.

Cluster Stability and the Use of Noise in Interpretation of Clustering. George S. Davidson, Brian N. Wylie, Kevin W. Boyack, Proc InfoVis 2001.

Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets. Jing Yang, Wei Peng, Matthew O. Ward and Elke A. Rundensteiner. Proc. InfoVis 2003.

Further Reading

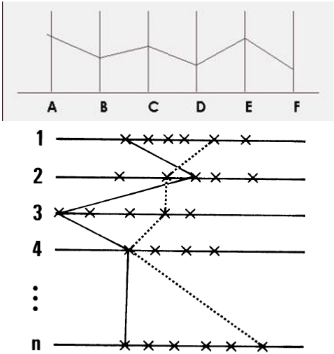
Visualizing the non-visual: spatial analysis and interaction with information from text documents. James A. Wise et al, Proc. InfoVis 1995

Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. Alfred Inselberg and Bernard Dimsdale, IEEE Visualization '90.

A Data-Driven Reflectance Model. Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. SIGGRAPH 2003. graphics.lcs.mit.edu/∼wojciech/pubs/sig2003.pdf

Parallel Coordinates

- only 2 orthogonal axes in the plane
- instead, use parallel axes!



[Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, 85(411), Sep 1990, p 664-675.]

PC: Correllation

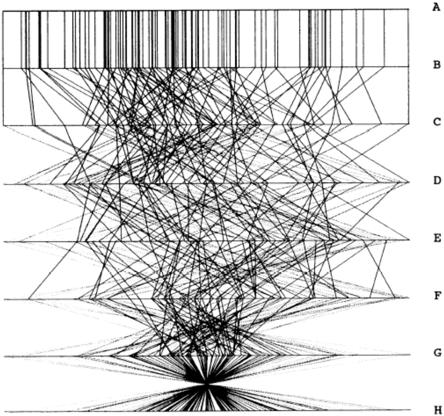
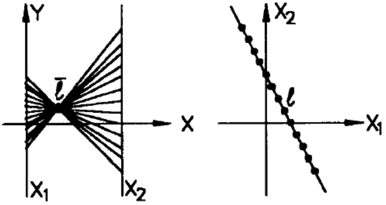


Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of ρ = 1, .8, .2, 0, −.2, −.8, and −1.

[Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, 85(411), Sep 1990, p 664-675.]

PC: Duality

- rotate-translate
- point-line
  - pencil: set of lines coincident at one point



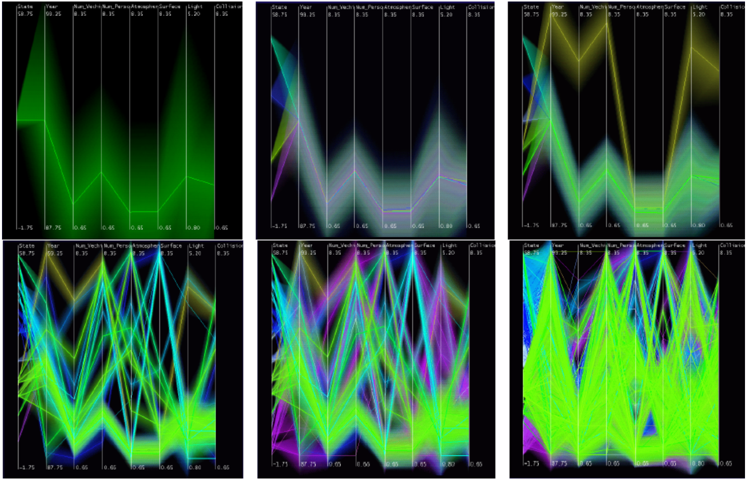
[Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. Alfred Inselberg and Bernard Dimsdale, IEEE Visualization '90.]

PC: Axis Ordering

- geometric interpretations
  - hyperplane, hypersphere
  - points do have intrinsic order
- infovis
  - no intrinsic order, what to do?
  - indeterminate/arbitrary order
    - weakness of many techniques
    - downside: human-powered search
    - upside: powerful interaction technique
- most implementations
  - user can interactively swap axes
- Automated Multidimensional Detective
  - Inselberg 99
  - machine learning approach

