High Dimensionality

Lecture 9 CPSC 533C, Spring 2004

9 Feb 2003

High Dimensionality

MDS

Themescapes/Galaxies

Cluster Stability

Dimension Ordering

project software

MDS

multidimensional scaling

high dimensional space: q dims

embed in much lower dimensional space: p dims
- often p is 2 or 3

need pairwise distances between points
- proximity data

minimize error/stress of low-dim wrt high-dim

Issues

which distance metric: Euclidean or other?

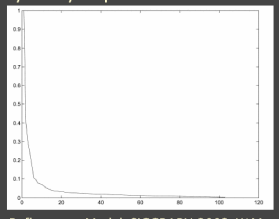
computation
- naive: O(n^3)
- better: O(n^2) Chalmers 96
- hybrid: O(n sqrt(n))

True Dimensionality: Linear

how many dimensions is enough? > 2 or 3?
- knee in error curve

example: measured materials from graphics

linear PCA: 25
- can get physically impossible intermediate points

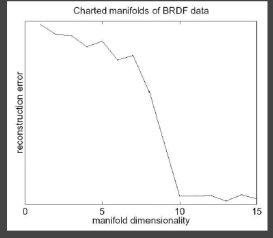


[A Data-Driven Reflectance Model, SIGGRAPH 2003, W Matusik, H. Pfister M. Brand and L. McMillan, graphics.lcs.mit.edu/~wojciech/pubs/sig2003.pdf]

True Dimensionality: Nonlinear

nonlinear MDS: 10-15
- all intermediate points possible

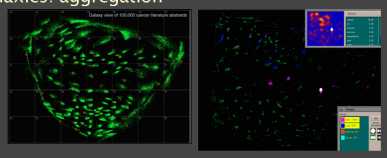
categorizable by people
- red, green, blue, specular, diffuse, glossy, metallic,
- plastic-y, roughness, rubbery, greasiness, dustiness...



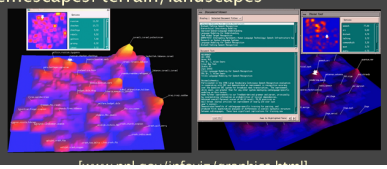
[A Data-Driven Reflectance Model, SIGGRAPH 2003, W Matusik, H. Pfister

Themescapes/Galaxies

MDS output: beyond just drawing points
- galaxies: aggregation



- themescapes: terrain/landscapes



[www.pnl.gov/infoviz/graphics.html]

Cluster Stability

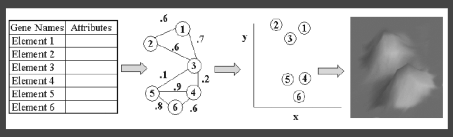
display
- also terrain metaphor

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

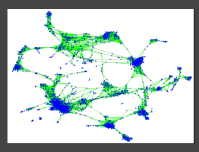
similarity metric
- discussion: Pearson's correllation coefficient

threshold value for marking as similar
- discussion: finding critical value

Graph Layout

criteria
- distance in layout 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

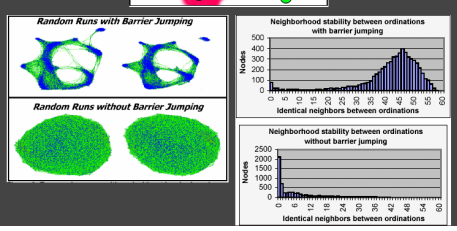
force-directed placement
- discussion: energy minimization
- others: gradient descent, etc
- discussion: termination criteria



Barrier Jumping

same idea as simulated annealing
- but compute directly
- just ignore repulsion for fraction of vertices

solves start position sensitivity problem

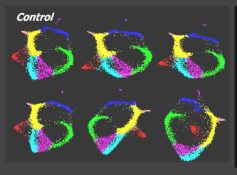


Results

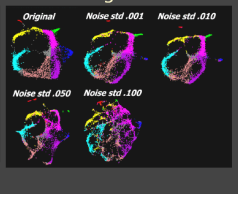
efficiency
- naive approach: O(V^2)
- approximate density field: O(V)

good stability
- rotation/reflection can occur

different random start



adding noise



Critique

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

Dimension Ordering

in NP, like most interesting infovis problems
- heuristic

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

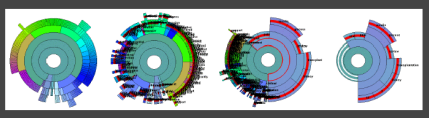
interaction

- manual intervention

- structure-based brushing
- focus+context, next week

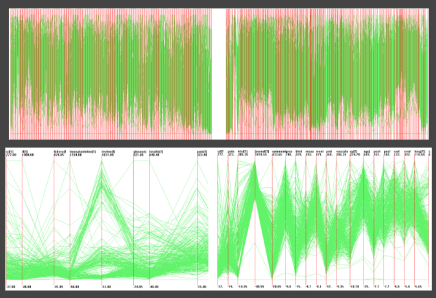
Results: InterRing

raw, order, distort, rollup (filter)



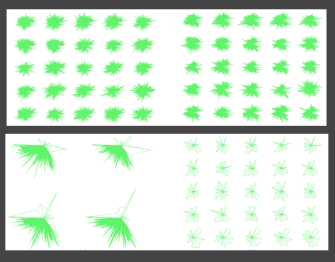
Results: Parallel Coordinates

raw, order/space, zoom, filter



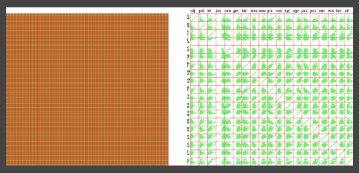
Results: Star Glyphs

raw, order/space, distort, filter



Results: Scatterplot Matrices

raw, filter



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?

Software

Tableau: commercial Polaris

vtk: scivis dataflow

xgobi

xmdv

data

- Klingner's online databases