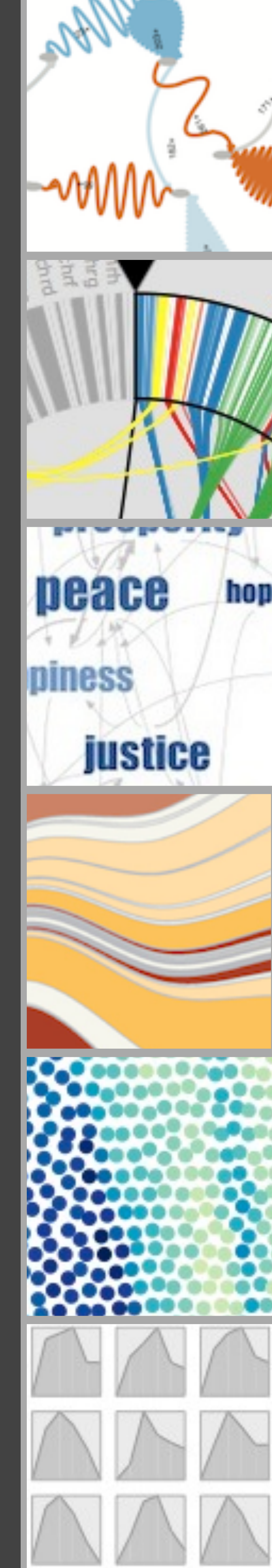
cs6964 | February 23 2012

TABULAR DATA

Miriah Meyer
University of Utah



cs6964 | February 23 2012

TABULAR DATA

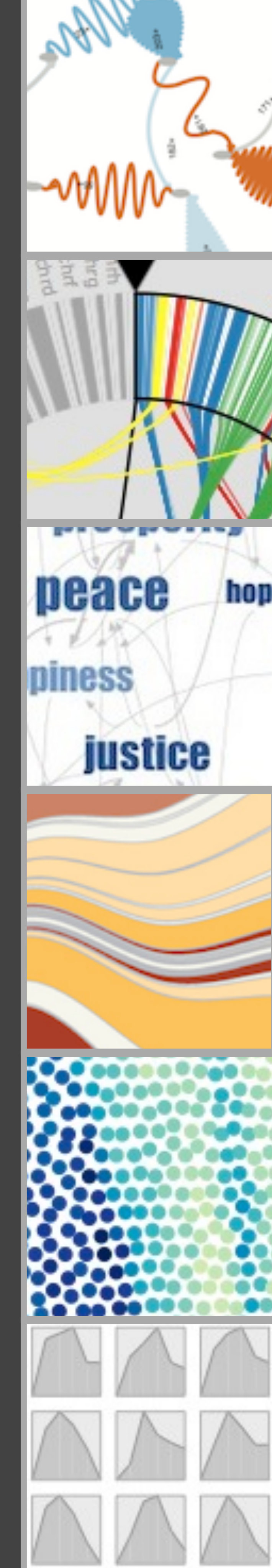
Miriah Meyer

University of Utah

slide acknowledgements:

John Stasko, Georgia Tech

Tamara Munzner, UBC

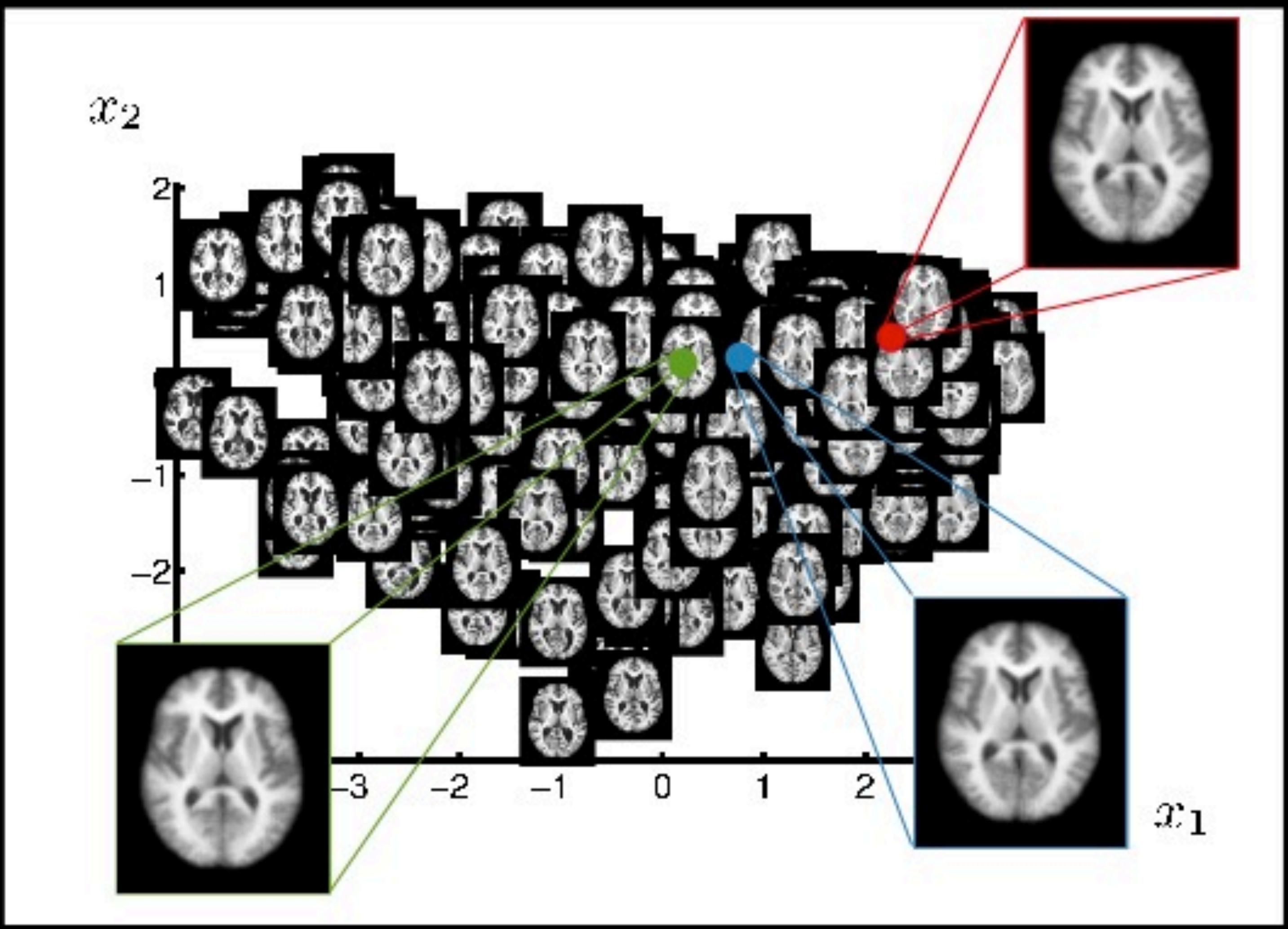


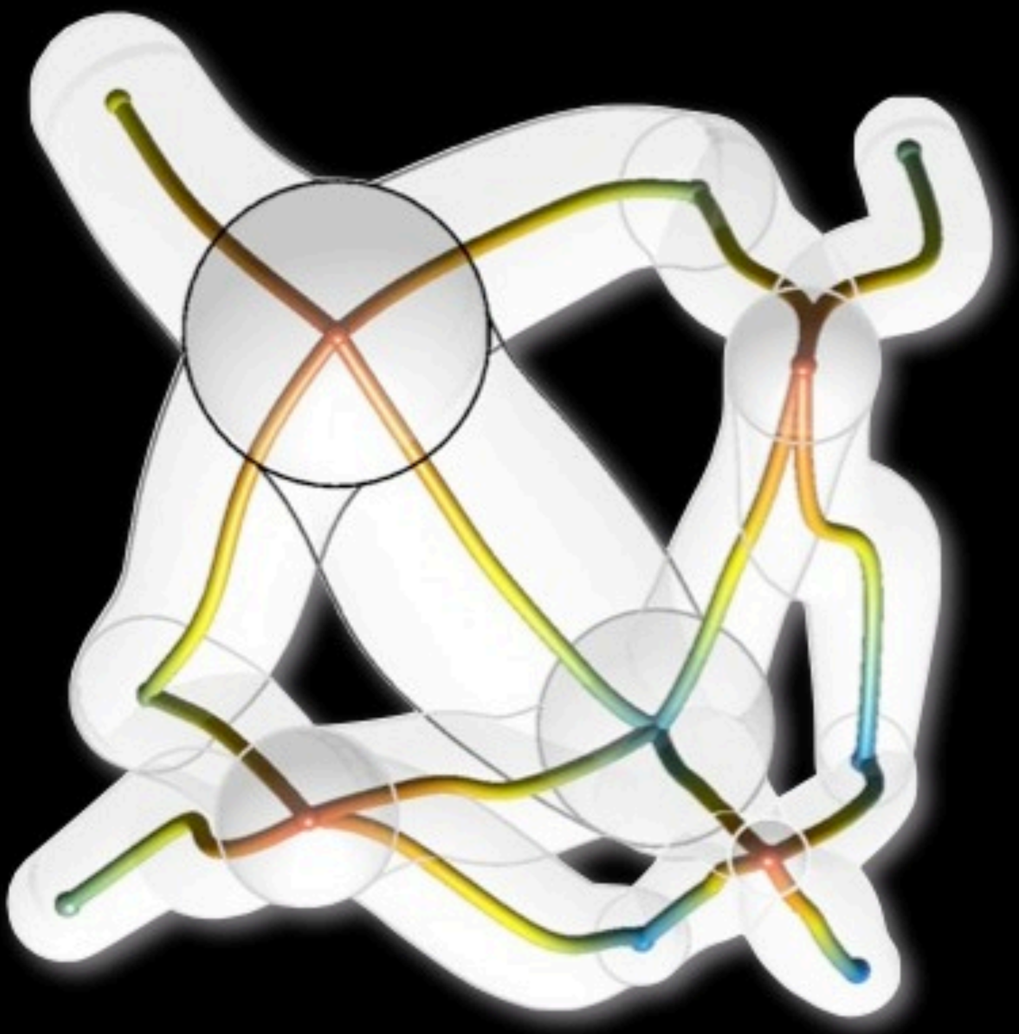
administrivia

- **for final projects that have been approved, email me:**
  - working title
  - group member names
  - two or three sentence description

LAST TIME

Dimension Reduction





Samuel Gerber, University of Utah

SCI INSTITUTE

- curse of dimensionality
- linear methods
- multidimensional scale
- manifold learning

WHERE ARE WE?

- **covered so far**
  - abstractions
  - methods
    - *visual representations*
    - *interactions*
- **next stage: use these ideas for analysis and design**
  - analyze previously proposed techniques and systems
  - design new techniques and systems
- **me: next couple of lectures as examples**
- **you: project proposal and topic presentations**

- multiscale scatterplots
- hierarchical parallel coordinates
- streamgraph

Metric-Based Network Exploration and Multiscale Scatterplot

Yves Chiricota*
Université du Québec à Chicoutimi, Canada

Fabien Jourdan, Guy Melançon†
LIRMM UMR CNRS 5506, Montpellier, France

ABSTRACT

We describe an exploratory technique based on the direct interaction with a 2D modified scatterplot computed from two different metrics calculated over the elements of a network. The scatterplot is transformed into an image by applying standard image processing techniques resulting into blurring effects. Segmentation of the image allows to easily select *patches* on the image as a way to extract sub-networks. We were inspired by the work of Wattenberg and Fisher [21] showing that the blurring process builds into a multiscale perceptual scheme, making this type of interaction intuitive to the user. We explain how the exploration of the network can be guided by the visual analysis of the blurred scatterplot and by its possible interpretations.

CR Categories: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction Techniques I.3.3 [Computer Graphics]: Picture / Image Generation—Viewing algorithms I.4.3 [Image Processing]: Enhancement—Smoothing

Keywords: Graph navigation, exploration, scatterplot, multiscale perceptual organization, clustering, filtering, blurring

1 INTRODUCTION

Part of the research activity in Information Visualization is devoted to exploratory techniques [4, 12]. Indeed, when designing a tool it is important to distinguish whether the user is facing familiar data and is actually using it for a specific task (annotating it or consulting it, for instance) or if she/he is exploring the data trying to find patterns

is to specify a threshold by moving the cursor down (or up) and filter out nodes or edges with a value above (or not exceeding) the threshold. This hiding method gains effectiveness when coupled with a colour map as the elements that are filtered out have a lighter hue and/or lesser intensity, are thiner, etc.

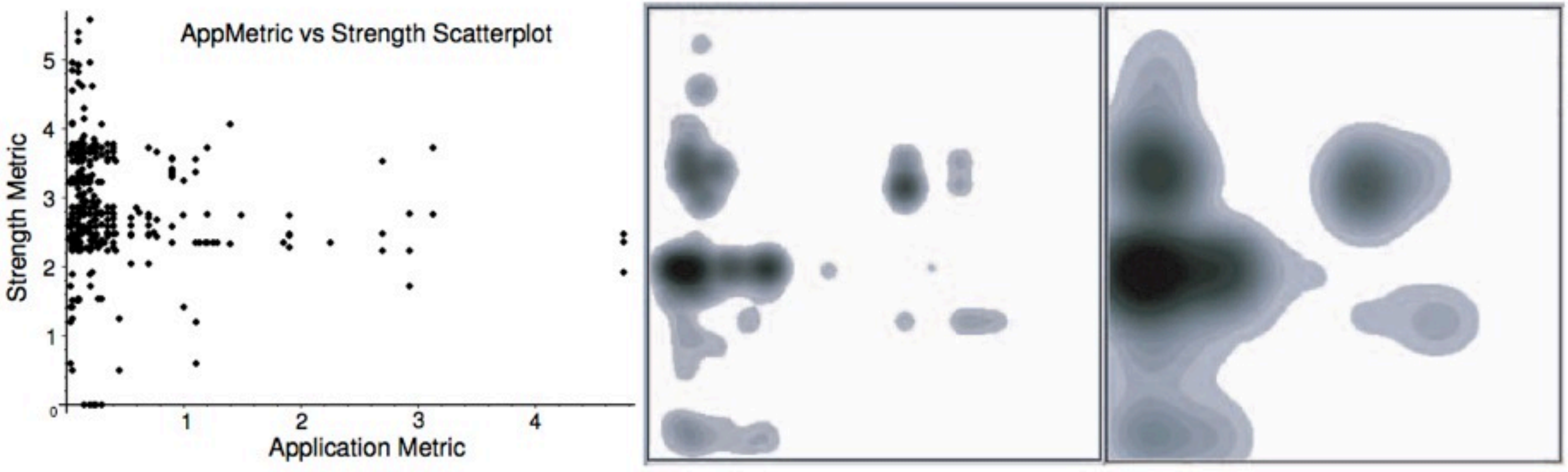
The use of multiple range sliders can help the exploration of a dataset by filtering elements based on a combination of criterion. Williamson and Schneiderman [24] have successfully applied this technique when exploring a real estate database, enabling a user to specify a price range and number of bedrooms, for instance. Barry Becker's MineSet [2] is a tool supporting the exploration of multidimensional databases, helping the user to navigate the data through the selection of range values on several dimensions.

It is unclear whether range selectors are as effective when dealing with less intuitive metrics. What if the values correspond to a *theoretical measure* computed over all nodes of the network, such as for example the so-called clustering index used to define small world networks [22, 23] or the pagerank index of web pages [17] ? What if the values are unevenly distributed over the range they cover ? How should a user manipulate the range selectors to correctly monitor the threshold (filter) ? These observations become even more relevant when dealing with two-dimensional metrics. Situations that are hardly predictable may appear where one slider requires finer tuning depending on the values that were selected using the other. Section 2 provides examples and a more detailed discussion on these issues that were one of the starting point of our work.

