Relational Redescription Mining

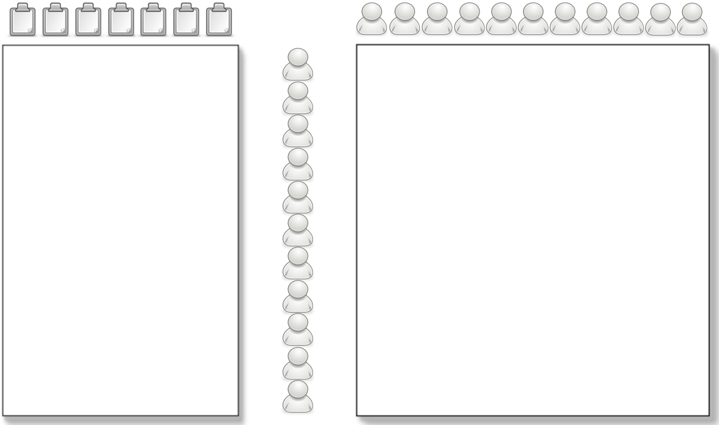
Esther Galbrun

joint work with
Pauli Miettinen and Angelika Kimmig

Helsinki Institute for Information Technology

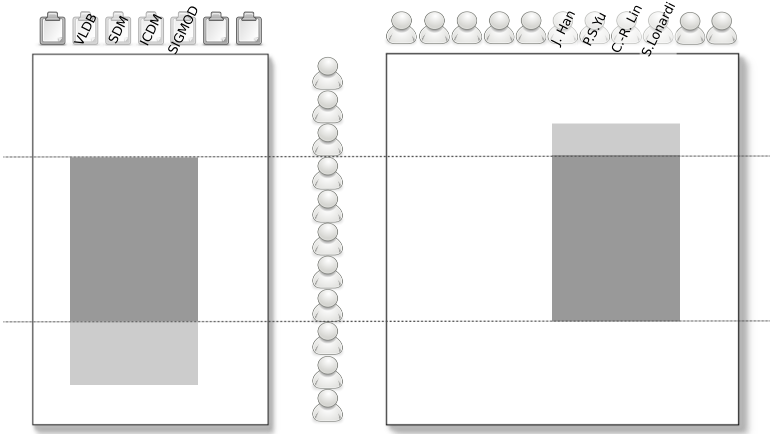
Department of Computer Science, University of Helsinki

Example: DBLP Data



Example: DBLP Data

$$\text{VLDB} \wedge \text{ICDM} \wedge \text{SDM} \wedge \text{SIGMOD}$$
$$(\text{J. Han} \wedge \text{P.S. Yu}) \vee \text{C.-R. Lin} \vee \text{S. Lonardi}$$



Definition

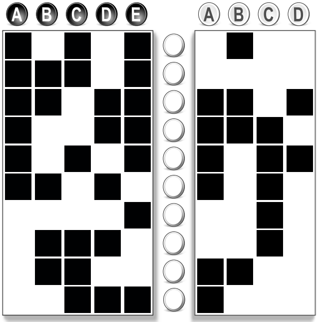
Redescription Given two datasets over the same entities, a **redescription** is a pair of queries $(q_{\mathbf{L}}, q_{\mathbf{R}})$ over the two dataset respectively, characterizing approximately the same sets of entities.

Aims

- Find coherent sets of objects
- Find sets of related attributes
- View the same objects under different perspectives

Boolean Redescriptions

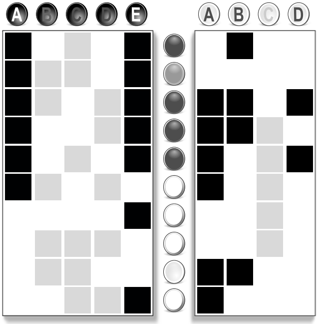
Dataset Boolean matrices



Boolean Redescriptions

$\boxed{A} \wedge \boxed{E}$

$(A \wedge D) \vee B$



Dataset Boolean matrices

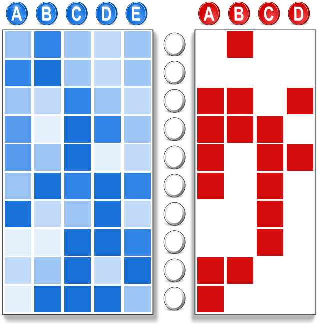
Queries Boolean formulae

Accuracy Jaccard coefficient

$$J(q_{\mathbf{L}}, q_{\mathbf{R}}) = \frac{|\mathrm{supp}(q_{\mathbf{L}}) \cap \mathrm{supp}(q_{\mathbf{R}})|}{|\mathrm{supp}(q_{\mathbf{L}}) \cup \mathrm{supp}(q_{\mathbf{R}})|}$$

