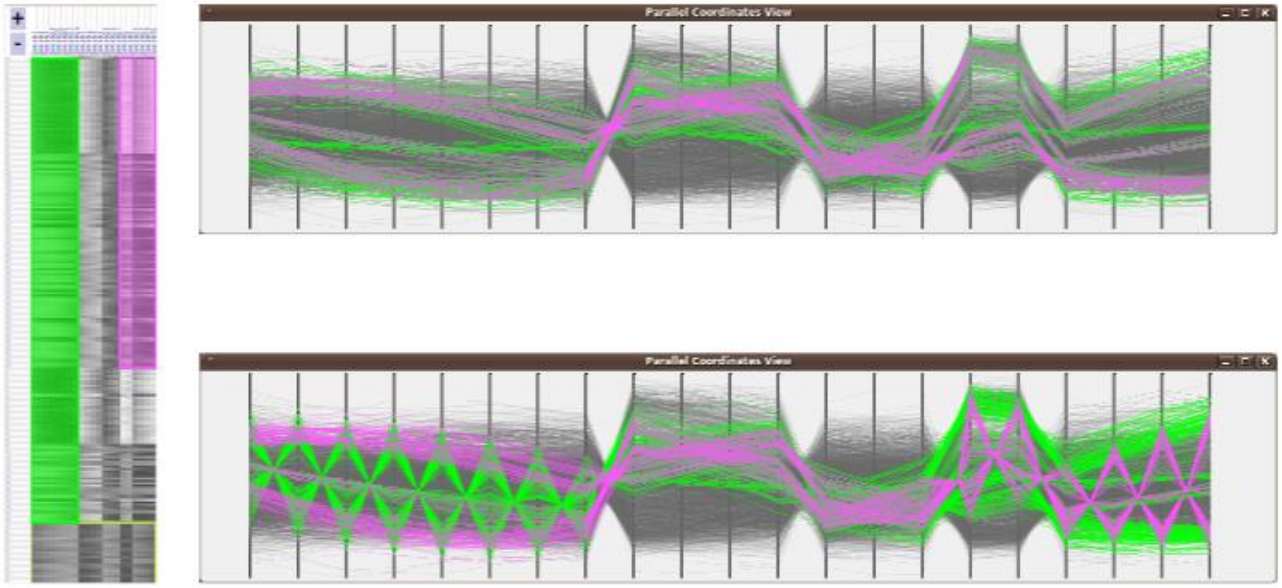
Visual Analysis of Biclusters



Outline

- Bi-Clustering: Definition and Algorithms
- Visualization of Biclusters
- Chaining Biclusters
- Summary & Outlook

Introduction

Biclustering[1]: data mining technique that searches for subsets in HD data that share the same values across different dimensions.

Examples:

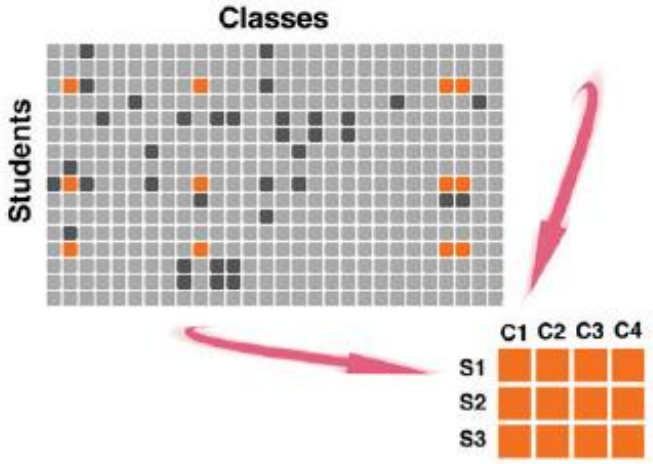
- A set of Students that take the same courses
- A set of students coming from the same city.
- These 2D biclusters are linked to each other related by the students involved.

→ Biclusters biclusters reveal relations.

In most applications, data are categorical, i.e. values are compared for being (exactly) equal.

[1]also called *co-clustering* or *block clustering*

Introduction



In a large table, it is difficult to recognize which elements in specific rows are related to elements in specific columns. This *n:m* relationship *R* between a subset of rows and columns *ExF* is detected by biclustering
(From: Sun, 2014)

Introduction

- Biclustering was introduced in statistics (Hartigan, 1972)
- The name biclustering was introduced later (Mirkin, 1996)
- Biclustering was generalized to objects with similar (numerical) values for use in gene expression data (Cheng, 2000)

  → Bicluster analysis revealed identical metabolic pathways
- Another essential application is text analysicts (a set of documents) where biclustering represents co-occurence of words (e.g. intelligence analysis)
- Clusters represent a partioning of the database (all objects belong to one cluster; clusters do not overlap).
- Biclusters (often heavily) overlap and many objects do not belong to biclusters (without being outliers).

Introduction

- Biclustering algorithms – like subspace clustering – reveals usually many (thousands) of biclusters
- Biclusters – again like subclusters – are often similar and thus redundant.
- We are interested in biclusters of maximum size (*closed biclusters*), e.g. the property of being a bicluster is violated if any row or column is added.
- Biclusters are a kind of abstraction of relations between data and serve as orientation for further more in-depth analysis.
- Biclusters are more flexible and fit biological behavior represented in gene expression data better than clusters (Santamariaa, 2008)

Definition and Algorithms

Definition: A bicluster (E‘, F‘) on a relation R(E,F) is a set E‘ ⊆E and a set F‘⊆F such that E‘xF‘⊆R.

Thus, there is a relation between every element of {E} and {F}.

In graph theory, a bicluster corresponds to a *complete bipartite graph*[1] – a graph where the set of vertices *N* may be split in two sets N_1 and N_2 where each node connects a node from N_1 and a node from N_2.

→ no edge connects vertices from N_1 and no edge connects vertices from N_2.

[1]also called Biclique (see Wikipedia)

Definition and Algorithms

- Algorithms are often parameterized w.r.t. minimum number of rows and/or columns that should be returned.
- CHARM (Closed Association Rule Mining, Zaki, 2002) searches for frequent itemsets.
- CHARM is more efficient than previous bottom-up search methods
- The output may be used for association rule mining and for defining biclusters.
- A number of bicluster algorithms are available as R plugins, e.g. (Kaiser, 2013)

Definition and Algorithm

- The exhaustive search for biclusters is NP complete.
- Heuristic algorithms often define biclusters bottom-up (Cheng, 2007)
- In every row possible clusters are determined first
- In the next step row clusters are analyzed and merged to form big biclusters when the number of rows does not „decrease too much“.
- If applied to numerical data, two types of comparison occur:
  - The absolute difference between two items (additive-related)
  - The relation between two items (multiplicative-related)
- Analytics tools should provide both

Algorithms

- When applied to numerical data, representing measurements, the presence of noise has to be considered.
- Biclustering algorithms estimate the noise level and consider a subset of the rows and columns as biclusters if the values differ only in an extent that is below the estimated noise level.
- Biclustering algorithms differ in assumptions and heuristics; they may be stochastic or deterministic (see [Tanay, 2004] and Wikipedia for surveys)
- The Bimax algorithm (Prelic, 2006) delivers a complete set of biclusters.