The technique we put forward in this paper gives the user direct access to the 2D set of values through a *modified* scatterplot view. More precisely, the view the user acts on is obtained from the actual

MULTISCALE SCATTERPLOTS

- **blur shows structure at multiple scales**
  - convolve with Gaussian
  - slider to control scale parameter interactively
- **easily selectable regions in quantized image**



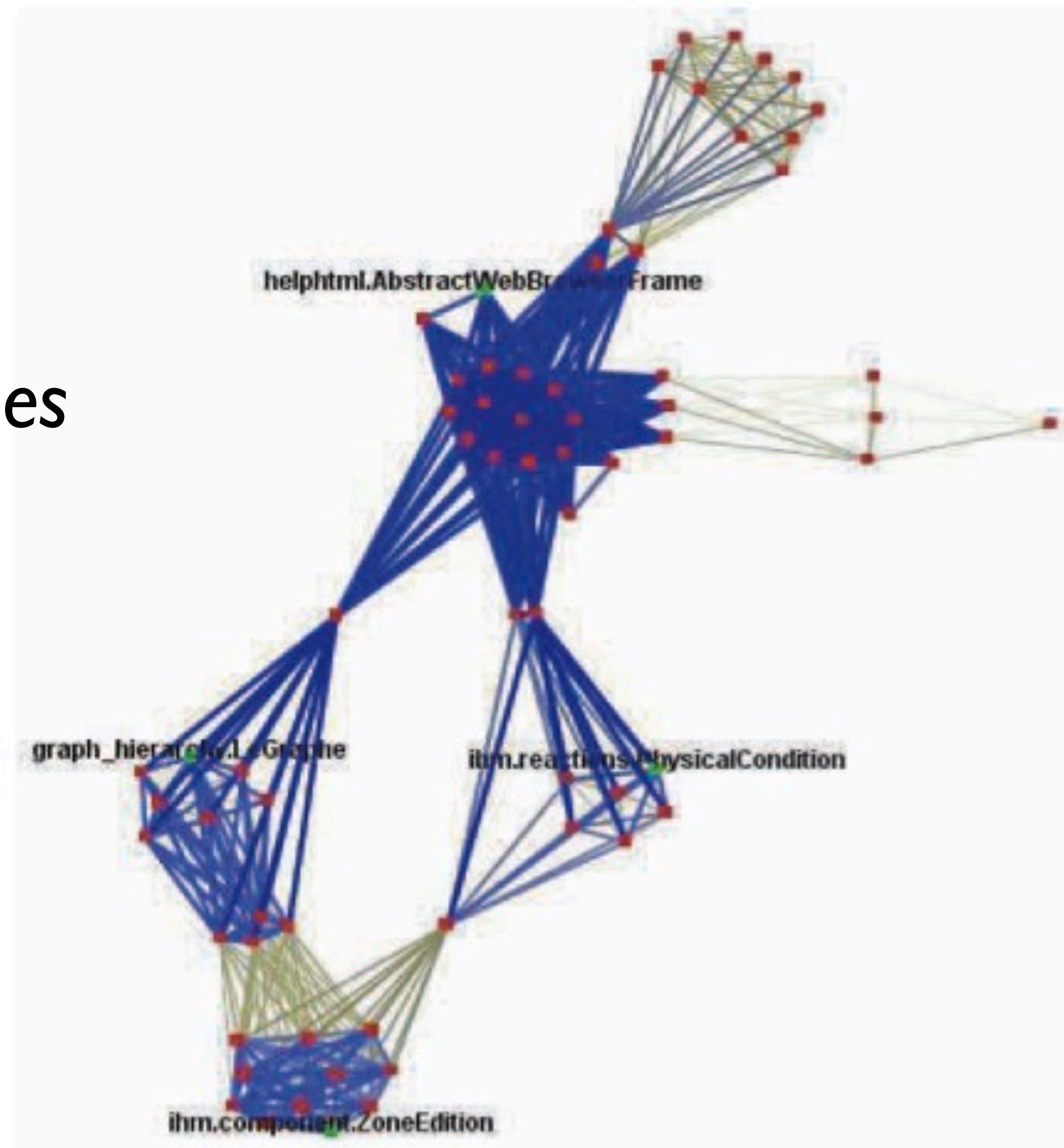
Chiricota 2004

MULTISCALE SCATTERPLOTS

- **problem characterization:**
  - generic network exploration
  - minimal problem context
    - *paper is technique-driven not problem-driven*
- **abstraction**
  - task
    - *selecting and filtering at different scales (within scatterplots)*

DATA ABSTRACTION

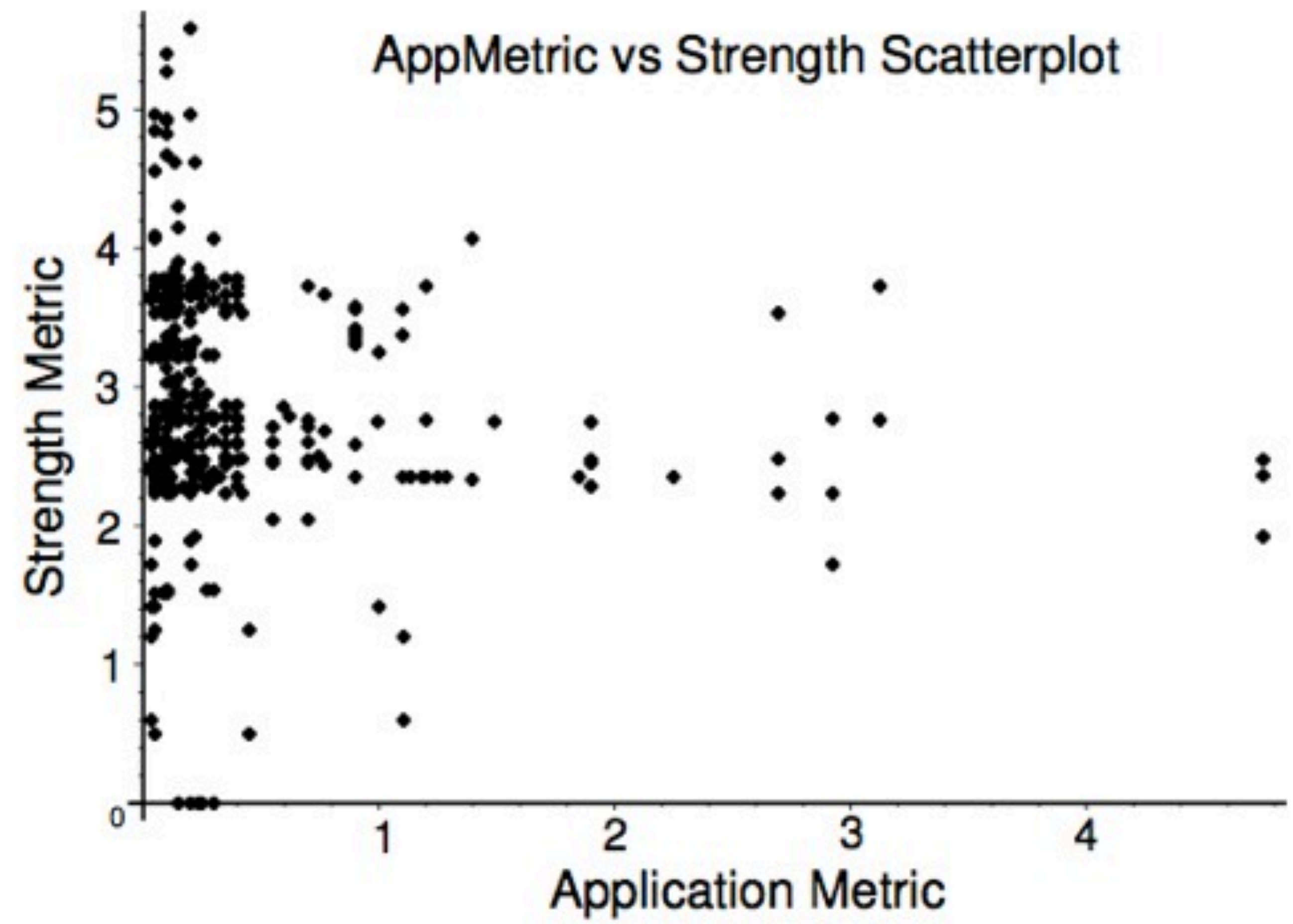
- **original data**
  - relational network
    - *such as links between Java classes*
- **derived attributes**
  - two structure metrics for network
    - *edge width: cluster cohesiveness*
    - *edge color: logical dependencies between classes*
- **thus, table of numbers!**



Chiricota 2004

DESIGN

- **basic solution**
  - visual representation: scatterplots
    - *mark type: points*
    - *channels: horizontal and vertical position*
  - interaction technique: range sliders
    - *filter max / min*
- **challenge**
  - interesting areas might not be easy to select as rectangular bounding box



Chiricota 2004

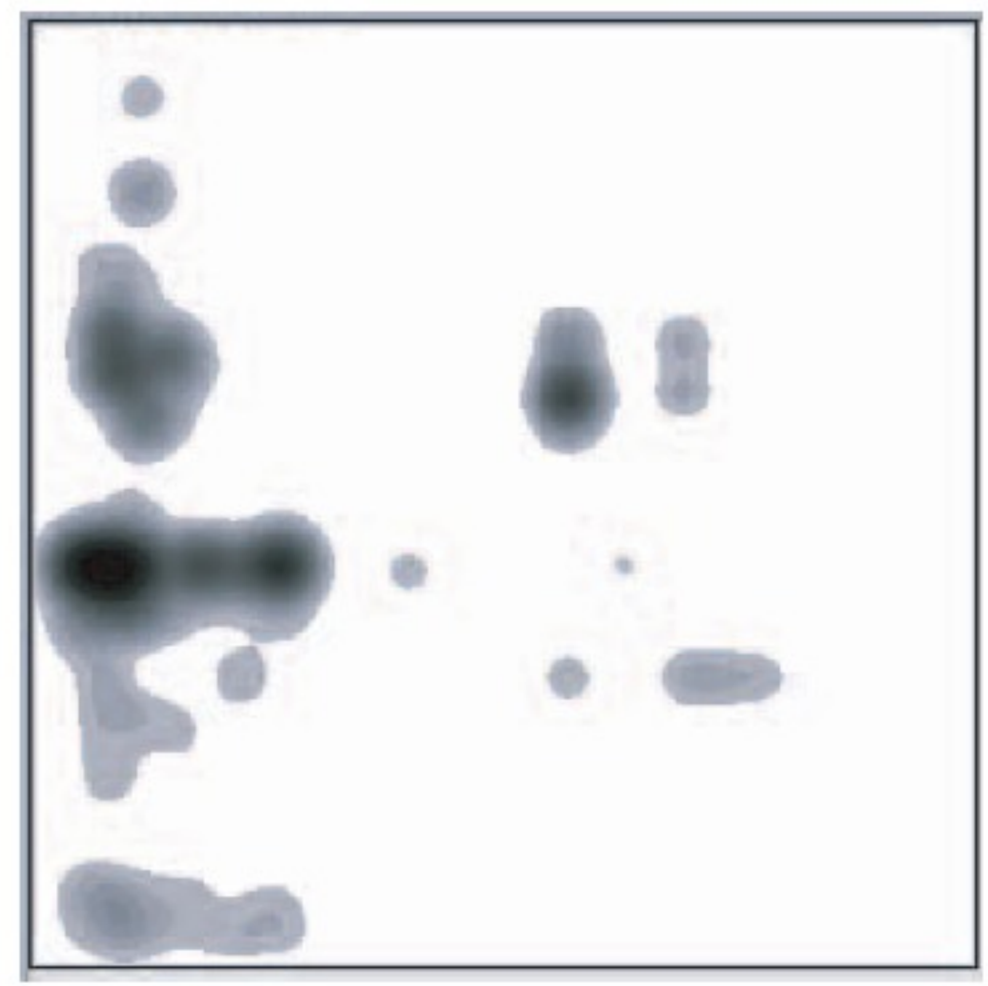
MULTISCALE SCATTERPLOT SELECTION TECHNIQUE

- **new representation**
  - derived space created from original scatterplot image
    - *greyscale patches forming complex shapes*
    - *enclosure of darker patches within lighter patches*
- **new interaction**
  - simple
    - *sliders for filtering size of patch and number of levels*
  - complex
    - *single click to select all items at and below the specified level*

Chiricota 2004

ALGORITHM

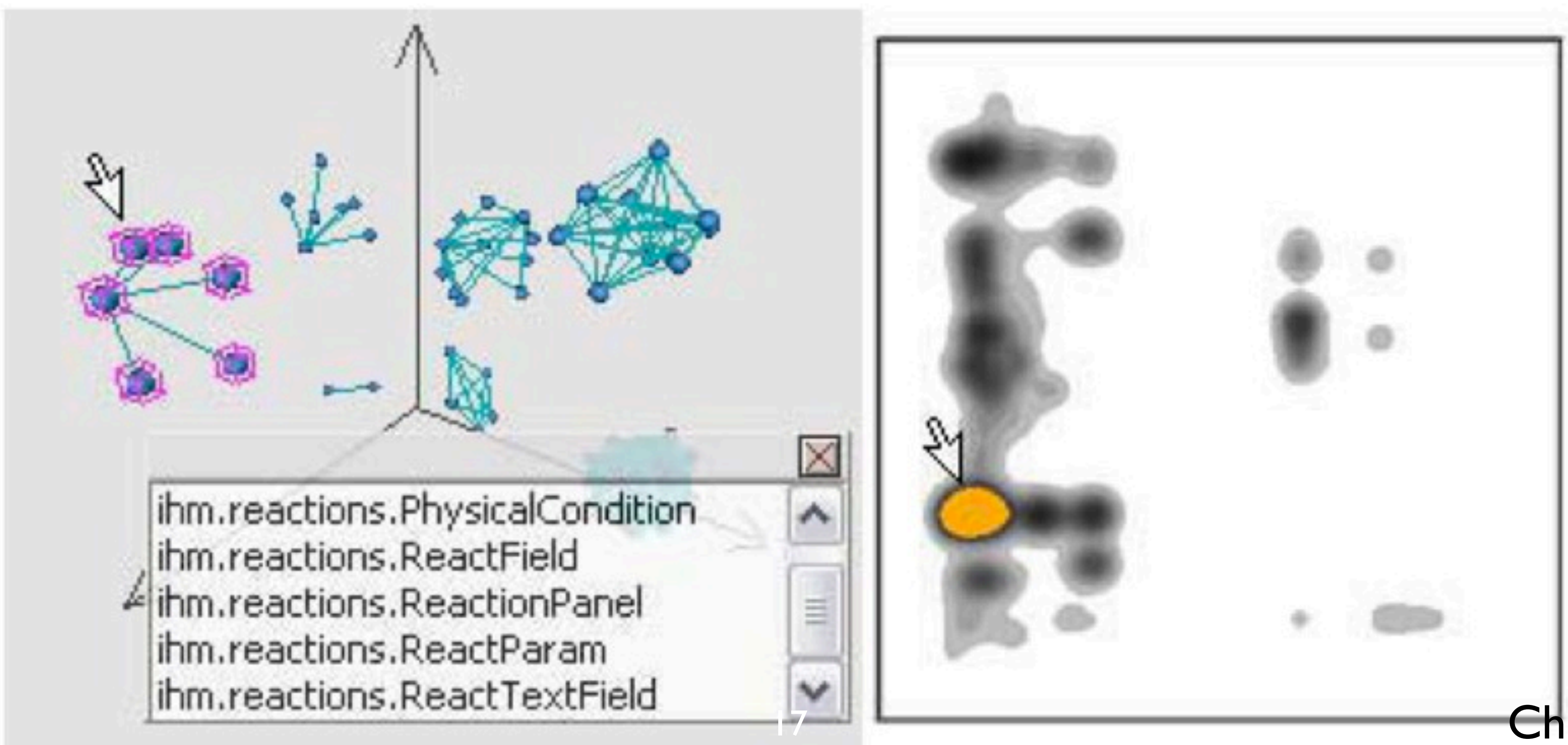
- **creating derived space**
  - greyscale intensity is combination of:
    - *blurred proximity relationships from original scatterplot image: convolve with Gaussian filter*
    - *point density in original scatterplot image*
    - *similar to splatting techniques*
  - quantize image into k levels



Chiricota 2004

METHOD: LINKED VIEWS

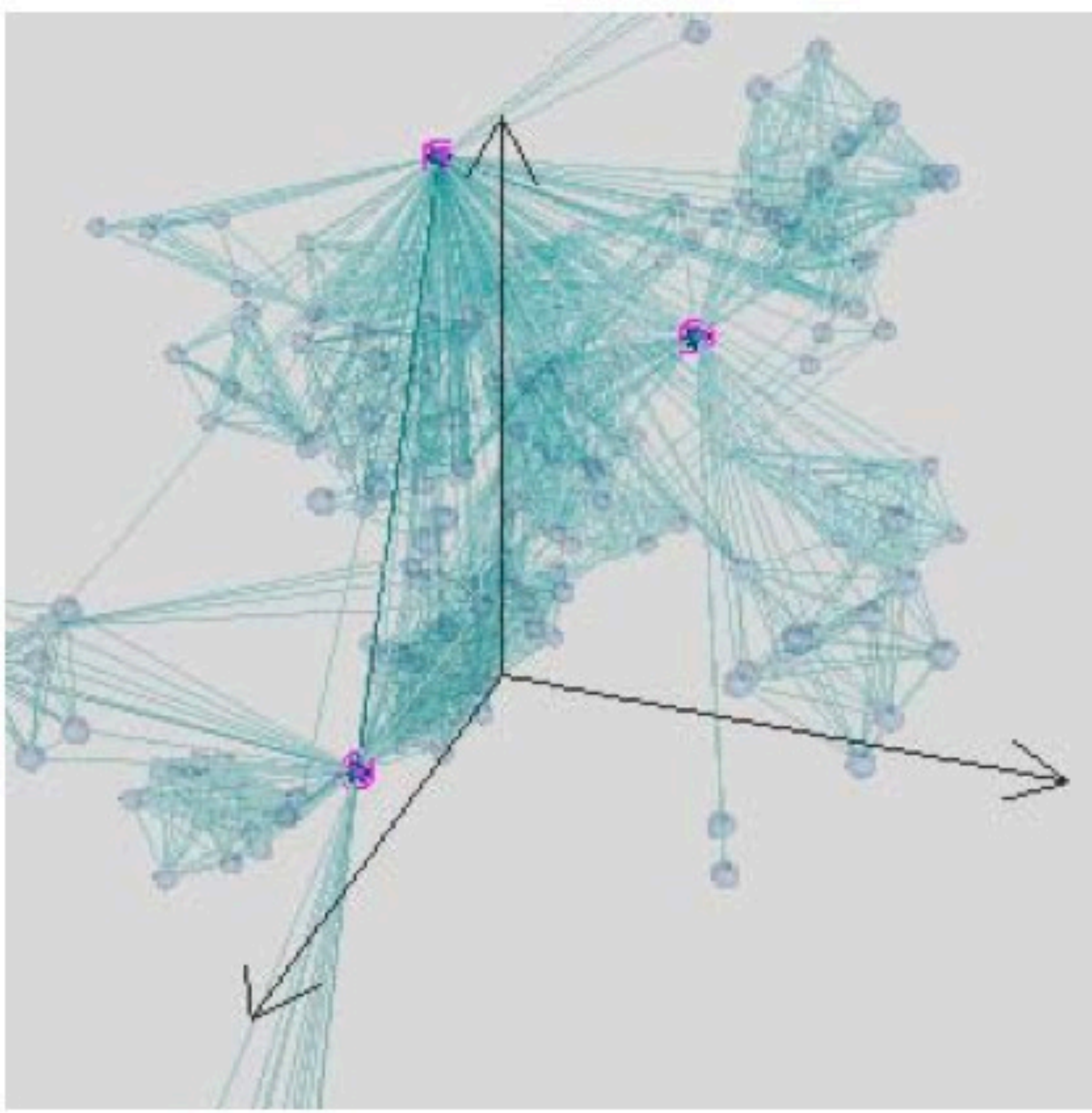
- **linked scatterplot and node-link network view**
  - linked highlighting
  - linked filtering

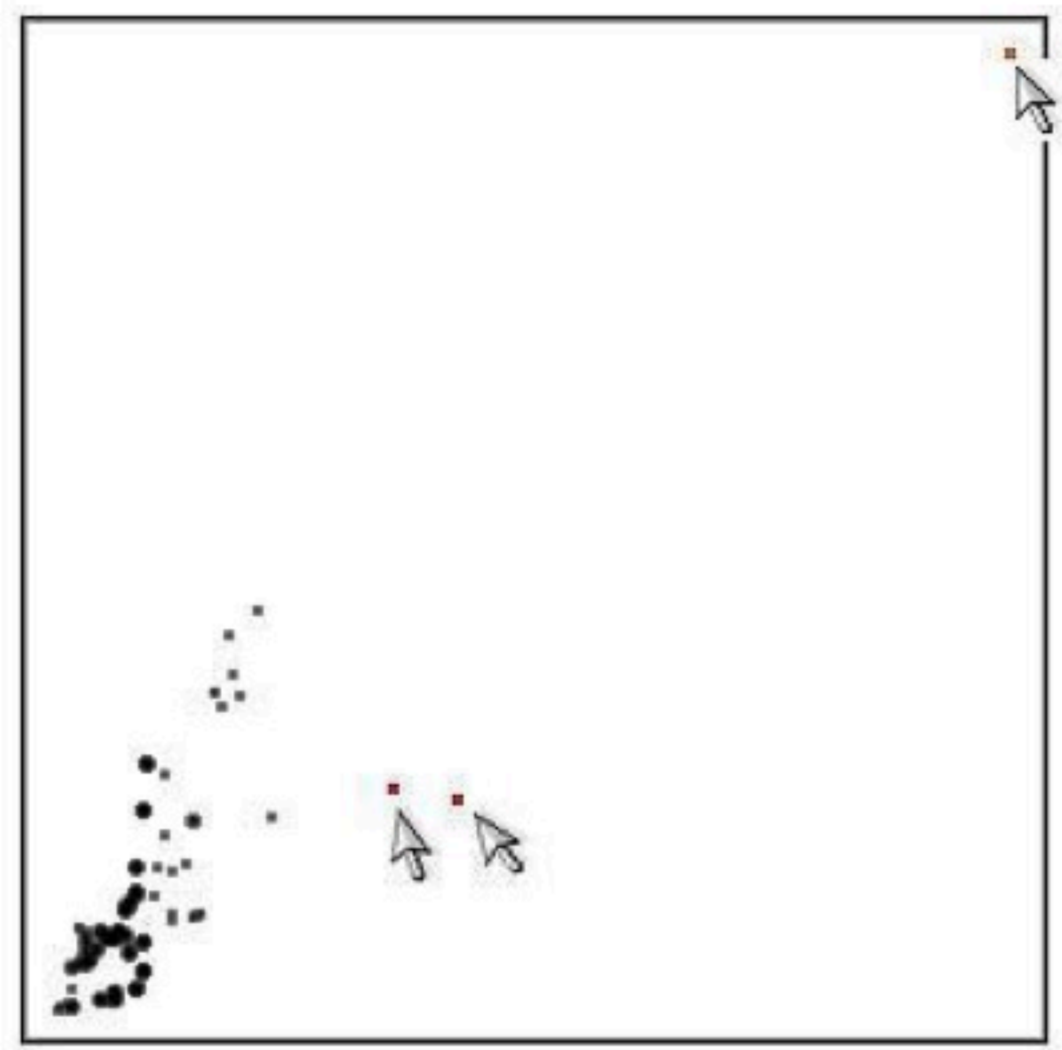


Chiricota 2004

RESULTS: IMDB

- **original data: IMDB graph of actors**
- **metrics: network centrality, node degree**
- **three hubs selected in network view**

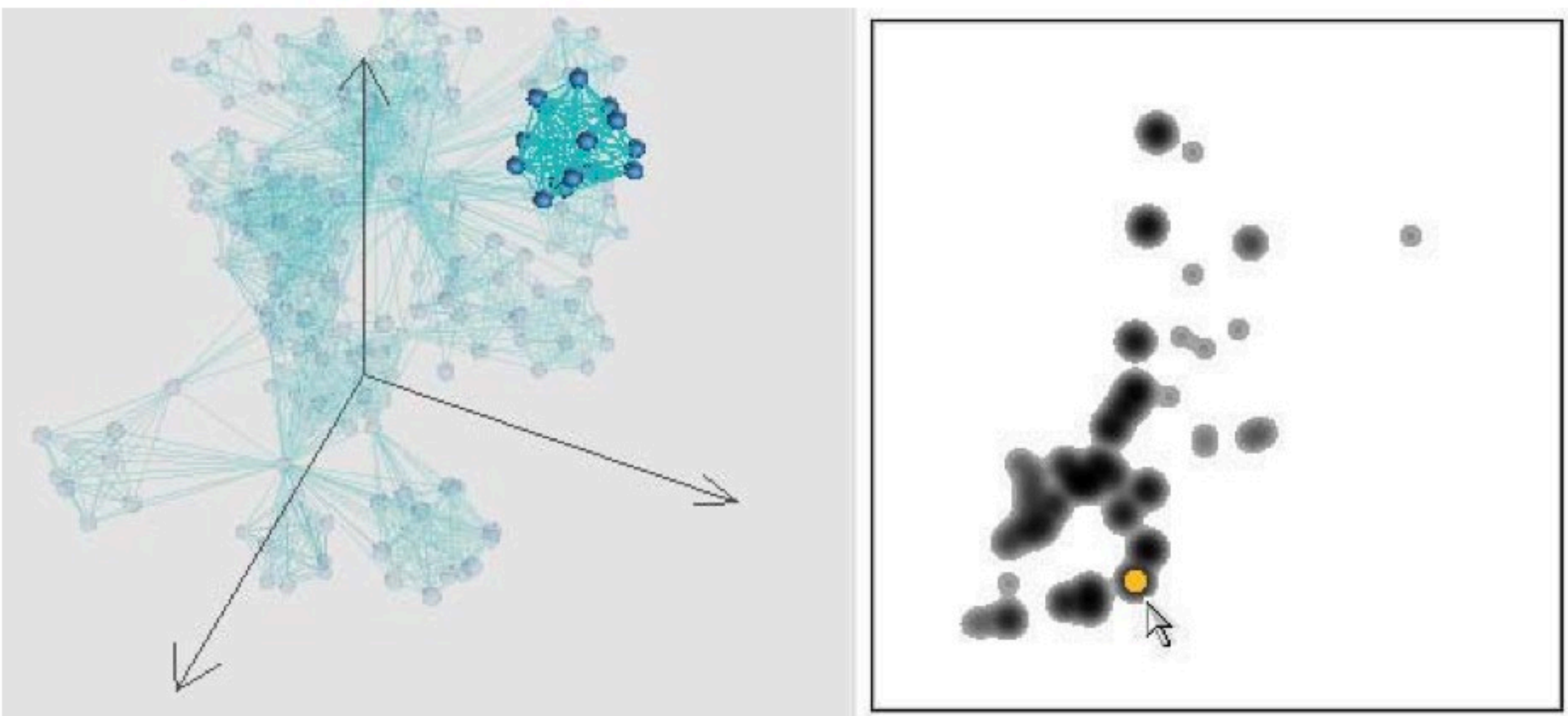




Chiricota 2004

RESULTS: IMDB

- **single click in blurred scatterplot view selects entire clique**



Chiricota 2004

CRITIQUE: what do you think?

CRITIQUE

- **strengths**
  - successful construction and use of derived space
  - appropriate validation
    - *qualitative discussion of result images to show new technique capabilities*
  - synergy between encoding and interaction choices
- **weaknesses**
  - tricky to follow thread of argument
    - *intro/framing focuses on network exploration*
    - *but, fundamental technique contribution more about scatterplot encoding and interaction*

Hierarchical Parallel Coordinates for Exploration of Large Datasets

Ying-Huey Fua, Matthew O. Ward and Elke A. Rundensteiner
Computer Science Department
Worcester Polytechnic Institute
Worcester, MA 01609
{yingfua,matt,rundenst}@cs.wpi.edu *

Abstract

Our ability to accumulate large, complex (multivariate) data sets has far exceeded our ability to effectively process them in search of patterns, anomalies, and other interesting features. Conventional multivariate visualization techniques generally do not scale well with respect to the size of the data set. The focus of this paper is on the interactive visualization of large multivariate data sets based on a number of novel extensions to the parallel coordinates display technique. We develop a multiresolutional view of the data via hierarchical clustering, and use a variation on parallel coordinates to convey aggregation information for the resulting clusters. Users can then navigate the resulting structure until the desired focus region and level of detail is reached, using our suite of navigational and filtering tools. We describe the design and implementation of our hierarchical parallel coordinates system which is based on extending the XmdvTool system. Lastly, we show examples of the tools and techniques applied to large (hundreds of thousands of records) multivariate data sets.

Keywords: Large-scale multivariate data visualization, hierarchical data exploration, parallel coordinates.

1 Introduction

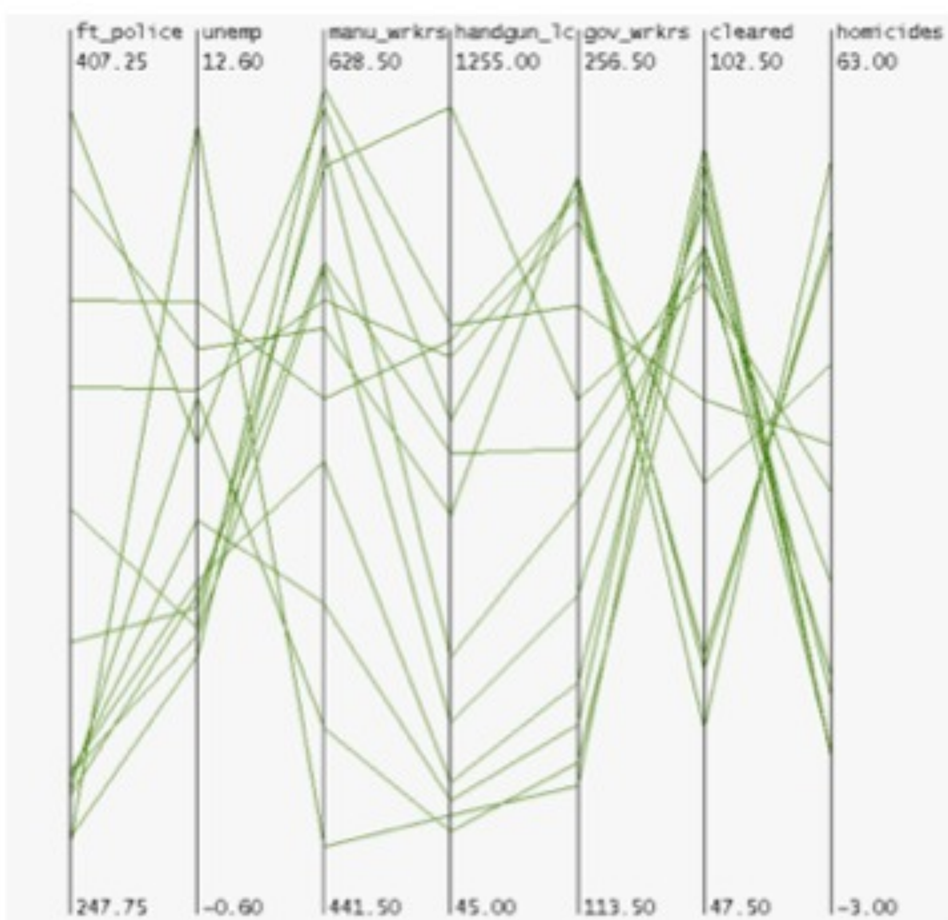
- Dimensional embedding techniques, such as dimensional stacking [16] and worlds within worlds [6].
- Dimensional subsetting, such as scatterplots [5].
- Dimensional reduction techniques, such as multidimensional scaling [20, 15, 29], principal component analysis [12] and self-organizing maps [14].

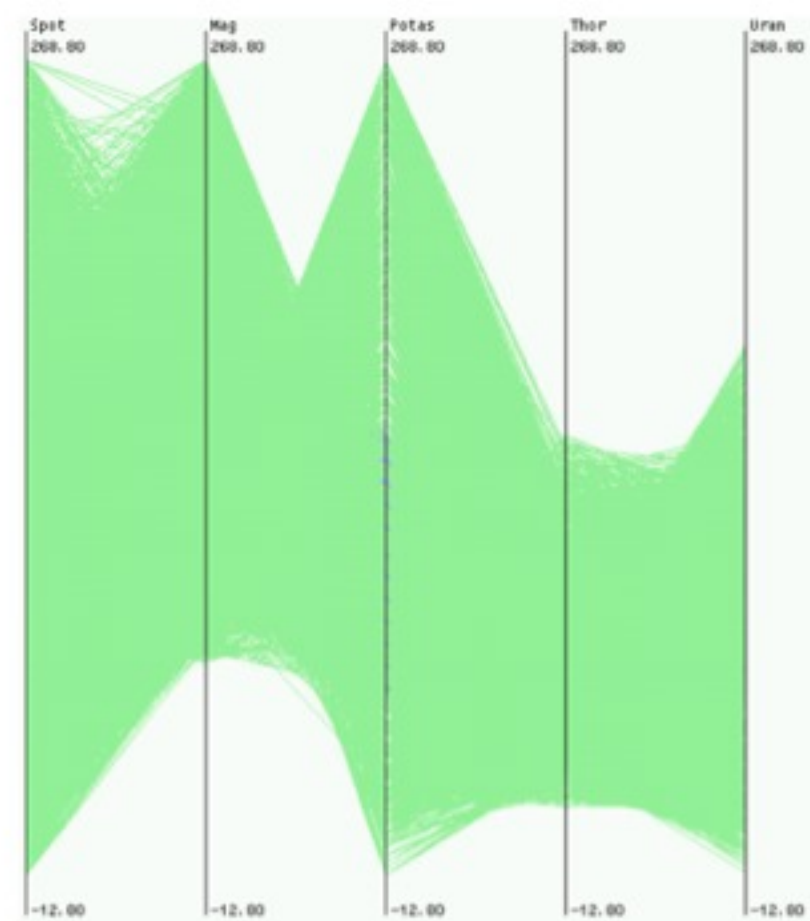
Most of these techniques do not scale well with respect to the size of the data set. As a generalization, we postulate that any method that displays a single entity per data point invariably results in overlapped elements and a convoluted display that is not suited for the visualization of large data sets. The quantification of the term "large" varies and is subject to revision in sync with the state of computing power. For our present application, we define a large data set to contain 10^6 to 10^9 data elements or more.

Our research focus extends beyond just data display, incorporating the process of data exploration, with the goal of interactively uncovering patterns or anomalies not immediately obvious or comprehensible. Our goal is thus to support an active process of discovery as opposed to passive display. We believe that it is only through data exploration that meaningful ideas, relations, and subsequent inferences may be extracted from the data. The major hurdles we need to overcome are the problems of display density/clutter (too

HIERARCHICAL PARALLEL COORDINATES

- **technique-driven paper**
  - no problem characterization
- **goal: scale up parallel coordinates to large datasets**
  - challenge: overplotting/occlusion

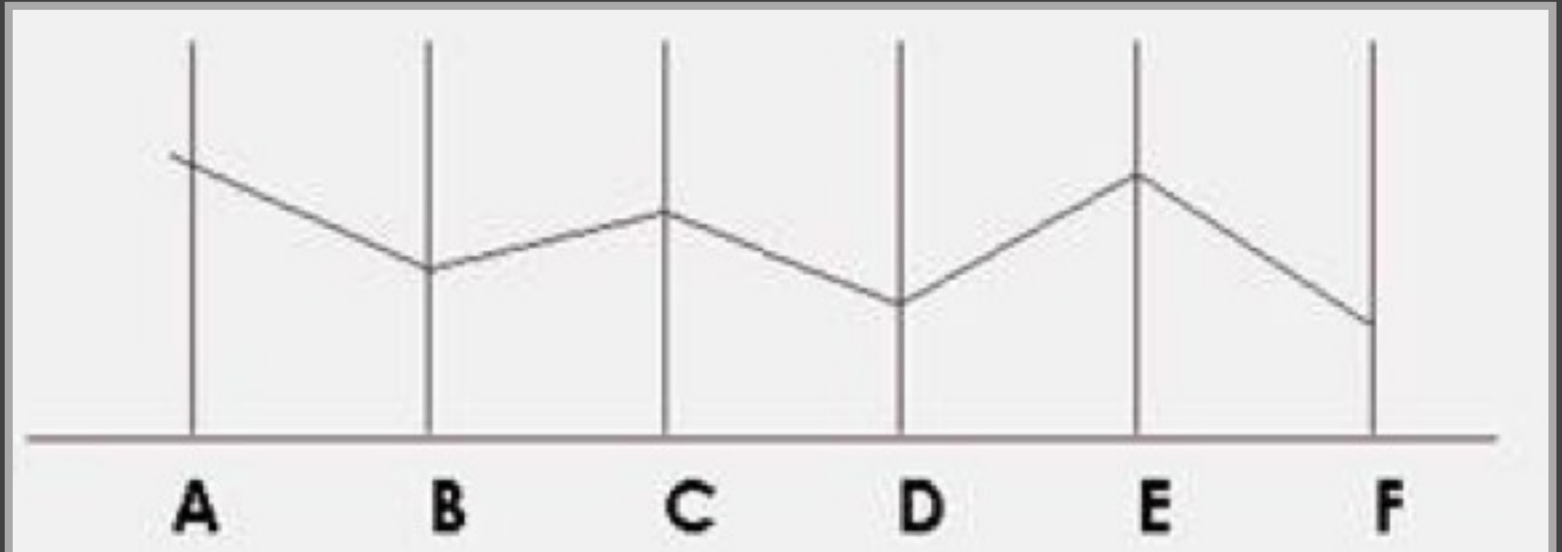




Fua 1999

PARALLEL COORDINATES

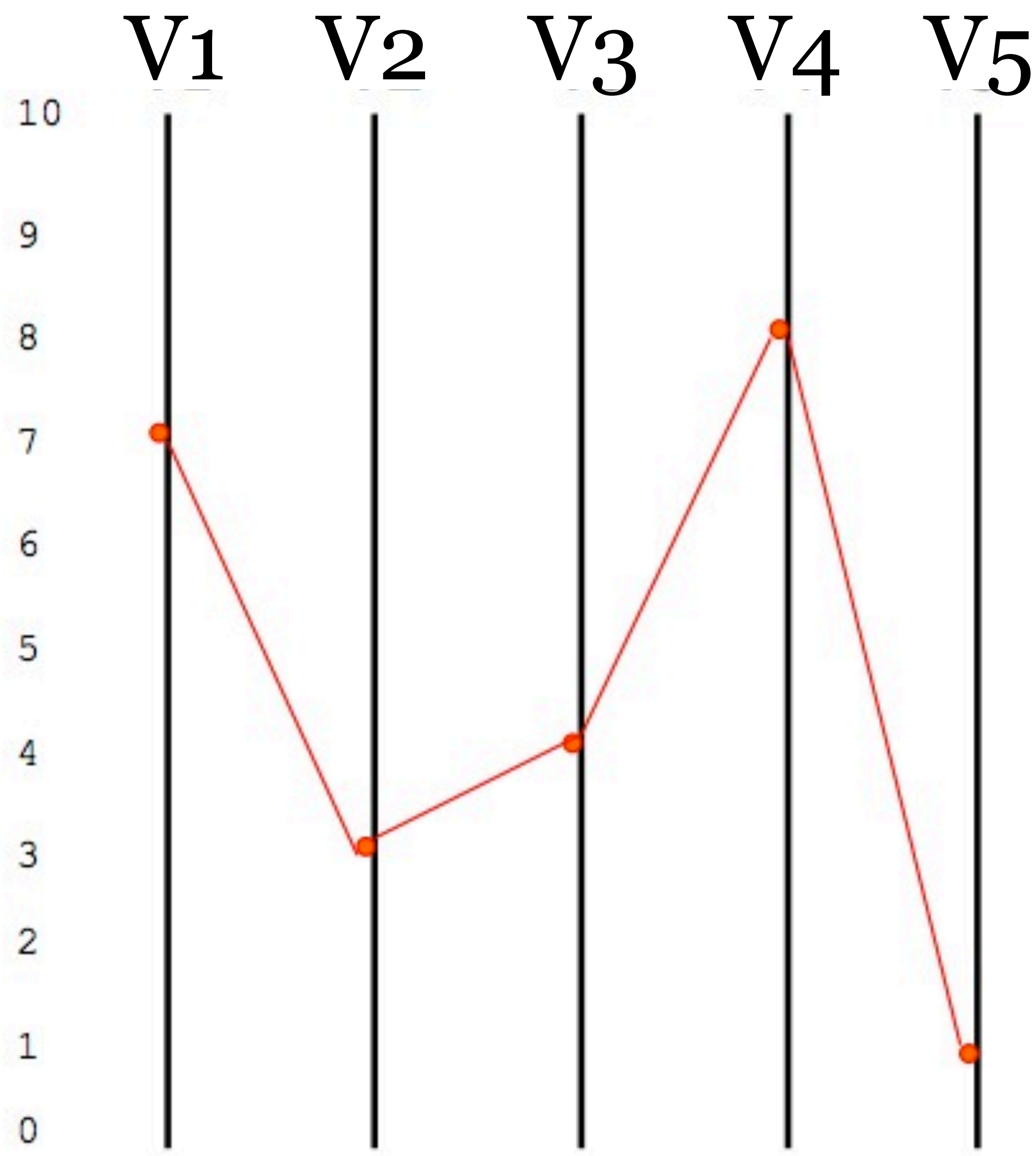
- **scatterplot limitation: visual representation with orthogonal axes**
  - can show only two attributes with spatial position channel
- **alternative: line up axes in parallel to show many attributes with position**
  - item encoded with a line with n segments
    - *n is the number of attributes shown*



EXAMPLE

	V1	V2	V3	V4	V5
D1	7	3	4	8	1
D2	2	7	6	3	4
D3	9	8	1	4	2

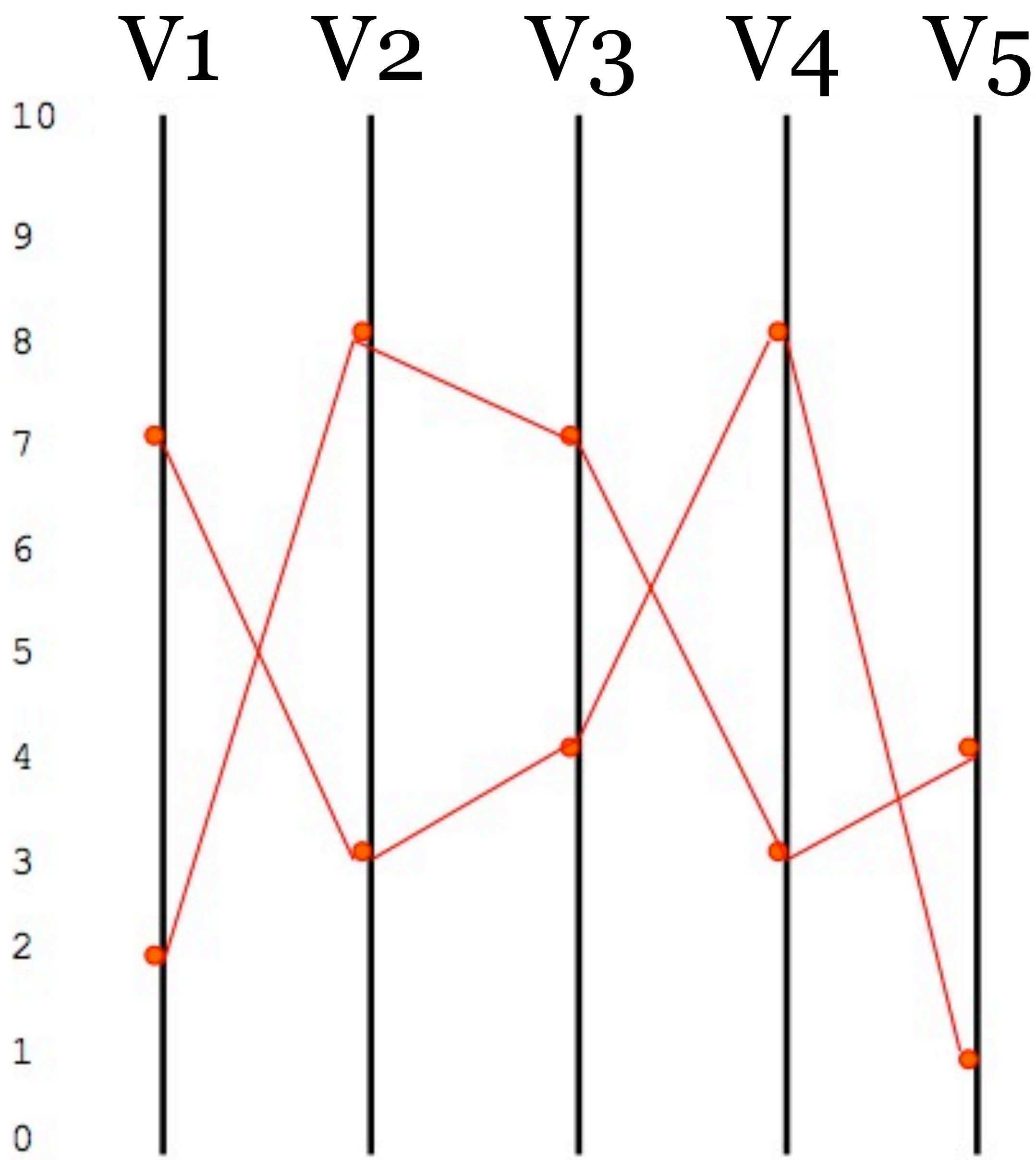
EXAMPLE



D1

7	3	4	8	1

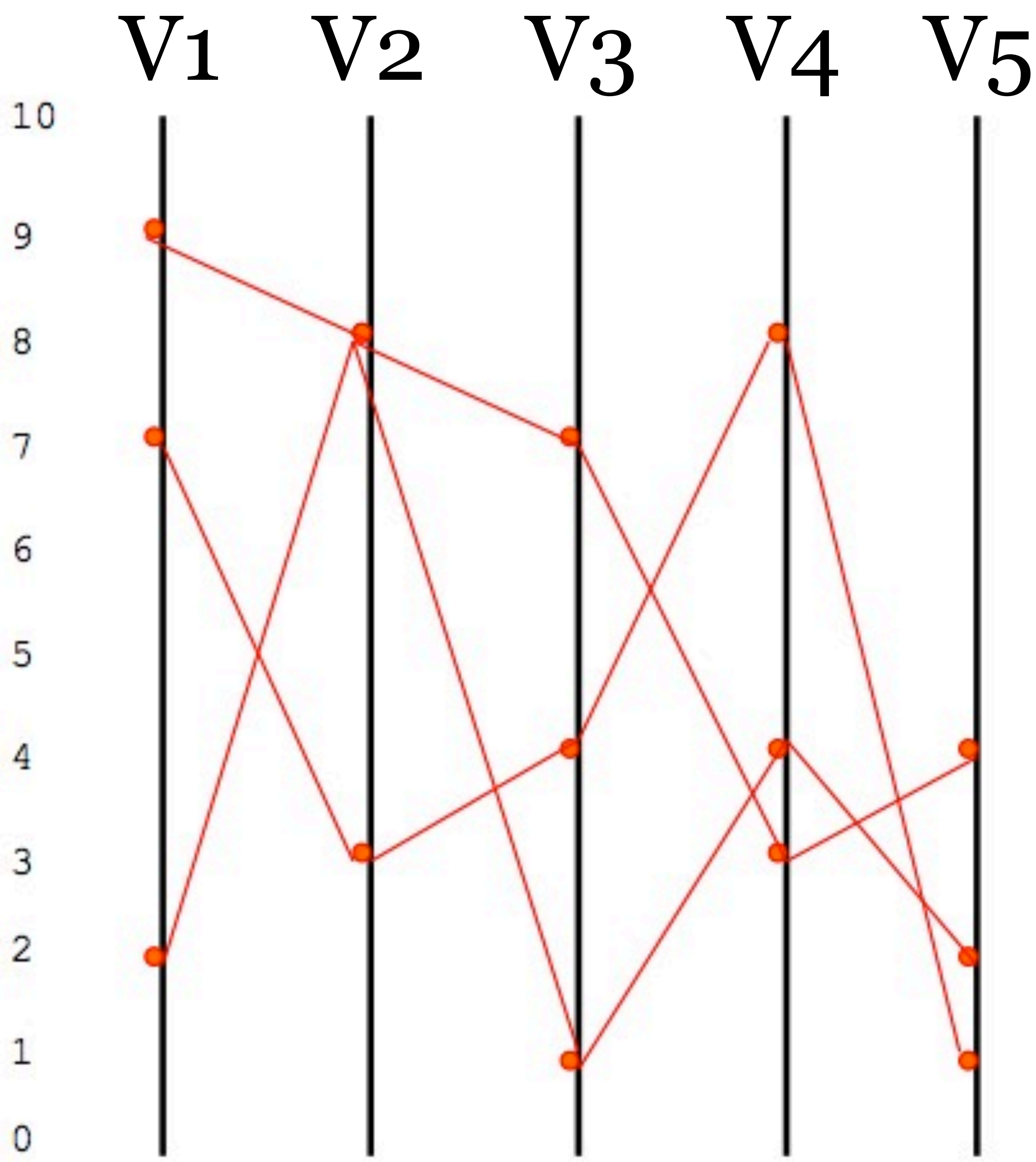
EXAMPLE



D2

2	7	6	3	4

EXAMPLE



D3

9 8 1 4 2

PARALLEL COORDINATES TASK

- **show correlation**
  - positive correlation: straight lines
  - negative correlation: all lines cross at a single pt