Hierarchical Parallel Coords: LOD

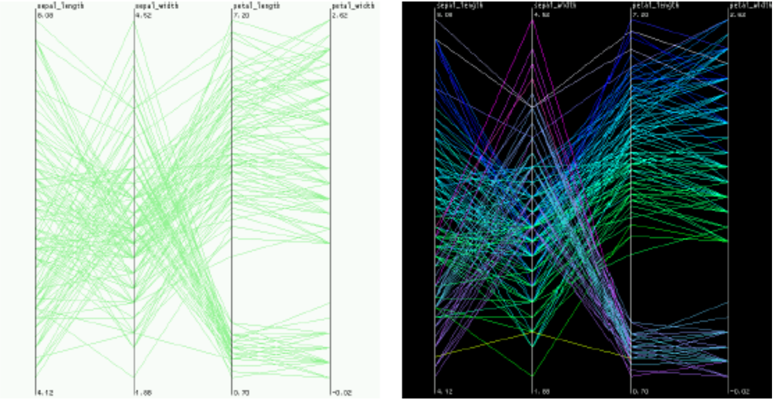
- variable-width opacity bands



[Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Fua, Ward, and Rundensteiner, IEEE Visualization 99.]

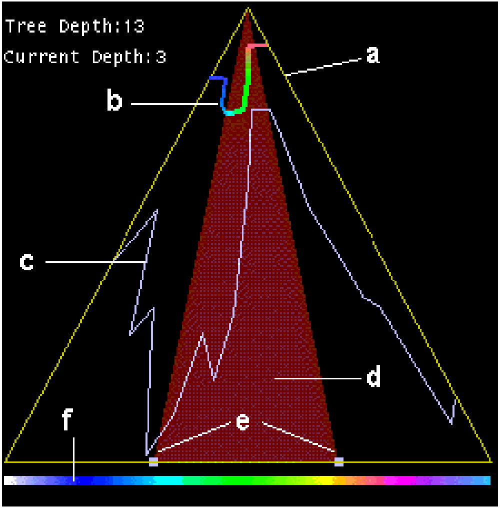
Proximity-Based Coloring

- cluster proximity



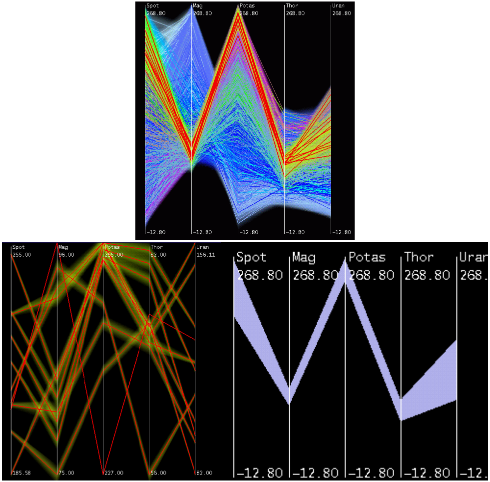
[Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Fua, Ward, and Rundensteiner, IEEE Visualization 99.]

Structure-Based Brushing



[Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Fua, Ward, and Rundensteiner, IEEE Visualization 99.]

Dimensional Zooming



[Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Fua, Ward, and Rundensteiner, IEEE Visualization 99.]

Critique

Critique

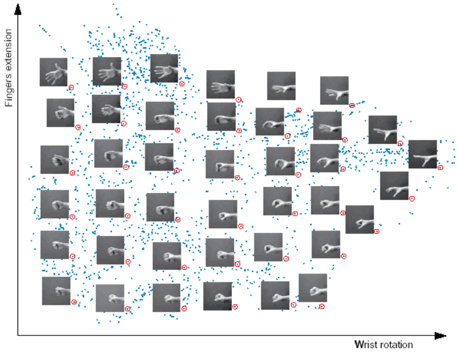
- not easy for novices
  - now used in many apps
- hier: major scalability improvements
  - combination of encoding, interaction

Dimensionality Reduction

- mapping multidimensional space into space of fewer dimensions
  - filter subset of original dimensions
  - generate new synthetic dimensions
- why is lower-dimensional approximation useful?
  - assume **true/intrinsic** dimensionality of dataset is (much) lower than measured dimensionality!
- why would this be the case?
  - only indirect measurement possible
    - fisheries ex: want spawn rates. have water color, air temp, catch rates...
  - sparse data in verbose space
    - documents ex: word occurrence vectors.10K+ dimensions, want dozens of topic clusters

