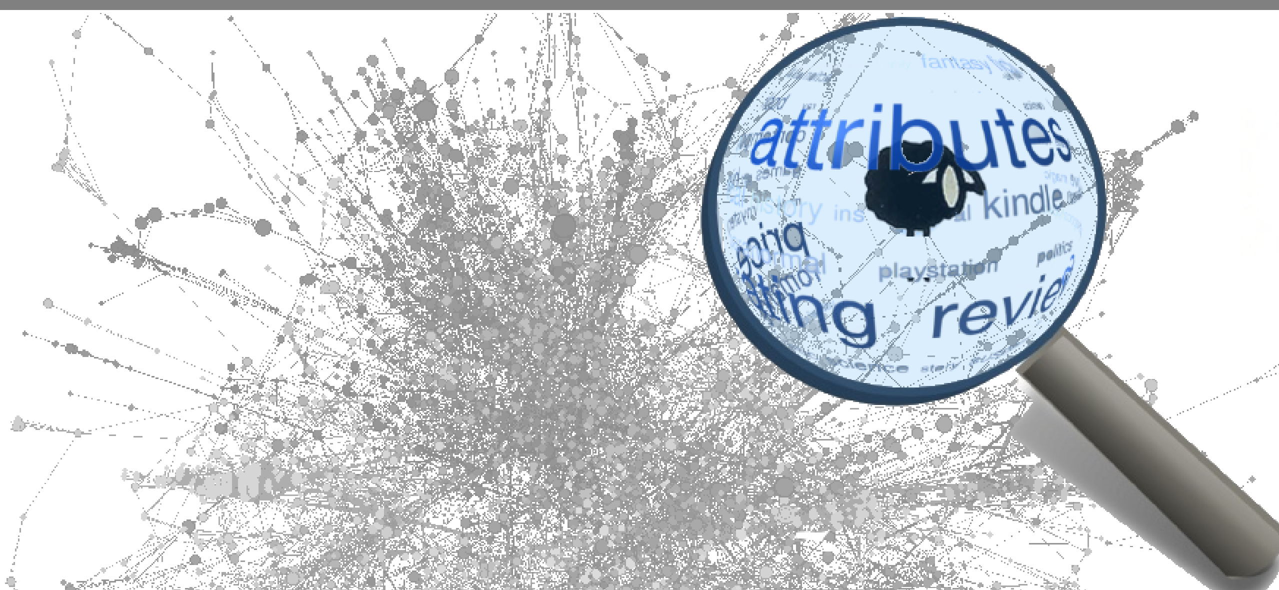
Karlsruhe Institute of Technology

Universiteit Antwerpen

Local Context Selection for Outlier Ranking in Graphs with Multiple Numeric Node Attributes

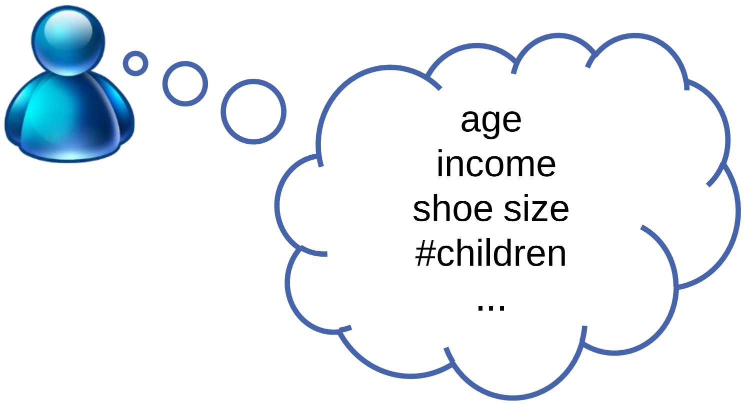
Patricia Iglesias, Emmanuel Müller, Oretta Irmler, Klemens Böhm

International Conference on Scientific and Statistical Database Management (SSDBM 2014)