Algorithms

- Most algorithms return hard clustering results (binary membership).
- For numerical data, fuzzy clustering better represents the membership of objects to a bicluster.
- FABIA (FABIA: factor analysis for bicluster acquisition, Hochreiter, 2010) is an essential fuzzy biclustering technique, e.g. for clustering gene expression data, drug discovery and recommender systems.
- FABIA assumes a sparse matrix where only few rows and few columns belong to a bicluster, e.g. representing a pathway with few genes and few samples involved.

Examples

Cohort study data:

- Persons that share a number of attributes, e.g. co-occurence of different diseases or different risk factors

Cerebral aneurysms

- Patients with ruptured/unruptured aneurysms and a certain morphology (bleb) or flow feature, e.g. inflow jet, embedded vortices

Cardiac diseases

- Patients with the same disease and severity and flow patterns

Preprocessing

Depending on the nature of the data, different preprocessing steps may be performed:

- Outlier removal
- Transformation of values (in gene expression data often a log transform of expression levels is performed)
- Normalization of data in case of different scales

Visualization of Biclusters: Tasks

Visualization tasks:

- Gain an overview of clusters
  - Size, involved rows/columns
- Show individual biclusters
  - Members, additional metadata per member
- Display a set of biclusters simultaenously
  - Enable comparison w.r.t. size and distribution of values
  - Indicate overlaps
- Explore properties of individual and sets of biclusters

For fuzzy biclusters (Streit, 2014):

- Display membership values and support transformation to hard clusters

Visualization of Biclusters

Requirements:

In addition to the vis. tasks, scalability requirements need to be considered:

Techniques should be scalable, w.r.t.

- Number of rows and columns,
- Number of biclusters,
- Number of rows and columns associated with several biclusters (overlap)

Visualization of Biclusters: Parameters

Parameters for Initial Generation or Filtering:

- Minimum number of rows/columns *(minimum support parameter)*
  - Users often use high values and decrease as long as the number of biclusters is manageable
- Percentage minimum (and/or maximum) number of rows/columns
- Noise threshold for numerical data (for the comparison)
- Overlap between biclusters (like with subspace clusters many biclusters are similar and may be redundant for the analyst)

A maximum support parameter may also be useful; otherwise biclusters are generated that indicate that many persons are married and have two children – not surprising.

Visualization of Biclusters

- Graph visualizations
  - Edge bundling for reducing clutter
- Table visualizations
  - Emphasis of Biclusters within tables
- Parallel coordinates
- Node-link diagrams for displaying chained biclusters
- Set-based visualizations

Visualization of Biclusters: Bixplorer

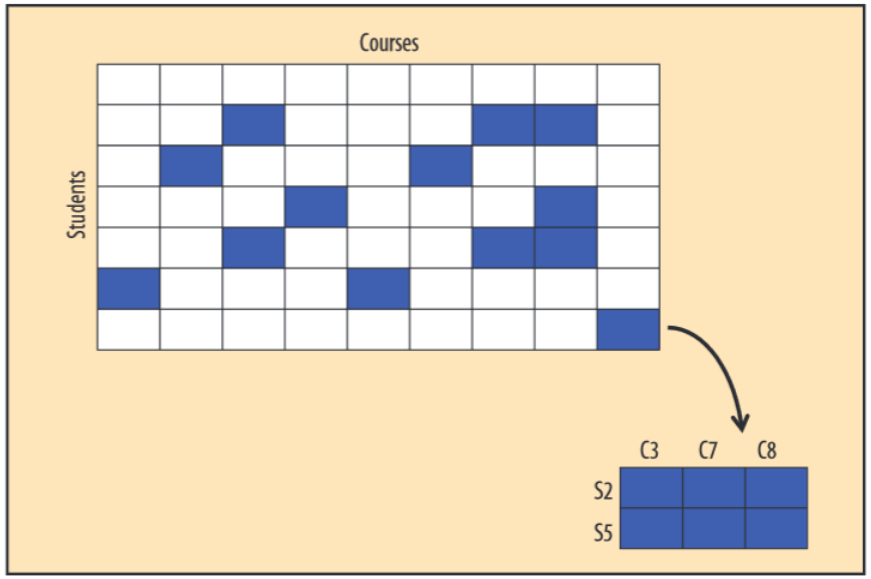
Bixplorer (Sun, 2014; Fiaux, 2013): Visual analytics prototype based on biclustering.

Bixplorer supports the following tasks:

- Biclusters and chains of biclusters serve as starting point for further analysis,
- Reveals connections,
- Biclusters serve to label „organized“ information

Bixplorer supports search in documents and biclusters. Bicluster may represent co-occurrence of words in documents.

Visualization of Biclusters



In a 2D matrix of courses and students, one bicluster (with at least two rows/columns) was identified (From: Fiaux, 2013).

The bicluster is labeled with the related items.

By reordering the matrix, the bicluster could be emphasized.

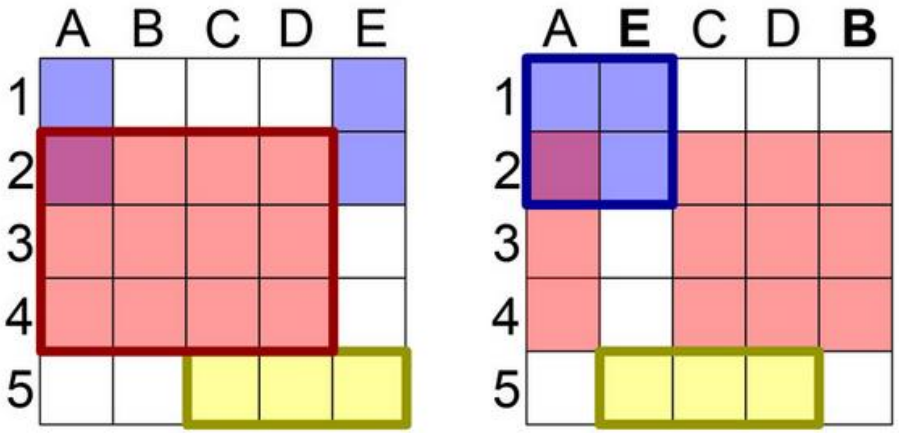
Visualization of Biclusters

Within one 2D subspace, several (overlapping) biclusters may occur.

The re-ordering of cells in a table display and an explicit visualization of the overlap visually represent this information.

Without duplication of rows or columns this technique is restricted to displaying two biclusters

Visualization of Biclusters

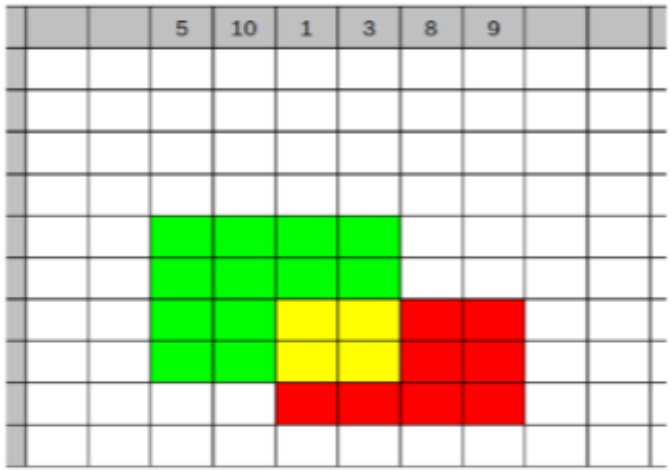


(From: Streit, 2014).

To display >2 biclusters in contiguous rows/columns requires to duplicate rows/columns.

Each bicluster should be mapped to a unique color.

Visualization of Biclusters



Duplicated entries should be tagged, e.g. with a transparent background (From: Heinrich, 2011).

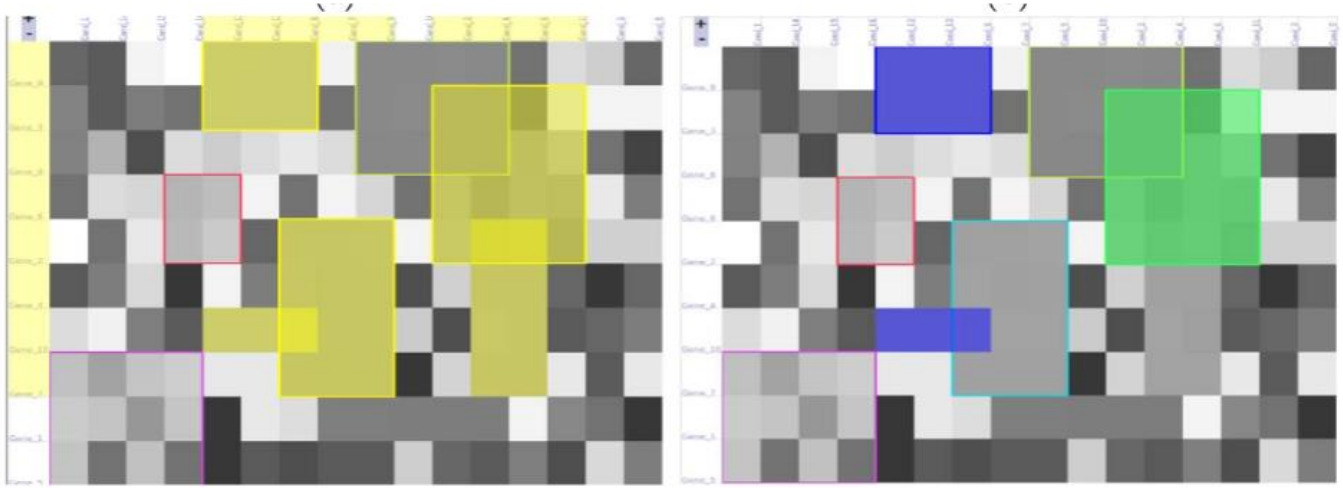
This duplication does not scale well: For HD data, many rows and columns need to be inserted.

Each bicluster should be displayed with a unique color → limited to a few dozens of biclusters

Visualization of Biclusters: Interaction

- Several algorithms compute a layout for overlapping biclusters with minimal duplication, e.g. (Grothaus, 2006).
- Interactive adjustment (Heinrich, 2011):
  - Users may select one bicluster as focus
  - Biclusters are added sorted by descending overlap to the focus bicluster
- Further interaction with the bicluster view
  - Select biclusters for temporary or permanent hightlighting
  - Sort biclusters, e.g. by bicluster size
  - Label biclusters with a descriptive term
  - Zoom in and out to see names of items
  - Show biclusters without duplicating rows and columns → use colors to reveal which separated regions form the same bicluster.

Visualization of Biclusters: Interaction



Bicluster viewer: Interactive Exploration of biclusters. Some biclusters are selected and emphasized. In the left image, the involved dimensions are also shown (From: Heinrich, 2011)

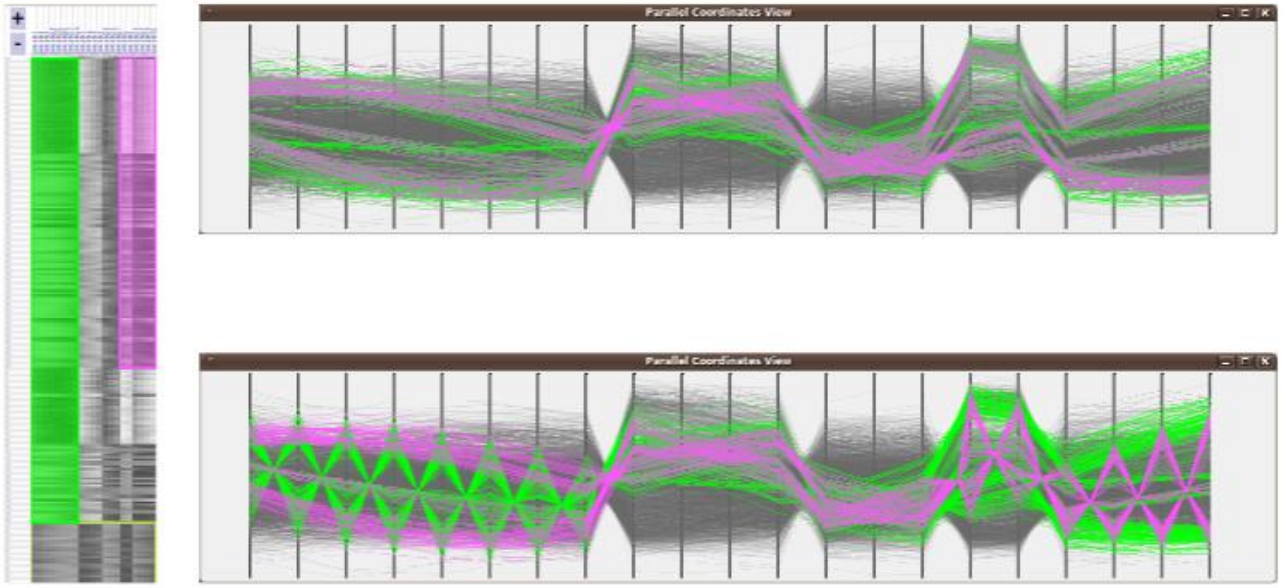
Visualization: Parallel Coordinates

- Items with all their attributes are represented as lines in Parallel coordinates (PC) (Cheng, 2007)
- Lines are color-coded to represent membership to a bicluster
  - Problem: Items that belong to several biclusters
- PC display is linked to a table-based heatmap to enable brushing and linking
- Colors in both displays are used equally.

Visualization: Parallel Coordinates

- In addition to displaying biclusters, single rows/columns (representing genes) may be selected and compared with the biclusters
- BiVisu (Cheng, 2007) employs PC plots as major visualization technique (From: Link). In their example, 2900 genes in 17 conditions were analyzed resulting in some 117 biclusters with the additive and 93 biclusters with the multiplicative model were detected (overlap < 80%, minimum 6 columns).

Visualization: Parallel Coordinates



Bicluster viewer: Integrated bicluster visualization with heatmaps (left) and parallel coordinates. In the bottom row, the centroids of the biclusters are emphasized
(From: Heinrich, 2011)

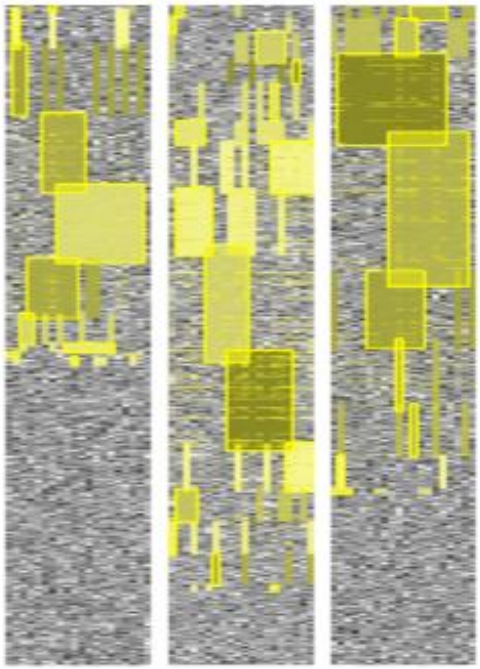
Visualization of Biclusters

- **Bixplorer**: user study revealed that users were interested in the frequency of items when selecting biclusters.
- They chose not the most frequent items, because if an item occurs everywhere (e.g. each student choses math), it is not interesting.
- Biclusters served as „lead“ information to guide the selection of individual documents that are interesting.
- Layouts (prev. Slide) were considered analytical results.
- Users employed a subset of the available biclusters (less than 10% from the 1001 biclusters)

Visualization of Biclusters: Layout

10, 9 and 15 biclusters are generated for three different datasets. Users are interested in the involved rows and columns, in the size of the clusters and the overlap (From: Heinrich, 2011).