Dimensionality Reduction: Isomap

- 4096 D: pixels in image
- 2D: wrist rotation, fingers extension



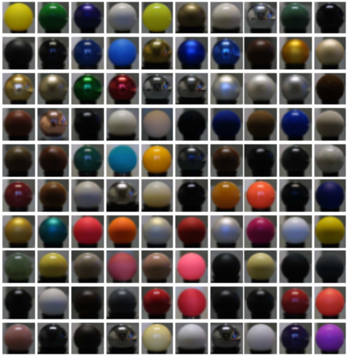
[A Global Geometric Framework for Nonlinear Dimensionality Reduction. J. B. Tenenbaum, V. de Silva, and J. C. Langford. Science 290(5500), pp 2319–2323, Dec 22 2000]

Goals/Tasks

- goal: keep/explain as much variance as possible
- find clusters
  - or compare/evaluate vs. previous clustering
- understand structure
  - absolute position not reliable
    - arbitrary rotations/reflections in lowD map
  - fine-grained structure not reliable
    - coarse near/far positions safer

Dimensionality Analysis Example

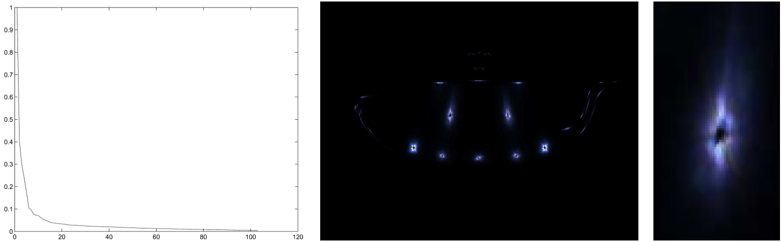
- measuring materials for image synthesis
  - BRDF measurements: 4M samples x 103 materials
  - goal: lowD model where can interpolate



[A Data-Driven Reflectance Model, SIGGRAPH 2003, W Matusik, H. Pfister M. Brand and L. McMillan, graphics.lcs.mit.edu/∼wojciech/pubs/sig2003.pdf]

Dimensionality Analysis: Linear

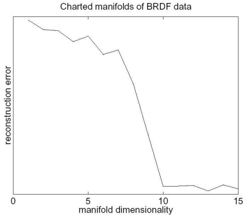
- how many dimensions is enough?
  - could be more than 2 or 3!
  - find knee in curve: error vs. dims used
- linear dim reduct: PCA, 25 dims
  - physically impossible intermediate points when interpolate

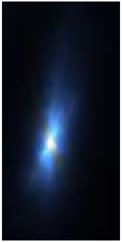


[A Data-Driven Reflectance Model, SIGGRAPH 2003, W Matusik, H. Pfister M. Brand and L. McMillan, graphics.lcs.mit.edu/∼wojciech/pubs/sig2003.pdf]

Dimensionality Analysis: Nonlinear

- nonlinear dim reduct (charting): 10-15
  - all intermediate points physically possible

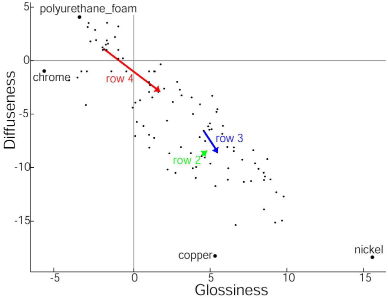




[A Data-Driven Reflectance Model, SIGGRAPH 2003, W Matusik, H. Pfister M. Brand and L. McMillan, graphics.lcs.mit.edu/∼wojciech/pubs/sig2003.pdf]

Meaningful Axes: Nameable By People

- red, green, blue, specular, **diffuse**, **glossy**, **metallic**, plastic-y, roughness, rubbery, greasiness, dustiness...



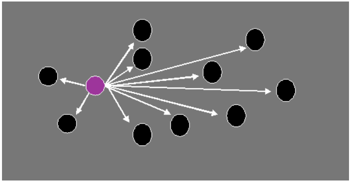
[A Data-Driven Reflectance Model, SIGGRAPH 2003, W Matusik, H. Pfister M. Brand and L. McMillan, graphics.lcs.mit.edu/∼wojciech/pubs/sig2003.pdf]

MDS: Multidimensional scaling

- large family of methods
  - minimize differences between interpoint distances in high and low dimensions
  - distance scaling: minimize objective function
  - $stress(D, \Delta) = \sqrt{\frac{\sum_{ij}\left(d_{ij}-\delta_{ij}\right)^2}{\sum_{ij}\delta_{ij}^2}}$
    - D: matrix of lowD distances
    - Δ: matrix of hiD distances δ_{ij}