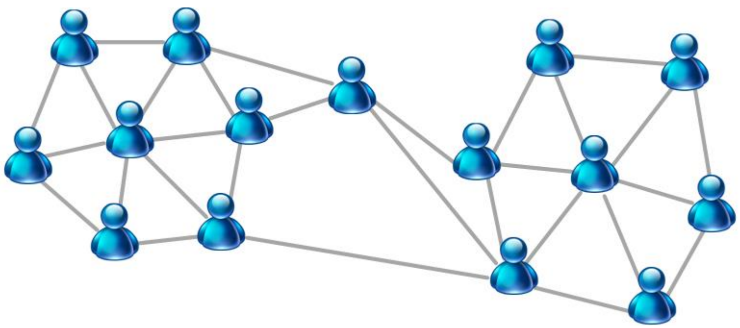


Attributed Graphs

- Complex databases: Attributed graphs



attributes



graph structure

- Several application domains:
  - Communication networks, co-purchased networks, social networks
  - Bibliographic networks, biological networks
- Outlier mining:
  - Fraud detection, network intrusion, data cleaning...

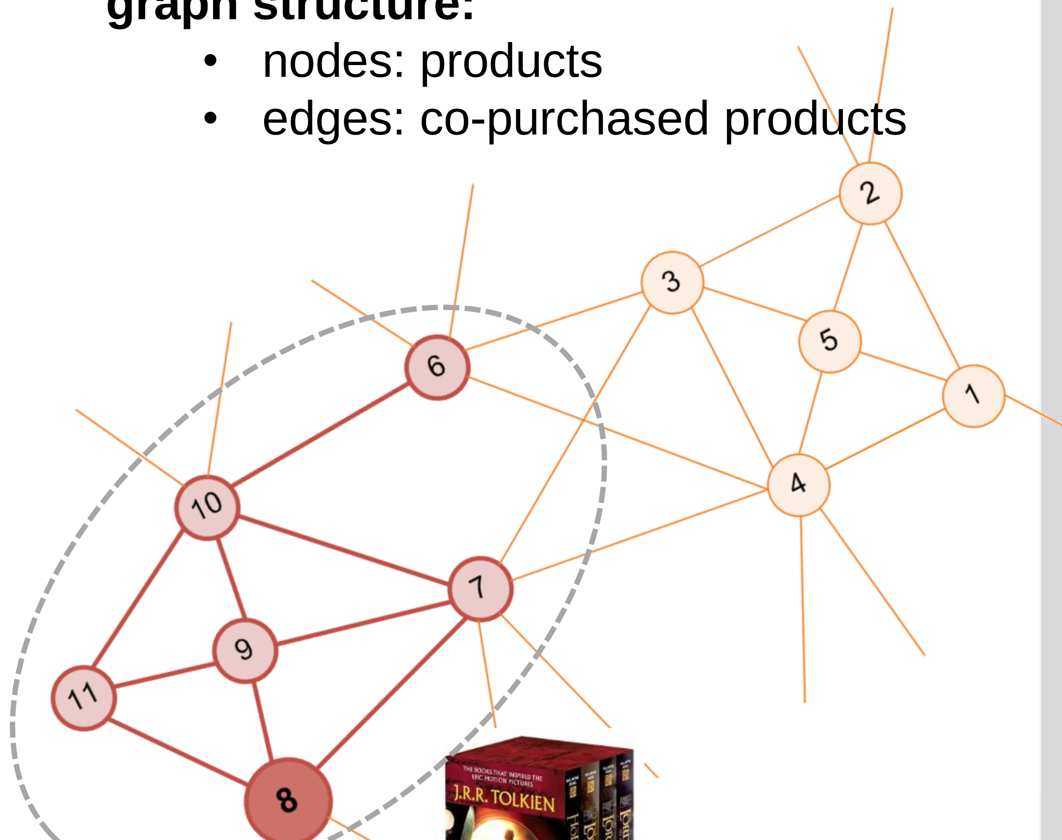
Problem Overview: Example

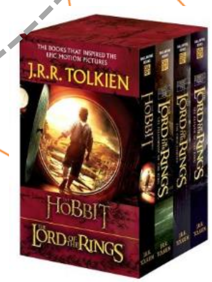
attributes: product description

Node	sales	#reviews	price
1	262	76	25
2	25	30	30
3	155	47	150
4	69	105	20
5	80	8	35
6	182	7	15
7	22	5	8
8	234	**28**	12
9	102	8	5
10	248	6	13
11	10	4	10
…	…	…	…

graph structure:

- nodes: products
- edges: co-purchased products



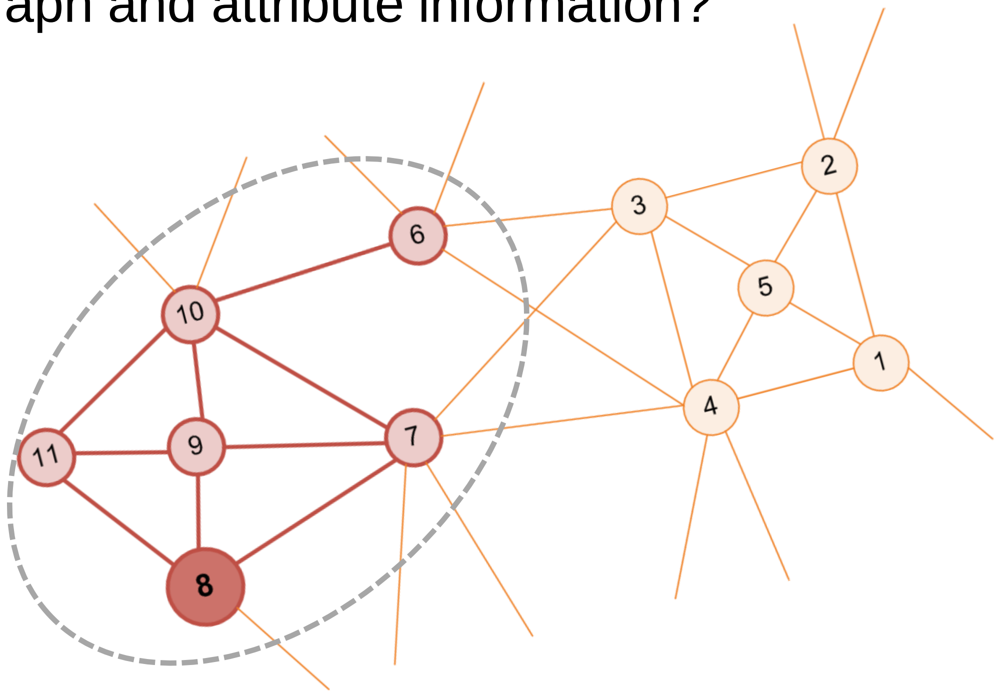


Tolkien Books

Challenges

- How to **define a local context** for each node?
- How to **efficiently** select only the **relevant attributes**?
- How to **rank** each node w.r.t. graph and attribute information?

Node	#reviews	price
6	7	15
7	5	8
8	**28**	12
9	8	5
10	6	13
11	4	10



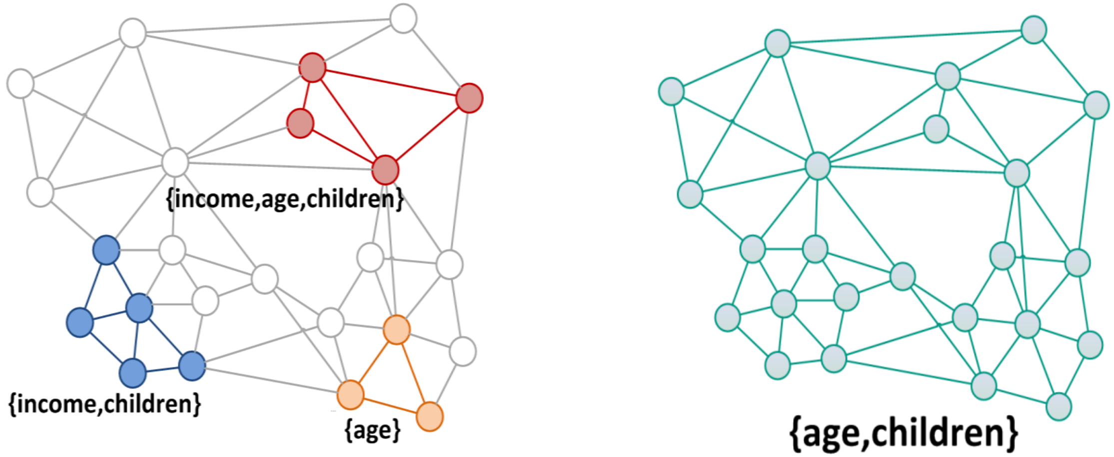
Comparison: Outlier Mining on Attributed Graphs

