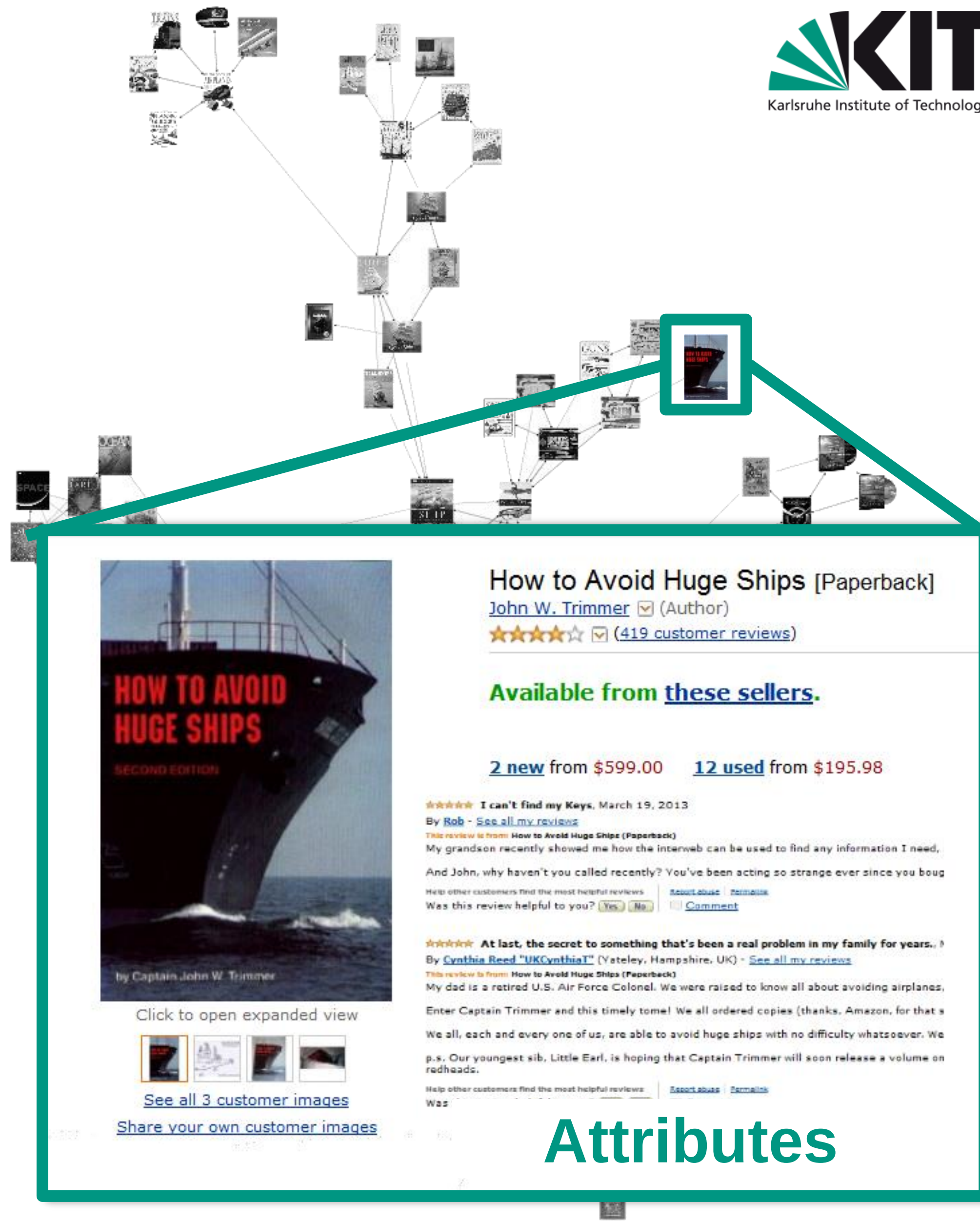
Ranking Outlier Nodes in Subspaces of Attributed Graphs