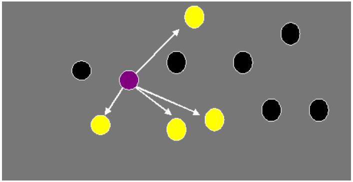
Spring-Based MDS: Naive

- repeat for all points
  - compute spring force to all other points
  - difference between high dim, low dim distance
  - move to better location using computed forces
- compute distances between all points
  - $O(n^2)$ iteration, $O(n^3)$ algorithm



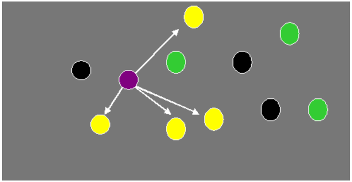
Faster Spring Model: Stochastic

- compare distances only with a few points
  - maintain small local neighborhood set



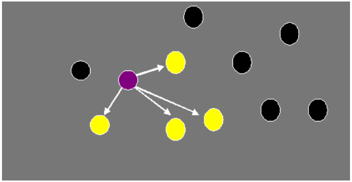
Faster Spring Model: Stochastic

- compare distances only with a few points
  - maintain small local neighborhood set
  - each time pick some randoms, swap in if closer



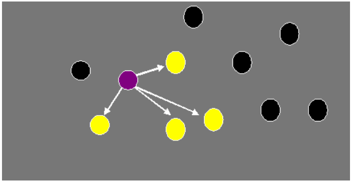
Faster Spring Model: Stochastic

- compare distances only with a few points
  - maintain small local neighborhood set
  - each time pick some randoms, swap in if closer

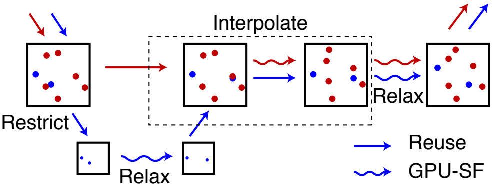


Faster Spring Model: Stochastic

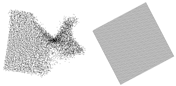
- compare distances only with a few points
  - maintain small local neighborhood set
  - each time pick some randoms, swap in if closer
- small constant: 6 locals, 3 randoms typical
  - $O(n)$ iteration, $O(n^2)$ algorithm



Glimmer Algorithm



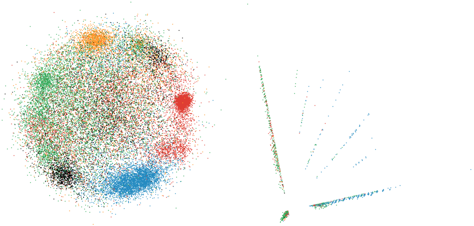
- multilevel, designed to exploit GPU
    - restriction to decimate
    - relaxation as core computation
    - relaxation to interpolate up to next level
- GPU stochastic as subsystem
    - poor convergence properties if run alone
    - low-pass-filter stress approx. for termination

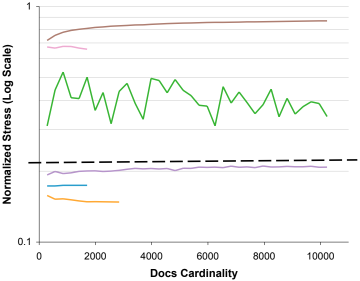


[Glimmer: Multilevel MDS on the GPU. Ingram, Munzner and Olano. IEEE TVCG, 15(2):249-261, Mar/Apr 2009.]

Glimmer Results

- sparse document dataset: 28K dims, 28K points



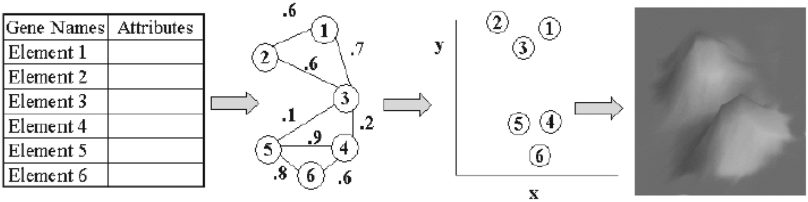


[Glimmer: Multilevel MDS on the GPU. Ingram, Munzner and Olano. IEEE TVCG, 15(2):249-261, Mar/Apr 2009.]