Algorithm	Local	Selection of attributes	Ranking	Time Complexity (#attributes)
CODA [Gao 2010]	✗	✗	✗	$O(d^2)$
CONSUB [Iglesias 2013]	✗	✓	✓	$O(2^d)$
GoutRank [Müller 2013]	✗	✓	✓	$O(2^d)$
ConOut	✓	✓	✓	$O(d)$

[Gao 2010] Gao et al. “On community outliers and their efficient detection in information networks" In ACM SIGKDD 2010
[Iglesias 2003] Iglesias et al. Statistical Selection of Congruent Subspaces for Mining Attributed Graphs. In IEEE ICDM. 2013
[Müller 2013] Müller et al. “Ranking outlier nodes in subspaces of attributed graphs” In GDM at IEEE ICDE 2013

Our Approach: ConOut

- Local vs. global



- Attribute projection vs. subspace selection
  - Avoid exponential runtimes w.r.t. the number of the attributes
  - Time complexity: Linear

- Ranking vs. Binary
  - Assessment of the outlierness w.r.t. both: graph and attributes

ConOut I: Context Definition

- **Local Context** of object o**:**
  - **C(o), R(o)**

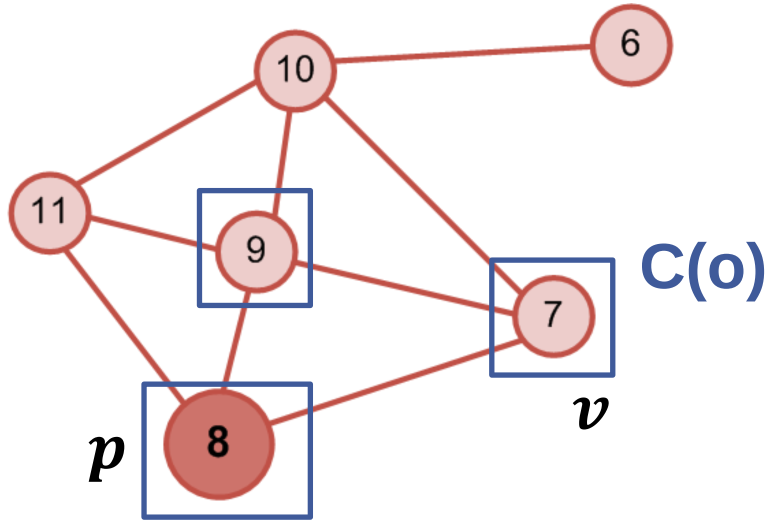
- **Graph Context** C(o)
  - Nodes similar w.r.t. graph structure
  - Graph similarity based on shared nearest neighborhood (SNN):

$$sim(v,p) = \frac{|Adj(v)| \cap |Adj(p)|}{\sqrt{|Adj(v)| \cdot |Adj(p)|}}$$

$$Adj(v) = \{p \in E \mid \exists (v,p) \in E\} \cup \{v\}$$

  - Other local graph context definitions possible

- **Relevant Attributes** R(o)?

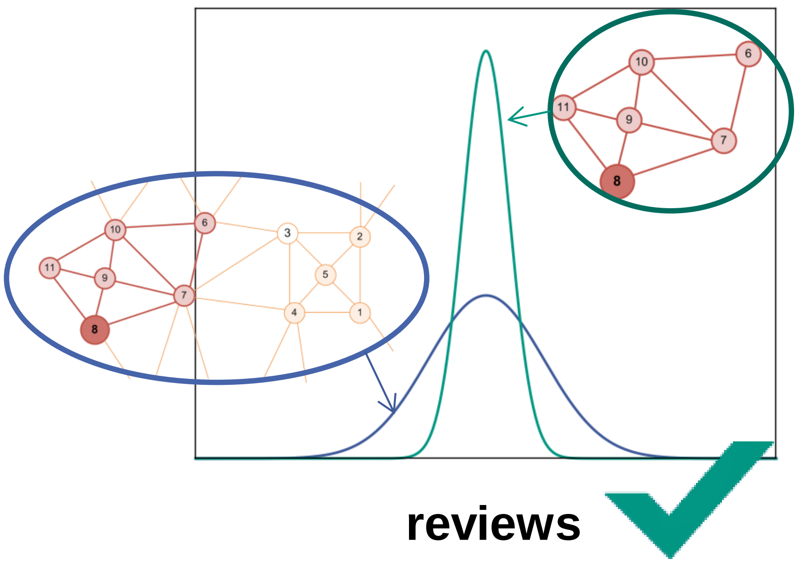


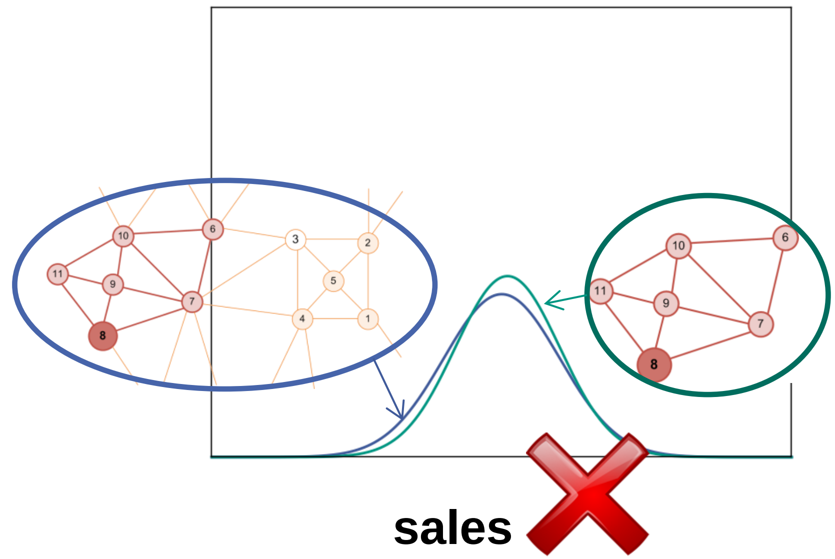
Node	#reviews	price
6	7	15
7	5	8
8	**28**	12
9	8	5
10	6	13
11	4	10

R(o)

ConOut II: Statistical Selection

- Attribute A_i has significantly lower variance in C(o) than the overall database





- Statistical test: Example of instantiation

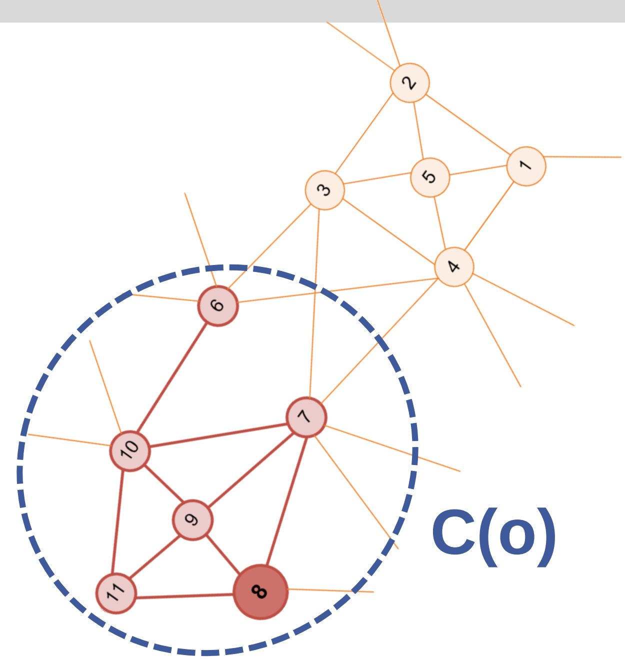
$$H_0 : \sigma^2_{local} = \sigma^2_{global}$$
$$\boldsymbol{H_1 : \sigma^2_{local} < \sigma^2_{global}}$$
$$P(H_o \, is \, rejected | H_o = true) \leq \alpha$$

ConOut III: Context Based Ranking

- Local context selection enables a **high contrast** between inliers and outliers

- **Goal:** Compare deviation of the attribute values and the graph density of each node to its local context
  - Local attribute deviation (LAD(o))
  - Local graph density (LGD(o))

- Context based score combines both:
  - $score(o) = LGD(o) \cdot LAD(o)$



Node	sales	#reviews	price
1	262	76	25
2	25	30	30
3	155	47	150
4	69	105	20
5	80	8	35
6	182	7	15
7	22	5	8
8	234	**28**	12
9	102	8	5
10	248	6	13
11	10	4	10
…	…	…	…

R(o)

Experimental Setup

- Synthetic data
- Real world data

Traditional Algorithms:

- LOF [Breunig 2001]: attributes
- SOF [Aggarwal 2001]: attributes
- SCAN [Xiu 2007]: graph

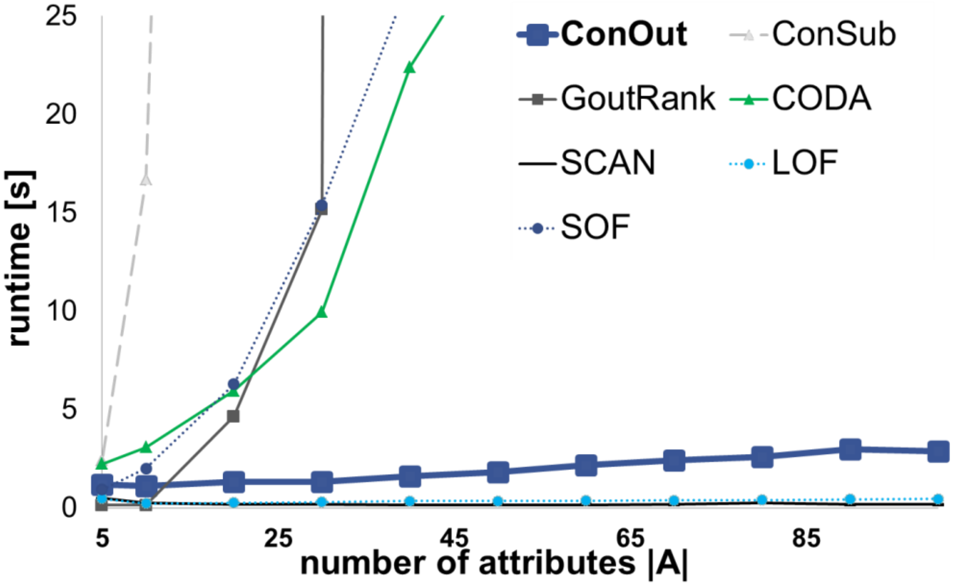
Algorithms for attributed graphs:

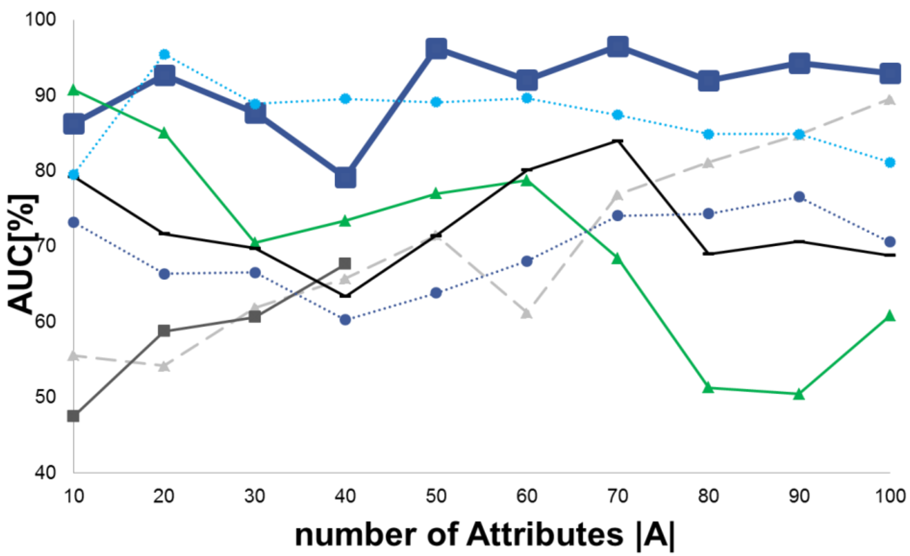
- CODA [Gao 2010]
- GOutRank [Müller 2013]
- ConSub [Iglesias 2013]

☑ **quality (AUC)** ☑ **runtime**

[Breunig 2001] Breunig et al. "LOF: identifying density-based local outliers." In *ACM SIGMOD* 2000
[Aggarwal 2001] Aggarwal et al. "Outlier detection for high dimensional data." In *ACM SIGMOD* 2001
[Xiu 2007] Xiu et al. "Scan: a structural clustering algorithm for networks." In ACM SIGKDD 2007

Synthetic Data





- Scalability w.r.t. increasing **number of attributes** and **graph size**
- **High quality** for the detection of contextual outliers

Real World Data

- Benchmark [Müller 2013]
  - 124 nodes, 333 edges and 28 attributes

		Algorithm	AUC [%]	run.[ms]
Attributes				
	full space	LOF	56.85	41
	subspace selection	SOF	65.88	825
Graph				
	graph clustering	SCAN	52.68	4
Both				
	full space	CODA	50.56	2596
	subspace cluster analysis	GOutRank	86.86	26648
	global subspace selection	ConSub	81.77	8930
	Local context selection	ConOut	81.21	199

Real World Data

- Benchmark [Müller 2013]
  - 124 nodes, 333 edges and 28 attributes

		Algorithm	AUC [%]	run.[ms]
Attributes				
	full space	LOF	56.85	41
	subspace selection	SOF	65.88	825
Graph				
	graph clustering	SCAN	52.68	4
Both				
	full space	CODA	50.56	2596
	subspace cluster analysis	**GOutRank**	**86.86**	26648
	global subspace selection	**ConSub**	**81.77**	8930
	Local context selection	**ConOut**	**81.21**	199

Real World Data

- Benchmark on a co-purchased network [Müller 2013]
  - 124 nodes, 333 edges and 28 attributes

		Algorithm	AUC [%]	run.[ms]
Attributes				
	full space	LOF	56.85	41
	subspace selection	SOF	65.88	825
Graph				
	graph clustering	SCAN	52.68	4
Both				
	full space	CODA	50.56	2596
	subspace cluster analysis	**GOutRank**	**86.86**	**26648**
	global subspace selection	**ConSub**	**81.77**	**8930**
	Local context selection	**ConOut**	**81.21**	**199**

Conclusions & Future Work

- Challenge: attributed graphs ✓ **Local context definition**
- Irrelevant Attributes ✓ **Statistical selection**
- Outlierness Scoring ✓ **Combined ranking functions**
- Algorithm ✓ **Efficiency**

Future Work

- Mixed attribute types
- Local correlations between attributes
- Other graph definitions (directed, weighted, …)

Thank you for your attention

Our datasets and parameter settings are available online:

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