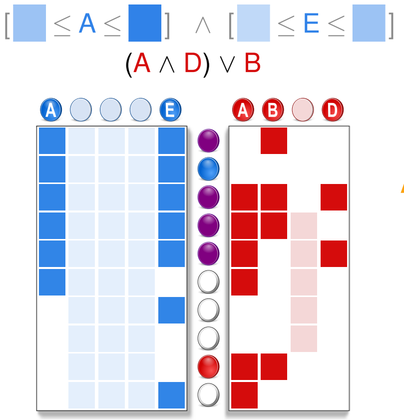
$$= \frac{\left|\boxed{E_{1,1}}\right|}{\left|\boxed{E_{1,0}}\right| + \left|\boxed{E_{1,1}}\right| + \left|\boxed{E_{0,1}}\right|}$$

Real-Valued Redescriptions



Dataset Real-valued matrices

Real-Valued Redescriptions



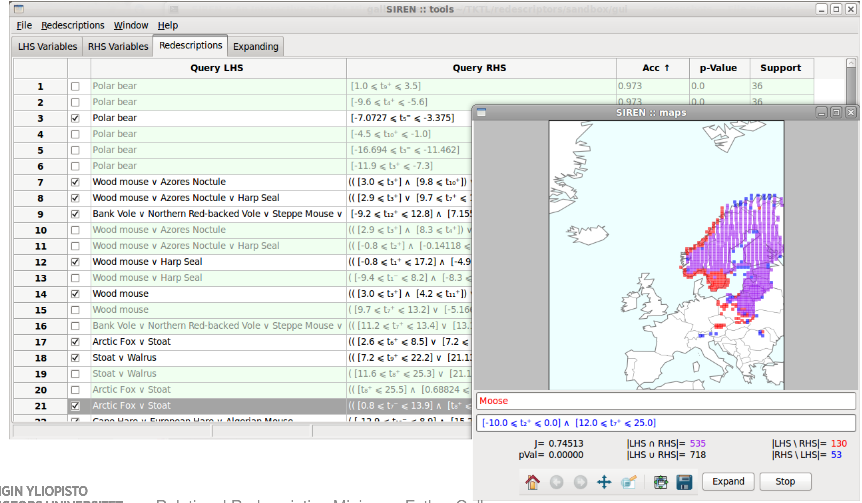
Dataset Real-valued matrices

Queries Intervals

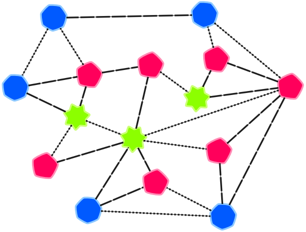
Accuracy Jaccard coefficient

$$J(q_\mathbf{L}, q_\mathbf{R}) = \frac{|\mathrm{supp}(q_\mathbf{L}) \cap \mathrm{supp}(q_\mathbf{R})|}{|\mathrm{supp}(q_\mathbf{L}) \cup \mathrm{supp}(q_\mathbf{R})|} = \frac{|E_{1,1}|}{|E_{1,0}| + |E_{1,1}| + |E_{0,1}|}$$