The overlap may be used to emphasize e.g. larger biclusters by drawing them on top (2 ½ D)



Visualization of Biclusters

Advanced filtering of biclusters. Heatmap as overview visualization. For the selected bicluster, the gene profile view (PC plot on the right) shows their expression levels in the involved conditions (From: Barkov, 2006).

Visualization of Biclusters: Graph Display

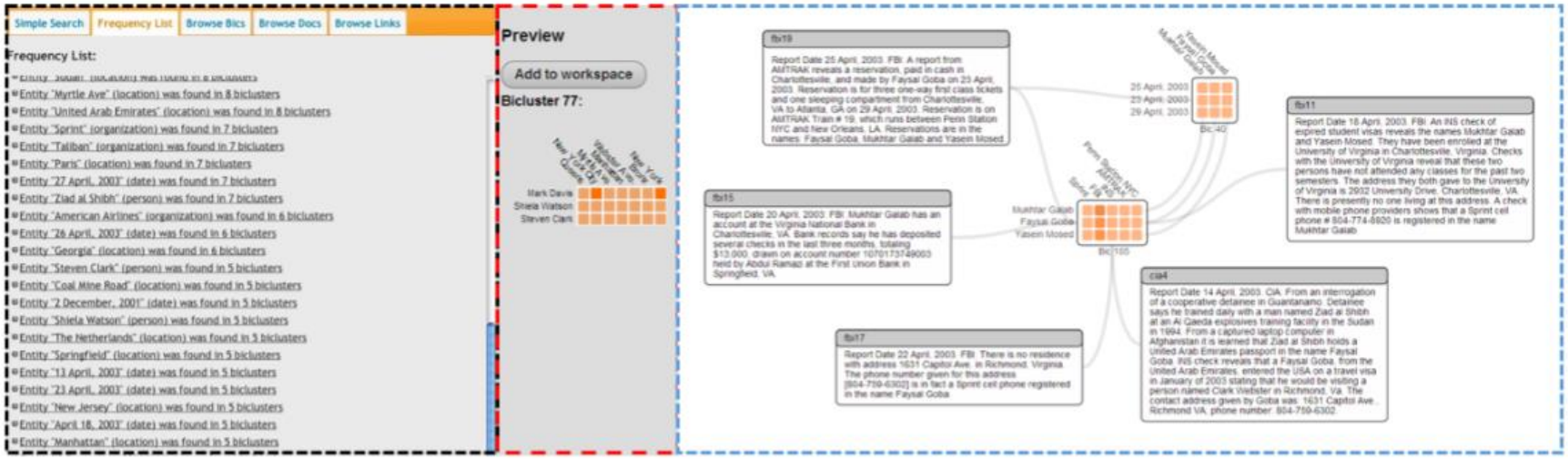
Chained biclusters:

If several documents or high-dimensional data are analyzed *m:n: .. : z* relations may occur that represent connections between a set of documents or dimensions.

These chained biclusters may be represented as node-link diagramms with biclusters represented by nodes.

Example: A number of patients share a symptom, a diagnostic procedure (e.g. a biopsy) and a treatment, e.g. with a drug.

Visualization of Biclusters: Graph Display

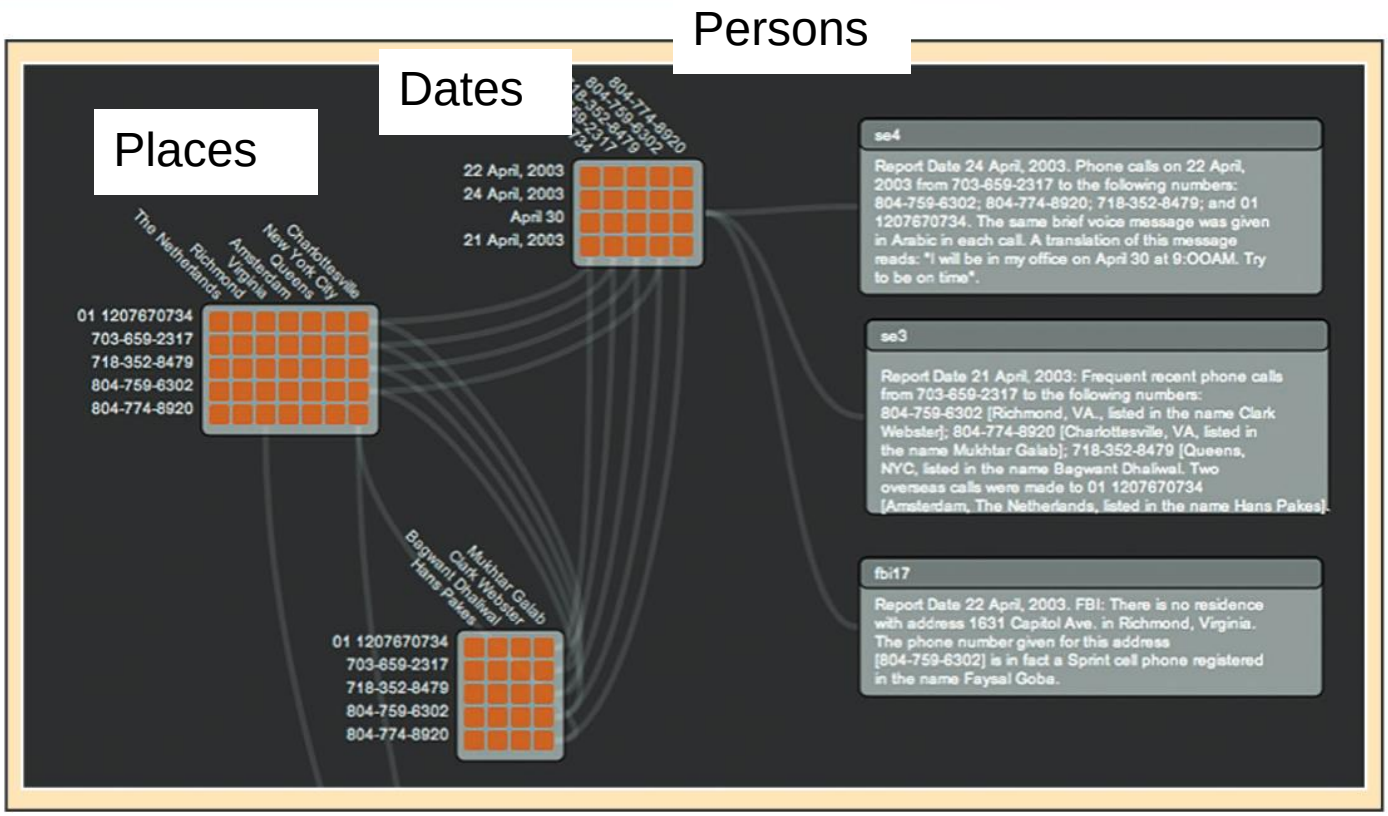


Bixplorer: From a set of entities that occur frequently (left) users may select a pair and identify biclusters that may be dragged to the workspace and labeled (From: Sun, 2014).

The darkness of the color decodes how similar the entries are (according to the additive or multiplicative model).

Chains of biclusters may be detected efficiently (Wu, 2014)

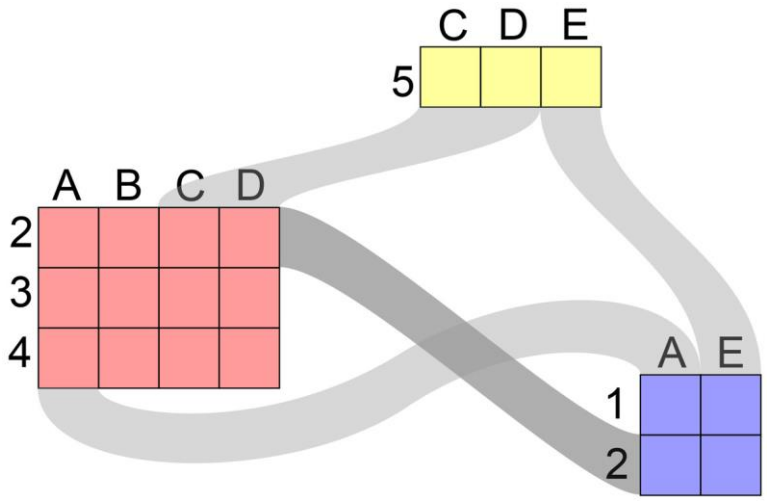
Visualization of Biclusters: Graph Display



With the Bixplorer different biclusters were identified and connected via joined attributes. The items are linked to (police) reports (From: Fiaux: 2013).

Chains may also be used for overlapping biclusters. Edges represent which (subset of) nodes of one bicluster are also part of another.

Visualization of Biclusters: Graph Display

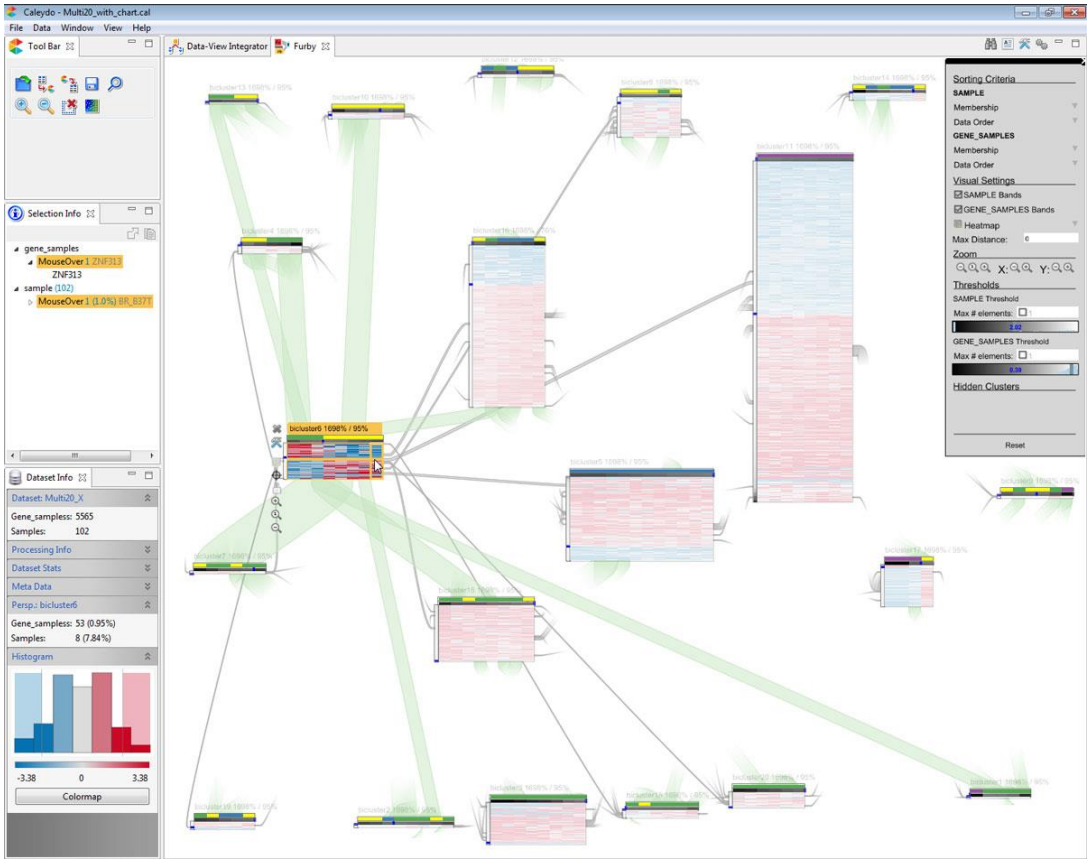


(From: Streit, 2014).

Edges between (overlapping) biclusters may be bundled and enable a good overview on biclusters. The width of the bands encodes the frequency of objects in the associated biclusters.

This technique scales well for data of a realistic complexity.

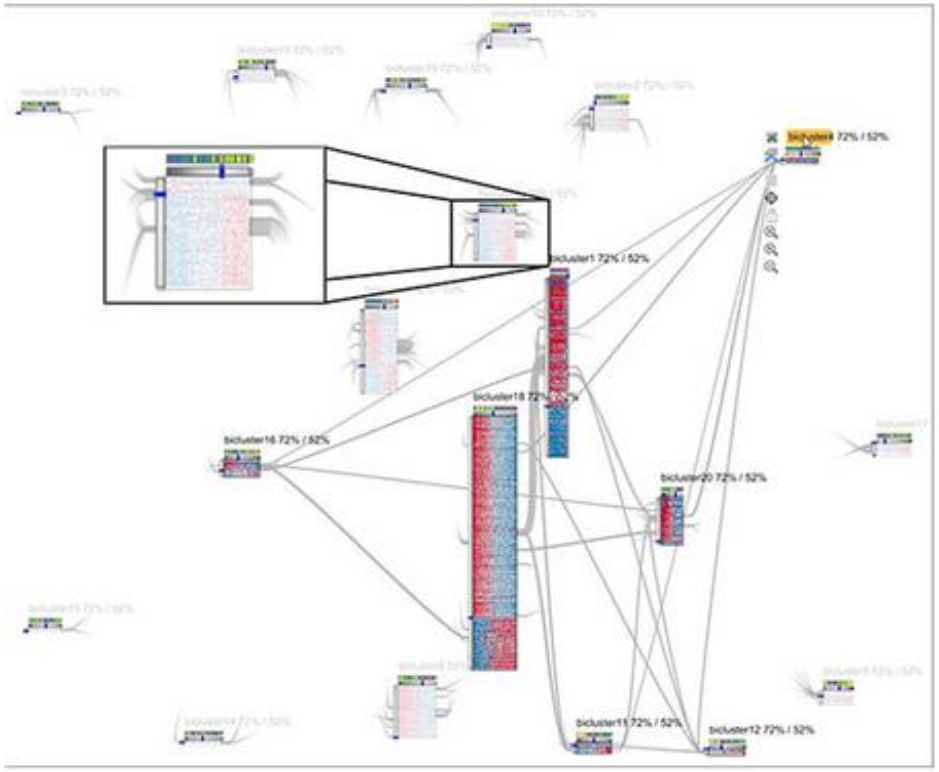
Visualization of Biclusters: Graph Display



(From: Streit, 2014).

Overview visualization of gene sample data analyzed with Biclustering. Layout (20 biclusters) is computed with a force-directed graph layout. Overlapping biclusters attract each other.

Visualization of Biclusters: Graph Display

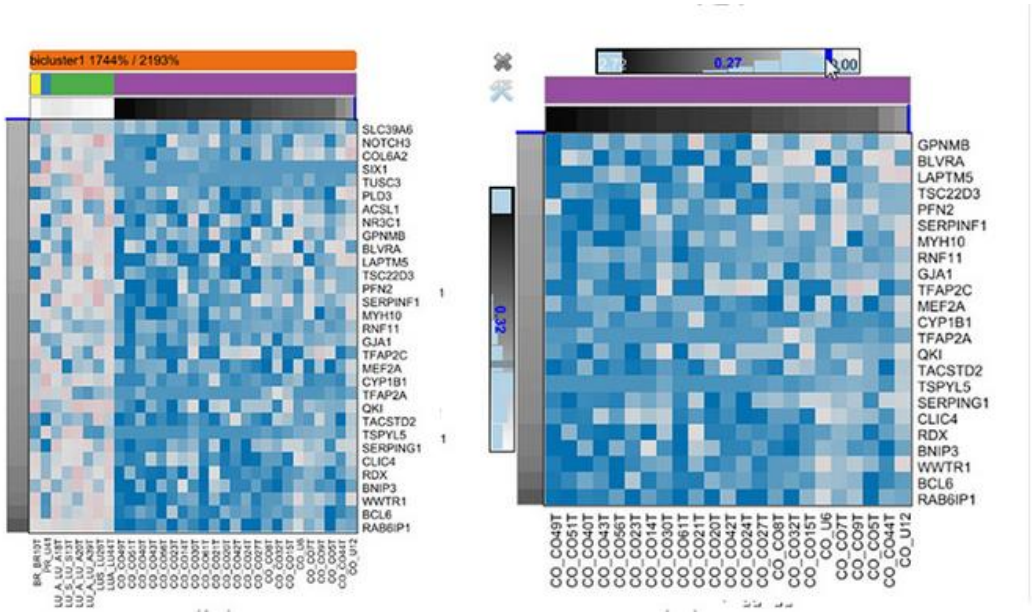


(From: Streit, 2014).

Detail visualization for a selected bicluster. To accommodate the additional information some biclusters faded out. Their edges are represented with stubs.

Visualization of Biclusters: Graph Display

The application to fuzzy biclustering means that membership thresholds are adapted. As a conseque, biclusters appear/disappear, grow or shrink.



(From: Streit, 2014).

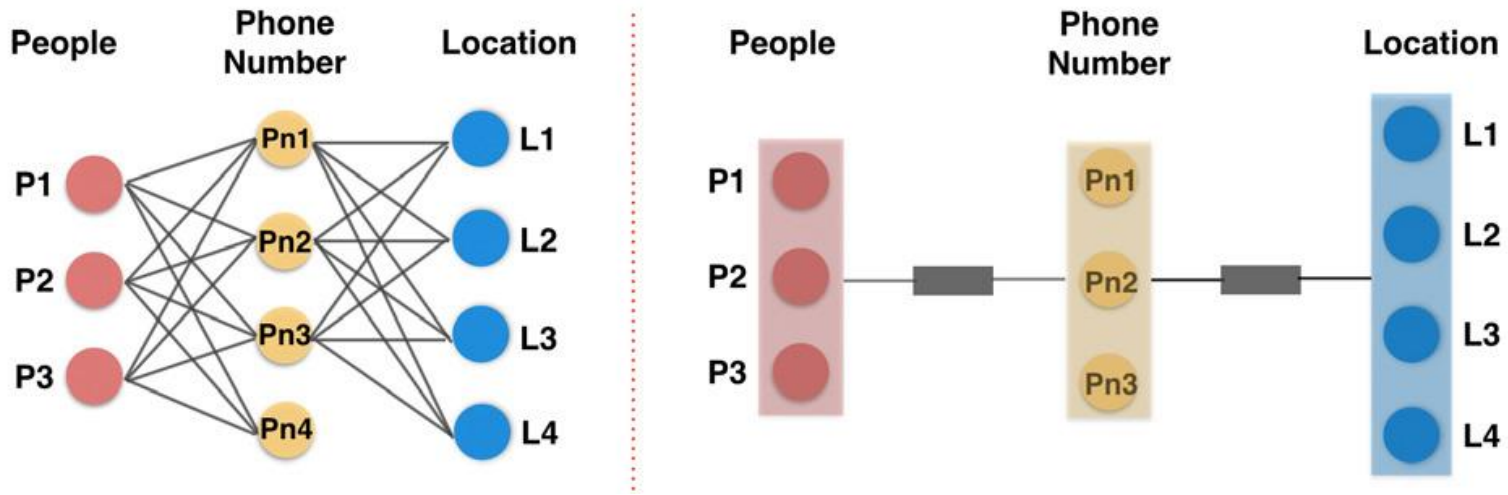
The detail visualization of a bicluster changed after manipulating the threshold.

Visualization of Biclusters: Set-based

Bi-Set: From left to right: a naive graph representation with edges between elements of both sets. More compact representations reducing clutter taking advantage of the complete connectivity between both sets (From: Sun, 2016).

The visualization shows that the individual connections may be *abstracted* in connections between sets.

Visualization of Biclusters: Set-based



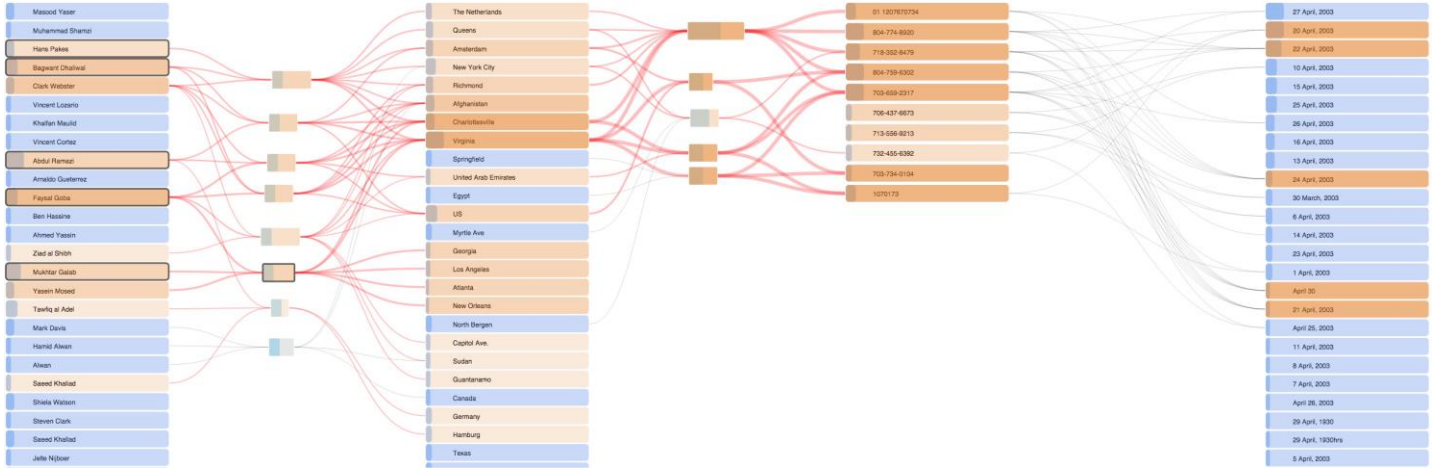
(From: Sun, 2016).

Bi-Set

- Visual abstraction and semantic edge bundling as major concepts for a set-based visualization of biclusters.
- Extension to bicluster chains is possible.
- Nodes represent entities and edges represent relations between sets of entities.

Visualization of Biclusters: Set-based

With entities drawn as lists connected by edge bundles, an „in-between“ layer may be added to display properties of the relation, e.g. how many items are involved.



Bi-Set: Chained biclusters with list views and an in-between layer. Two color hues distinguish items part of a bicluster or not. Different shades of orange represent to how many biclusters an item belongs. Some items (and their associated edges) are highlighted (From: Sun, 2016)

Visualization of Biclusters: Set-based

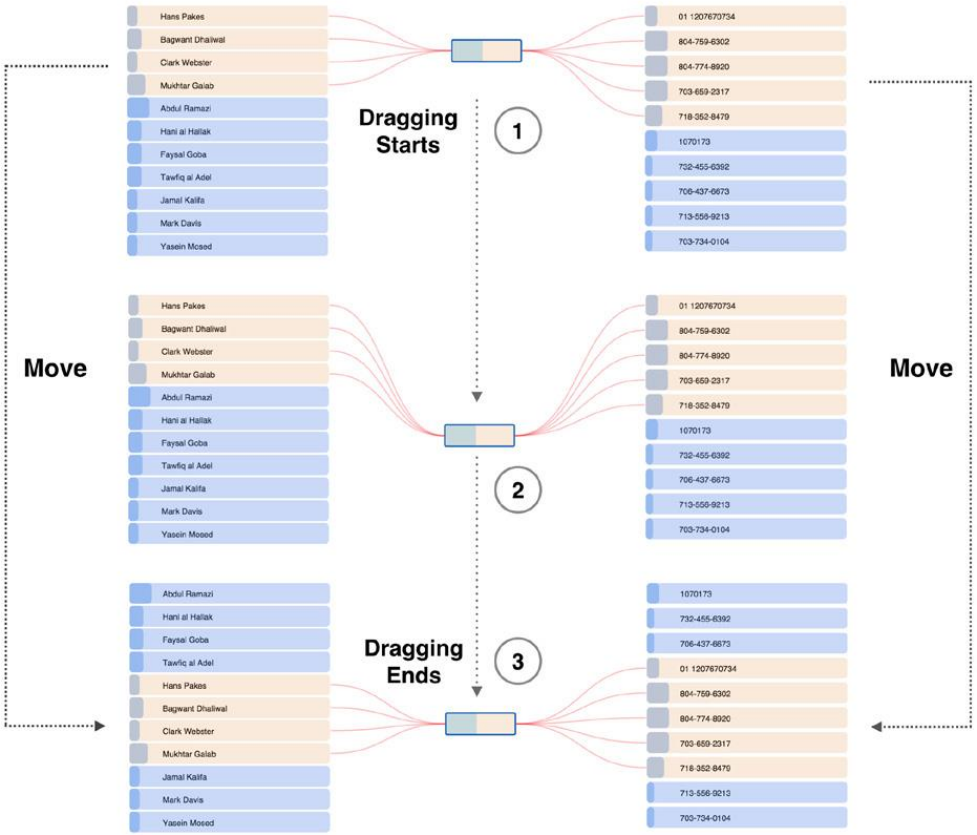
Further encodings:

- Length of the bundle represents size of the bicluster.
- Small rectangles in the item list view represents frequency in the dataset (e.g. in text documents, where words/names may occur multiple times)

Interaction:

- Items may be ordered automatically (based on lexicographic order, frequency, or association to biclusters)
- Items may be interactively placed

Visualization of Biclusters: Set-based

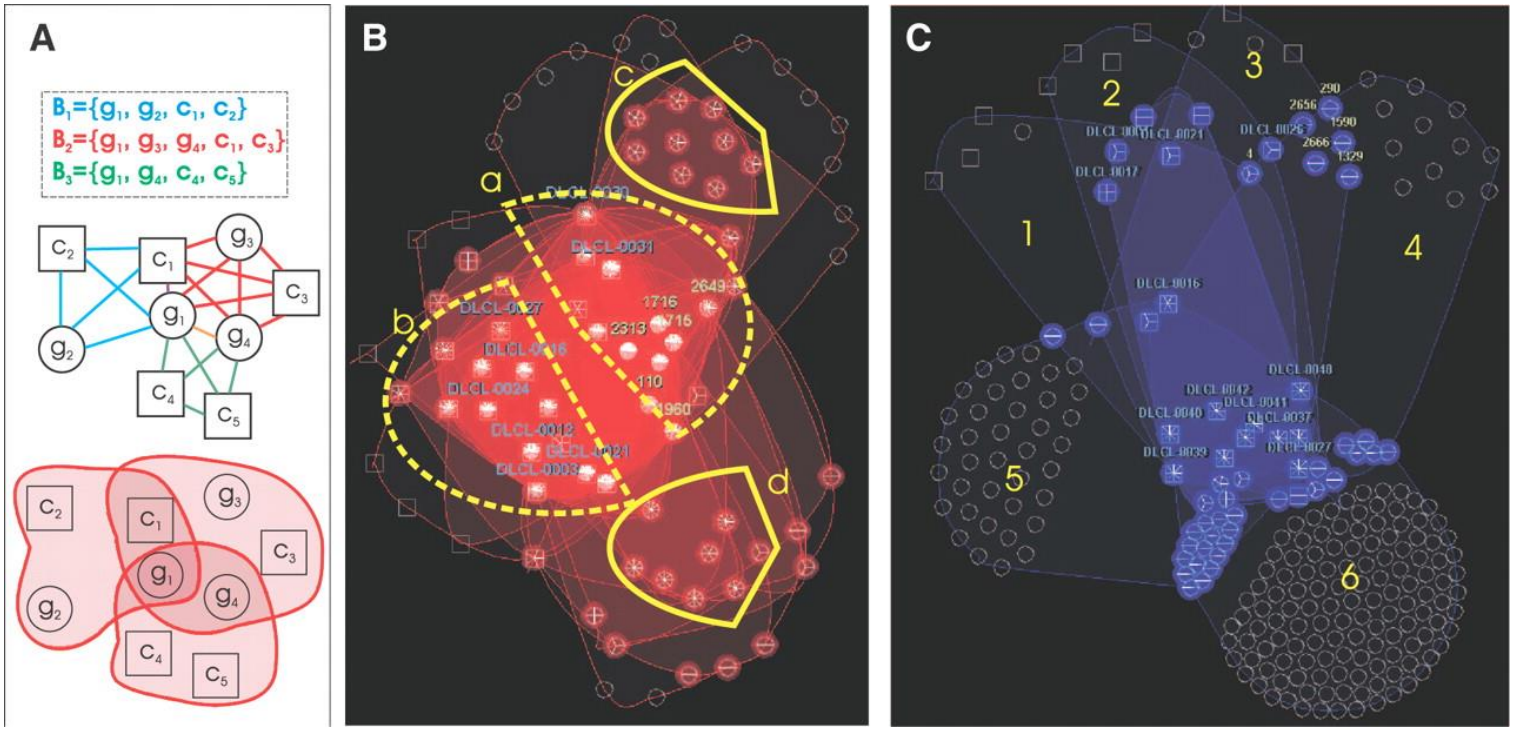


Interactive adjustment of items. The movement also affects the associated edge bundles (From: Sun, 2016). Edge bundling has a semantic basis; not spatial proximity. Thus it is stable after changes.

Visualization of Biclusters: BicOverlapper

- The use of node-link diagrams for chained biclusters does not scale well to large amounts of biclusters.
- The multitude of links (edges) leads to clutter that may be reduced with the general techniques (curved lines, edge bundling).
- Some techniques avoid edges at all, e.g. the BicOverlapper.
- Biclusters are wrapped in hulls that are drawn transparently to reveal overlaps.

Visualization of Biclusters: BicOverlapper



BicOverlapper: Biclusters are shown as set of genes and conditions, as bipartite graphs and as Euler diagrams revealing their overlap.

Nodes representing genes and conditions have different shape.

In „real“ data (B and C) more biclusters occur. Within a node link diagramm sets of genes that are part of many biclusters are recognizeable (From: Santamaría, 2008)

Visualization of Biclusters: BicOverlapper

Interactive exploration:

- Search for nodes (genes, conditions),
- Highlight of all connections to selected nodes
- Adapt parameters of the force-directed layout algorithm,
- Fix node positions,
- Navigation through the graph

Applications of Biclusters

- Most publications in this area are driven by and focused on gene expression data.
- Biclusters help to identify genes that express similar proteins along different conditions, e.g. when being in different types of tissues (normal, cancer, different subtypes) or after treatment. → This similarity likely reflects the same metabolic pathways.
- Other publications (Sun, 2014 – Sun, 2016) aim at intelligence analysis based on text occurence in documents.

Systems and Techniques

System	Techniques
BicOverlapper (Santamaría, 2008)	Multiple coordinated view framework with Euler diagrams, transparent overlays, Parallel coordinates, comprehensive interaction techniques
Bixplorer (Sun, 2014; Fiaux, 2013)	Node link diagramms, annotations. Only system that is NOT intended for microarray analysis.
Bicluster viewer (Heinrich, 2011)	Comprehensive interaction techniques and highlighting
BiVisu (Cheng, 2007)	Different cluster models, simple parallel coordinates plot
BicAT: a biclustering analysis toolbox (Barkow, 2006)	First comprehensive tool providing 5 bicluster and 2 cluster algorithms. Gene Profile Plots (parallel coordinates) and matrix visualizations (heat map), extensive preprocessing (normalization schemes)
BiSet (Sun, 2016)	List views for items (rows, columns) and an in-between layer that represents relations
Furby (Streit, 2014)	Visualization aiming for scalability (many biclusters), representation and manipulation of fuzzy memberships

Some Useful Links

Software

FABIA Biclustering Software for gene expression data

Videos

- Fuzzy Force-Directed Bicluster Visualization
- Biclustering multivariate data for correlated subspace mining

Summary

- Biclusters are an essential structure in HD data
- In some applications, in particular gene analysis, they are more expressive than conventional (subspace) clusters
- Biclustering employs algorithms also used for association rule mining
- Visualization include node-and-link diagrams for chains of biclusters, table-based heatmaps and parallel coordinates.
- Similar to (subspace) clustering, individual clusters and sets of clusters need to be explored and compared.
- An overview first, ..., details on demand strategy is promising.

Summary

Open research (see also Sun, 2014 and Streit, 2014):

- Good overview visualizations for large amounts of biclusters are missing. Layout computation is (also computationally) challenging.
- More research is needed how individual views should be combined and how this affects sensemaking.
- How much guidance is reasonable to support the exploration process related to large sets of biclusters.
- Biclustering should be enhanced with other data mining and statistics tools, e.g. w.r.t. variance and skewness of raw data.

References

Barkow S, Bleuler S, Prelic A, Zimmermann P, Zitzler E. “BicAT: a biclustering analysis toolbox”, *Bioinformatics*. 2006, 15;22(10):1282-3.

Cheng, Y., Church, G.: Biclustering of expression data. In: *Proc. of International Conference on Intellegent Systems for Molecular Biology*. (2000) 93-103

Cheng, K., Law, N., Siu, W., Lau, T.: BiVisu: Software tool for bicluster detection and visualization. *BMC Bioinformatics* 23 (2007) 2342-2344

Patrick Fiaux, Maoyuan Sun, Lauren Bradel, C. North, N. Ramakrishnan, A. Endert. 2013. Bixplorer: Visual Analytics with Biclusters. *Computer* 46, 8 (2013), 90-94.

Grothaus, G., Mufti, A., Murali, T.: Automatic layout and visualization of biclusters. *Algorithms for Molecular Biology* 1 (2006)

J. A. Hartigan. Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, 67(337):123–129, Mar. 1972

Julian Heinrich, Robert Seifert, Michael Burch, Daniel Weiskopf: BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data. *ISVC* (1) 2011: 641-652

Hochreiter S, Bodenhofer U, Heusel M, et al. (2010). "FABIA: factor analysis for bicluster acquisition". *Bioinformatics* 26 (12): 1520–1527

Kaiser, S., Santamaria, R., and Others (2013). biclust: BiCluster Algorithms. R package version 1.0.2.

References (II)

Mirkin, Boris (1996). *Mathematical Classification and Clustering*. Kluwer Academic Publishers

Prelic, A. et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22: 1122–1129

Santamaría R, Therón R, Quintales L. BicOverlapper: a tool for bicluster visualization. *Bioinformatics*. 2008 1;24(9):1212-3.

Santamaría, R.; Therón, R. and Quintales, L. BicOverlapper 2.0: Visual Analysis for Gene Expression. *Bioinformatics* 30(9), 1785-86, 2014

Maoyuan Sun, Lauren Bradel, Chris L. North, and Naren Ramakrishnan. 2014. “The role of interactive biclusters in sensemaking”, In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI '14).

Maoyuan Sun, Peng Mi, Chris North, Naren Ramakrishnan: BiSet: Semantic Edge Bundling with Biclusters for Sensemaking. IEEE Trans. Vis. Comput. Graph. 22(1): 310-319 (2016**)**

Marc Streit, Samuel Gratzl, Michael Gillhofer, Andreas Mayr, A. Mitterecker, S. Hochreiter: Furby: fuzzy force-directed bicluster visualization. *BMC Bioinformatics* 15(S-6): S4 (2014)

A.Tanay. R. Sharan, and R. Shamir, "Biclustering Algorithms: A Survey", In *Handbook of Computational Molecular Biology*, Edited by Srinivas Aluru, Chapman (2004)

H. Wu, J. Vreeken, N. Tatti, and N. Ramakrishnan. Uncovering the plot: detecting surprising coalitions of entities in multi-relational schemas. *Data Mining and Knowledge Discovery*, 28(5-6):1398–1428, 2014.

Zaki, M. J., and Hsiao, C.-J. Charm: An efficient algorithm for closed itemset mining. In *SDM*, vol. 2 (2002), 457–473