Line Graph Visual Clutter

Last update: 2015 / 08 / 03

Outline

- Introduction
- Visual Design
- Related Work
- Expected Results
- Remaining Tasks

Introduction – Logic Flow

Introduction

Explain the main objective:

- Line graph importance: the most fundamental visualization for time series data and scientific measurement
  - Time series data (the fluctuation of stocks, climatic changes, etc.)
  - High dimensional data (temperature, humidity, microarray data, etc.)
- What we use line graph for:
  - Observe trends / directions
  - Observe absolute values
  - Relationship of some dependent variables and an in-reorderable independent variable
  - Comparison between lines
- Line graphs' insufficiency when dealing with multiple line:
  - ("The main disadvantage is that it only works well with very few time series at the same time if multiple line graphs are compared with each other") (Ziegler, 2010)
- Importance of solving line graph's visual clutter:
  - The number of line graphs in a given study (i.e. the subject of analysis) can be quite large, with hundreds to thousands of samples.(Kincaid, 2006)

->> Summarize the objective.

To alleviate the over-plotting problem in line graphs when having multiple lines simultaneously drawn, and make their expressiveness more comparable to those with small number of lines.

Introduction

Summaries of Approach

Explain considerations for hierarchical partial clustering:

- Clustering's necessity
  - Techniques other than clustering could overlook line graph's features
    - e.g., Small multiples – not good at comparison, query – lack of overview for existing trends
- Partial clustering's importance
  - Two lines are unlikely to follow the exact same x-y relationship throughout the whole x domain.
    - Stock price raise together in a short period only due to certain events
    - Users with various watching behaviors response similarly to a specific short video frame only
  - Partial trends in only a period might be important.
    - (Hochheiser, 2004) Find genes with similar expression patterns during key period in programmed cell death
- Hierarchical clustering's importance
  - Some less distinguishable feature could be visually unrecognizable due to clutter reduction, and will require more detailed observation.

->> Hierarchical partial clustering is needed to emphasize all these kinds of relations.

Visual exploration based on clustering results:

- For observing and comparing the otherwise cluttered lines based on the partial similarities discovered.

->> Visual design is closely connected with clustering and works effectively for clutter reduction

Contributions

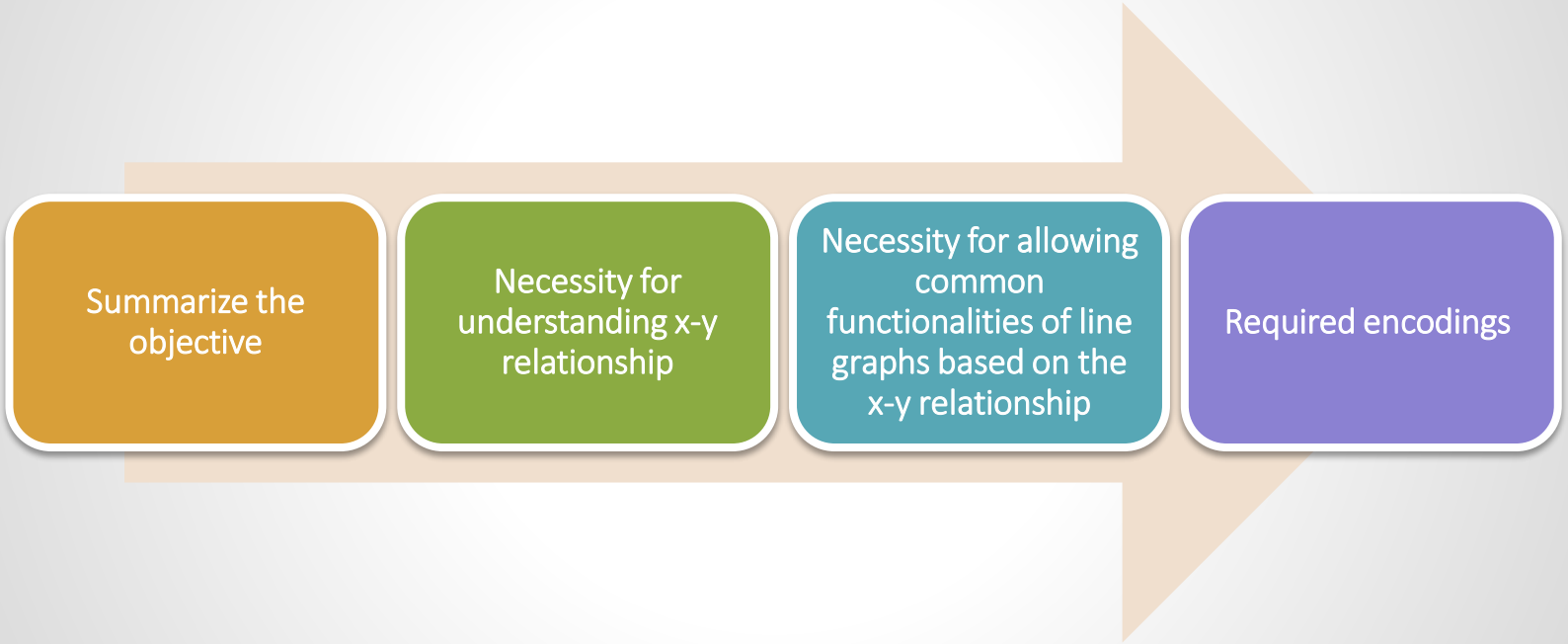
The major contributions of this paper are

(1) Study the line graph visual clutter problem

(2) Develop a hierarchical visual clustering algorithm specifically for line graphs

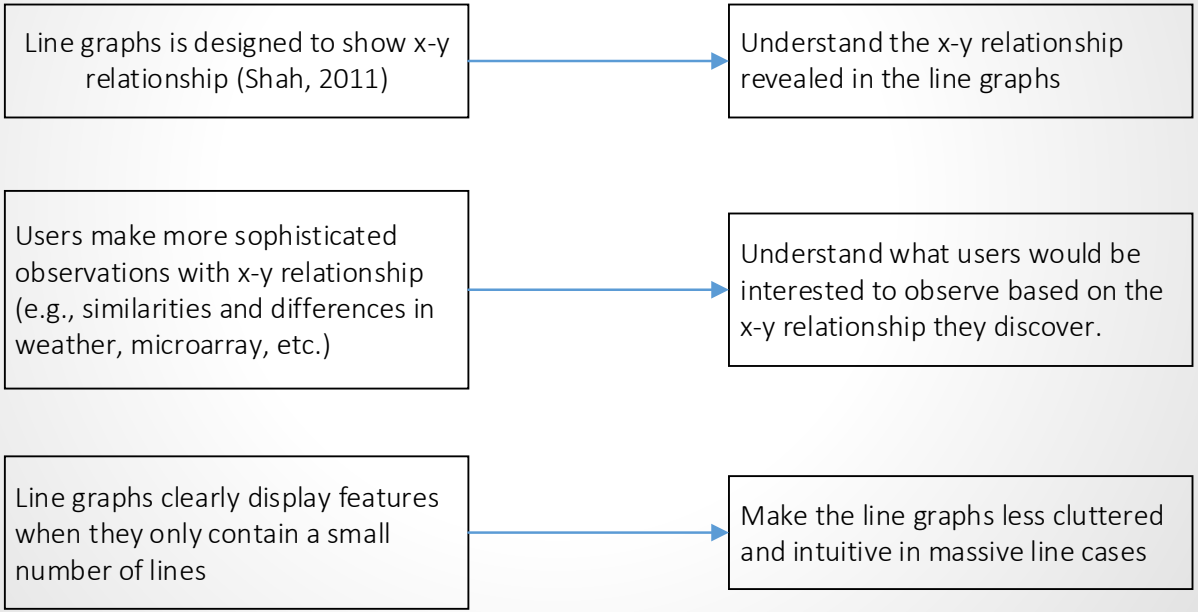
(3) A new design that deals with line graph visual clutter issue, and allows hierarchical exploration.

Requirement Analysis – Logic Flow



Requirement Analysis – Overview

Motivation re-visited



Requirement Analysis – x-y relationship

x-y relationship regards of: This is the primal observation from the line graphs and the basis for users' more complicated usage of line graphs. Should eventually lead to the design consideration of Algorithm.

- **Absolute positions:** differences on y (Kelley, 1997)
- **Relative changes:** gradients or directions of a line in a certain interval (Kelley, 1997)
- **Un-reorderable independent value / independent value correspondence:** Both the afore mentioned two are with respect of x (independent variable)'s specific range. (Kincaid, 2006)

->> need to

- **Find the leading x-y relationships that:**
  - preserve the aforementioned three features
  - Emphasis on partial similarities (as mentioned in INTRODUCTION)
- **Observe How do the obtained x-y relationship distribute**

Requirement Analysis – Line graph functionalities

Users use line graph to: To demonstrate that with the x-y relationship being obtained, how we would enable other functionalities of line graphs in our design. Should be closely connected to visual designs.

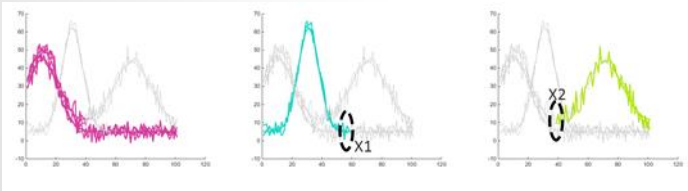
Assess trends:

- [FIND DIRECTION] For a specific range of x, how would y vary with it? What are the direction they go?
- [Direction alignment] If the trend is commonly shared / frequently occurred?

Compare between multiple lines:

"We expect that some line graphs will show similar features and others will not. This similarity or dissimilarity is what we seek in the analysis. (Kincaid, 2006). "

- Similarities:
  - [Representativeness of a cluster] Do several lines grow similarly / stay close in a certain period?
  - [Centrality of a cluster] How similar are them in the period?
  - [Target exploration]Track the similarity of lines : *If several lines have practical correlation in real world, are they clustered together?*
- Dissimilarities:
  - [Outliers] Are there any lines that develop significantly differently from others? , either on the whole time period or during a interval?
  - [Turning points] Are there any lines that certainly start to develop uniquely? When do these differences start?

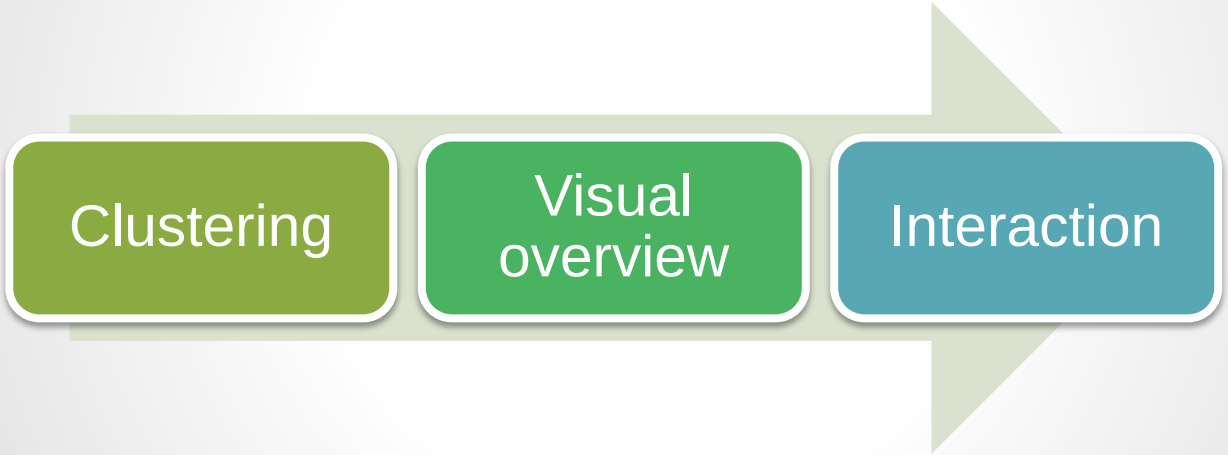


A clustering result generated by the current algorithm. X1 and x2 might be the turning points that worth more investigation.

- [Hierarchy] Explore details of some specific lines in a certain period based on their interests

Your Methods

Basic workflow:



Algorithm – partial clustering

Requirements:

- LoD: Allow hierarchical exploration
- Emphasis on global features:
  - Eliminate local details: Piecewise linear representation (Douglas Peucker)
  - Priority for consecutive segments from one line
- Meaningful comparison - independent value correspondence
  - Settings in distance functions: Do not accept similarity of two lines that are horizontally too far away.
- Emphasis on x-y relationship features
  - Absolute value: no normalization
  - Relative change: consider Perpendicular, parallel, angle distance

Algorithm – partial clustering

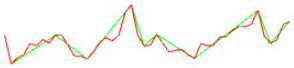
n input lines

Piecewise linear representation

Eliminate local noise

m_i segments for line l_i

M line segments in total



Compute distance between line segments based on some criteria

Basis for clustering. Consider :

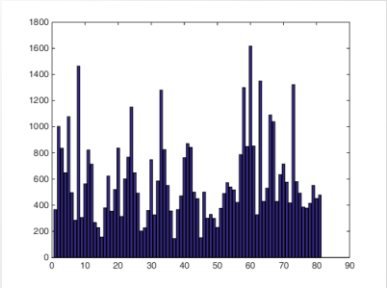
- **Perpendicular, parallel, angle distance**
- **Priority for consecutive segments from one line:** avoid over-emphasis on local detail
- **Segments with no overlap x interval AND is not consecutive segment of a line:** distance should be NaN – independent variable are meaningful in line graphs, and similarity of two lines being horizontally too far away is useless.

$M \times M$ distance matrix

Build the Optics reachability plot

Consecutively ordered line segments are potentially members of the same cluster

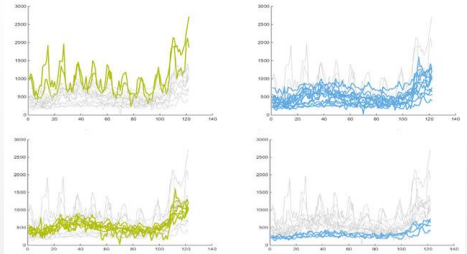
*T*he order of *M* segments in the reachability plot, and its distance



Extract hierarchical clustering results

Group similar lines together w.r.t different level of details

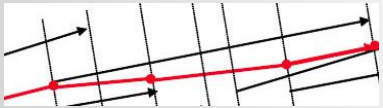
Clustering result

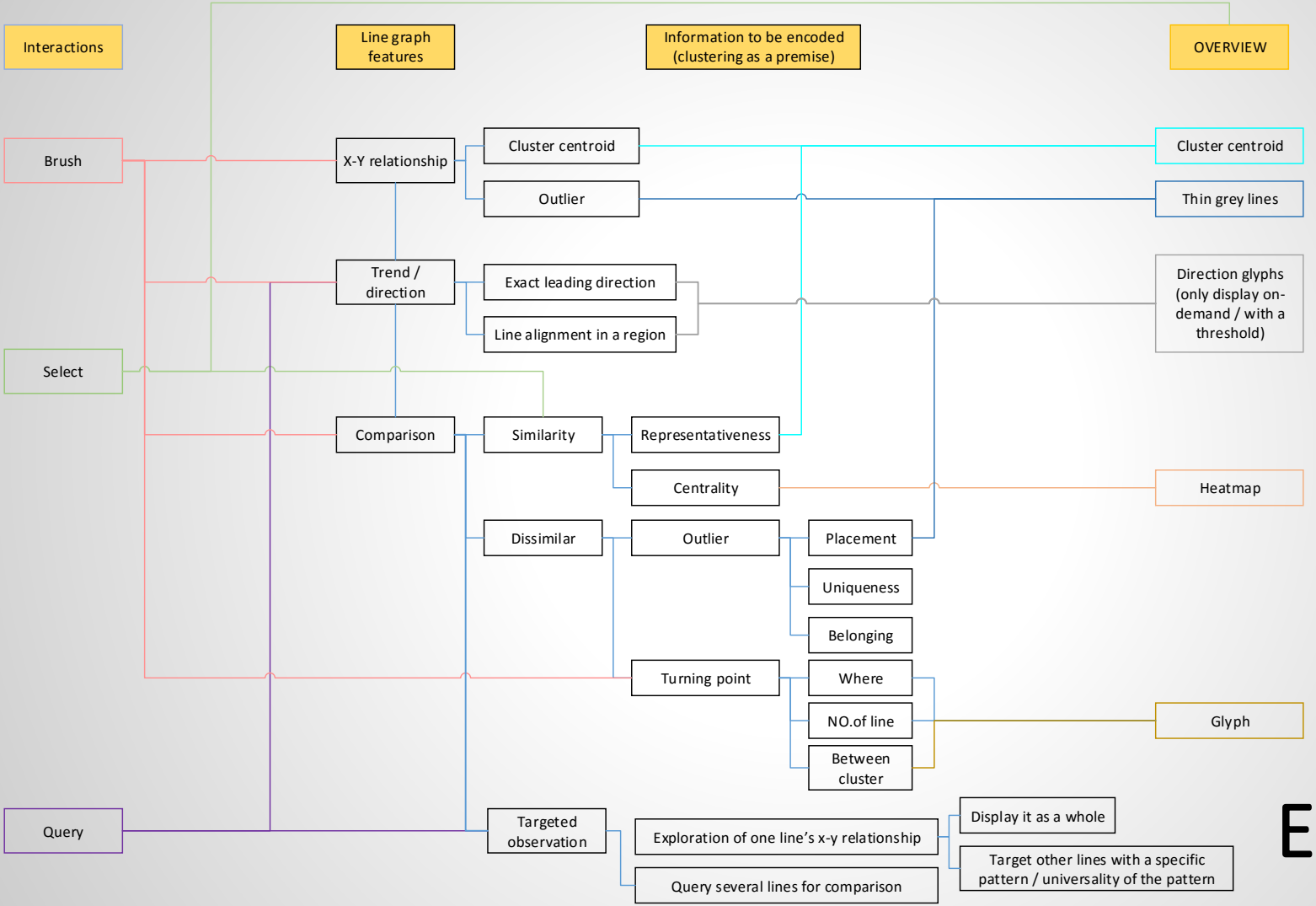


Centroid extraction

Used for representing a set of lines in visualization

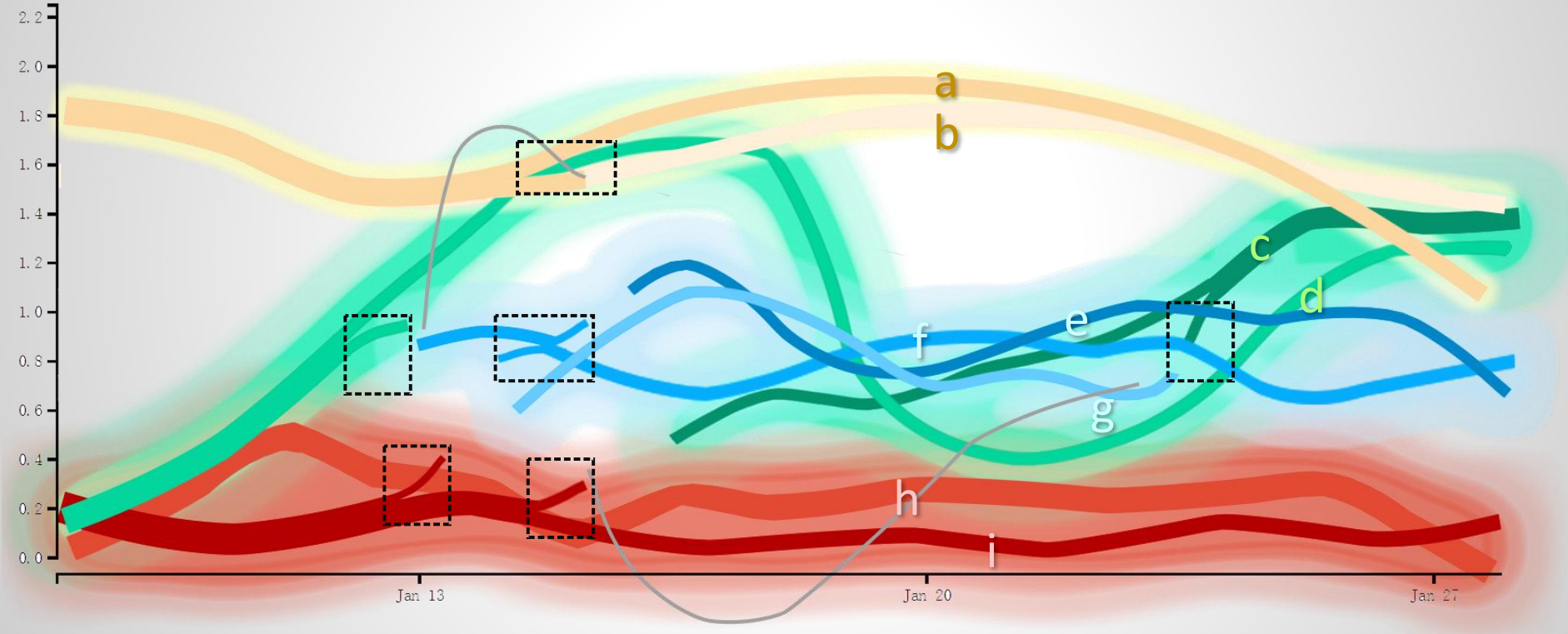
Centroid of each cluster (i.e., main x-y relationships in every level)





Visual Design Encoding

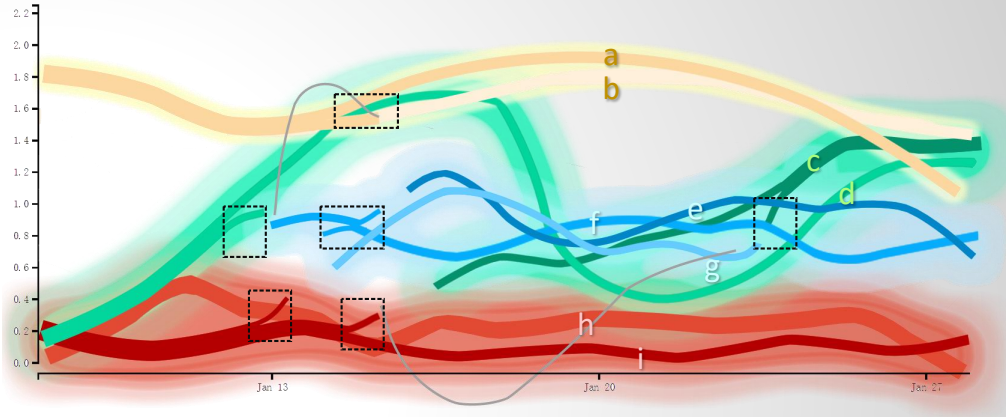
Visual design – Overview



Visual design – Overview

a – i: cluster centroid

- **Width:** No. of line in the cluster at the specific time point
- **Color :** The cluster.
- **Similar color:** will be clustered together in the next lower level. (e.g., a,b; c,d; e,f,g; h,i)
- **REAL MEANING**: there are several lines distributing similarly,(e.g., all in blue), they have several leading x-y relationship (e, f, g).
- **X-y R as indication for LoD:** they are close if a region has been well-clustered (a, b), and would be very dissimilar (offset or very different angle) if a region worth more detailed exploration.



Heatmap - the **distribution** of lines follow some x-y relationship

- **Color:** the cluster (cluster with similarly colored centroids will share a heatmap)
- **Density of the color:** No. of lines distributed in a specific region

Glyph packing - detailed **direction,** which is not available in heatmap

- See the algorithm later
- Only display the most extreme cases / or only display on demand (see in the interaction)

Turning points (lines entering or exiting a cluster)

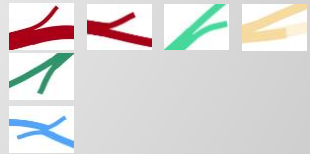
- **Width**: No. of lines turning in / out
- **Placement**: where the entering / exiting happen (only draw the leading placements, approximated based on the turning points of every line in a cluster)

Examples of turning points

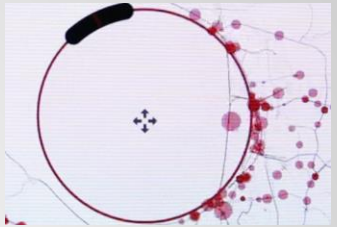
Lines exiting the original x-y relationship

Lines entering a x-y relationship

A place that both have line entering and exiting



Visual design – Possible interactions - Brush

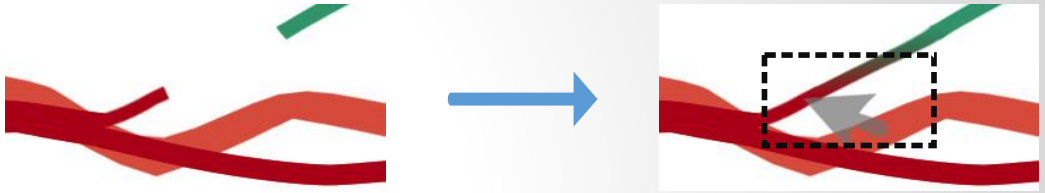


With the encodings in the overview, brushing helps to explore more specific information. It works as a guideline for selecting regions / clusters for further exploration.

- Basic idea: take the brushing as lens / fisheye to enlarge or shrink the target region, only that some distortion would be based on hierarchical clustering results.

An in / out turning point:

Connect the cluster before and after transfer

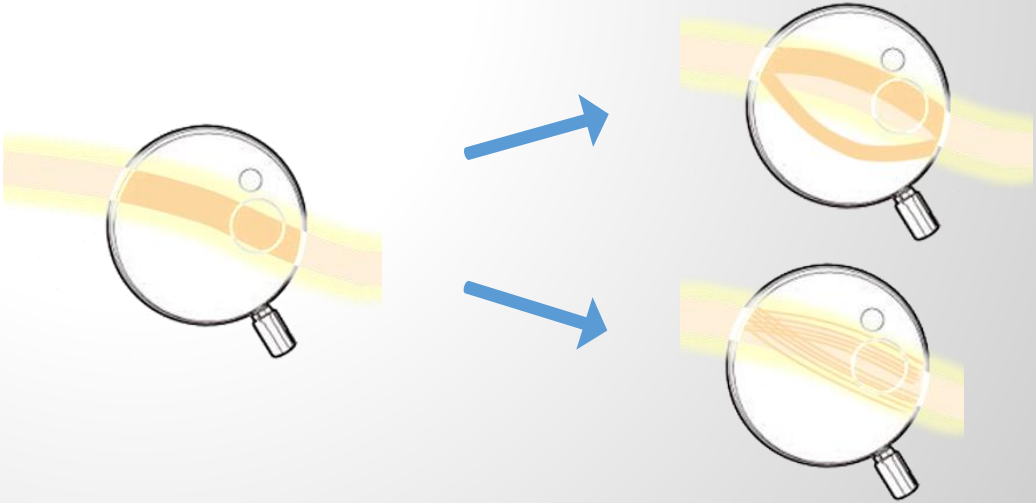


A cluster:

Lens view for previewing clusters' splitting behavior at other levels

OR

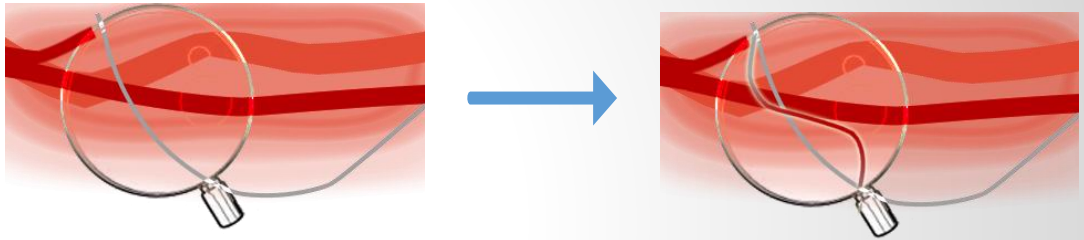
Its involved lines' real placement (could further select the shown line)



Visual design – Possible interactions – Brush (Cont')

An outlier:

- Distort the outlier to make it closer to the cluster it will later join
  - degree of distortion encode number of levels later that it will be clustered
- and re-color part of it as the to-be-clustered color.



A random region:

Display the glyphs showing the direction of each sub-region (see possible algorithm)

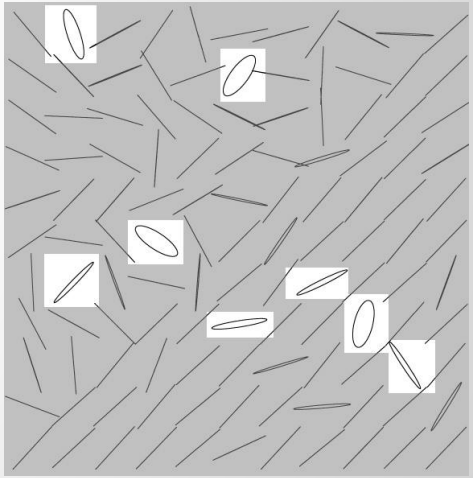
Visual design – Possible interactions – Select & Query

Select:

- Select a cluster
  - More precise heatmap showing the line distribution
  - Show the glyphs for the directions at all the regions where the lines belonging to the cluster exist
- Select a region
  - Zoom-in the overview to explore some region

Query (to link the individual features with the overview)

- Query a line
  - What to query:
    - By its name
    - By the line observed in brushing
  - Display it as a whole (connect its segments in different clusters)
  - Target other lines with a specific pattern
- Query a trend pattern
  - What to query:
    - The degree of alignment of the lines
    - The lines' main directions
  - Visual cue
    - Directions observed in brushing
    - Distributions of the main directions (slope: -90 to 90 degree)and the alignment of the lines (a score)

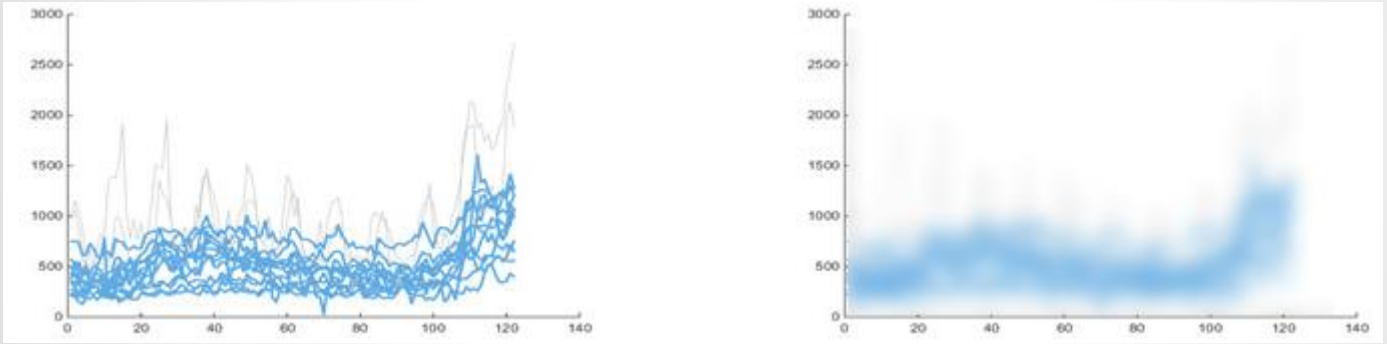


Algorithm – Heatmap for density (TBD)

Encode: spatial distribution of lines with similar x-y Relationships (i.e., the centroid for them should be in similar color)

Possible algorithm

1. Scale Space



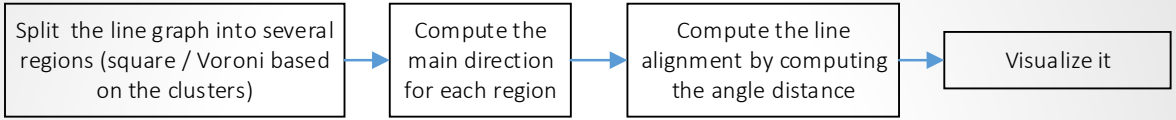
2. Contour map: No. of line in a specified range around the centroids

Algorithm – Glyph packing for directions (TBD)

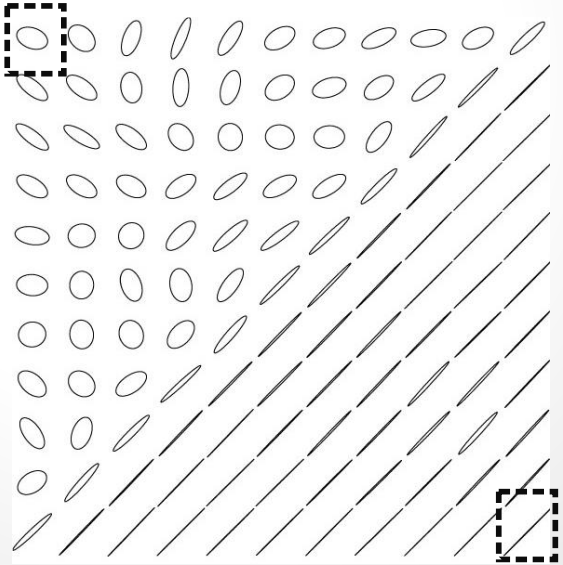
Encode: The main direction and the lines' alignments in a small region

Possible algorithm

This could allow different LoD



No apparent alignment, roughly -20 degree



All lines follow an approximately 45 degree direction

Glyphs (ellipse)

- Thinness represent the alignment: thinner -> more aligned
- Direction of the major axis: direction of the region

Related Work – Line Graph

Time Series Visualization Comparion

Javed W, McDonnel B, Elmqvist N. Graphical perception of multiple time series[J]. Visualization and Computer Graphics, IEEE Transactions on, 2010, 16(6): 927-934.

https://engineering.purdue.edu/~elm/projects/multilinevis/multilinevis.pdf

- Comparison between line graph, Braided graph, small multiples, horizon graphs with Maximum, Slope, Discrimination tasks.
- Synthetic data.

Heer J, Kong N, Agrawala M. Sizing the horizon: the effects of graph size and layering on the graphical perception of time series visualizations[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2009: 1303-1312.

http://vis.berkeley.edu/papers/horizon/2009-TimeSeries-CHI.pdf

- Use position encodings for both time and value
- Synthetic data.

Related Work – Line Graph

Characteristics of line graph (trend, metric information about absolute scales)

Playfair W. The commercial and political atlas: representing, by means of stained copper-plate graphs, the progress of the commerce, revenues, expenditure and debts of england during the whole of the eighteenth century[M]. T. Burton, 1801.

- The invention of line graph

Shah, Priti, and Eric G. Freedman. "Bar and Line Graph Comprehension: An interaction of top-down and bottom-up processes." Topics in Cognitive Science 3.3 (2011): 560-578.
http://onlinelibrary.wiley.com/doi/10.1111/j.1756-8765.2009.01066.x/epdf

- Line graphs reflects trends because lines connect discrete entities and directly represent slope.
- Line graphs reflect x-y relationship

Shah P, Hoeffner J. Review of graph comprehension research: Implications for instruction[J]. Educational Psychology Review, 2002, 14(1): 47-69.
http://deepblue.lib.umich.edu/bitstream/handle/2027.42/44452/10648_2004_Article_363437.pdf?sequence=1

- Num of data points – influence interpretation
- Number of trend reversals is the main determinant of comprehension difficulty as measured by study times.

Kelley D L. Graphing Statistics and Data: Creating Better Graphs[J]. Technometrics, 1997, 39(4): 429-430.

- Line graph is suitable for observing the development, judge the gradients, find significant changes and turning points
- Variations: index graph, sparkline

Related Work – Line Graph

Line graph scalability

Kincaid R, Lam H. Line graph explorer: scalable display of line graphs using Focus+ Context[C]//Proceedings of the working conference on Advanced visual interfaces. ACM, 2006: 404-411.

http://www.heidilam.com/doc/KincaidLam_LGE_AVI06.pdf

- Line graph use for scientific measurements
- Visualization system with a gel-like overview.
- Data: (1) time series climate data (2) Wheat Strains (3) Cancer data (All available online)
- Showed similar view: Lam, Heidi, Tamara Munzner, and Robert Kincaid. "Overview use in multiple visual information resolution interfaces." *Visualization and Computer Graphics, IEEE Transactions on* 13.6 (2007): 1278-1285.

Hochheiser H, Shneiderman B. Visual queries for finding patterns in time series data[J]. University of Maryland, Computer Science Dept. Tech Report, CS-TR-4365, 2002.

http://hcil2.cs.umd.edu/trs/2002-06/2002-06.pdf

- Interactive system for querying time series among multiple time lines.
- Data: microarray data

Berry L, Munzner T. Binx: Dynamic exploration of time series datasets across aggregation levels[C]//Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on. IEEE, 2004: p2-p2.

https://www.cs.ubc.ca/~tmm/papers/binx/BinXPoster2004.pdf

- Display a single line graph that focuses scalability on the x-dimension

Related Work – Line Graph

Line graph scalability – Cont'

Ziegler, Hartmut, et al. "Visual market sector analysis for financial time series data." Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on. IEEE, 2010.
http://www.cs.ubc.ca/~tmm/courses/cpsc533c-06-fall/readings/vanwijk99cluster.pdf

- Small multiple for financial data: clusters a selection of large amounts of financial time series data according to their similarity, and analyzes the distribution of the assets among market sectors

Bernard, Jürgen, et al. "A visual digital library approach for time-oriented scientific primary data."International Journal on Digital Libraries 11.2 (2010): 111-123.
https://kops.uni-konstanz.de/bitstream/handle/123456789/28470/Bernard_284706.pdf?sequence=1&isAllowed=y

- Small multiple for scientific primary data (time-oriented research data). Cluster with SOM map, and then show the centrality of the clusters with opacity-based overlaying view.

Related Work – Line Graph

Line graph visual improvements

Heer J, Agrawala M. Multi-Scale Banking to 45º[J]. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 2006, 12(5): 2.

http://vis.pku.edu.cn/course/Visualization_2011F/materials/2006-Banking-InfoVis.pdf

- The orientations of the line segments in the graph
- **Data**: Monthly atmospheric CO_2, daily download counts of the prefuse visualization toolkit, EEG reading

Kong N, Agrawala M. Perceptual interpretation of ink annotations on line graphs[C]//Proceedings of the 22nd annual ACM symposium on User interface software and technology. ACM, 2009: 233-236.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.185.2602&rep=rep1&type=pdf

- Identify and highlight perceptual parts in line graphs
- **Data**: Annual number of sunspots, Annual number of goals scored by leading scorer in the National Hockey League

Van Wijk, Jarke J., and Edward R. Van Selow. "Cluster and calendar based visualization of time series data." Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on. IEEE, 1999.

http://www.cs.ubc.ca/~tmm/courses/cpsc533c-06-fall/readings/vanwijk99cluster.pdf

- Cluster similar daily data patterns, and to visualize the average patternsas graphs and the corresponding days on a calendar, to identify patterns and trends on multiple time scales.
- **Data:** power demand by ECN (electronic communications network)

Zhao, Jian, et al. "Exploratory analysis of time-series with chronolenses." Visualization and Computer Graphics, IEEE Transactions on 17.12 (2011): 2422-2431.

- http://www.cs.toronto.edu/~jianzhao/papers/chronolens.pdf
- Perform transformation of the data points in their focus area, visualizing derived values, identifying correlations, or discovering anomalies Data: visiting **Data**: population of two P2P video-on-demand channels, ALMA Line Length Correction Stretcher Voltage plots for four antennas

Related Work – Visual Clutter

Summary

Ellis G, Dix A. A taxonomy of clutter reduction for information visualisation[J]. Visualization and Computer Graphics, IEEE Transactions on, 2007, 13(6): 1216-1223.
http://eprints.lancs.ac.uk/12942/1/Ellis_%26_Dix_clutter_reduction_taxonomy_5.3.pdf

More related paper: clustering-based, hierarchical / aggregated / filtering, visual enhancement.

Mostly pure technical improvements. Some real-world problem motivated paper:
Clustering:
Zhang L, Tang C, Song Y, et al. VizCluster and its application on classifying gene expression data[J]. Distributed and Parallel Databases, 2003, 13(1): 73-97.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.4866&rep=rep1&type=pdf
Filtering:
Brodbeck D, Chalmers M, Lunzer A, et al. Domesticating bead: adapting an information visualization system to a financial institution[C]//Information Visualization, 1997. Proceedings., IEEE Symposium on. IEEE, 1997: 73-80.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.4938&rep=rep1&type=pdf

Related Work

Discuss the feasibility of visual clutter reduction techniques: To assess if a certain common techniques could be used in our case for line graphs. (and MOST IMPORTANTLY prove clustering is a must).

- Current clutter reduction techniques for lines graphs

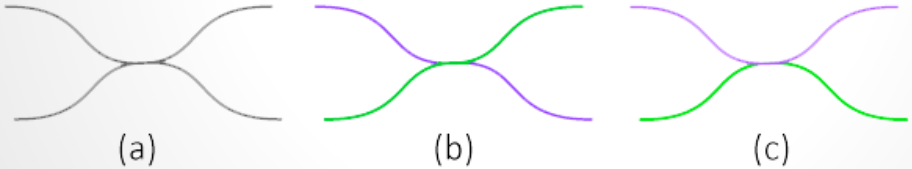
  - Small multiples (SOM clustering (Bernard, 2010), serial density map (Kincaid, 2006), etc.)
    - This separation would require users to move back and forth to understand lines in different sub-graphs, which greatly sacrifices the comparability of line graphs, and make it difficult to observe the overall trend of all lines.
  - Queries (Timebox (Hochheiser, 2004), query with scratch / example (Wattenberg, 2001), etc.)
    - Lack an overview for guiding the users for further interaction

->> Draw with a common baseline

Related Work

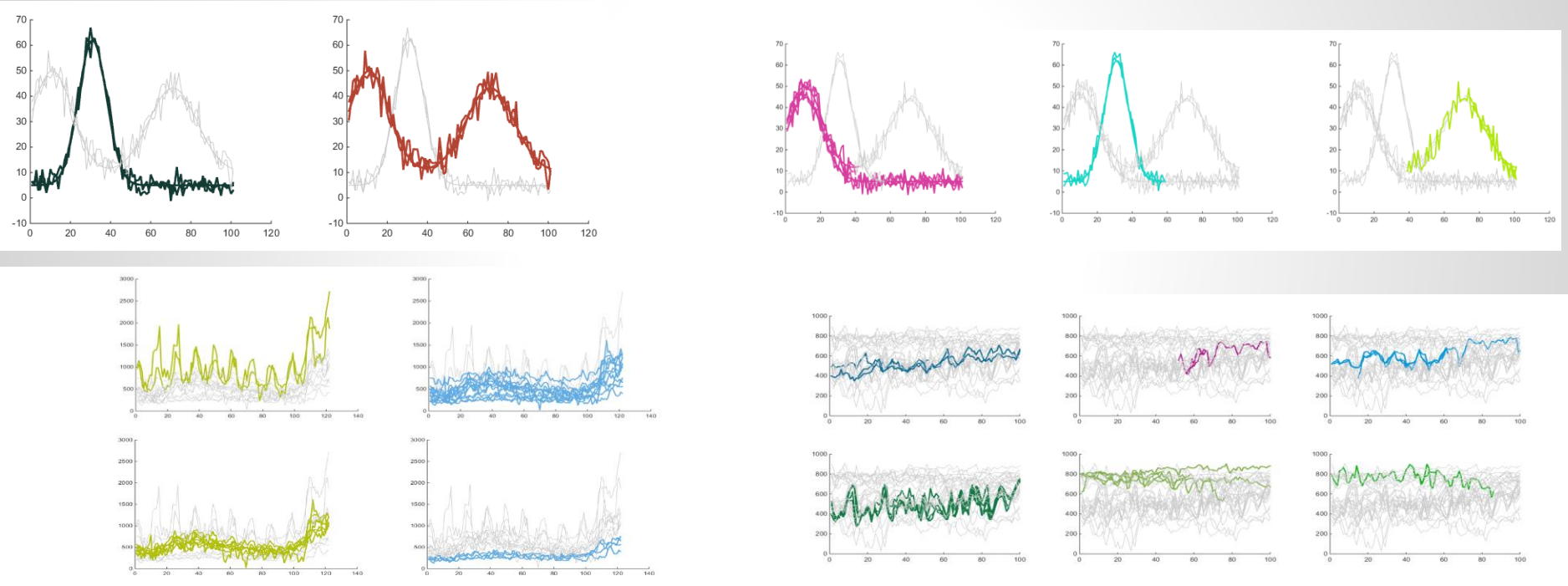
- Other related techniques
  - Edge bundling, e.g., Kernel Density Estimation (KDE) (Hurter, 2012)
    - The main pattern extracted by KDE is undirected, and the trend in line graph could not be revealed. (see the figure below.)
  - Reordering for parallel coordinates
    - Line graph's independent variable are not re-orderable.
  - etc.

->> Clustering is the one that could reduce clutter problems without much loss on the emphasis of line graphs' primal features.



Whereas KDE may summarize the patterns as (a), given that line graphs have the development direction, the real case could be (b) or (c).

Initial Results from the clustering



Expected Results

- Implementation: javascript + d3.js + WebGL
- Evaluation: two case study + a user study
  - Emphasis on user study

Evaluation - Case Study

Ideal cases

- General and understandable data
- Use data that partial similarities are very meaningful

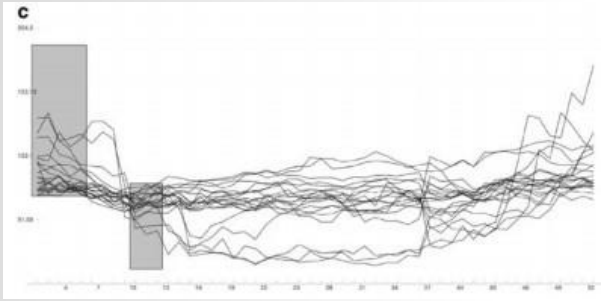
Otherwise only summarize the phenomenon without specific explanation. (e.g., micro-array data)

Data	Pro	Con	Source
Synthetic Control Graph Time Series	Classical data; partial similarity is meaningful;	Not easy to explain triggers behind the patterns	https://archive.ics.uci.edu/ml/datasets/Synthetic+Control+Graph+Time+Series (BESDOK, 2000)
Weather	Easy to understand; time series data; data range is reasonable	The change is annual-based and may not discover a lot of partial similarity patterns	Kincaid, 2006
DNA sequence	Widely used in studies	Not explanable	Kincaid, 2006
Stock	Widely used in studies; partial similarity is meaningful	Relative changes are much more important than absolute values in reality; most studies would perform normalization	Hochheiser, 2004 http://www.cs.umd.edu/hcil/timesearcher/
Search log data	Easy to understand, partial similarity is meaningful	Data may vary in a too wide range and difficult to observe very meaningful result without normalization	

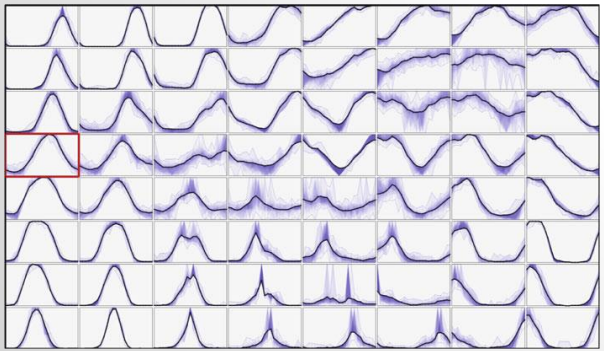
Evaluation - User Study

Baselines

- Query — Hochheiser, 2004



- Small multiples — Bernard, 2010



Potential tasks

- Point out main x-y relationships;
- Point out outliers;
- Which x-y relationship tend to be stable throughout the whole time period? Which ones do lines easily break away from?
- Lines in which regions are most similar?
- Find lines that follow several specified x-y relationship;
- Etc.

(Mainly follow the tasks analysis. TBD when the system is done, and when we understand what it is most suitable for)

Appendix:

User study task or TimeSearcher (Hochheiser, 2004) with Stock data:

- During days 22–23, are there more stocks between 69–119, 59–109, or 49–99
- Find stocks that traded in a 20 range for at least three consecutive time periods.

Discussions

A List of Expected Figures

Remaining Tasks

- Coding
  - Clustering (Mostly done. Compute the centroid)
  - Visual system Working on
- Evaluation
  - Data collection
  - Case study
  - User study
- Writing
  - W1 abstraction
  - W2 introduction
  - W3 related work *Collecting related paper*
  - W4 problem characterization
  - W5 system design
  - W6 case studies
  - W7 user study
  - W8 discussion
  - W9 conclusion and future work