Geospatial Redescriptions

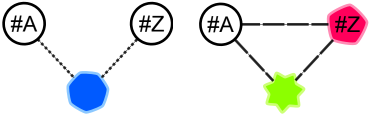


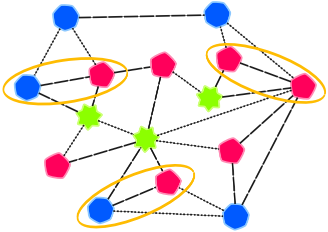
Relational Redescriptions



Dataset A network with node and egde attributes

Relational Redescriptions

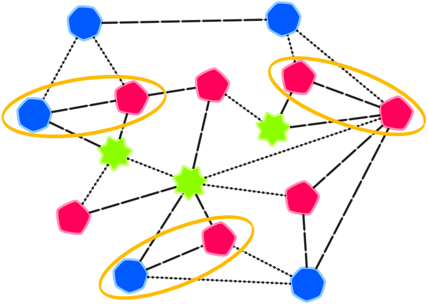




Dataset	A network with node and egde attributes
Queries	Connection patterns
Accuracy	Jaccard coefficient

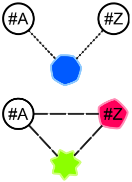
$$J(q_{\mathbf{L}}, q_{\mathbf{R}}) = \frac{|\mathrm{supp}(q_{\mathbf{L}}) \cap \mathrm{supp}(q_{\mathbf{R}})|}{|\mathrm{supp}(q_{\mathbf{L}}) \cup \mathrm{supp}(q_{\mathbf{R}})|}$$

Definition



Dataset A network with node and egde attributes

Task Find structurally different patterns covering (almost) the same pairs of nodes.



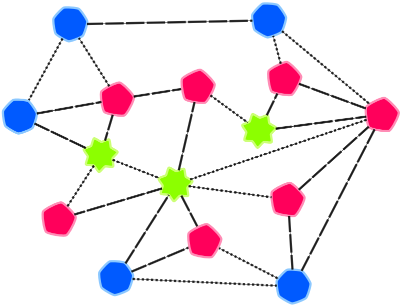
Entities and queries

✓ **pairs of nodes** and their connections.

✗ *individual nodes* and surrounding relations.

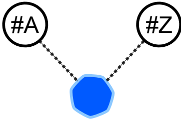
✗ *a transactional graph* and occuring subgraphs.

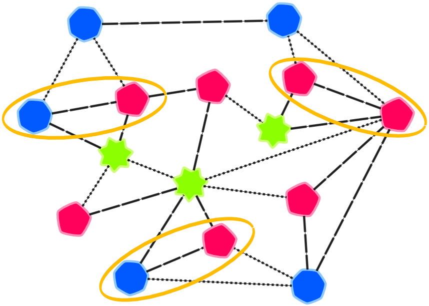
Alternating Scheme



Alternating Scheme

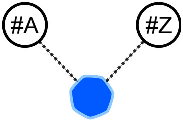
1. Fix a pattern to obtain examples

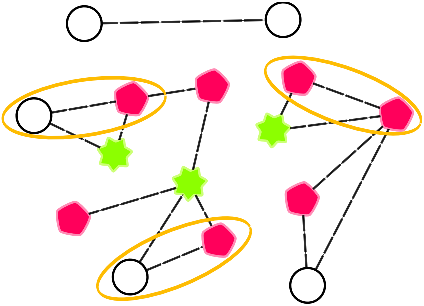




Alternating Scheme

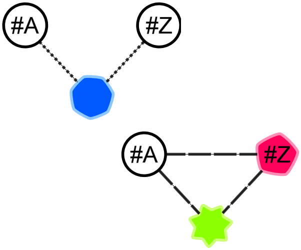
1. Fix a pattern to obtain examples
2. Consider remaining attributes

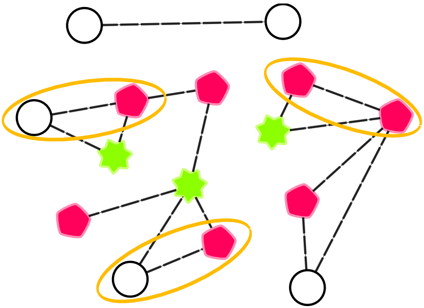




Alternating Scheme

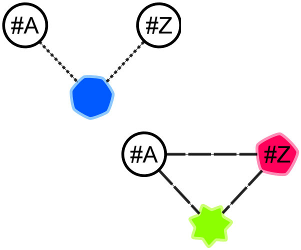
1. Fix a pattern to obtain examples
2. Consider remaining attributes
3. Find a matching pattern