Emmanuel Müller, Patricia Iglesias, Yvonne Mülle, Klemens Böhm

International Workshop on Graph Data Management (GDM 2013)
in conjunction with IEEE International Conference on Data Engineering (ICDE 2013)

Motivation

- Networks
  - Communication networks
  - Social networks
  - Auction networks
  - Co-purchased networks



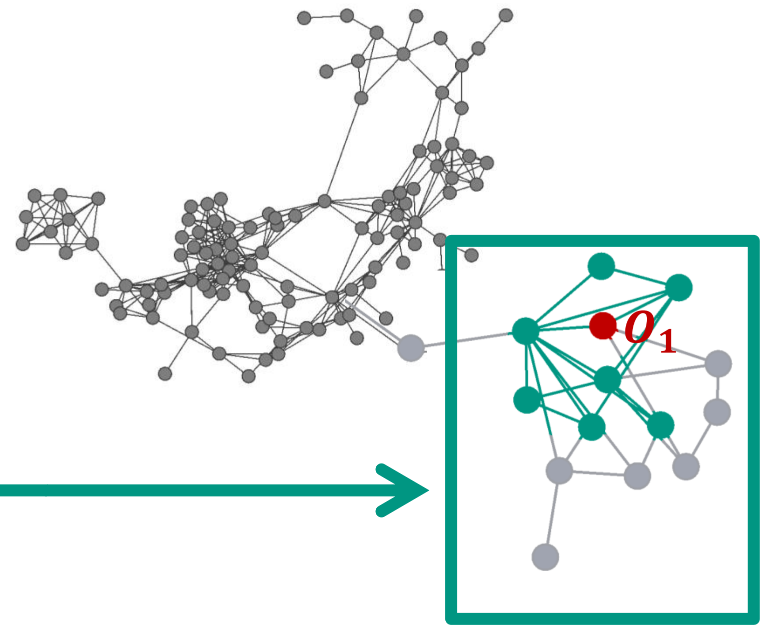
- Application
  - Fraud detection
  - Spam detection
  - Network intrusion analysis

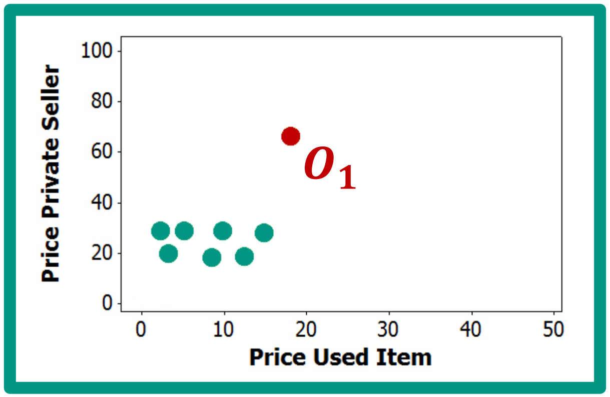
Example: Outlier Mining on Attributed Graphs

- **Input:**

Node Attributes

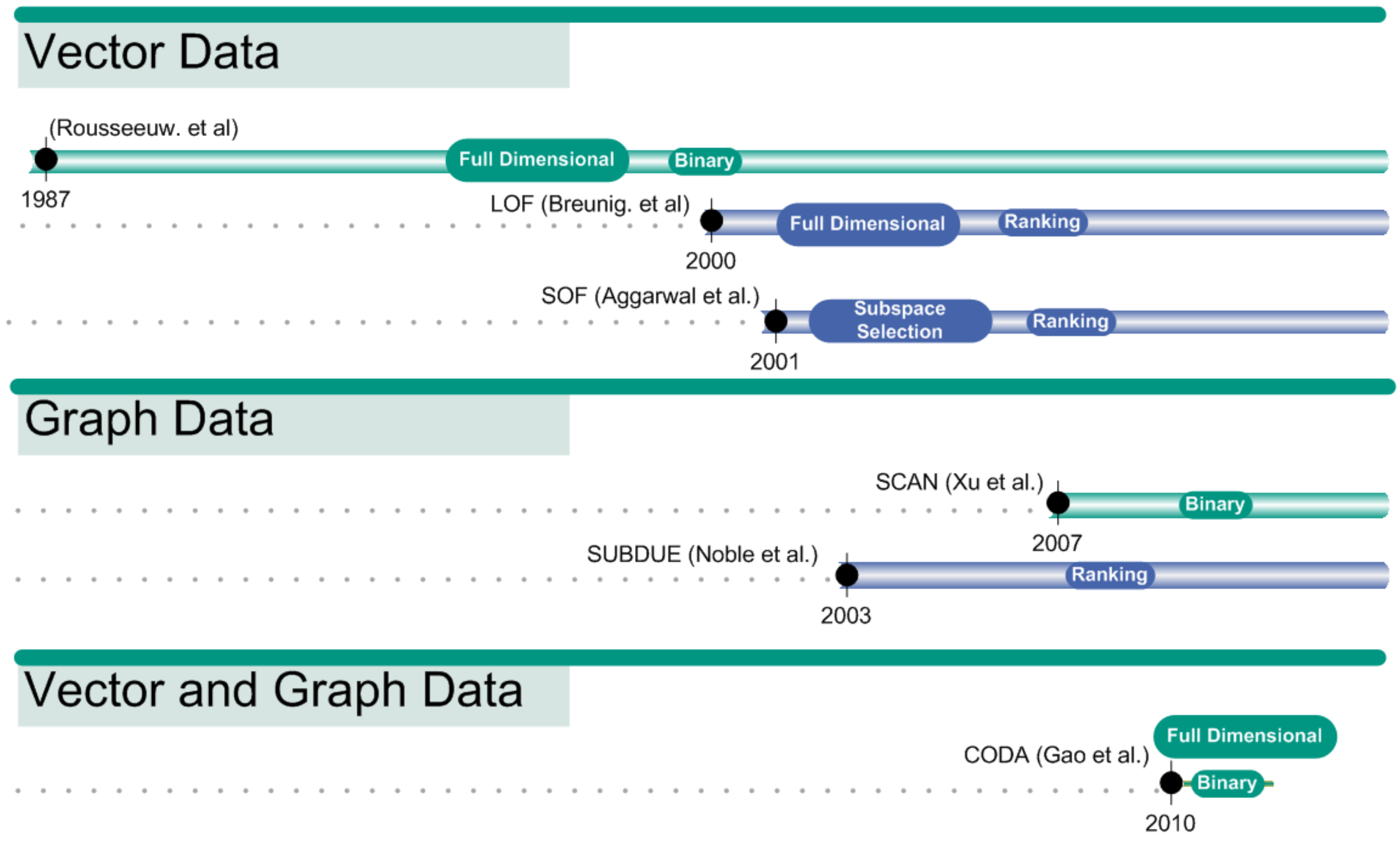
Graph Structure





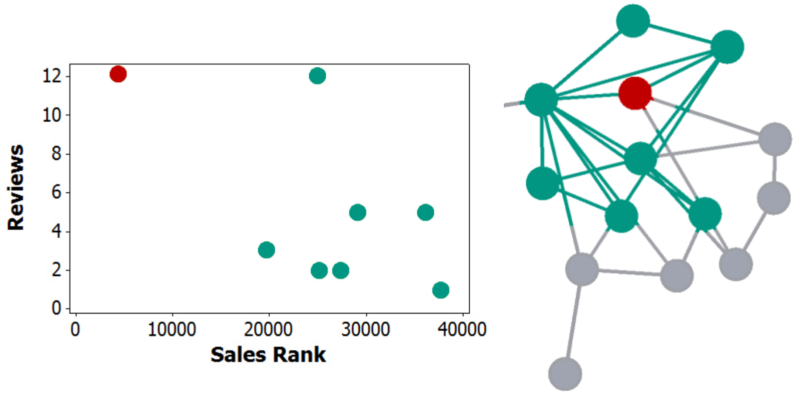
- **Output:** Is a **ranking of all nodes** ordered by **deviation** w.r.t. **subgraph** and **relevant attribute subspaces**

Related Work: Outlier Mining

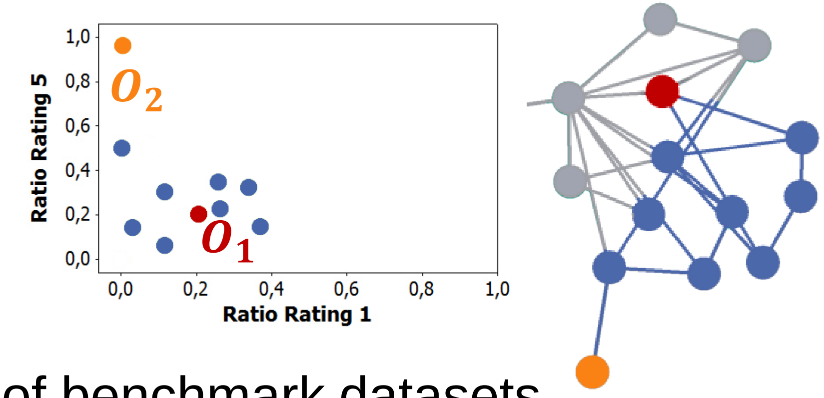


Challenges

- (1) Selection of relevant subspaces and subgraphs



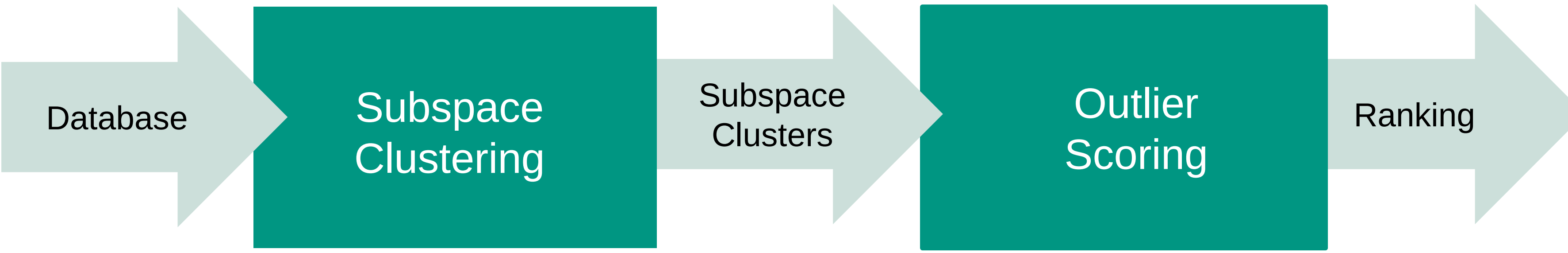
- (2) Scoring of objects in multiple subspace clusters



- (3) Availability of benchmark datasets

Our GOutRank Framework

- We propose a **decoupled process:**



(1) Selection:

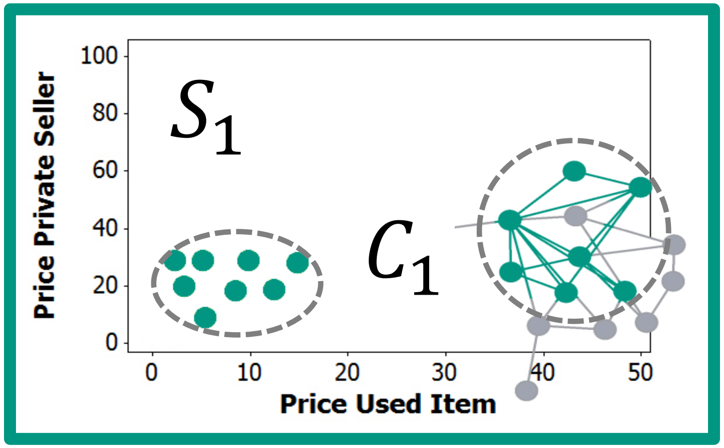
- **subgraphs**
- **relevant subspaces**

(2) Scoring:

- **multiple subspace clusters**

(1) Selection of Subspaces and Subgraphs

- Subspace clustering on attributed graphs
  - **Input:** graph *(V,E)* and attributes *A*
  - **Output:** $Res = \{ (C_1, S_1) \dots (C_n, S_n)\}$ with $C_i \subseteq V$ and $S_i \subseteq A$



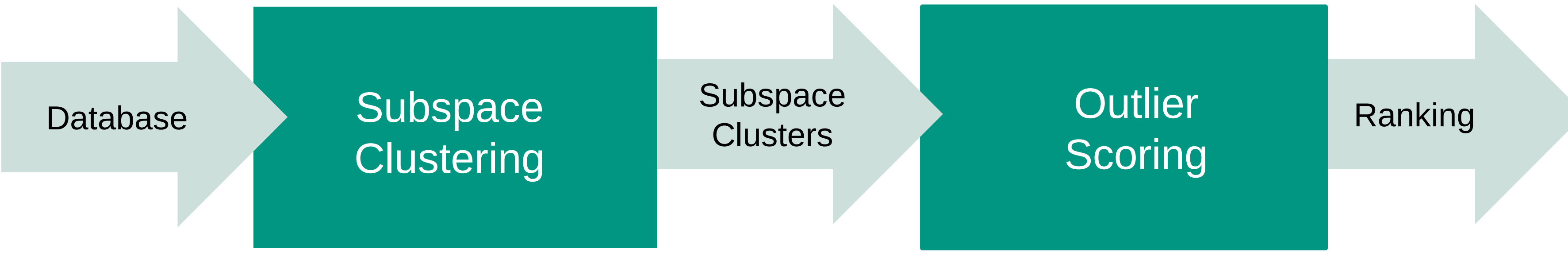
- Algorithmic solutions:
  - GAMer[1]
  - Cocain[2]
  - CoPam[3]
  - ...

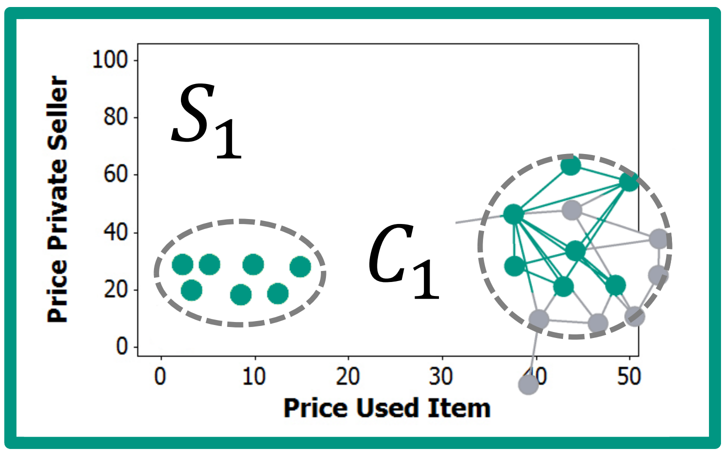
- Provide models for groups of similar nodes

[1] Günnemann et al. "Subspace clustering meets dense subgraph mining: A synthesis of two paradigms." In IEEE ICDM 2010
[2] Zeng et al. "Coherent closed quasi-clique discovery from large dense graph databases." In ACM SIGKDD 2006
[3] Moser et al. "Mining cohesive patterns from graphs with feature vectors." In SIAM SDM 2009

Our GOutRank Framework

- We propose a **decoupled process:**





Scoring:

- **multiple subspace clusters**

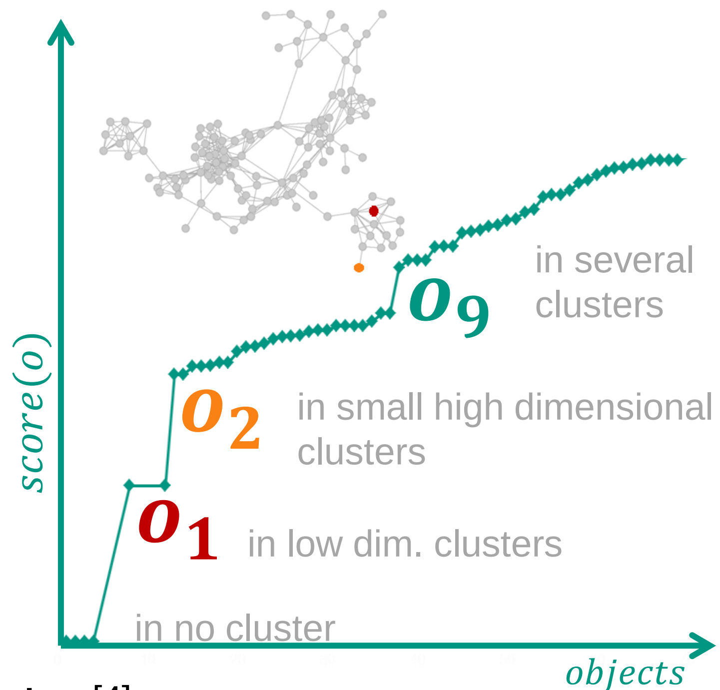
- How to derive an outlier score based on subspace cluster results?

(2) Scoring with Multiple Subspace Clusters

- Properties of subspace clusters:
  - Overlap (i.e. objects belong to several clusters in different subspaces)
  - Different cluster sizes and dimensionality

Res:

$$(C1, S1) = (\{o_3, o_4, o_5, o_7, o_8, \mathbf{o_9}, o_{10}\}, \{d_1, d_2\})$$
$$(C_2, S_2) = (\{\mathbf{o_1}, o_6, o_7, \mathbf{o_9}, o_{10}, o_{11}, o_{12}, o_{13}, o_{14}\}, \{d_3\})$$
$$(C_3, S_4) = (\{\mathbf{o_2}, o_5, \mathbf{o_9}, o_{13}, o_{14}\}, \{d_1, d_2, d_4, d_5, d_6\})$$



- Scoring function considering cluster properties[4]

$$score(o) = f(Res)$$

⇒ **Information loss**

[4] Müller et al.: "Outlier Ranking via Subspace Analysis in Multiple Views of the Data." In IEEE ICDM 2012

Combined Scored Function

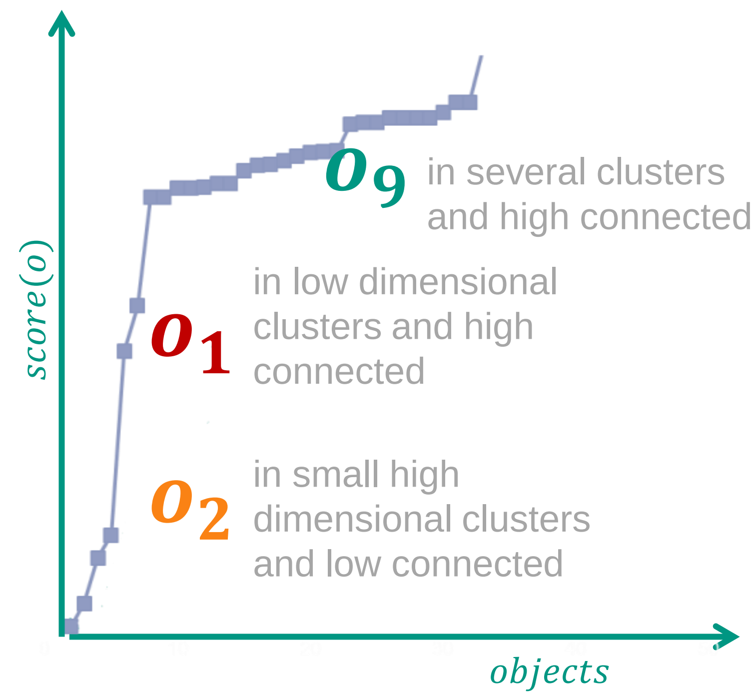
- Properties from the **graph structure:**
  - **centrality of a node**
  - Edge density of the subgraph (ongoing work)
  - Analysis of neighboring subspace clusters (ongoing work)

Res:

$$(C1, S1) = (\{o_3, o_4, o_5, o_7, o_8, \mathbf{o_9}, o_{10}\}, \{d_1, d_2\})$$
$$(C_2, S_2) = (\{\mathbf{o_1}, o_6, o_7, \mathbf{o_9}, o_{10}, o_{11}, o_{12}, o_{13}, o_{14}\}, \{d_3\})$$
$$(C_3, S_4) = (\{\mathbf{o_2}, o_5, \mathbf{o_9}, o_{13}, o_{14}\}, \{d_1, d_2, d_4, d_5, d_6\}$$

+

Graph:



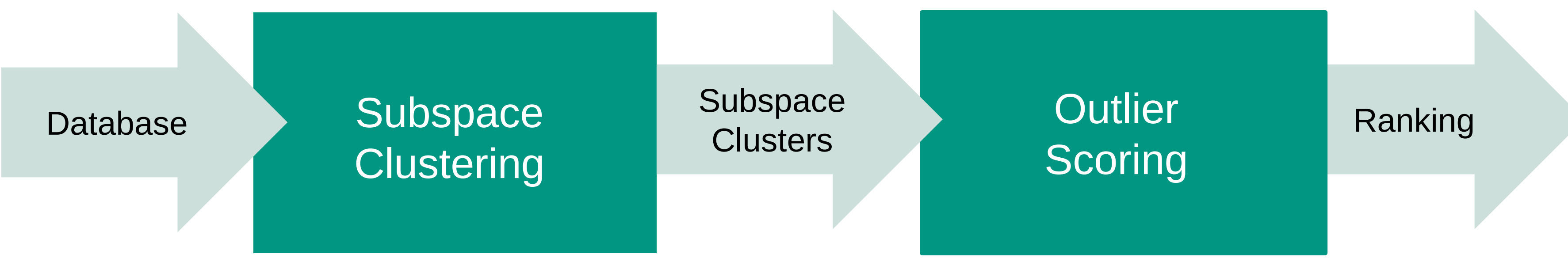
- **Combine** both sources of information:

$$score(o) = f(Res, Graph)$$

Experimental Setup

- Competitors
  - Only on **vector data**: full dimensional vs. subspace selection
  - Only on **graph data**: node outliers as by-product of graph clustering
  - **On vector and graph data**: community outlier detection

- Instantiation of different **cluster models** and **scoring functions**



- All experiments on:
  - subgraph of the Amazon co-purchase network

Outlier Identification

- Setting of our user experiment
  - Users (high school students)
    - **No prior knowledge** on outlier mining
    - **Expertise** by domain knowledge
  - Attributed graph:
    - Disney DVDs (as Amazon products)
    - Presentation of co-purchased products (i.e. pre-computed graph clusters)

Product Visualization

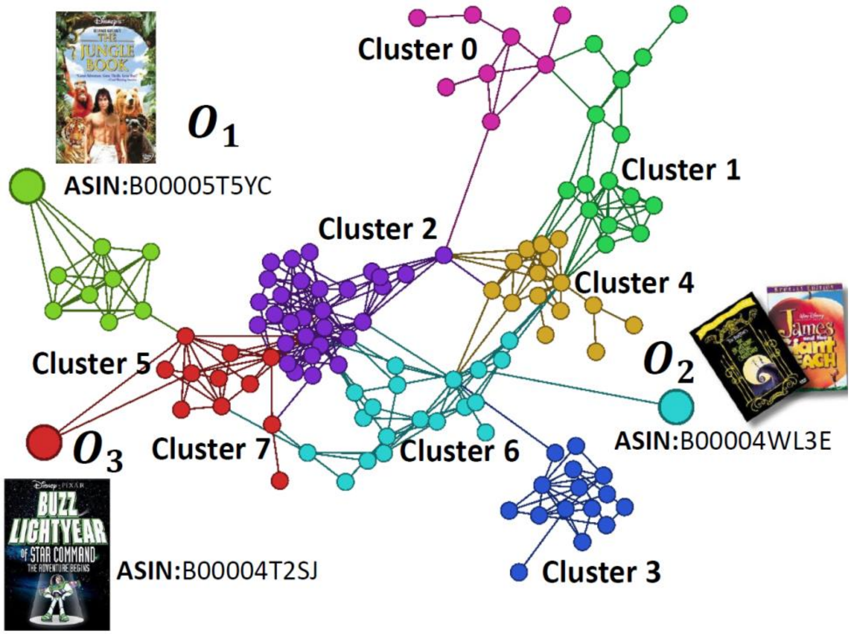
- Tasks:
  1. **Select outliers** in each set of co-purchased products
  2. **Write an explanation** for the deviation of outliers

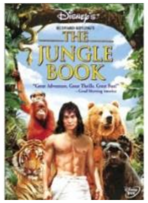
Form for outlier

Our Benchmark Database

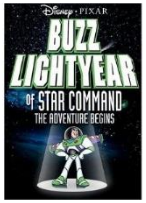
- Disney subgraph with **124 products**, **334 edges.**
- Each product is labeled as outlier iff selected by >50% of the students

Examples:





Price: 100$
Suggested price: **14,99$**
(2003)



High 1 Rating Rating and low 5 Rating Ratio w.r.t. Pixar Films

Evaluation w.r.t. Competitors

- Comparison w.r.t. several outlier mining paradigms

Database	Paradigm	Algorithm	AUC [%]
Vector data	full data space	LOF[5]	56,85
	Subspace selection	SOF[6]	65,88
Graph structure	graph clustering	SCAN[7]	52,68
Attributed Graph	full data space	CODA[8]	50,56
	selected subspaces	**GOutRank**	**86,86**

[5] Breunig et al. "LOF: identifying density-based local outliers." In *ACM SIGMOD Record*. Vol. 29. No. 2. 2000
[6] Aggarwal et al. "Outlier detection for high dimensional data." In *ACM SIGMOD Record* Vol 30 No. 2 2001
[7] Xu et al. "Scan: a structural clustering algorithm for networks." In ACM SIGKDD 2007
[8] Gao et al. "On community outliers and their efficient detection in information networks." In *ACM SIGKDD* 2010

Internal Evaluation

- Comparison of ***Res*** from different **subspace clustering models**
- Comparison of different **scoring functions**

Res	**Graph**	**AUC [%]**
GAMer[1]	--	75,28
	$degree(o)$	82,91
	$eigenvalue(o)$	86,86
Extension of Cocain[2]	--	75,85
	$degree(o)$	76,97
	$eigenvalue(o)$	77,96
CoPaM[3]	--	58,61
	$degree(o)$	69,49
	$eigenvalue(o)$	72,45

Karlsruhe Institute of Technology

Conclusion & Outlook

- Selection of subgraphs and subspaces
  - ✓ Decoupled processing scheme exploiting subspace clusters

> ➢ **Scalability to large attributed graphs**
> ➢ Integration of outlier ranking into graph clustering algorithms

- Scoring of objects in multiple subspace clusters
  - ✓ Ranking combining graph structure and subspace cluster analysis

> ➢ **Improvement of the scoring functions**
> ➢ Extraction of more graph subspace cluster properties

- Availability of benchmark datasets
  - ✓ First benchmark on a subgraph from the Amazon co-purchased network

> ➢ **Complete benchmark graph** (>300,000 nodes) with large user experiment (> 200 users)

Thank you for your attention

Our benchmark database is available online:

http://www.ipd.kit.edu/~muellere/GOutRank/