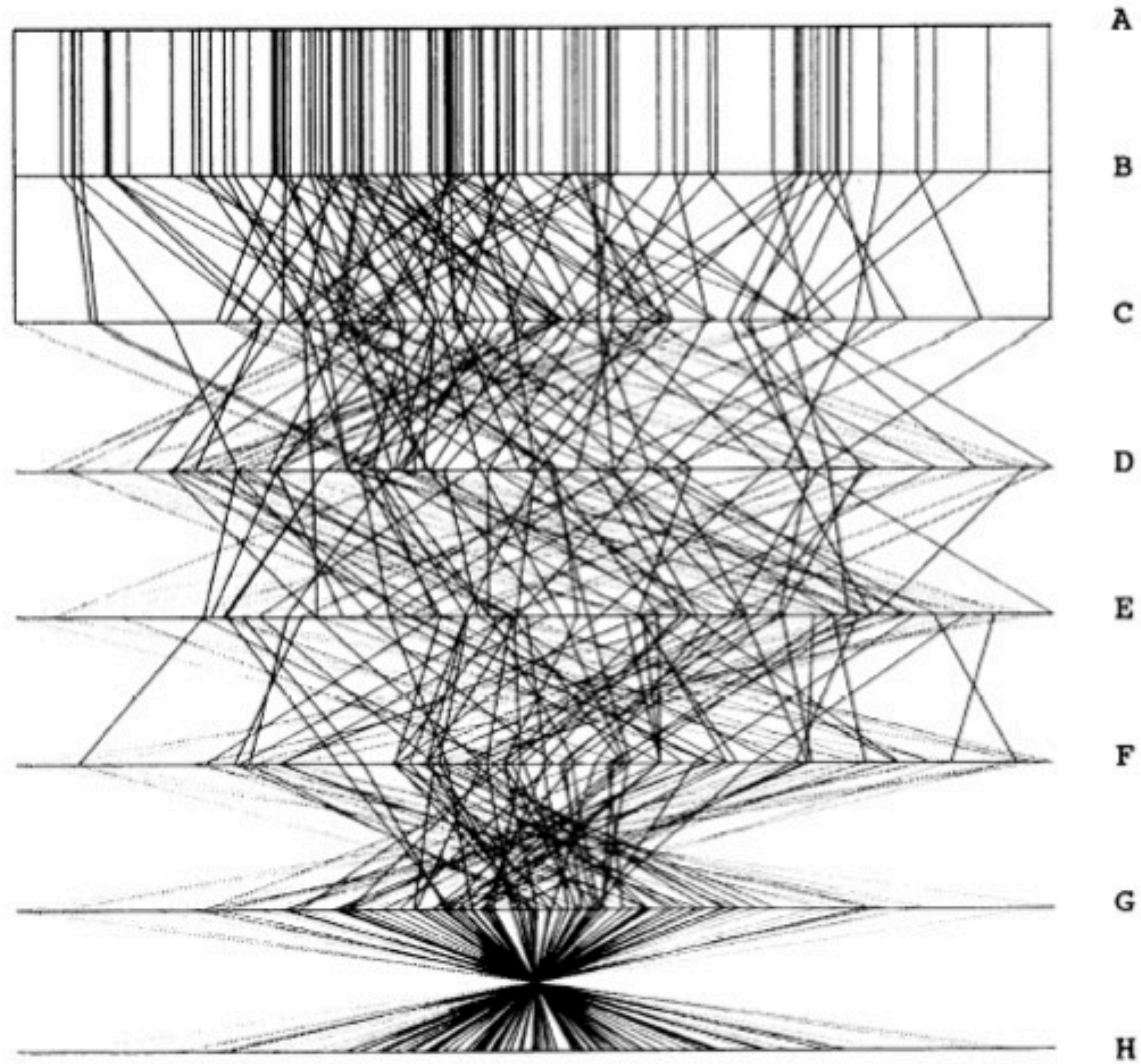
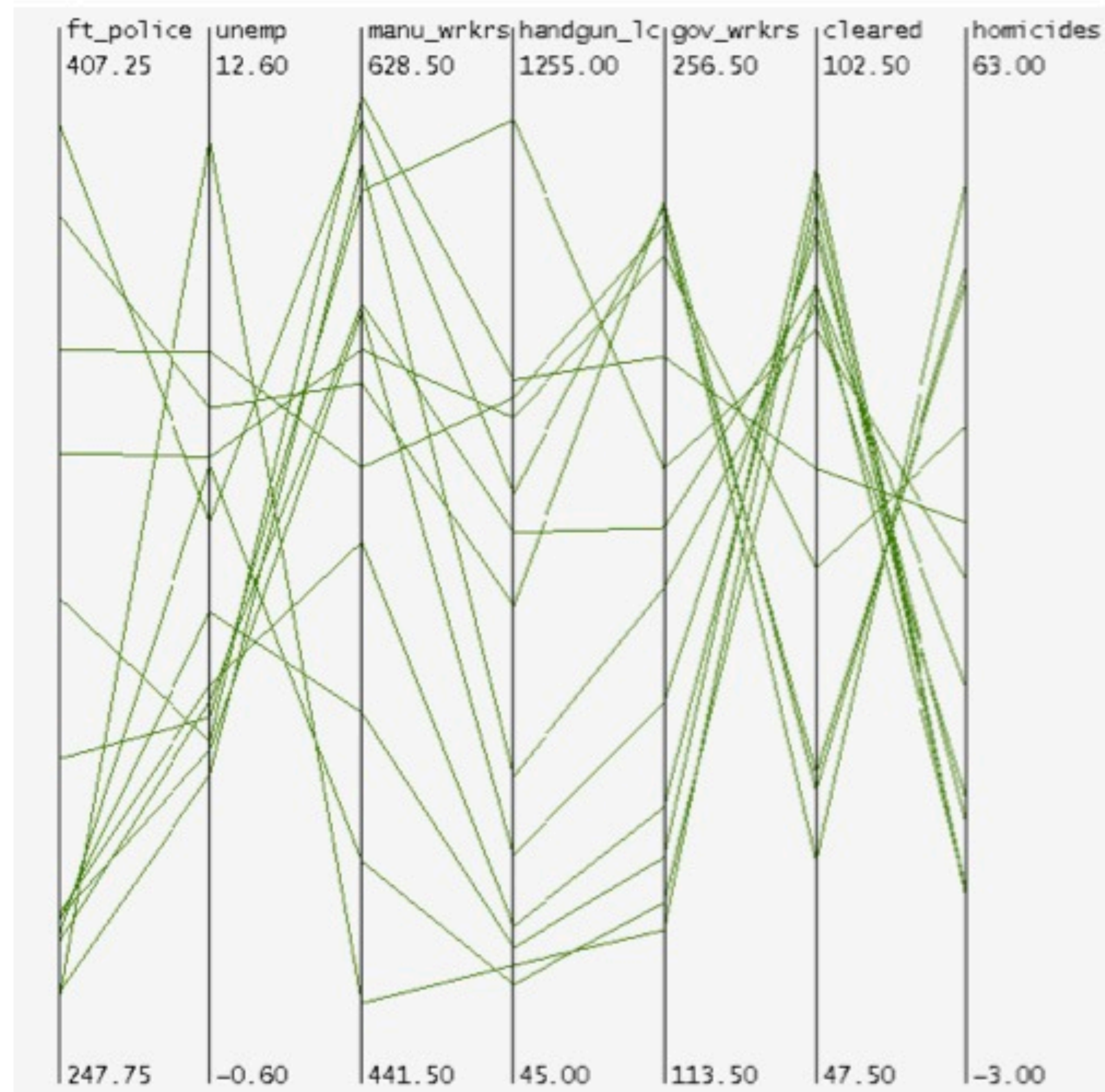


Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of $\rho = 1, .8, .2, 0, -.2, -.8$, and -1.

Wegman 1990

PARALLEL COORDINATES TASK

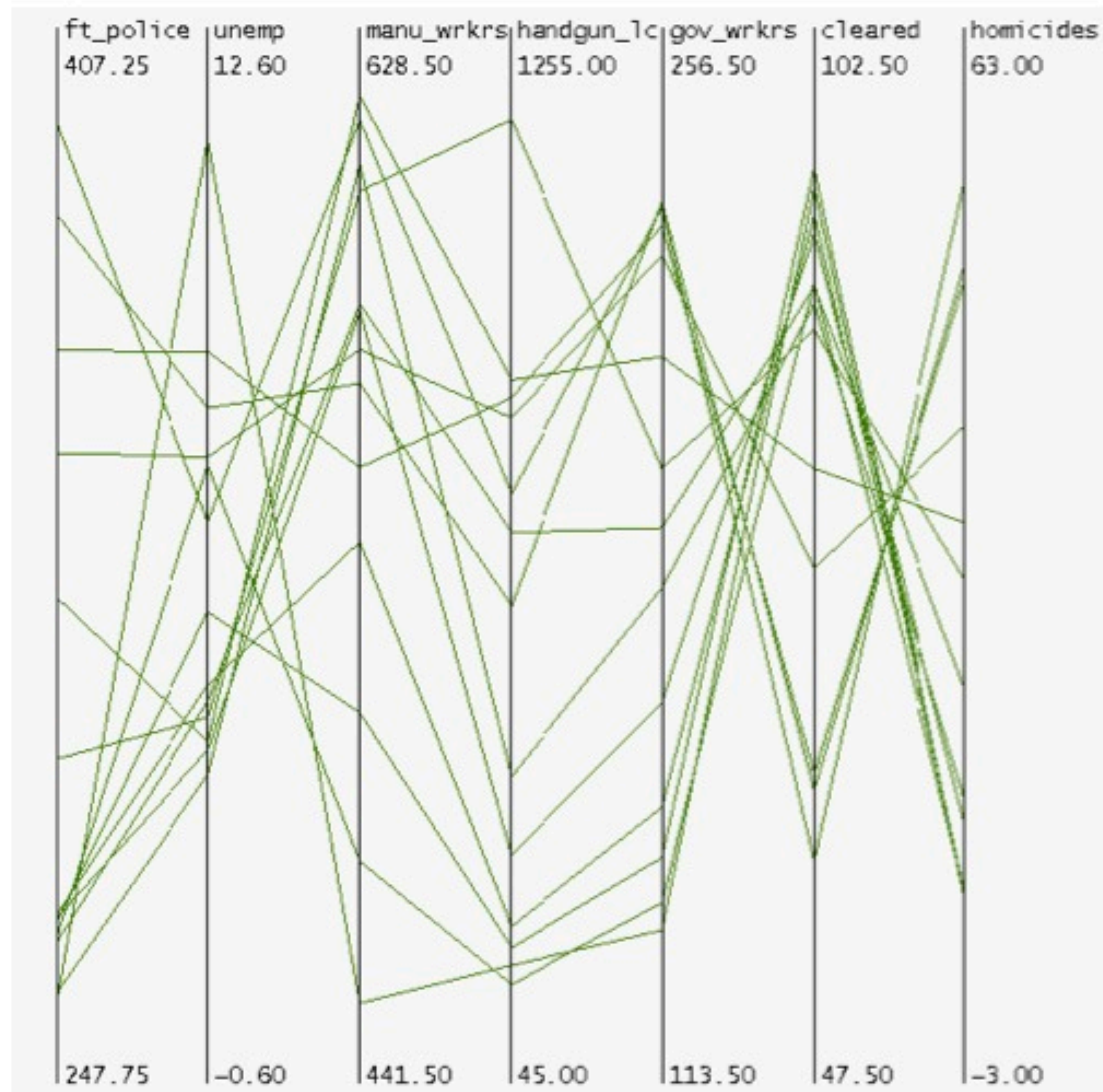
do you see any correlations?



Fua 1999

PARALLEL COORDINATES TASK

- **visible patterns only between neighboring axis pairs**
- **how to pick axis order?**
  - usual solution: reorderable axes, interactive exploration
    - *same weakness as many other techniques*
    - *downside: human-powered search*
  - not directly addressed in HPC paper



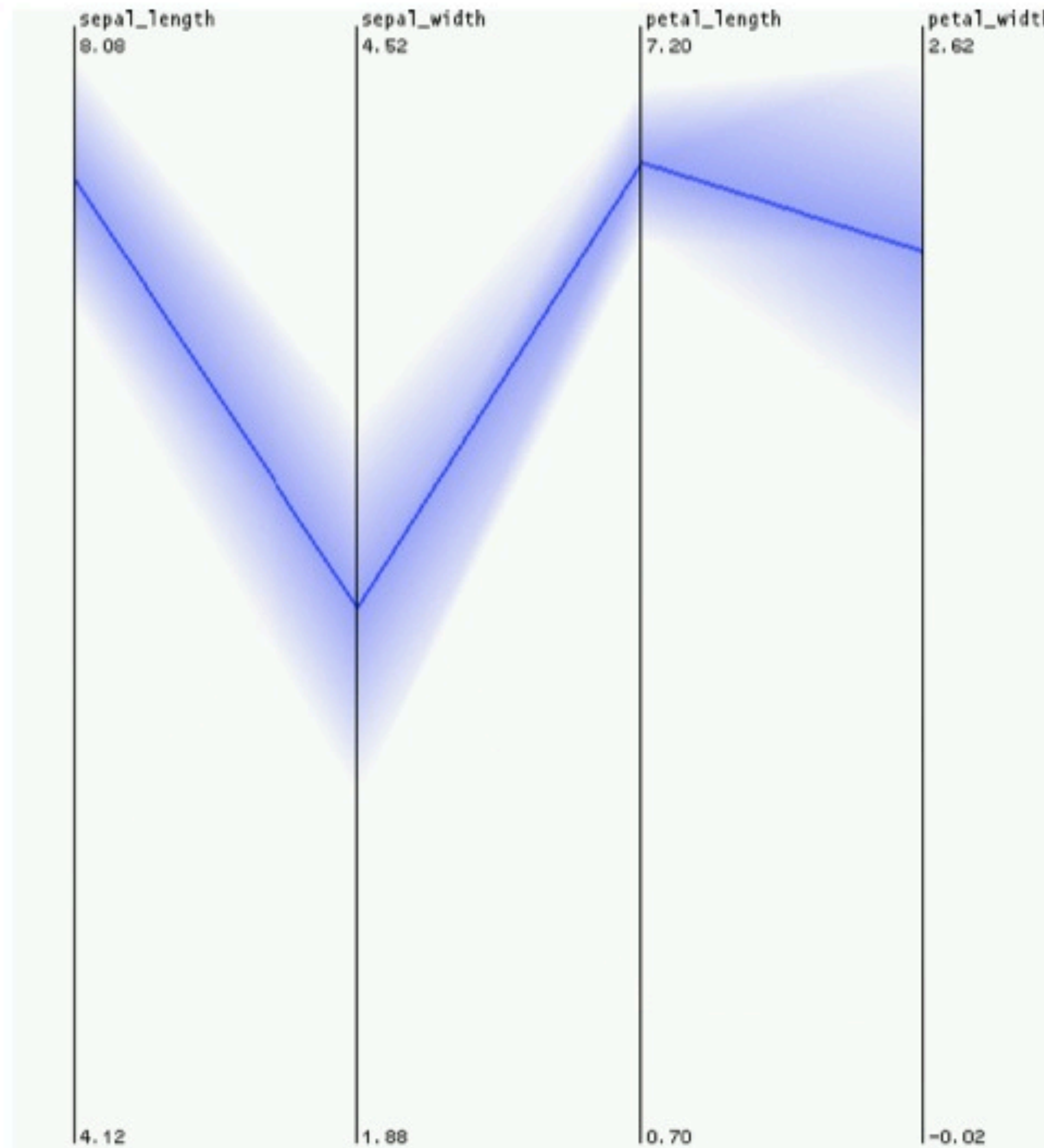
Fua 1999

HIERARCHICAL PARALLEL COORDINATES

- **data abstraction**
  - original data
    - *table of numbers*
  - derived data
    - *hierarchical clustering of items in table*
    - *clustering stats: # of points, mean, min, max, size, depth*
    - *cluster density: points/size*
    - *cluster proximity: linear ordering from tree traversal*
- **task abstraction**
  - find correlations
  - find trends and outliers at multiple scales

HPC: ENCODING DERIVED DATA

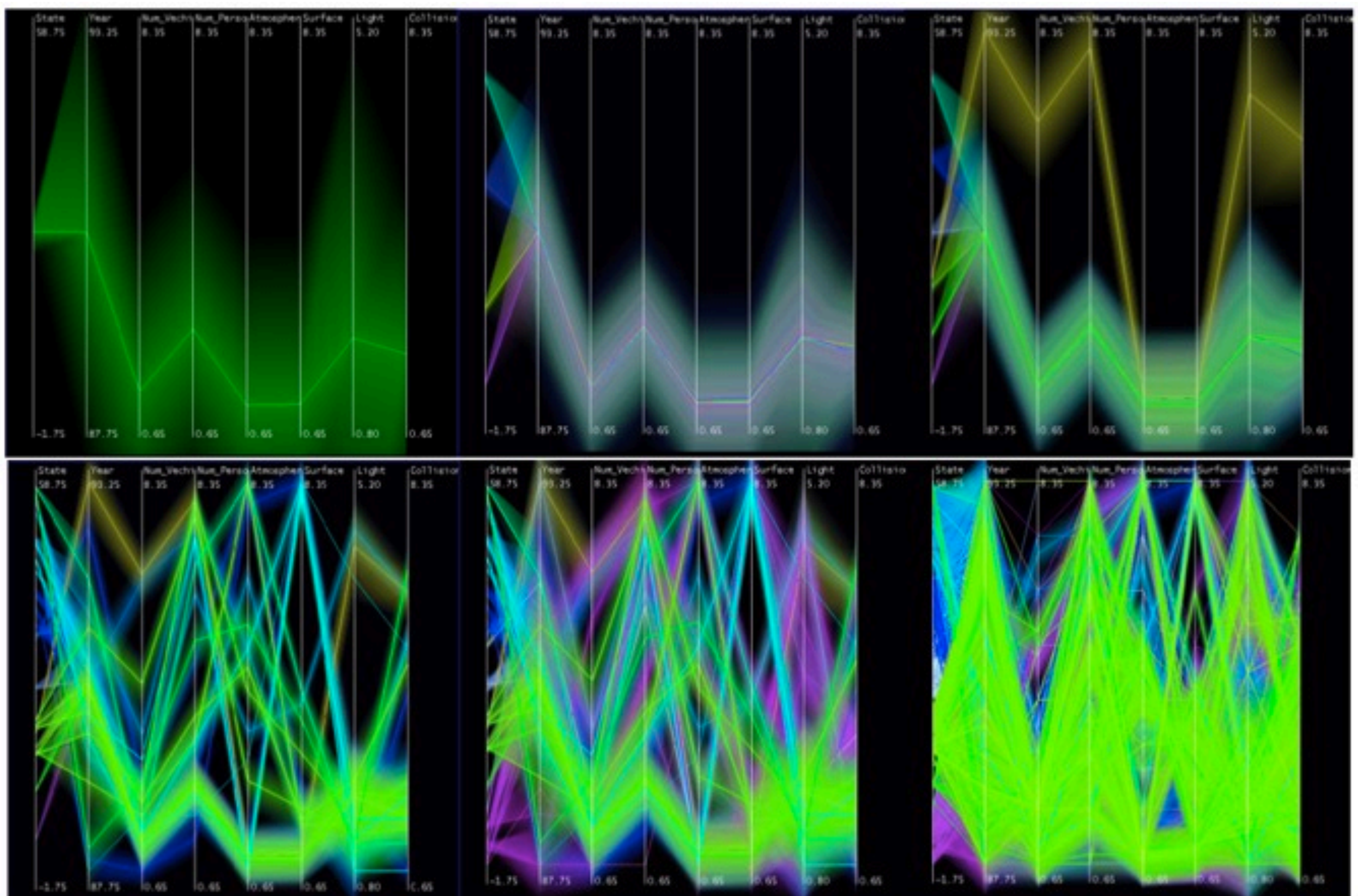
- **visual representation: variable-width opacity bands**
  - show whole cluster, not just single item
  - min / max: spatial position
  - cluster density: transparency
  - mean: opaque



Fua 1999

HPC: INTERACTING WITH DERIVED DATA

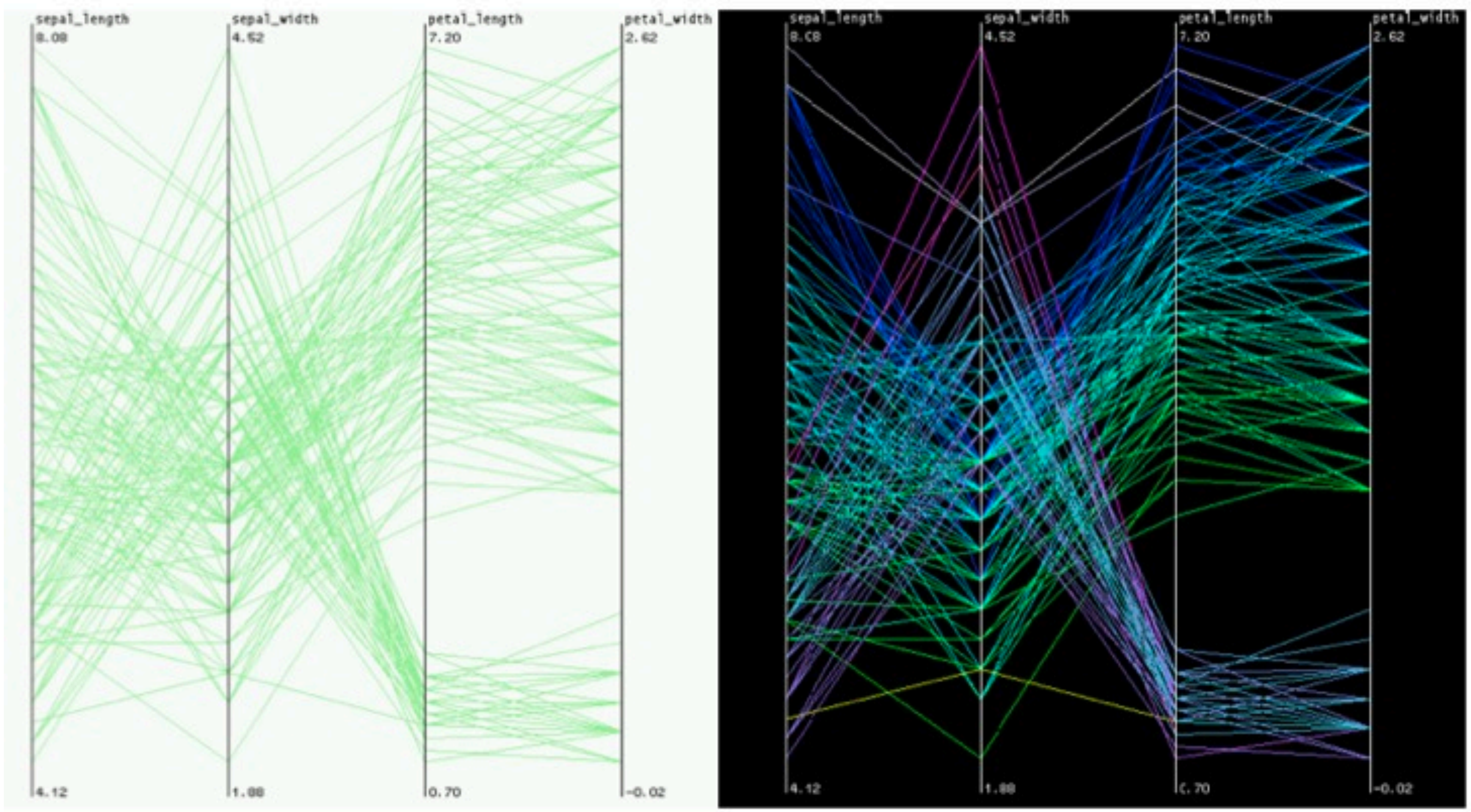
- **interactively change level of detail to navigate cluster hierarchy**



Fua 1999

HPC: ENCODING DERIVED DATA

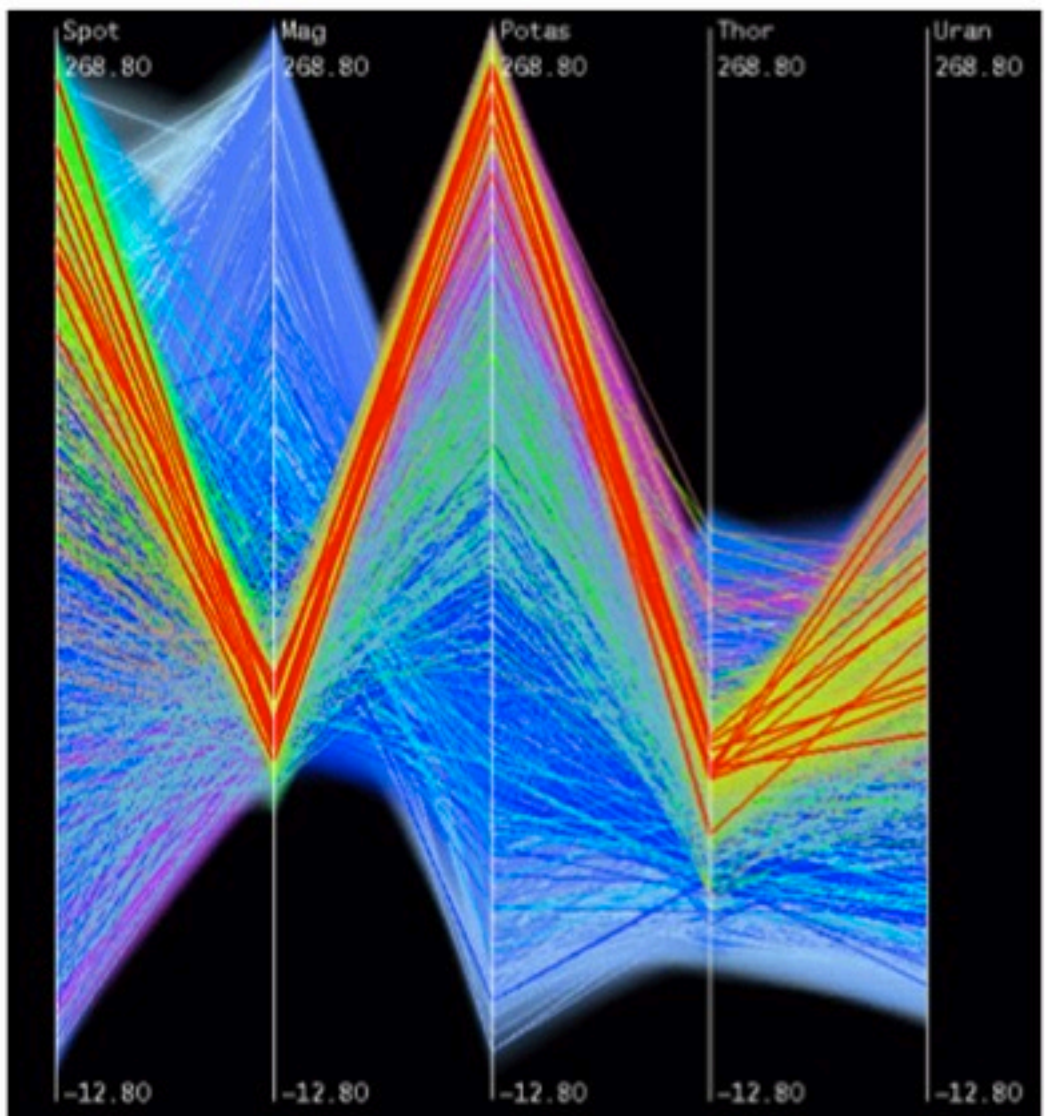
- **visual encoding: color based on cluster *proximity derived attribute***
  - *resolves ambiguity from crossings, clarifies structure*

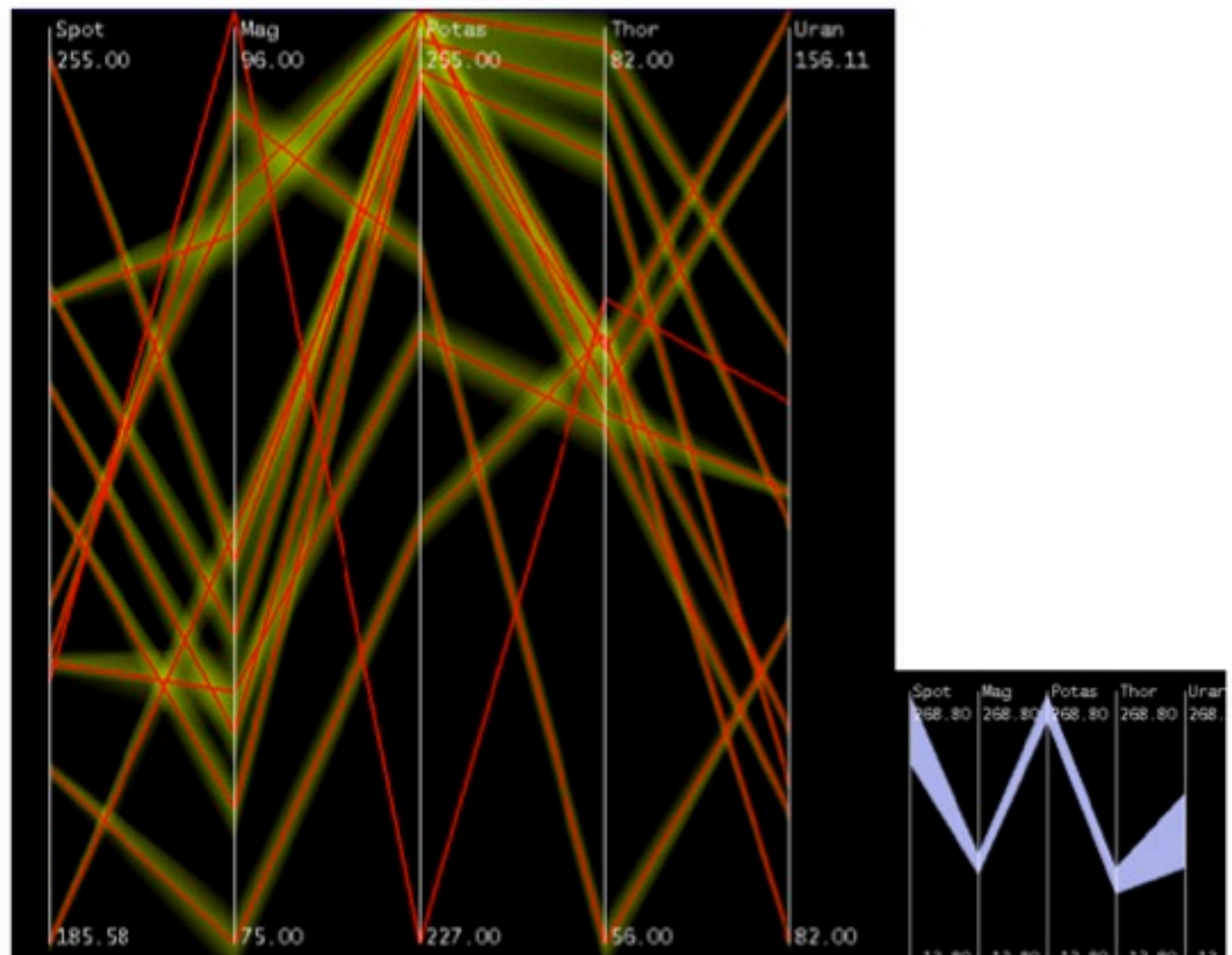


Fua 1999

HPC: MAGNIFICATION INTERACTION

- **dimensional zooming: use all available space**
  - methods
    - *linked vies to show true extent*
    - *overview + detail to maintain context*





Fua 1999

CRITIQUE: what do you think?

CRITIQUE

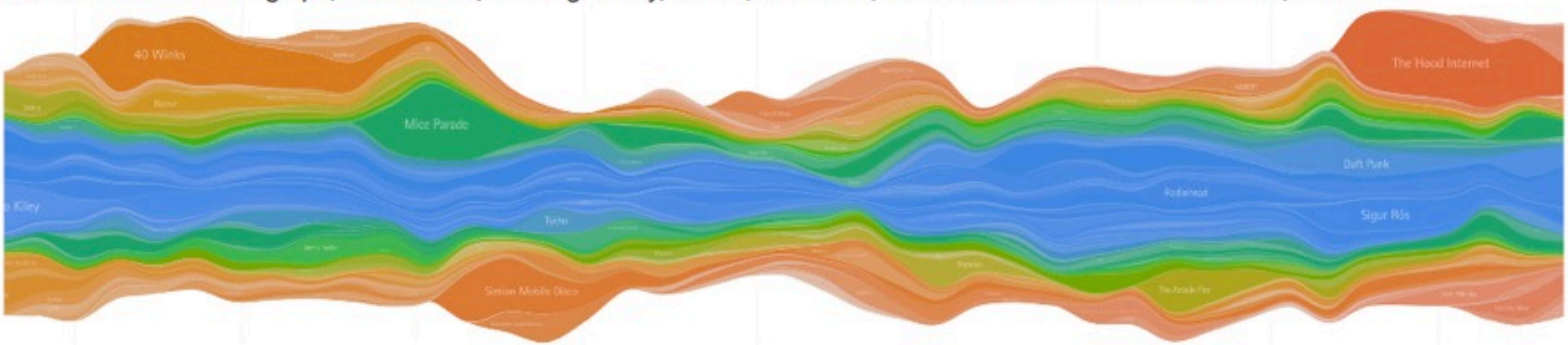
- **parallel coordinates**
  - strengths
    - *can be a useful additional view*
    - *(rare to use completely stand-alone)*
    - *now popular, many follow-on techniques*
  - weakness
    - *major learning curve, difficult for novices*
- **hierarchical parallel coordinates**
  - strengths
    - *success with major scalability improvement*
    - *careful construction and use of derived space*
    - *appropriate validation (result image discussion)*
  - weakness
    - *interface complexity (structure-based brushing)*

Stacked Graphs – Geometry & Aesthetics

Lee Byron & Martin Wattenberg

Abstract — In February 2008, the New York Times published an unusual chart of box office revenues for 7500 movies over 21 years. The chart was based on a similar visualization, developed by the first author, that displayed trends in music listening. This paper describes the design decisions and algorithms behind these graphics, and discusses the reaction on the Web. We suggest that this type of complex layered graph is effective for displaying large data sets to a mass audience. We provide a mathematical analysis of how this layered graph relates to traditional stacked graphs and to techniques such as ThemeRiver, showing how each method is optimizing a different "energy function". Finally, we discuss techniques for coloring and ordering the layers of such graphs. Throughout the paper, we emphasize the interplay between considerations of aesthetics and legibility.

Index Terms — Streamgraph, ThemeRiver, listening history, last.fm, aesthetics, communication-minded visualization, time series.



1 Introduction

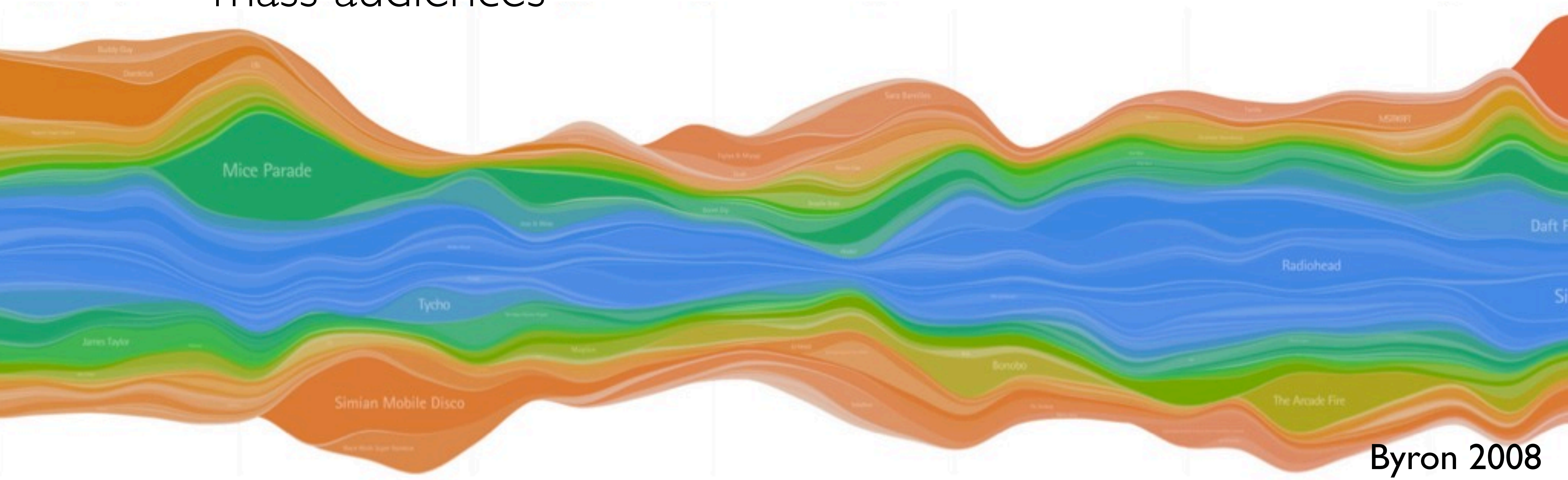
In February 2008, The New York Times stirred up a debate. The famous newspaper is no stranger to controversy, but this time the issue was not political bias or anonymous sources—it was an unusual graph of movie ticket sales. On information design blogs, opinions of the chart ranged from "fantastic" to "unsavory." Meanwhile, on other online forums and blogs, hundreds of people posted insights and questions spurred by the visualization.

graphic and accompanying online interactive visualization of the box office revenue for 7500 movies over a 21-year period.

In this paper we first provide a case study of the New York Times and last.fm visualizations. We pay special attention to the response on the web and the role of aesthetics in the appeal of visualizations. Second, we perform a detailed analysis of the algorithms that define these graphs. A key theme is the role of aesthetics in visualization

STREAMGRAPH

- **problem-driven paper**
  - development of new technique to solve a specific problem
- **challenge**
  - convey a large amount of data in a way that engages mass audiences



Byron 2008

STREAMGRAPH

- **problem: show personal last.fm history**
  - want to visually embody personal connection that listeners have with their music
- **design considerations**
  - use stacked graph
    - *focus on legibility and aesthetics*
- **abstraction**
  - task: engage audience

DATA ABSTRACTION

- **original data**
  - set of time series
- **derived data**
  - layer silhouette
    - *consider baseline*
    - *consider deviation*
    - *consider wiggliness*

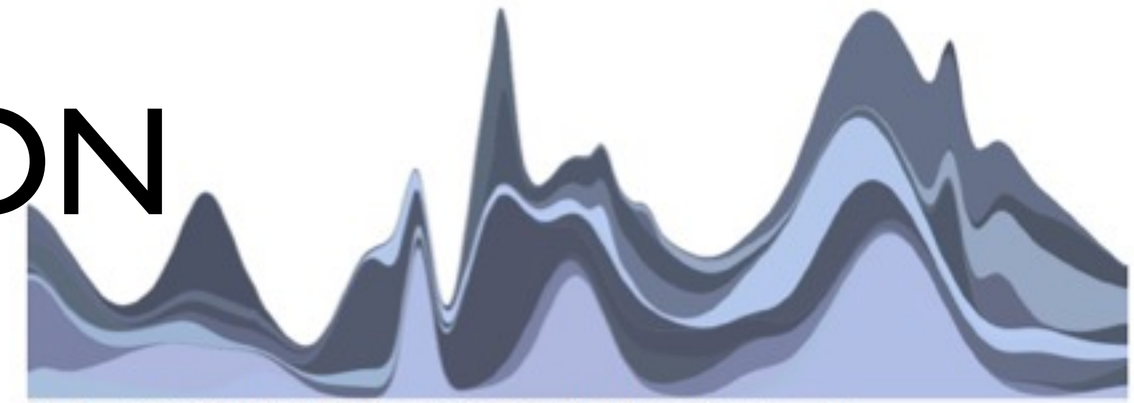


fig 5 – A traditional stacked graph with a baseline $g_0 = 0$

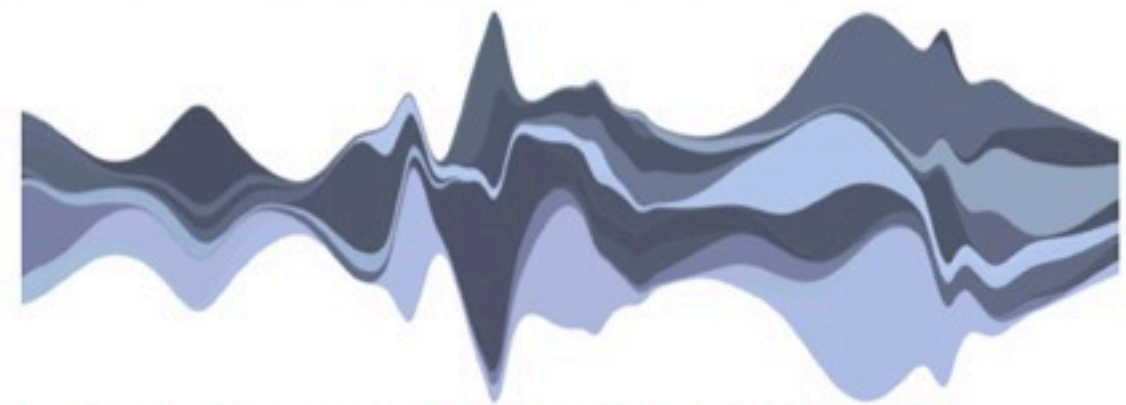


fig 6 – the same data set using the ThemeRiver layout algorithm

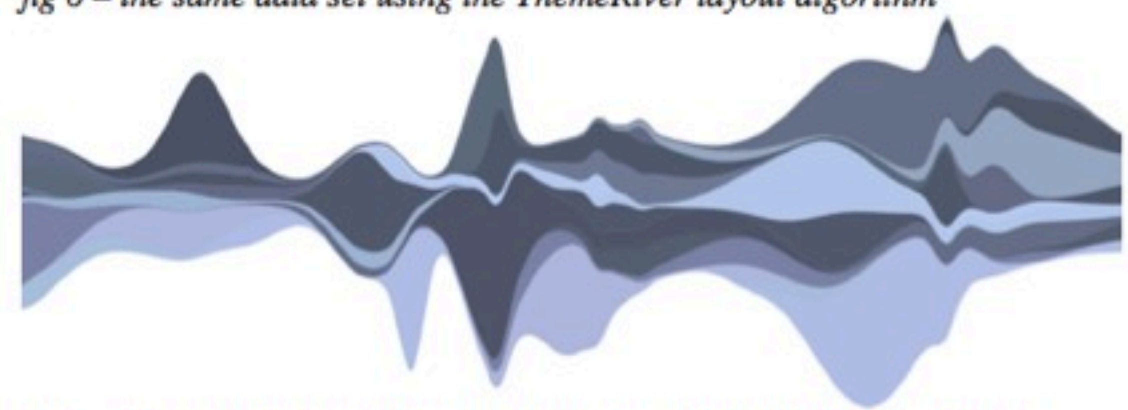


fig 7 – the same data set optimized to reduce the "wiggle" function, or overall variation in slope

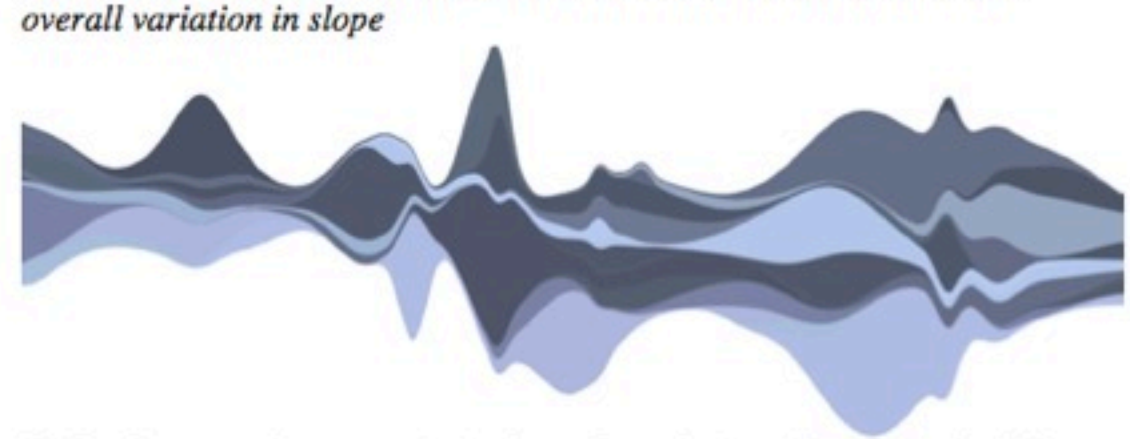
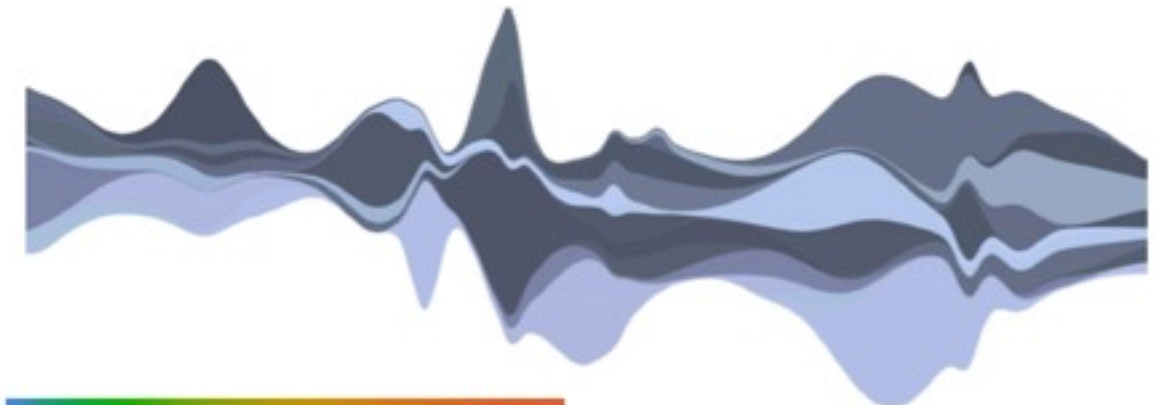


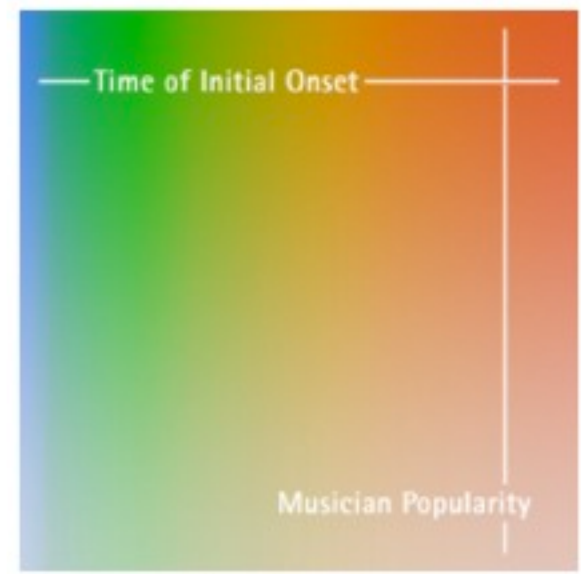
fig 8 – the same data set optimized to reduce the "weighted wiggle," the algorithm used in Streamgraph

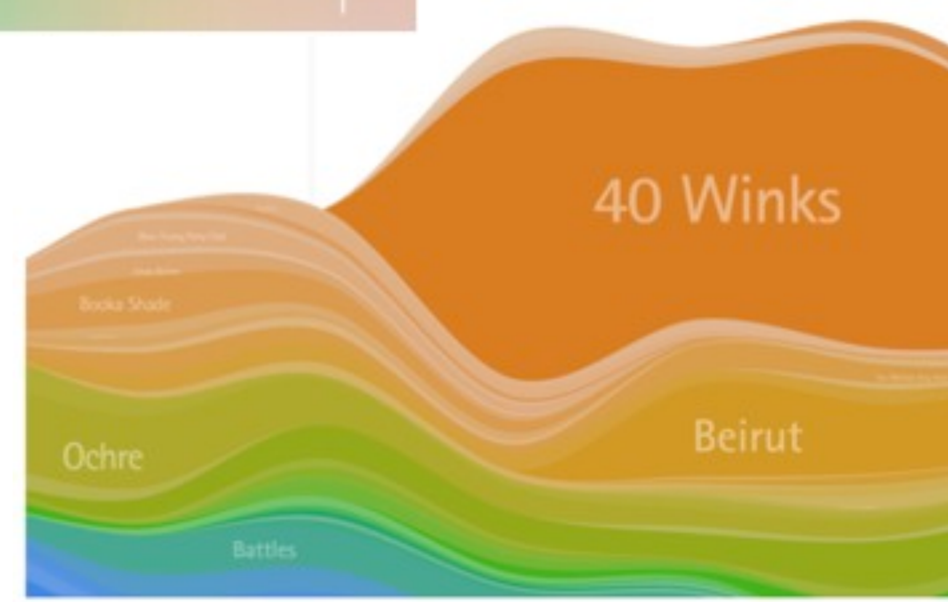
Byron 2008

DESIGN

- **visual representation: stacked graph**
  - new technique for minimizing wiggle of layers
- **color: 2D colormap**
  - hue: time of onset
  - saturation: popularity
- **labels**
  - placed where embedded labels can be largest







Byron 2008

DESIGN

- **layer ordering**
  - inside-out ordering
    - *avoid diagonal striping effect*
    - *burst are on outside which minimizes effect on other layers*
    - *prevents drift away from x-axis*

fig 12 – an unsorted data set, exhibiting the type of "burstiness" apparent in last.fm and box office data sets

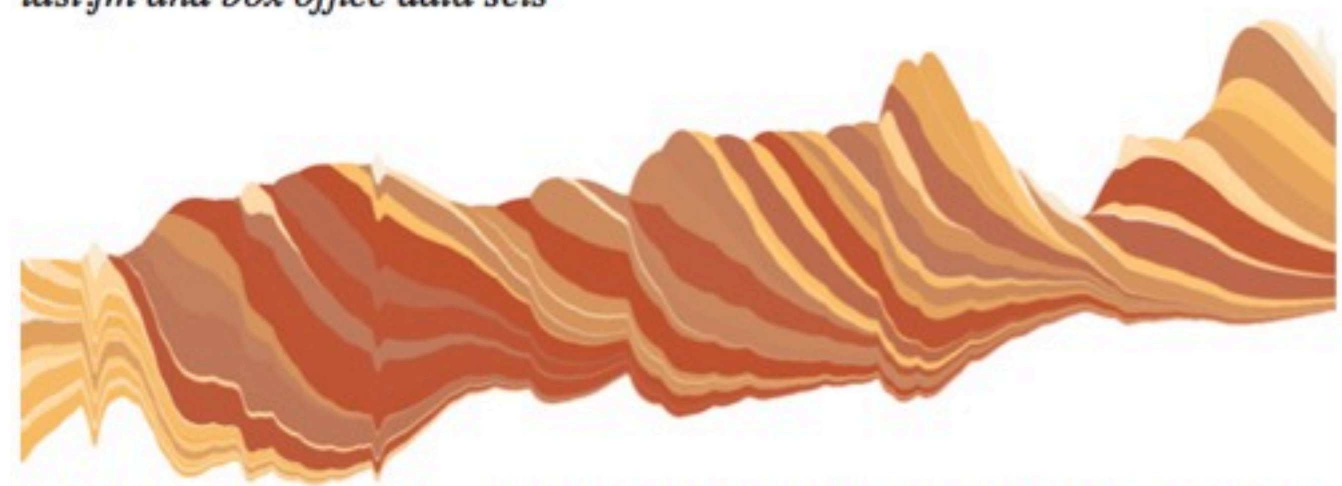


fig 13 – the same data set, naively sorted in order of "onset time" exhibiting the distracting diagonal striping effect

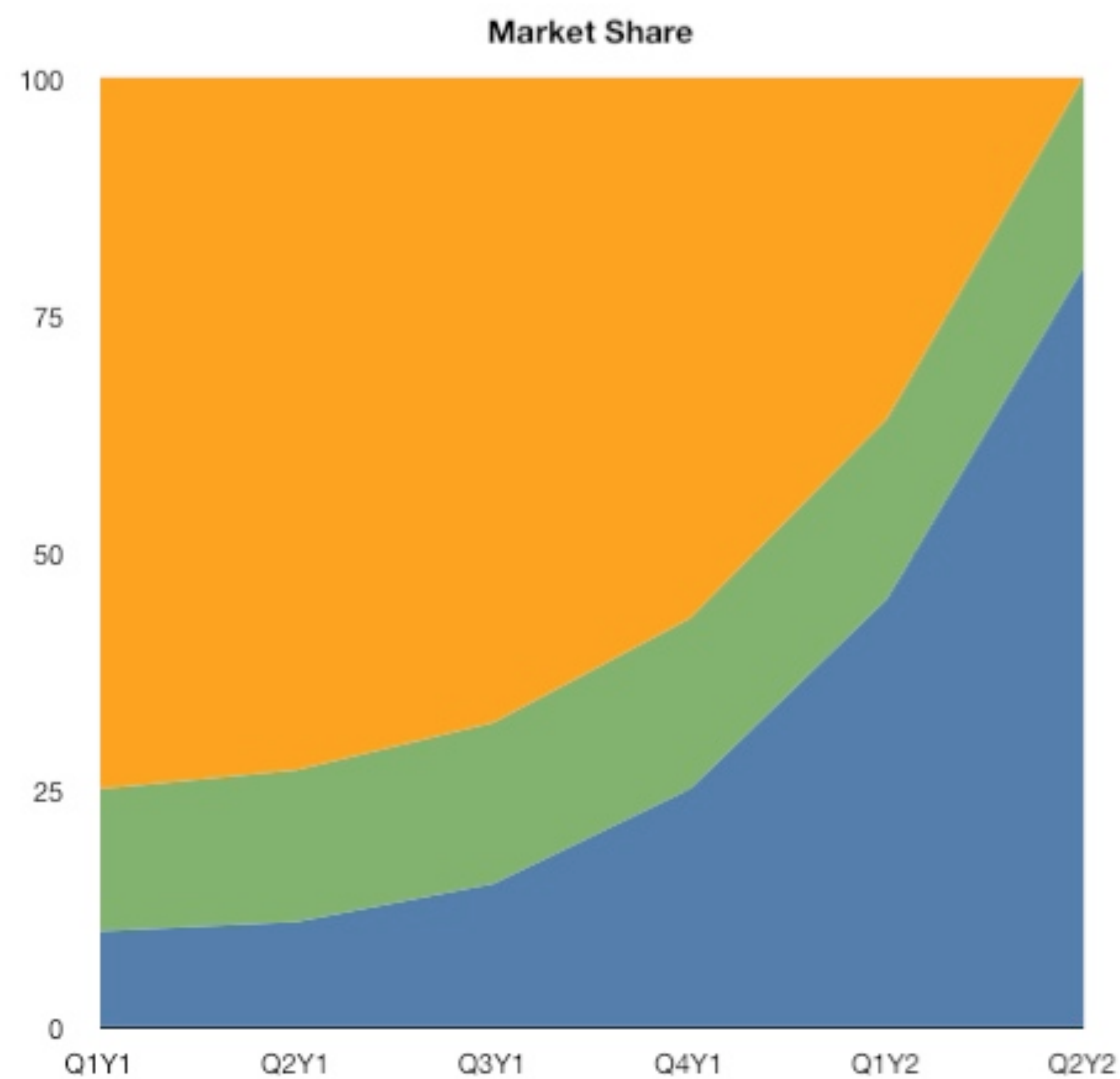
fig 14 – the same data set sorted using the weighted "inside out" strategy to highlight the initial onset of each time series

Byron 2008

EVALUATION

- **case study of NYTimes graphic**
  - gathered many comments from social media sites
  - categorized comments
    - *legibility issues*
    - *engagement*
    - *aethestics*

COMMENTS



The Ebb and Flow of Movies – Box Office Receipts 1986 — 2008 – Interactive Graphic – NYTimes.com

http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html

Apple Yahoo! Google Maps YouTube Wikipedia News (1,819) Popular Google Scholar

The New York Times

Movies

February 23, 2008

E-MAIL | FEEDBACK

The Ebb and Flow of Movies: Box Office Receipts 1986 — 2008

Summer blockbusters and holiday hits make up the bulk of box office revenue each year, while contenders for the Oscars tend to attract smaller audiences that build over time. Here's a look at how movies have fared at the box office, after adjusting for inflation.

Find Movie Harry Potter and the Order of the Phoenix Go

CRITIQUE: what do you think?

CRITIQUE

- **strengths**
  - clear target problem
  - thorough evaluation of aesthetic and legibility issues
  - reached and engaged a large audience
- **weaknesses**
  - stacked graphs make between comparisons difficult
    - *both between time points and time series*

THE SMARTPHONE CHALLENGE

part 6

- get back into large groups
- share sketches
- turn in abstraction and individual sketches

L14: Graphs and Trees

REQUIRED READING

Graph Visualisation and Navigation in Information Visualisation

I. Herman, G. Melan�on, M. S. Marshall

Centre for Mathematics and Computer Sciences (CWI)

Kruislaan 413

1098 SJ, Amsterdam, The Netherlands

{I.Herman, G.Melancon, M.S.Marshall}@cwi.nl

CWI Information Visualization Home Page

[The preprint of an updated version of this article, published in IEEE Transactions on Visualization and Computer Graphics, *6*(1), pp. 24-43, 2000. is also available here in PDF.]

Abstract. This is a survey on graph visualisation and navigation techniques, as used in information visualisation. Graphs appear in numerous applications, like web browsing, state-transition diagrams, computer data structures, etc. The ability to visualise and to navigate in these potentially very large, abstract graphs is often a crucial part of an application. Information visualisation has specific requirements, which means that this survey approaches the results of traditional graph drawing from a different perspective than the traditional surveys; as such it is a useful complementary survey to those.

Keywords: information visualisation, graph visualisation, graph drawing, navigation, focus+context, fish-eye, clustering.

1998 Computing Reviews Classification System: G.2.2., H.3.3, H.4.m, H.m, I.3.4, I.3.m, J.m

Table of Content

mailto:mwatten@us.ibm.com

Visual Exploration of Multivariate Graphs

Martin Wattenberg
Visual Communication Lab, IBM Research
1 Rogers St., Cambridge MA 02142
mwatten@us.ibm.com

Figure 1. *A PivotGraph visualization of a large graph rolled up onto two categorical dimensions*

Author Keywords
information visualization, graph drawing

ACM Classification Keywords

ABSTRACT

ABySS-Explorer: Visualizing Genome Sequence Assemblies

Cydney B. Nielsen, Shaun D. Jackman, Inanç Birol, and Steven J.M. Jones

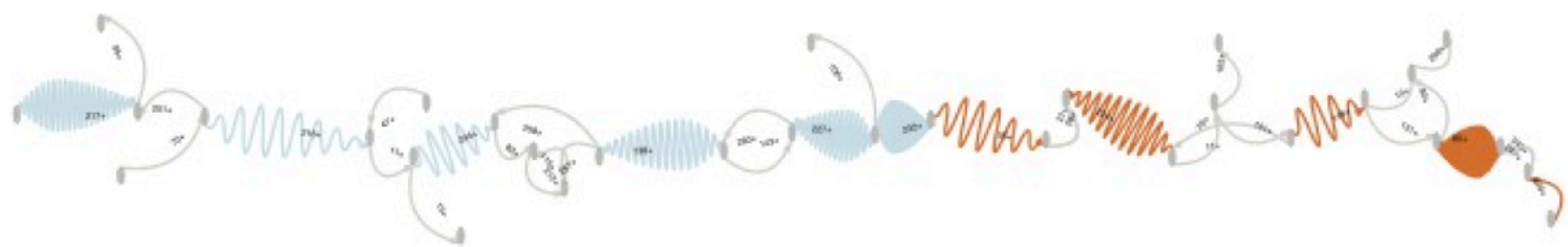


Fig. 1. ABySS-Explorer employs a novel graph representation enabling biologists to examine the global structure of a genome sequence assembly.

Abstract—One bottleneck in large-scale genome sequencing projects is reconstructing the full genome sequence from the short subsequences produced by current technologies. The final stages of the genome assembly process inevitably require manual inspection of data inconsistencies and could be greatly aided by visualization. This paper presents our design decisions in translating key data features identified through discussions with analysts into a concise visual encoding. Current visualization tools in this domain focus on local sequence errors making high-level inspection of the assembly difficult if not impossible. We present a novel interactive graph display, ABySS-Explorer, that emphasizes the global assembly structure while also integrating salient data features such as sequence length. Our tool replaces manual and in some cases pen-and-paper based analysis tasks, and we discuss how user feedback was incorporated into iterative design refinements. Finally, we touch on applications of this representation not initially considered in our design phase, suggesting the generality of this encoding for DNA sequence data.

Index Terms—Bioinformatics visualization, design study, DNA sequence, genome assembly.

1 Introduction

Data generation used to be the expensive and time consuming step in biology research. Recent innovations in high-throughput techniques have transformed it into a cost-effective and rapid process, pushing the bottleneck of discovery into the analysis phase. There is increasing recognition in the field that improvements in visualization tools will be essential for understanding our growing wealth of data. This paper presents one such tool for a genome analysis problem.

The term "genome" refers to the genetic material of a cell and can

subjected to many rounds of automated improvement, but ultimately it is visually inspected and manually edited by specialists.

Our work was motivated by the needs of genome analysts and the shortcomings of existing visualization tools in this domain. A genome assembly consists of long contiguous sequences, called contigs, assembled from short sequencing reads. An analyst integrates diverse data types used by the assembly algorithm together with external metadata to make final judgements about whether an assembly is correct