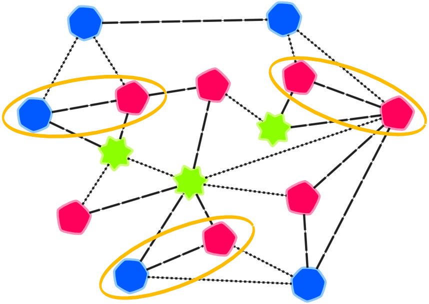




Alternating Scheme

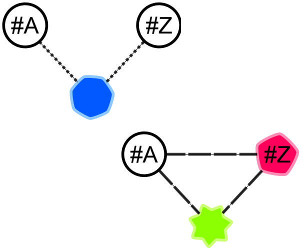
1. Fix a pattern to obtain examples
2. Consider remaining attributes
3. Find a matching pattern
4. Swap roles and iterate

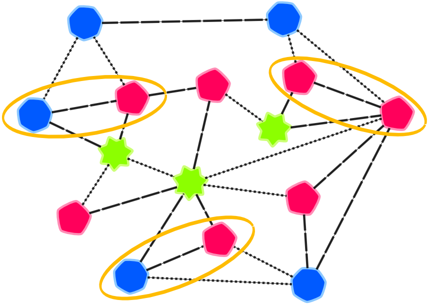




Alternating Scheme

1. Fix a pattern to obtain examples
2. Consider remaining attributes
3. **Find a matching pattern**
4. Swap roles and iterate

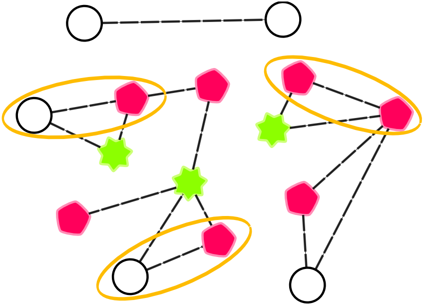




Subproblem: Query mining

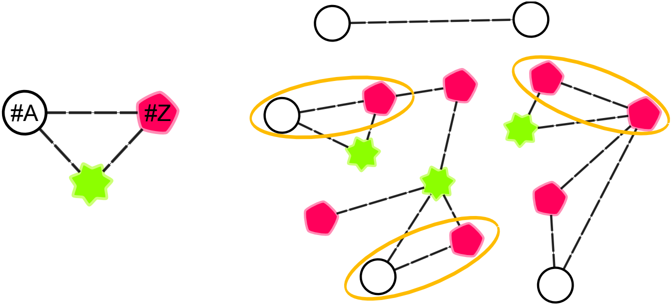
✓ Given a set of examples
  and a subset of attributes

- **Find a matching pattern**



Subproblem: Query mining

- ✓ Given a set of examples
  and a subset of attributes
- **Find a matching pattern**



FpQm: Stepwise Approach

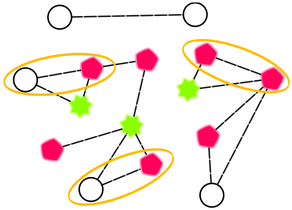
1. Enumerate connecting paths
   and mine frequent path patterns
2. Build graph patterns from path patterns
3. Select a subset of graph patterns

1. Find path patterns

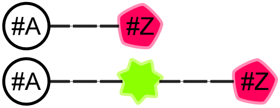
Starting with paths of length $k = 1$

1. Enumerate connecting paths
2. Mine frequent path patterns
3. Increase k by one and iterate

Until all examples are connected or k exceeds a chosen threshold



Outcome a set of frequent path patterns



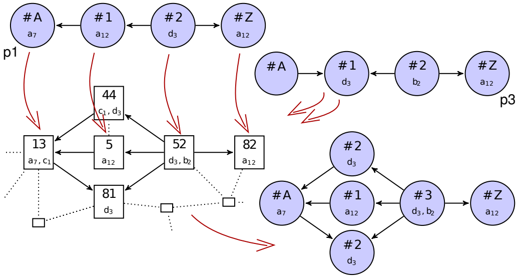
2. Build graph patterns

- ