KIT
Karlsruhe Institute of Technology

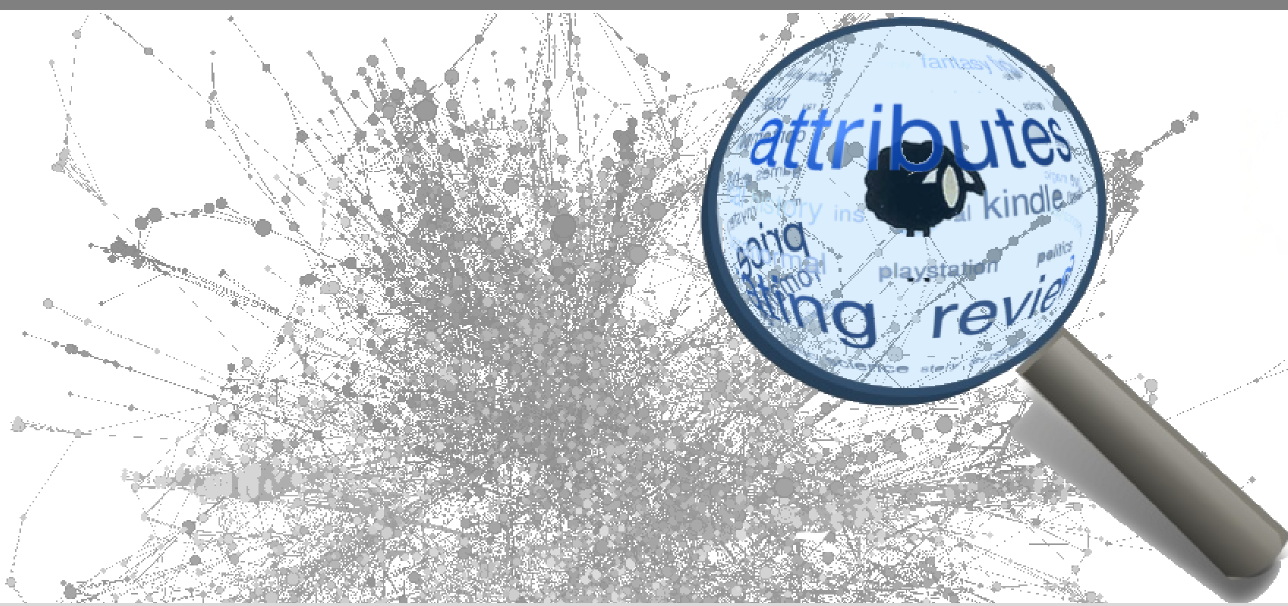
Universiteit Antwerpen

# Statistical Selection of Congruent Subspaces for Mining Attributed Graphs


**Patricia Iglesias**, Emmanuel Müller, Fabian Laforet, Fabian Keller, Klemens Böhm

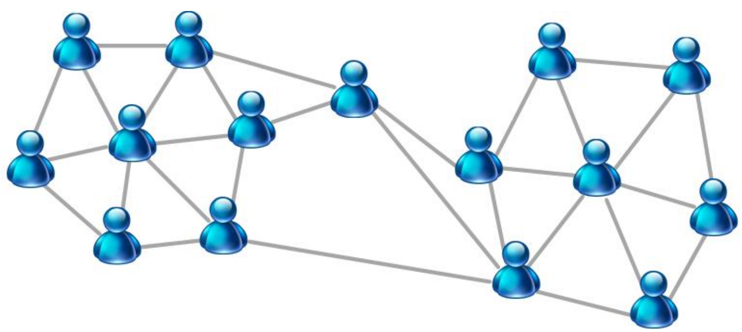


IEEE International Conference on Data Mining (ICDM 2013)

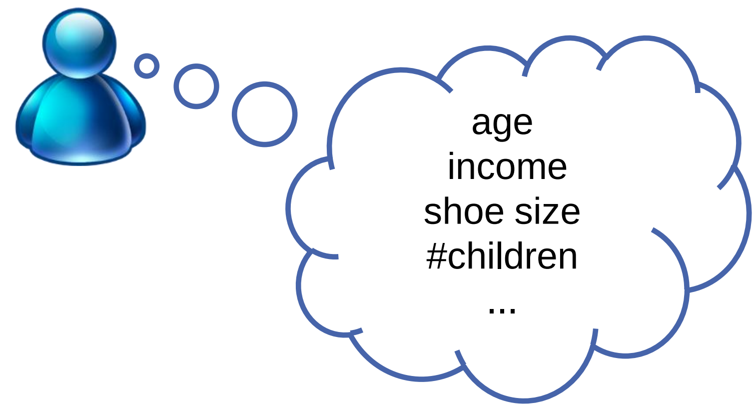



# Attributed Graphs

- Several application domains
  - Communication networks, co-purchased networks, social networks



graph structure

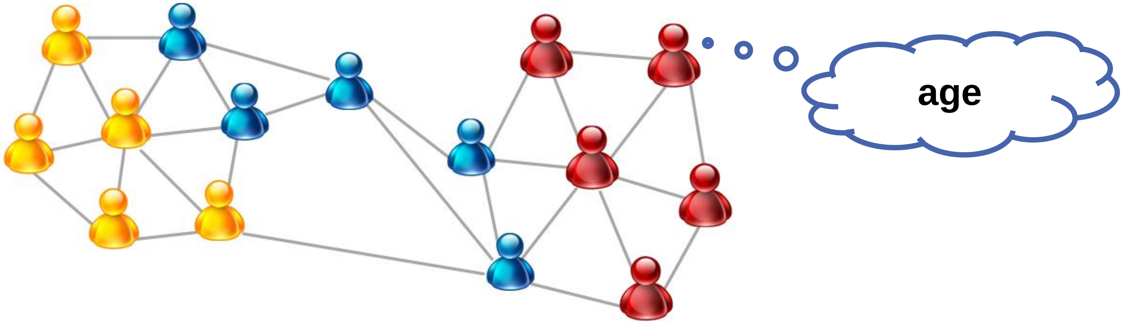



attributes

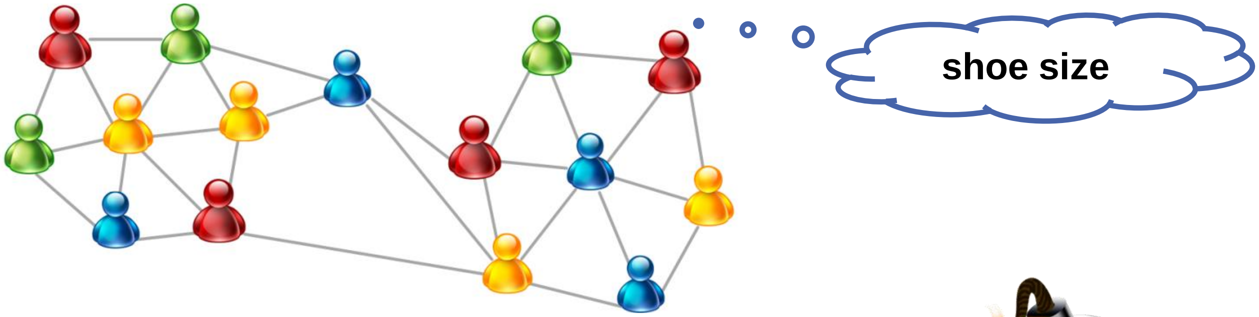
- Novel problems on attributed graphs

# Commonly Used Assumption

- **Homophily:** *„birds of a feather flock together"*



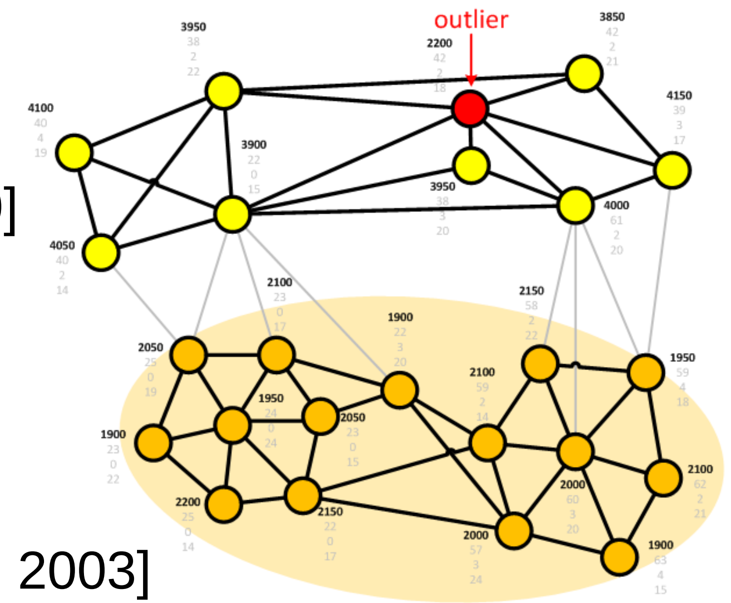

- **Homophily:** not fullfilled for all attributes




➡ **deterioration** of mining techniques on attributed graphs

# Mining Attributed Graphs

- Different graph mining techniques
  - Clustering
  - **Community outlier detection** [Gao 2010]
- Used assumption: **Homophily**
  has to be fulfilled for **all** the attributes
- Problem: **disassortative mixing** [Newman 2003]
  hinders the detection of communities
  (i.e. similarity assessment of nodes)

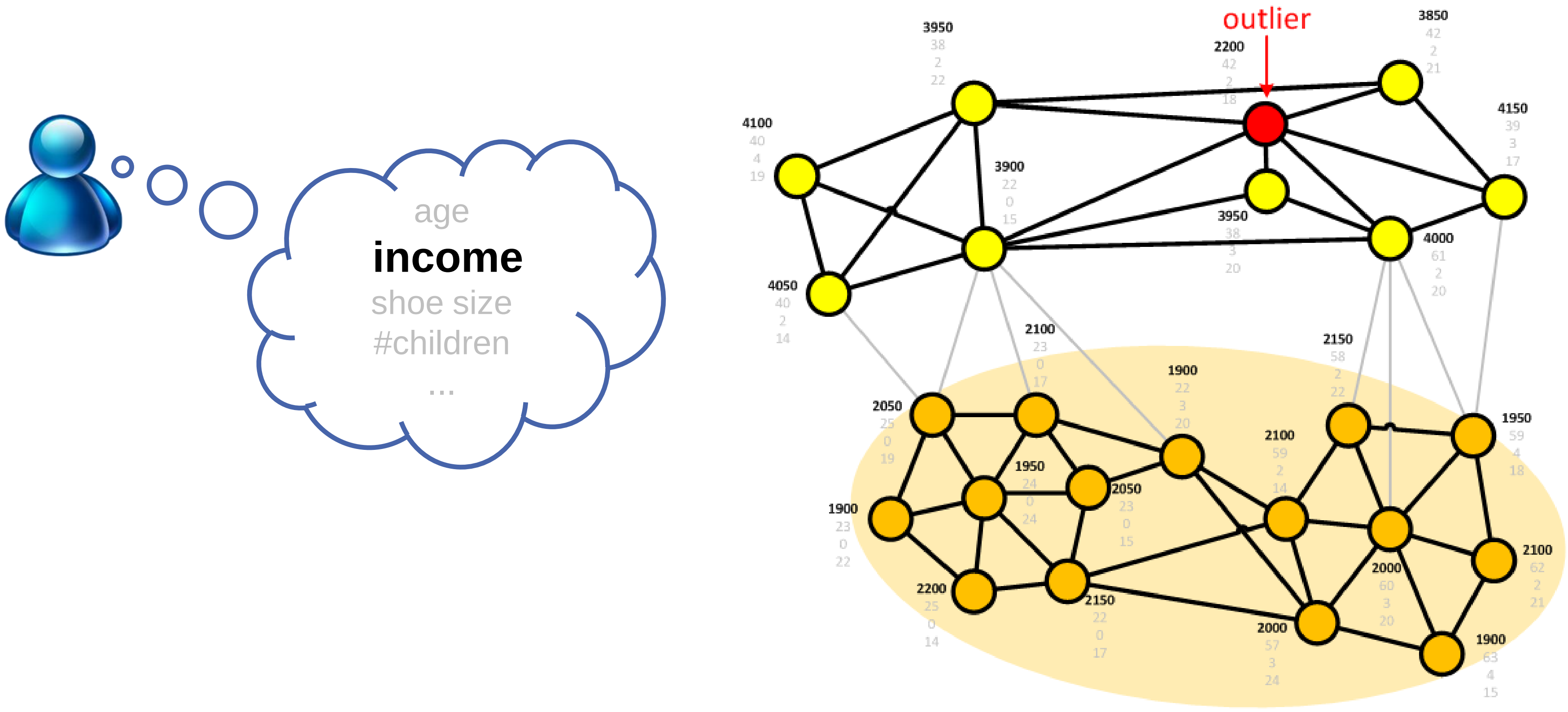



➡ **Solution: pre-processing techniques ensuring homophily**

[Gao 2010] Gao et al. “On community outliers and their efficient detection in information networks" In ACM SIGKDD 2010
[Newman 2003] M.E. Newman. Mixing patterns in networks. Physical Review, 2003

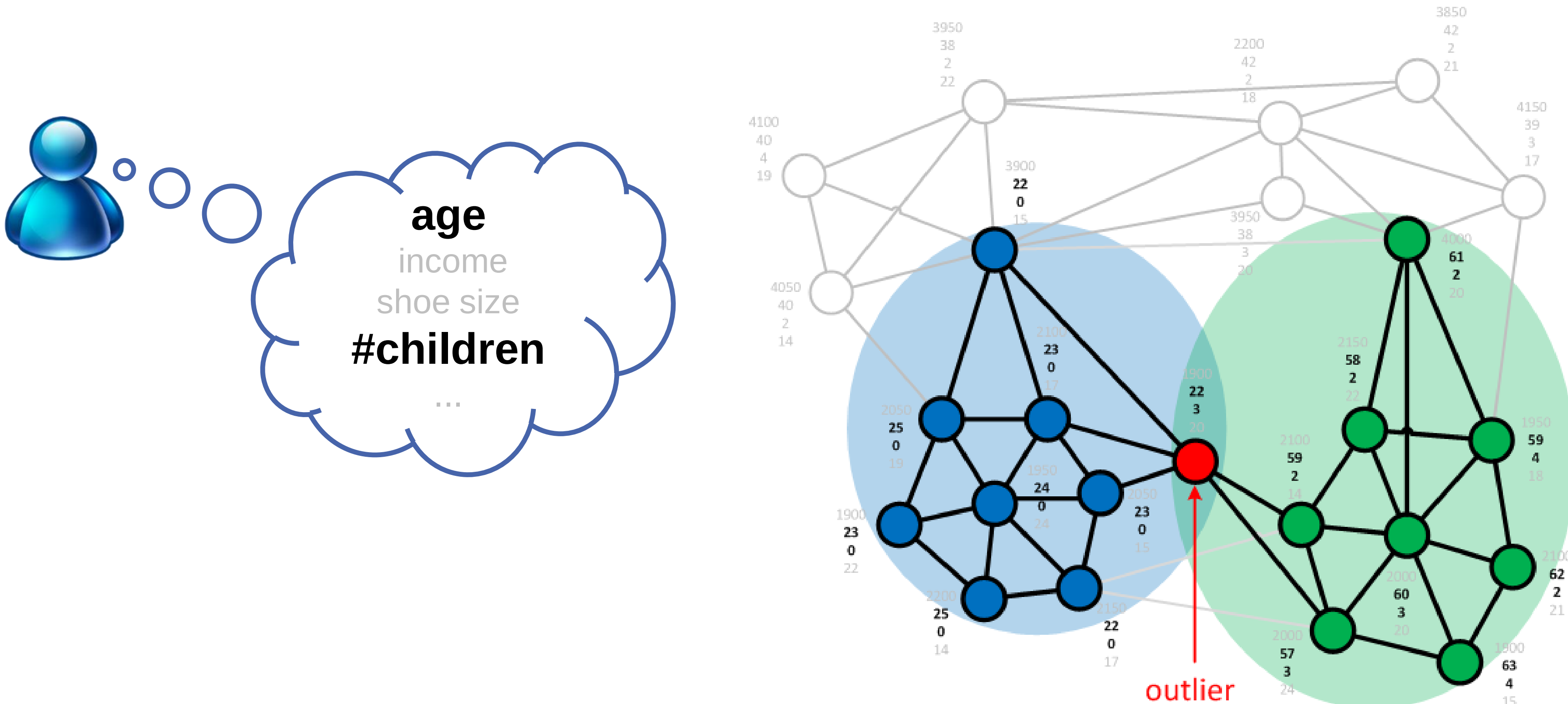
# Multiple Views in Attributed Graphs

- Different structures depending on the subset of attributes



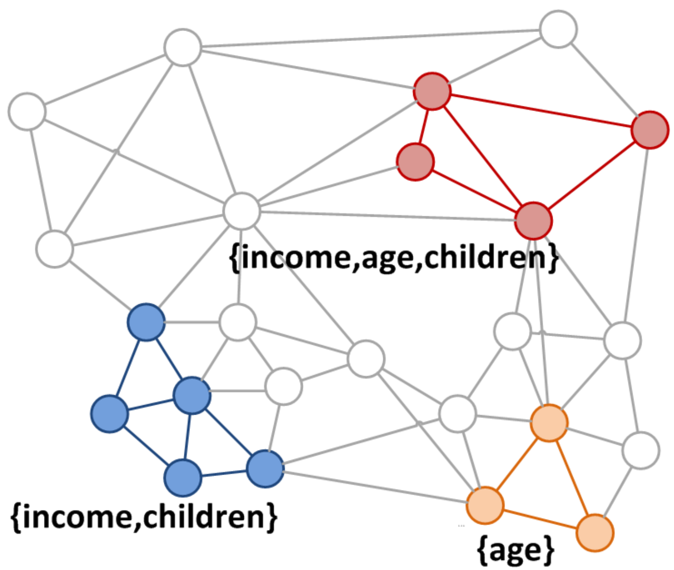
# Multiple Views in Attributed Graphs

- Different structures depending on the subset of attributes



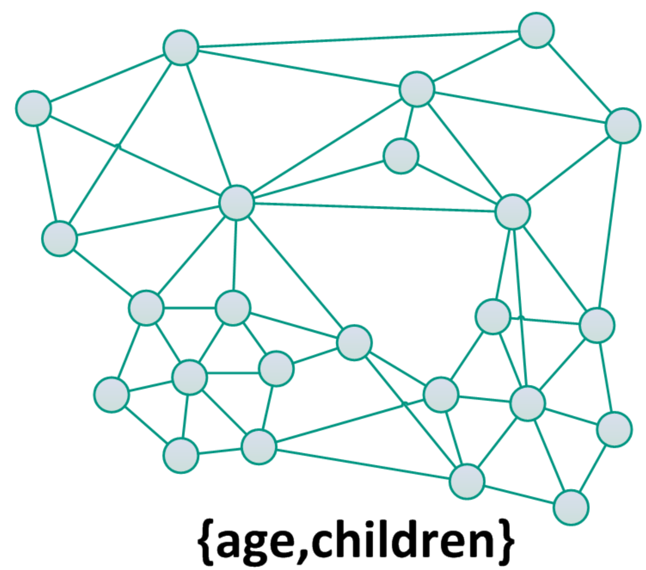
# Specialized Approaches (Related Work I)

- Frequent subgraph mining, graph partitioning, subspace clustering ...
  - Local selection of the attributes
  - Individual subgraphs




➔ not designed as **pre-processing step** for other graph mining methods

**In contrast, we aim at:**



# General Approaches (Related Work II)

- Assortative mixing coefficient [Newman 2003]
  - Correlation between an attribute and the graph structure
  - **For a single attribute only**

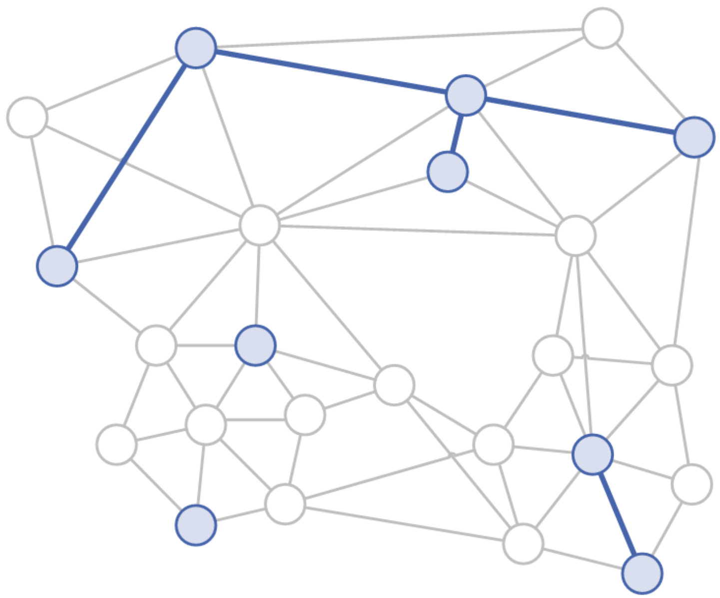
- Unsupervised feature selection LUFS [Tang 2012]
  - Improvement of traditional feature selection
    by incorporating additional information from the graph structure
  - **No selection of multiple view possible**

[Tang 2012] Tang et al. “Unsupervised feature selection for linked media data" In ACM SIGKDD 2012

# ConSub I

- Congruent subspaces
  - **Mutual similarity** between attribute values in subspace *S*
  - **Significantly more edges** than expected by a random distribution
- Constraint Subgraph $G_{C,S}$
  - Set of constraints formed by all the pairs $(\boldsymbol{I_j} = [low_j, high_j],\ \boldsymbol{A_j} \in S)$

S = {shoe size}
nodes with **8 ≤ *shoe size* ≤ 9**



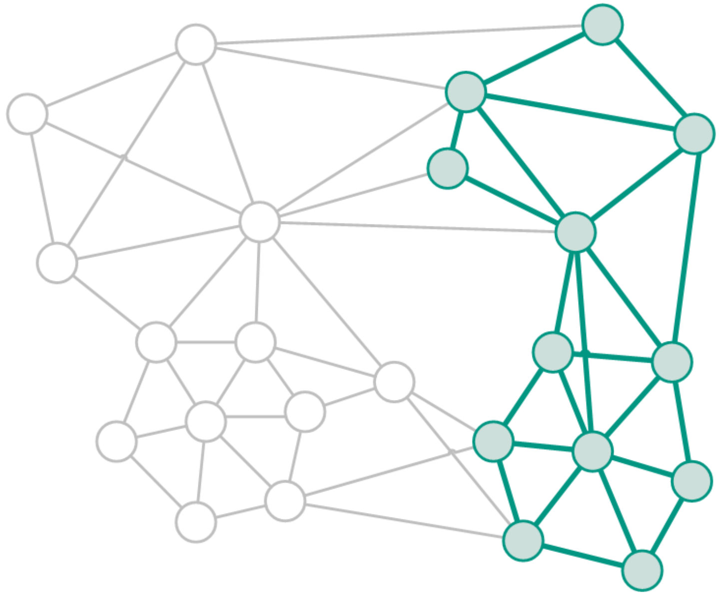
➡ **small number of edges**

# ConSub II

- Congruent subspaces
  - **Mutual similarity** between attribute values in subspace *S*
  - **Significantly more edges** than expected by a random distribution
- Constraint Subgraph $G_{C,S}$
  - Set of constraints formed by all the pairs $(\boldsymbol{I_j} = [low_j, high_j],\ \boldsymbol{A_j} \in S)$

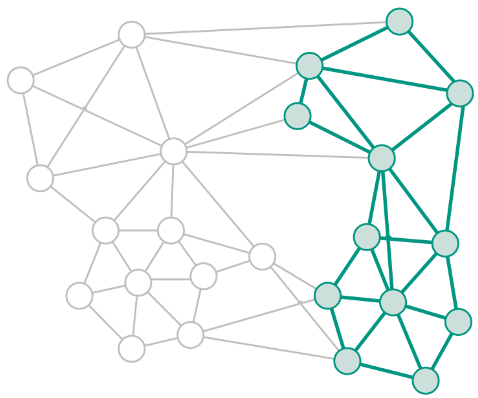
S ={age,income}
nodes with **45 ≤ *age* ≤ 60** and
**1900 ≤ *income* ≤ 4500**

➡ **high number of edges**

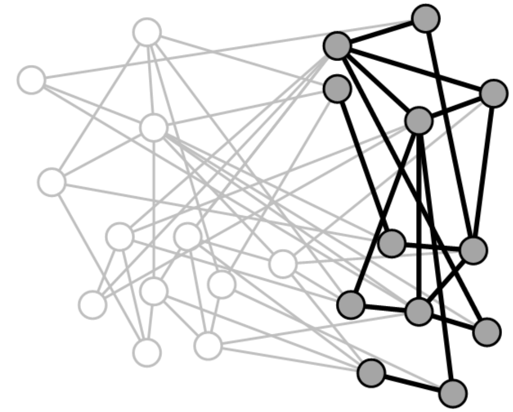


# ConSub III

- Edge count (constraint subgraph $G_{C,S}$)



observed edges: $|E_{C,S}|$ vs. expected edges: $E_{exp}(G_{C,S})$
(w.r.t. some given null model)



- Statistical test

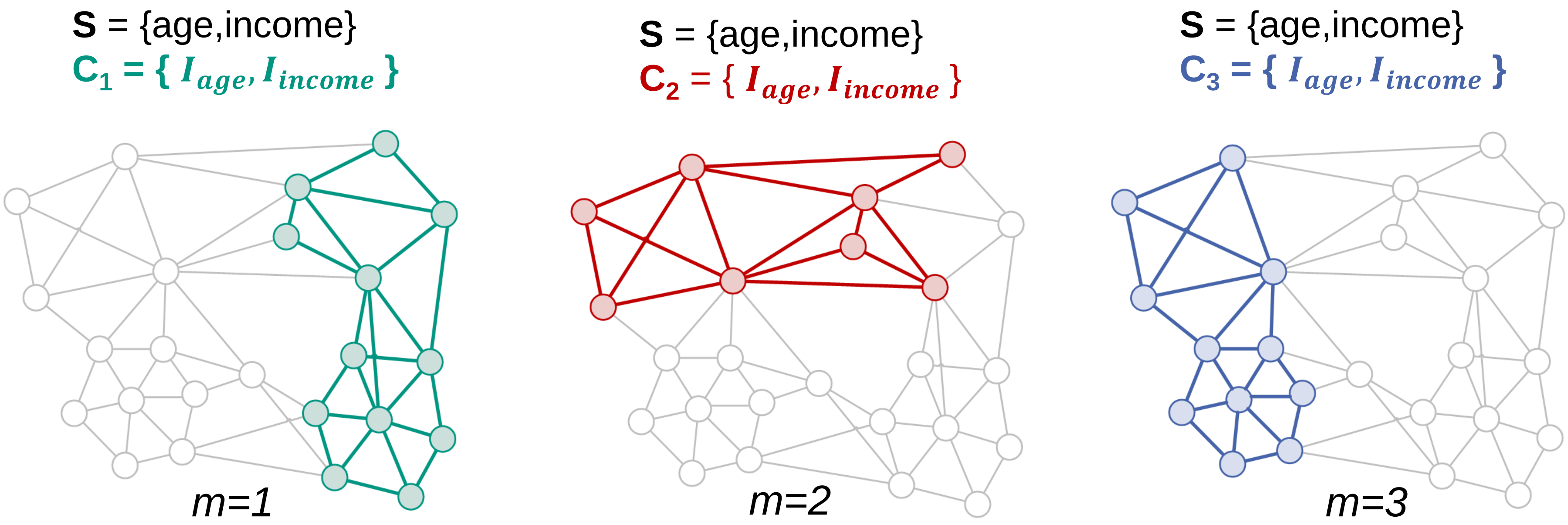
$$H_0 : |E_{C,S}| = E_{exp}(G_{C,S})$$

$$H_1 : |E_{C,S}| > E_{exp}(G_{C,S})$$ ⇐ **congruent**

**Statistical evidence for the congruence of the entire graph?**

# ConSub IV

- **Monte Carlo** approach
  - Random generation of constraint subgraphs in each iteration




$$congruence(S) \equiv \frac{1}{M} \sum_{m=1}^{M} deviation(|E_{C,S}^{m}|, E_{exp}(G_{C,S}^{m}))$$

# Experimental Setup

- Synthetic data
- Real world data

**Preprocessing**

- Fullspace
- LUFS [Tang 2012]
- **ConSub**

+

**Outlier Mining**

- CODA [Gao 2010]
- **DistOut**

☑ **quality**

AUC for known outliers

☑ **runtime**

# Experiments on Real World Networks

| | #nodes | #edges | #attributes | ground truth |
|---|---|---|---|---|
| **Amazon: Disney** | **124** | 333 | 28 | **Benchmark [Müller 2013]**<br>(external human knowledge for evaluation) |
| **Amazon: Books** | **1,418** | 3,695 | 28 | **tag: amazonfail**<br>(external human knowledge for evaluation) |
| **Enron** | **13,533** | 176,987 | 20 | **spammers**<br>(external labels used for evaluation) |


[Müller 2013] Müller et al. “Ranking outlier nodes in subspaces of attributed graphs” In GDM at IEEE ICDE 2013

# Experiments on Real World Networks

| Disney | | AUC [%] | Runtime [s] |
|---|---|---|---|
| | ConSub + DistOut | 81.77 | 8.93 |
| | ConSub + CODA | 67.97 | 152.66 |
| | LUFS + CODA | 44.44 | 3.46 |
| | Fullspace + CODA | 50.00 | 6.05 |
| **Books** | | | |
| | ConSub + DistOut | 60.02 | 2.15 |
| | ConSub + CODA | 53.53 | 14.81 |
| | LUFS + CODA | - | - |
| | Fullspace + CODA | 53.35 | 36.14 |
| **Enron** | | | |
| | ConSub + DistOut | 74.80 | 840.50 |
| | ConSub + CODA | 60.80 | 1130.78 |
| | LUFS + CODA | 48.30 | 472.60 |
| | Fullspace + CODA | 45.70 | 397.33 |

# Experiments on Real World Networks

| Disney | | AUC [%] | Runtime [s] |
|---|---|---|---|
| | ConSub + DistOut | 81.77 | 8.93 |
| | ConSub + CODA | 67.97 | 152.66 |
| | LUFS + CODA | 44.44 | 3.46 |
| | Fullspace + CODA | 50.00 | 6.05 |
| **Books** | | | |
| | ConSub + DistOut | 60.02 | 2.15 |
| | ConSub + CODA | 53.53 | 14.81 |
| | LUFS + CODA | - | - |
| | Fullspace + CODA | 53.35 | 36.14 |
| **Enron** | | | |
| | ConSub + DistOut | 74.80 | 840.50 |
| | ConSub + CODA | 60.80 | 1130.78 |
| | LUFS + CODA | 48.30 | 472.60 |
| | Fullspace + CODA | 45.70 | 397.33 |

# Experiments on Real World Networks

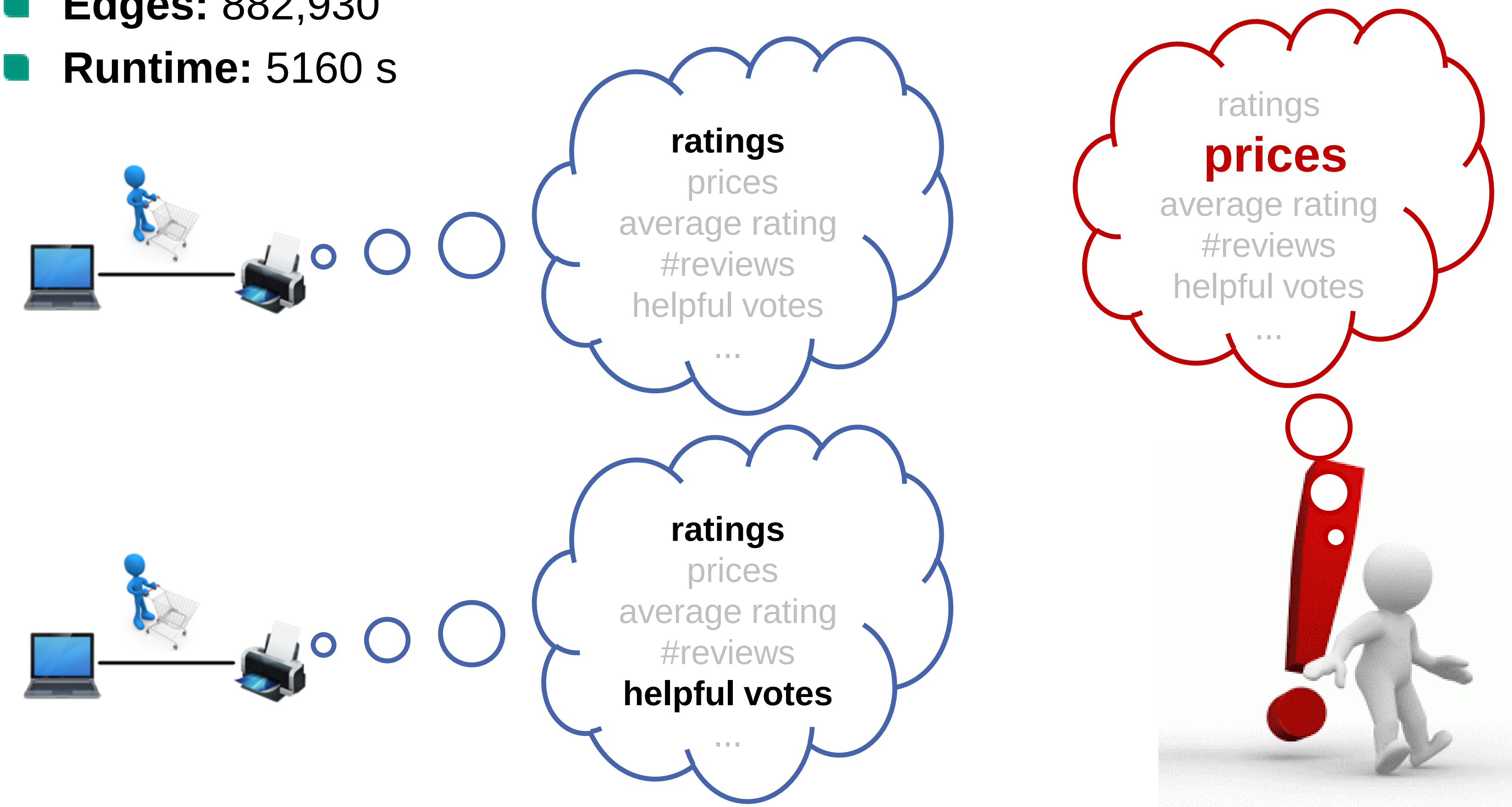
| Disney | | AUC [%] | Runtime [s] |
|---|---|---|---|
| | ConSub + DistOut | 81.77 | 8.93 |
| | ConSub + CODA | 67.97 | 152.66 |
| | LUFS + CODA | 44.44 | 3.46 |
| | Fullspace + CODA | 50.00 | 6.05 |
| **Books** | | | |
| | ConSub + DistOut | 60.02 | 2.15 |
| | ConSub + CODA | 53.53 | 14.81 |
| | LUFS + CODA | - | - |
| | Fullspace + CODA | 53.35 | 36.14 |
| **Enron** | | | |
| | ConSub + DistOut | 74.80 | 840.50 |
| | ConSub + CODA | 60.80 | 1130.78 |
| | LUFS + CODA | 48.30 | 472.60 |
| | Fullspace + CODA | 45.70 | 397.33 |

# Experiments on Real World Networks

| Disney | | AUC [%] | Runtime [s] |
|---|---|---|---|
| | ConSub + DistOut | 81.77 | 8.93 |
| | ConSub + CODA | 67.97 | 152.66 |
| | LUFS + CODA | 44.44 | 3.46 |
| | Fullspace + CODA | 50.00 | 6.05 |
| **Books** | | | |
| | ConSub + DistOut | 60.02 | 2.15 |
| | ConSub + CODA | 53.53 | 14.81 |
| | LUFS + CODA | - | - |
| | Fullspace + CODA | 53.35 | 36.14 |
| **Enron** | | | |
| | ConSub + DistOut | 74.80 | 840.50 |
| | ConSub + CODA | 60.80 | 1130.78 |
| | LUFS + CODA | 48.30 | 472.60 |
| | Fullspace + CODA | 45.70 | 397.33 |

# Subspaces Provide Novel Insights

- Giant component of the Amazon co-purchased network
  - **Nodes:** 314,824
  - **Edges:** 882,930
  - **Runtime:** 5160 s



# Conclusions & Future Work

- Challenge: attributed graphs — ✓ **Congruent subspaces**
- Homophily measure — ✓ **Congruence measure** based on statistical selection of subspaces
- Subspace selection algorithm — ✓ First algorithm: **ConSub**
- Applications — ✓ **Pre-processing of existing methods**
  ✓ **Design of novel graph mining methods**
  ✓ **Knowledge discovery in attributed graphs**

- **Future Work**
  - Mixed attribute types
  - Extensions for semi-supervised tasks


# Thank you for your attention

Our benchmark databases are available online:

**http://www.ipd.kit.edu/~muellere/consub/**