A Network-Based Interface for the Exploration of High-Dimensional Data Spaces

Zhiyuan Zhang[1], Kevin T. McDonnell[2], Klaus Mueller[1]

[1] Computer Science Department, Stony Brook University

[2] Dowling College

Motivation

- The navigation of high-dimensional data spaces remains challenging.
- Valuable insight: inter-attribute relationships.
  - Business scenario
- Unfortunately
  - High-dimensional space exceeds human comprehension
  - Curse of dimensionality

Motivation - Parallel Coordinates

- Good:
  - Present overviews of the whole, raw data set
  - Show relationships among the dimensions in a sequential way
- However:
  - Ordering of the dimensions in the PC display has a great impact on the visual relationship.
  - Worst case: O(n!) possible dimension orderings.

Motivation – Dimension Ordering

- Automatic Ordering.
  - Similarity: Ankerst et al.; Tatu et al.
  - Rating: Johansson and Johansson
  - Correlation: Artero et al.
  - Appearance: Dasgupta and Kosara .
  - Subspace clustering: Ferdosi and Roerdink.
- There is rarely just one perfect dimension ordering.
- Depends on current data exploration goal.
- Need effective tools for exploratory dimension ordering
  - Suitable "map"
  - Navigation & Interaction: familiar

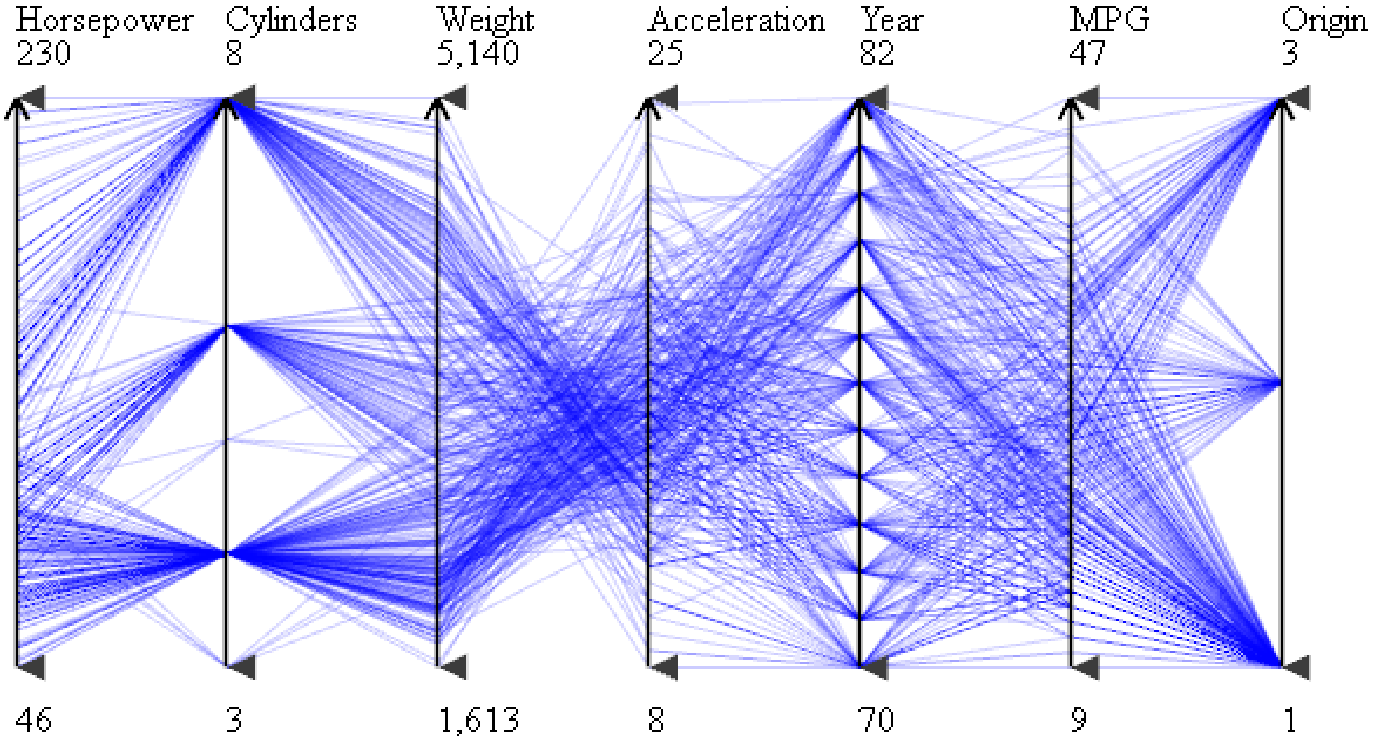
Framework Overview

- Two coordinated displays:
  - Parallel coordinates display + Network display.
- Operations in either display are reflected in the other.
- Parallel coordinates (PC) display
  - Shows the raw data in an undistorted way.
  - Brush, filter, select data for network display
- Network display
  - Overview of all dimensions in terms of their pairwise correlations
  - Interface to drive dimension ordering in PC display

Demonstration Dataset

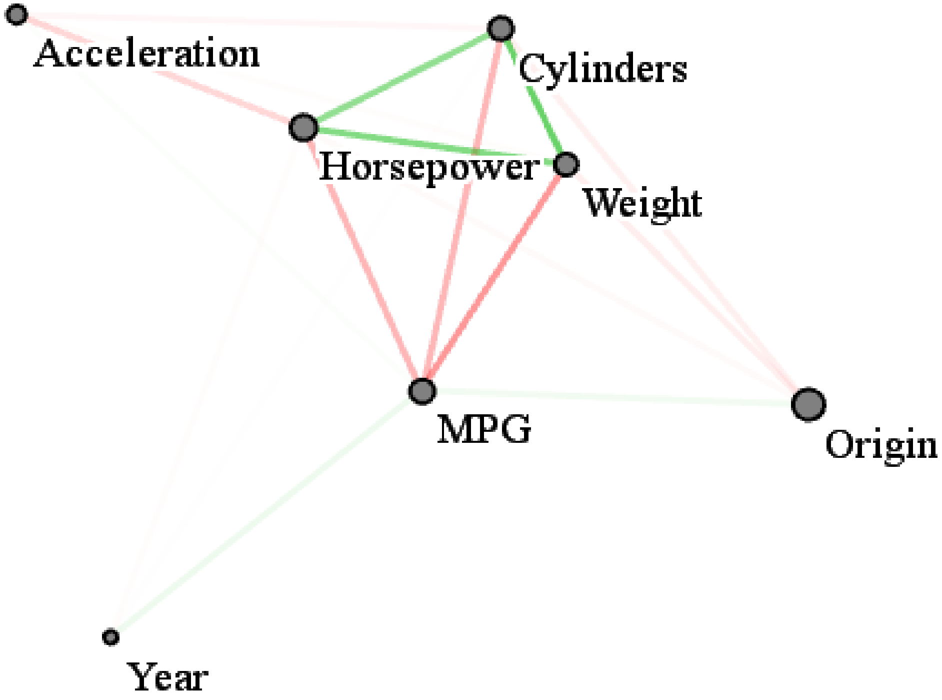
The ever popular "cars" dataset

- 392 cars
- 7 attributes: MPG, #cylinders, horsepower, weight, acceleration time, year and origin.



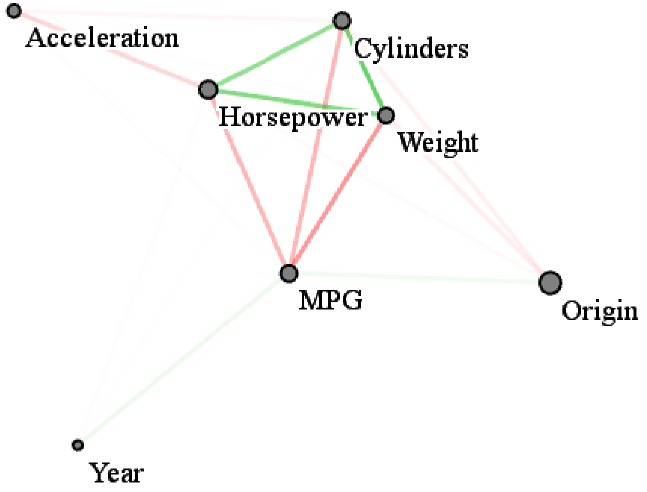
Network Display

- Vertices: dimensions
- Edges: relationships between dimensions



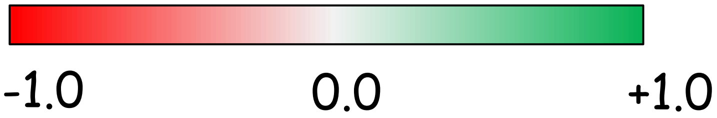
Network Semantics – Vertex

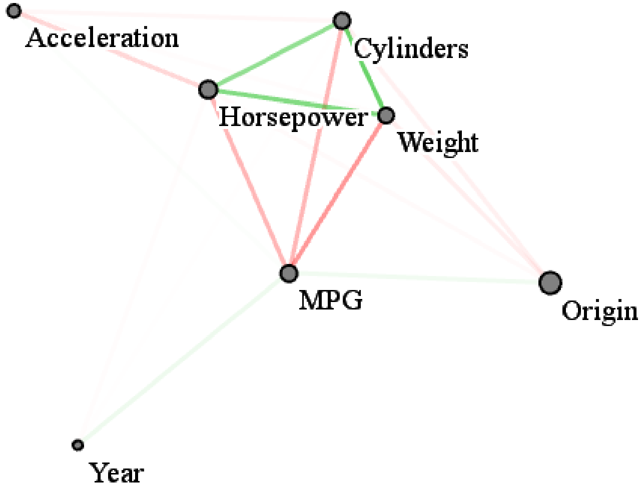
- Size
  - More significant dimensions -> larger vertices
- Dimension significance determined by diversity
  - Range
  - Standard deviation σ
  - Coefficient of variation σ/|μ|
- Allow users to build new metric or alter significance.



Network Semantics – Edge

- Edge is weighted by the correlation.
- The correlation is encoded by color.
  - Green : correlation = 1.0
  - Red : correlation = -1.0
  - Linear interpolation computes the colors in between.





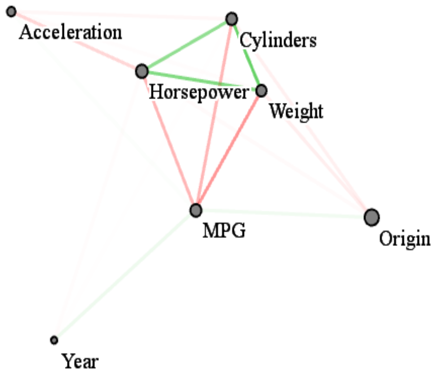
Build the Network

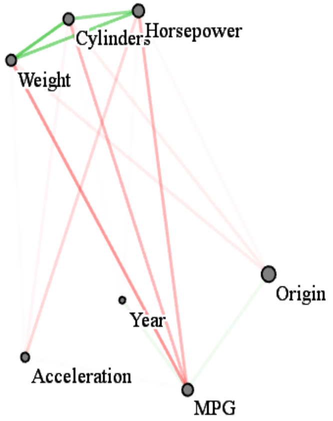
- Compute Pearson's correlation for each pair of dimensions.
- Layout the points using a mass-spring model.
    - Forces between dimensions are computed from the correlations

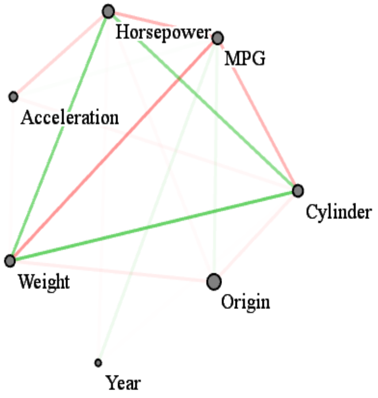
Mass-Spring Model Layout

- Highly correlated dimensions will be drawn close to one another.
  - Distances of the nodes: correlation

- Three configurations:
  - Correlation strength;
  - Positive correlation preferred;
  - Negative Correlation Preferred.







Network-Driven Dimension Ordering

- A path in the network display

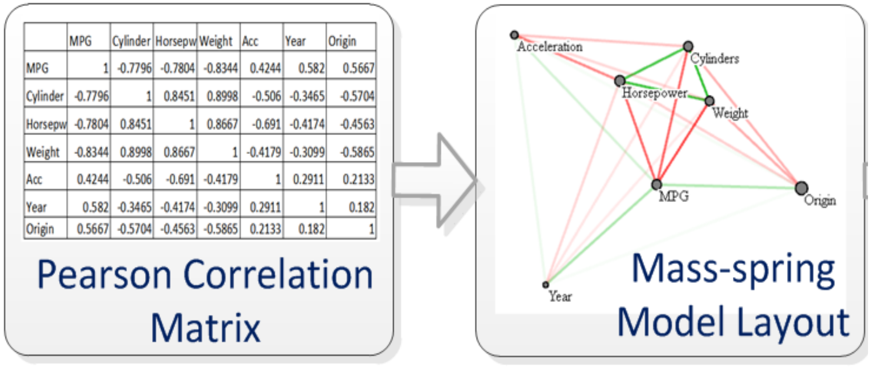
  -> an ordering of the PC dimensions

- The shortest path
  - Approximates maximum of total correlation.

- Travelling sales man (TSP) problem!

- A genetic-algorithm-based TSP solver
  - Balance between performance and accuracy.

Network-Driven Dimension Ordering



	MPG	Cylinder	Horsepw	Weight	Acc	Year	Origin
MPG	1	-0.7796	-0.7804	-0.8344	0.4244	0.582	0.5667
Cylinder	-0.7796	1	0.8451	0.8998	-0.506	-0.3465	-0.5704
Horsepw	-0.7804	0.8451	1	0.8667	-0.691	-0.4174	-0.4563
Weight	-0.8344	0.8998	0.8667	1	-0.4179	-0.3099	-0.5865
Acc	0.4244	-0.506	-0.691	-0.4179	1	0.2911	0.2133
Year	0.582	-0.3465	-0.4174	-0.3099	0.2911	1	0.182
Origin	0.5667	-0.5704	-0.4563	-0.5865	0.2133	0.182	1

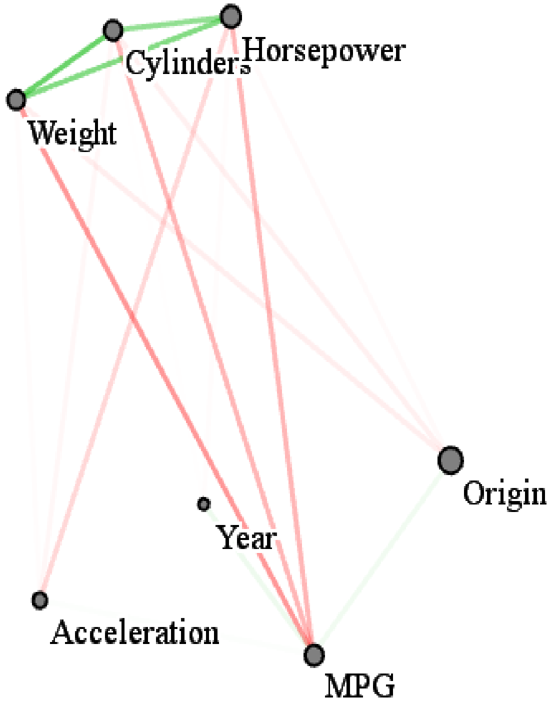
Pearson Correlation Matrix

Mass-spring Model Layout

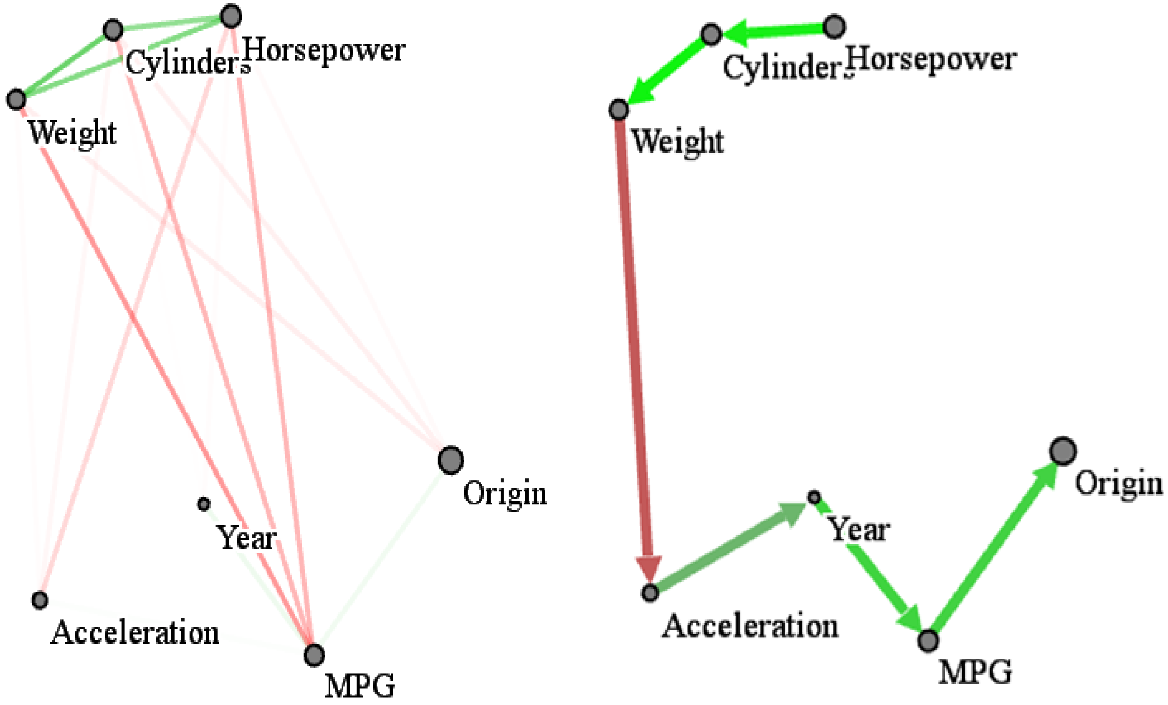
- Determine a default TSP route.
- Interactive route modification via mouse interaction
  - the interface then computes a new optimized route.

- Familiar interaction paradigm
  - Route planning interactions similar to Google Map interface

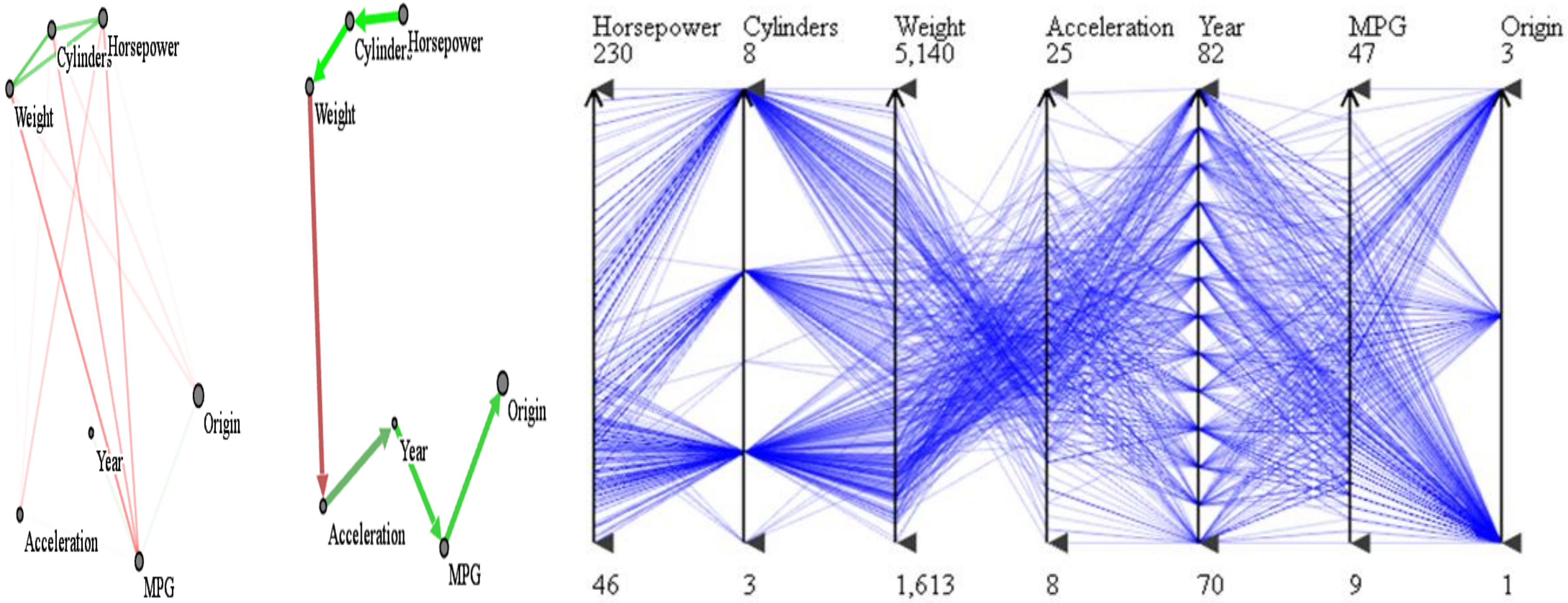
Network-Driven Dimension Ordering



Network-Driven Dimension Ordering

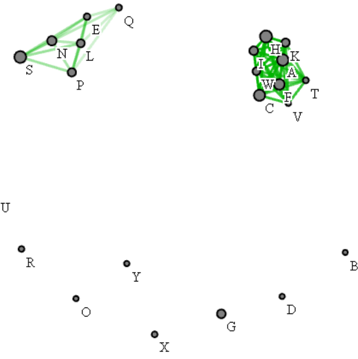


Network-Driven Dimension Ordering

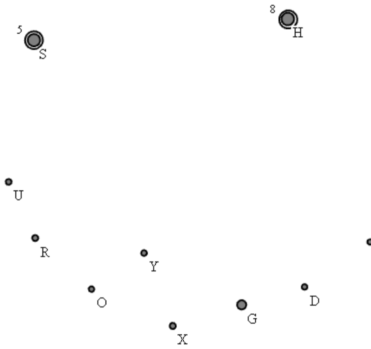


Multi-Scale Zooming in the Network Display

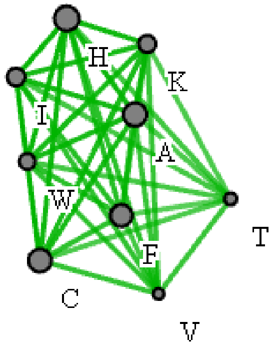
- High-dimensional datasets support:
  - Multi-scale zooming in PC and Network Display



- Correlation-related metric for this purpose.
  - Close dimensions: high correlation.
  - Zoom out: nearby dimensions will merge into one
  - Zoom in: merged dimensions split into the original ones.
  - Similar to popular map exploration.



- Mouse-clicks on merged dimension
  - merge or collapse.



- Zooming extends to the parallel coordinate display

Application – Story Telling

- The PC display provides a sequential view of the data.
  - Dimension – shot/story slice.
  - Reading the plot from left to right -> reading a story from beginning to the end.

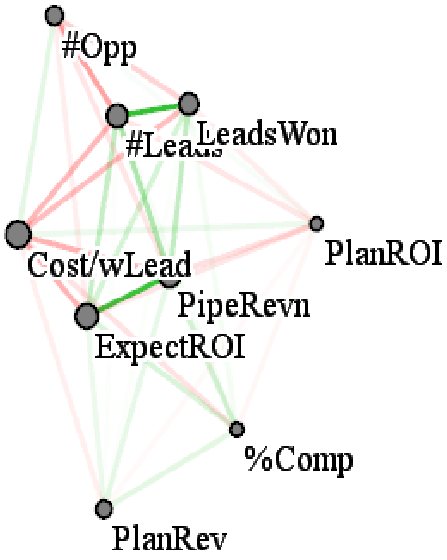
- Network display: Story building board.
  - Enables users to script insightful and informative 'movies' in parallel coordinates.
  - Reveals insightful shot (dimension) sequences
  - Allows automated and manual interactive arrangement of these clips.
  - Arrange a sequence that tells the story

Demonstration Dataset – Sales

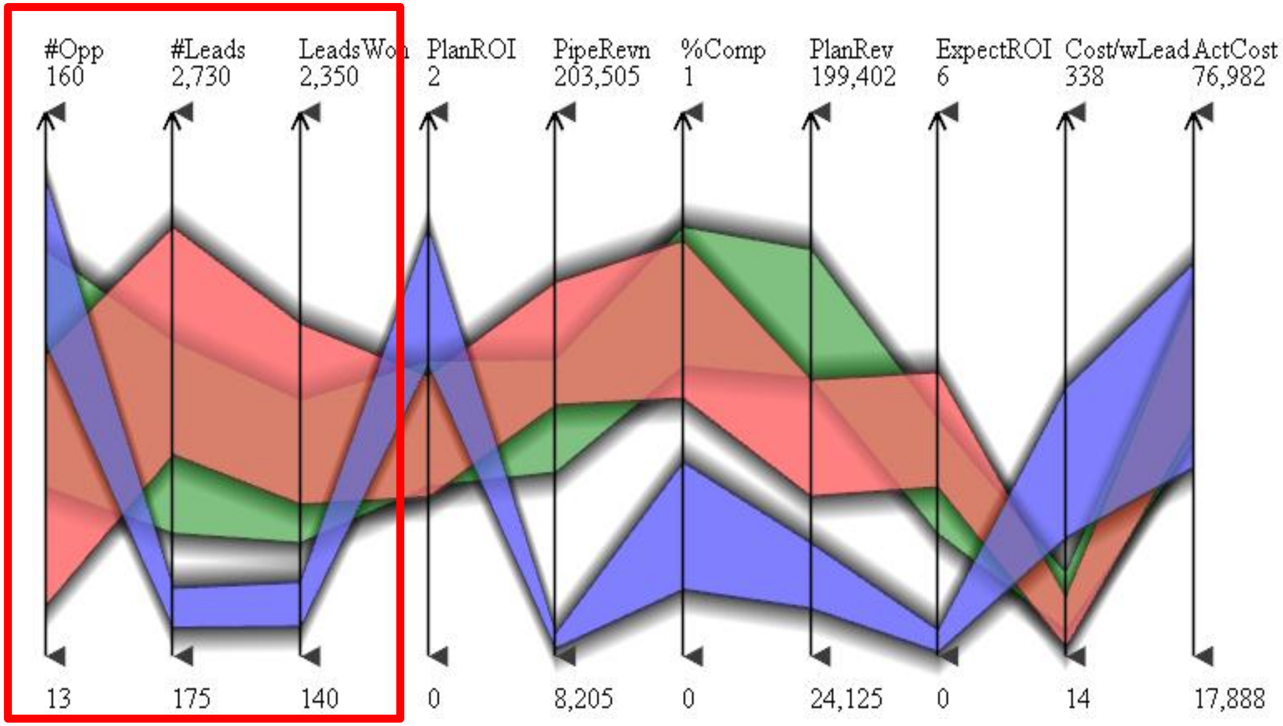
Consists of

- Data for 3 sales teams
- 900 data points (one per sales person)
- 10 attributes, among these:
  - # sales leads generated (# leads)
  - # won leads (# leadsWon)
  - # opportunities resulting from won leads (# opportunities)
  - Cost/Won Lead

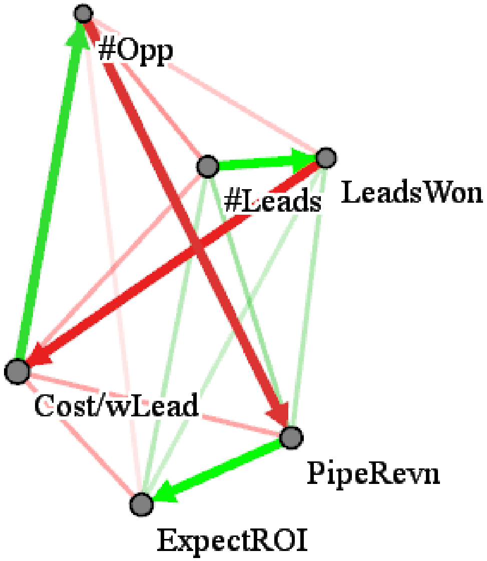
Automatic Dimension Ordering



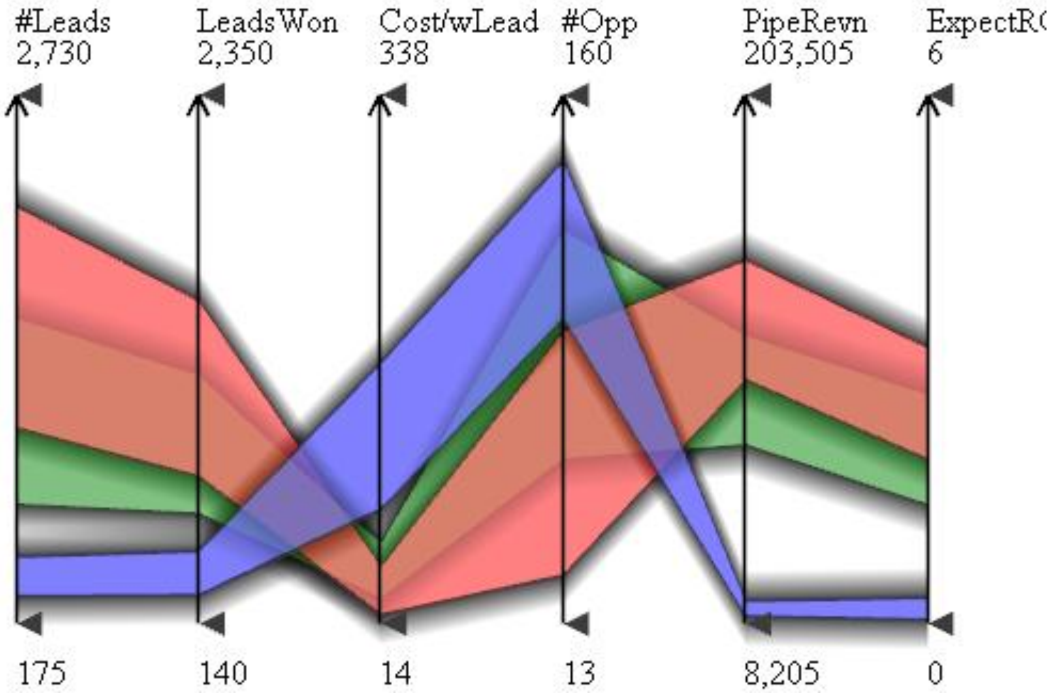
Does not tell the story



Expertise-Guided Ordering



<u>Does</u> tell the story



Sales Dataset

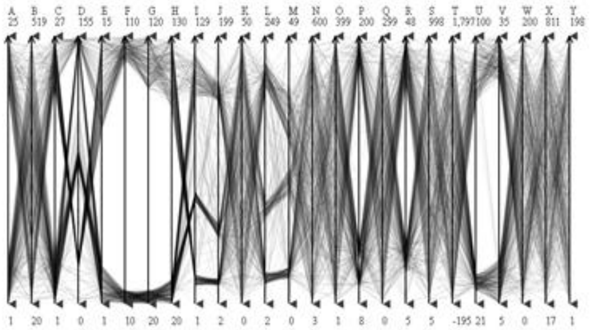
Case analysis with sales campaign dataset

Evaluation – Dimension Ordering

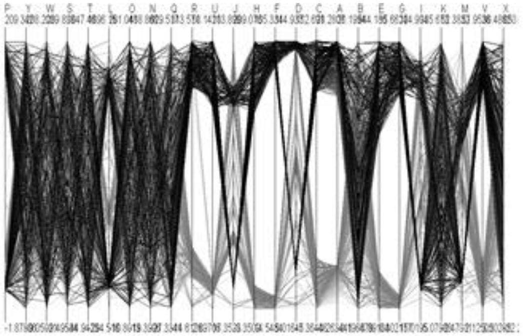
- Synthetic dataset: 25 dims(A,B,C,…,Y) and 1k points

Evaluation – Dimension Ordering

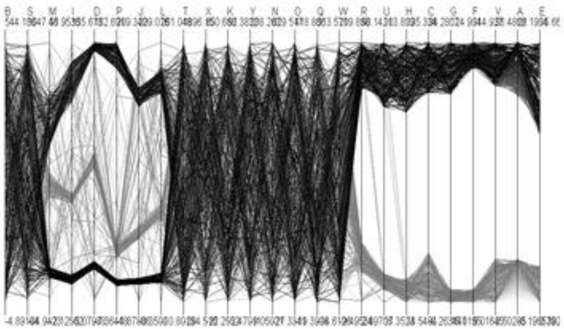
- Synthetic dataset: 25 dims(A,B,C,…,Y) and 1k points



Original



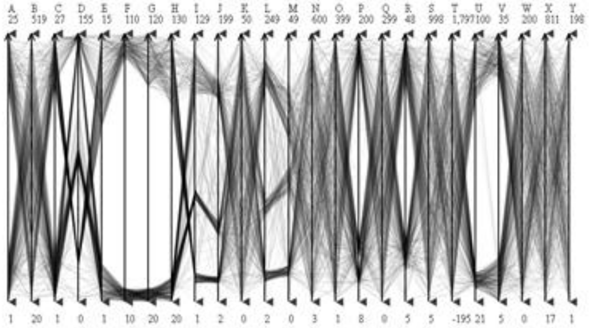
Clutter-based
[Peng et al.]



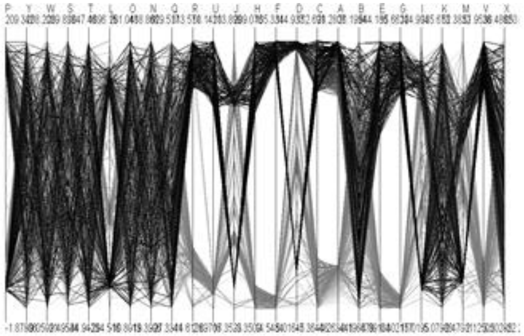
Similarity-based TSP
[Ankerst]

Evaluation – Dimension Ordering

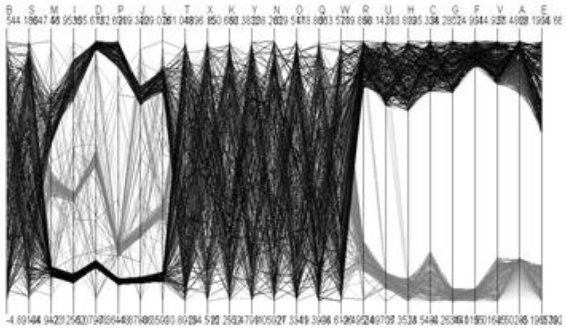
- Synthetic dataset: 25 dims(A,B,C,…,Y) and 1k points



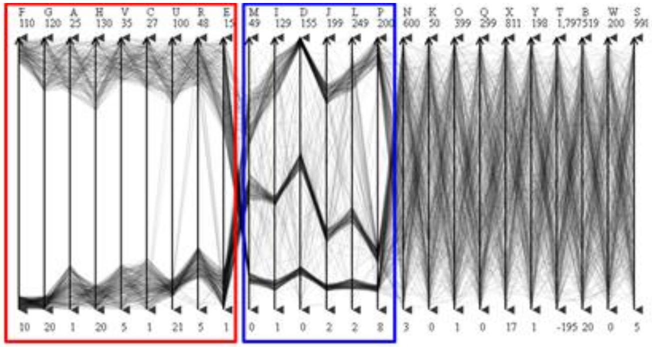
Original



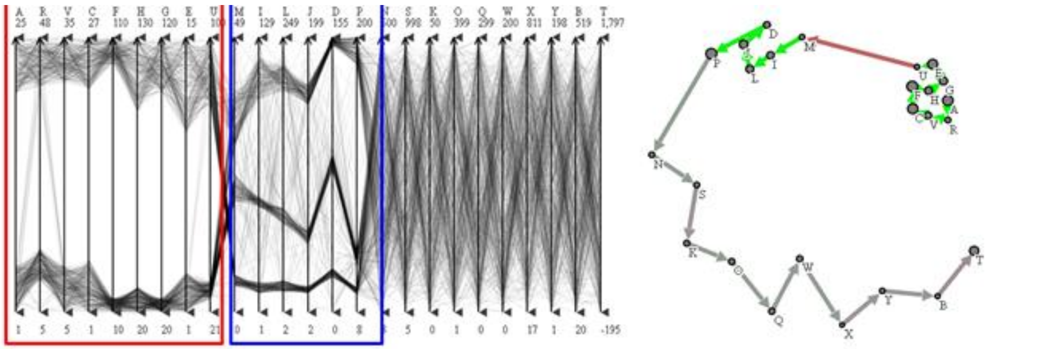
Clutter-based
[Peng 04]



Similarity-based TSP
[Ankerst 98]



Sub-space method
[Ferdosi 11]



Our method
more continuous ordering of the subspace dimensions

Evaluation - Usability

- Sales dataset was used with 18 subjects
- Tested two hypotheses :
  - **H1**. With the help of our network-based display, users are able to find the relationship more accurately.
  - **H2**. With the help of our network-based display, users are able to find the relationship faster.
- Outcome
  - **H1**: Our interface boosted correctness by 2x
  - **H2**: Our interface boosted speed by 2x
- For details please see paper.

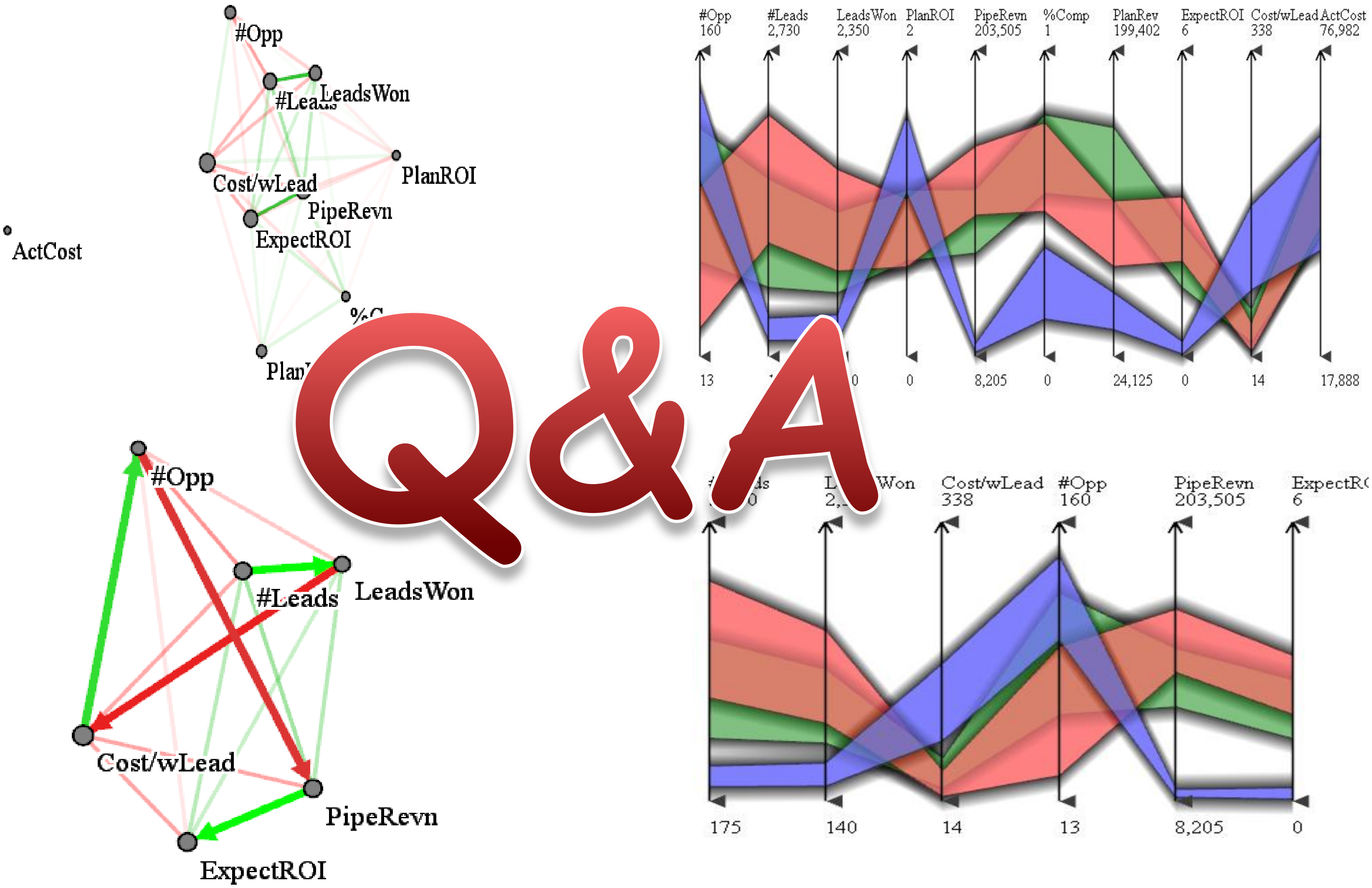
Current/Future Work

Still much work to do:

- Support for categorical data.
- Handle
    - outliers
    - non-linear relationships
    - multicollinearity (related combinations of variables)
    - heteroskedasticity (fanning out)
- Subspace analysis
    - dimensions may appear in more than one subspace.
- Animated PC axes transitions when the order changes.

Demo

Thank you!


#Opp
#Leads
LeadsWon
PlanROI
Cost/wLead
PipeRevn
ExpectROI
ActCost
#Opp 160
#Leads 2,730
LeadsWon 2,350
PlanROI 2
PipeRevn 203,505
%Comp 1
PlanRev 199,402
ExpectROI 6
Cost/wLead 338
ActCost 76,982
13
0
0
8,205
0
24,125
0
14
17,888
#Opp
#Leads
LeadsWon
Cost/wLead
PipeRevn
ExpectROI
Cost/wLead 338
#Opp 160
PipeRevn 203,505
6
175
140
14
13
8,205
0

Q&A

User Study - Setup

- 18 graduate students (none majored in business).
- First we spent about 20 minutes to give them an introduction to our framework. We used the Cars dataset because this domain is the most generally familiar.
- We made sure that after this period all subjects knew the concepts of parallel coordinates and the network display and knew all the interactions supported by our framework.
- Then we randomly split the subjects into two equal-sized groups:
  - Group1: only used the PC display along with the raw data table
  - Group2: PC display + Network display
- We then asked each subject to select the attribute in the sales dataset that best explained the scenario elaborated on in Section 5.

User Study - Result

In Group1, 3 students found the correct answer, i.e. *cost/wonLead.* In Group2, 7 students picked *cost/wonLead* because this attribute is the closest one with a dark red edge to *#leads* and *#leadsWon.* 1 student picked 3 attributes (*cost/wonLead, pipelineRev, and plannedROI)* which are nearby and said the scenario might be caused by the combination of them (regarded as 1/3 correct). So in this case we observed 7.33 (7+1/3) students with the right answer, more than twice than in Group1. Therefore our network display clearly helped. The corresponding p-value is 0.039, which means Hypothesis 1 is confirmed.

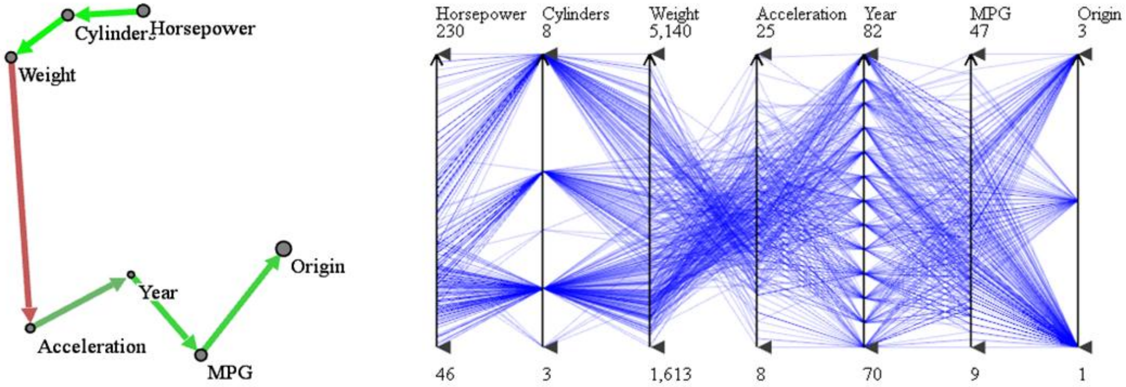
To test Hypothesis 2, we used an independent two-sample t-test based on equal sample sizes and equal variance. On average, participants spent more time to find the answers in Group1 (*Mean* = 20.22 seconds) than those in Group2 (*Mean* = 11.56 seconds). The corresponding t-value is 2.85 and p-value=0.018. For 18 participants (degree of freedom = 16), t must be at least 2.12 to reach $p < 0.05$, so this difference was statistically significant.

Also, among the 18 students, 11 of them claimed that it was the first time they had seen a parallel coordinate display. It was interesting to notice that these 11 students asked more questions and spent more time on learning the parallel coordinate system than on the network display. They stated that the network display was quite easy to understand since they had seen similar displays before. Some mentioned that the network display reminded them of the "Get direction" feature in Google Maps. This insight suggests that our network-based navigation interface is quite accessible, even to novice users.

Interactions on the Network

- Familiar paradigms and metaphors
  - Route planning interactions
- Assign constraints by simple mouse interactions.
  - Specifying edges that should to be maintained or avoided on the route,
  - Vertices that should be avoided.
- Constrained TSP Solver
  - An edge is maintained: all paths that do not pass through the edge will be penalized.
  - An edge is avoided: all paths that pass through this edge will be penalized.
  - A vertex is avoided: remove the vertex from the TSP computation.
  - Starting dimension: the genetic TSP-generated paths will contain only those starting with the specified dimension.

Illustrative Correlation Display in PC



(a) (b)

- Aid less experienced users see correlations on PC.
- A bounding hull of the line bundles
  - Lines' centers and standard deviations.
  - Positive correlation: band-shape bounding hull
  - Negative correlation: bow-tie shape bounding hull
- Saturation: correlation strength