Cluster Stability

- display
  - also terrain metaphor
- underlying computation
  - energy minimization (springs) vs. MDS
  - weighted edges
- do same clusters form with different random start points?
- "ordination"
  - spatial layout of graph nodes

Approach



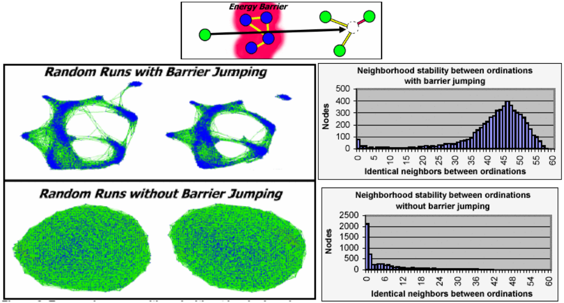
- normalize within each column
- similarity metric
  - discussion: Pearson's correllation coefficient
- threshold value for marking as similar
  - discussion: finding critical value

Graph Layout

- criteria
  - geometric distance matching graph-theoretic distance
    - vertices one hop away close
    - vertices many hops away far
  - insensitive to random starting positions
    - major problem with previous work!
  - tractable computation
- force-directed placement
  - discussion: energy minimization
  - others: gradient descent, etc
  - discussion: termination criteria

Barrier Jumping

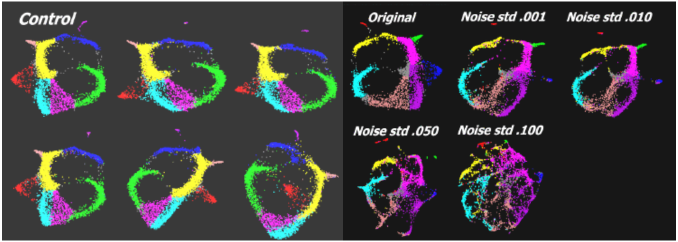
- same idea as simulated annealing
  - but compute directly
  - just ignore repulsion for fraction of vertices
- solves start position sensitivity problem



Results

- efficiency
  - naive approach: $O(V^2)$
  - approximate density field: $O(V)$
- good stability
  - rotation/reflection can occur

different random start adding noise



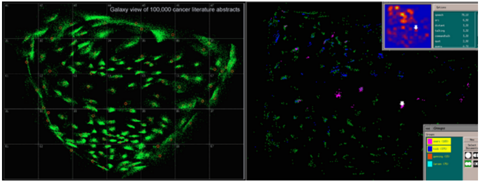
Critique

Critique

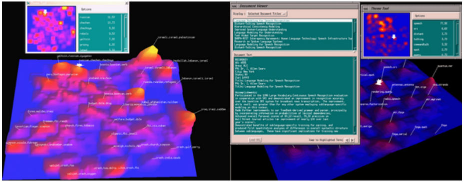
- real data
  - suggest check against subsequent publication!
- give criteria, then discuss why solution fits
- visual + numerical results
  - convincing images plus benchmark graphs
- detailed discussion of alternatives at each stage
- specific prescriptive advice in conclusion

MDS Beyond Points

- galaxies: aggregation



- themescapes: terrain/landscapes
  - studies: less effective than points alone [Tory 07, 09]



[www.pnl.gov/infoviz/graphics.html] [Visualizing the non-visual: spatial analysis and interaction with information from text documents. James A. Wise et al, Proc. InfoVis 1995]

Dimension Ordering

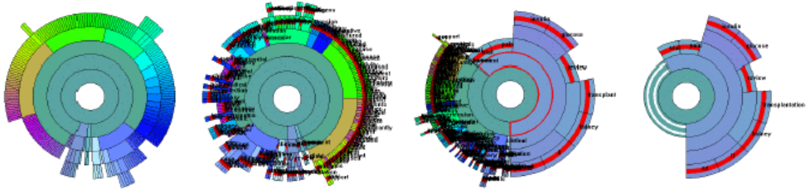
- in NP: heuristic, like most interesting infovis problems
- divide and conquer
  - iterative hierarchical clustering
  - representative dimensions
- choices
  - similarity metrics
  - importance metrics
    - variance
  - ordering algorithms
    - optimal
    - random swap
    - simple depth-first traversal

Spacing, Filtering

- same idea: automatic support
- interaction
    - manual intervention
    - structure-based brushing
    - focus+context

Results: InterRing

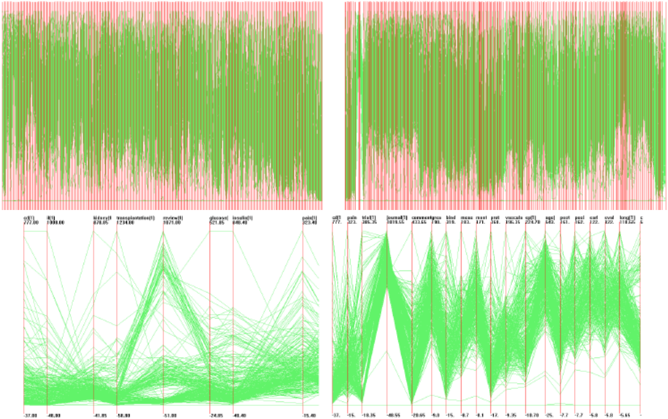
- raw, order, distort, rollup (filter)



[Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets. Yang Peng, Ward, and Rundensteiner. Proc. InfoVis 2003]

Results: Parallel Coordinates

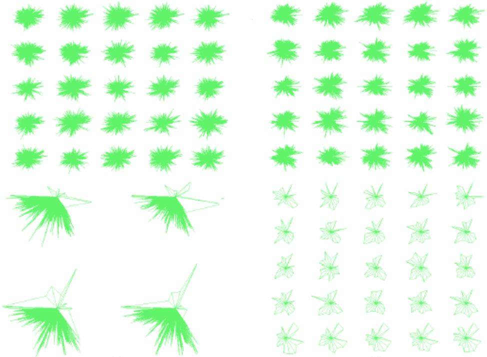
- raw, order/space, zoom, filter



[Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets. Yang Peng, Ward, and Rundensteiner. Proc. InfoVis 2003]

Results: Star Glyphs

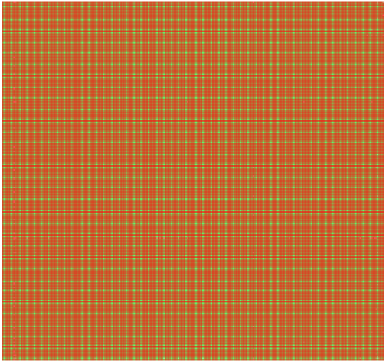
- raw, order/space, distort, filter

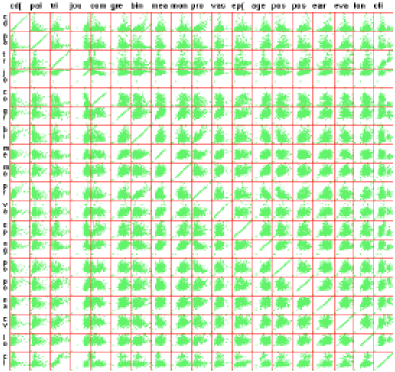


[Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets. Yang Peng, Ward, and Rundensteiner. Proc. InfoVis 2003]

Results: Scatterplot Matrices

- raw, filter





[Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets. Yang Peng, Ward, and Rundensteiner. Proc. InfoVis 2003]

Critique

Critique

- pro
  - approach on multiple techniques,
  - real data!
- con
  - always show order then space then filter
    - hard to tell which is effective
    - show ordered vs. unordered after zoom/filter?

Reminders

- meet with me before end of week!
- presentation topics also due Friday
  - your call whether presentation and project topics match
  - submit: 3 topic choices, veto day
- project data/task ideas on resources page
  - VAST/InfoVis Contests!

Readings Next Week

Graph Visualisation in Information Visualisation: a Survey. Ivan Herman, Guy Melancon, M. Scott Marshall. IEEE Transactions on Visualization and Computer Graphics, 6(1), pp. 24-44, 2000. http://citeseer.nj.nec.com/herman00graph.html

change:
Configuring Hierarchical Layouts to Address Research Questions. Adrian Slingsby, Jason Dykes, and Jo Wood. IEEE Transactions on Visualization and Computer Graphics 15 (6), Nov-Dec 2009 (Proc. InfoVis 2009).

Multiscale Visualization of Small World Networks. David Auber, Yves Chiricota, Fabien Jourdan, Guy Melancon, Proc. InfoVis 2003.
http://dept-info.labri.fr/∼auber/documents/publi/auberIV03Seattle.pdf

Topological Fisheye Views for Visualizing Large Graphs. Emden Gansner, Yehuda Koren and Stephen North, IEEE TVCG 11(4), p 457-468, 2005.
http://www.research.att.com/areas/visualization/papers_videos/pdf/DBLP-conf-infovis-GansnerKN04.pdf

IPSep-CoLa: An Incremental Procedure for Separation Constraint Layout of Graphs. Tim Dwyer, Kim Marriott, and Yehuda Koren. Proc. InfoVis 2006, published as IEEE TVCG 12(5), Sep 2006, p 821-828.
http://www.research.att.com/∼yehuda/pubs/dwyer.pdf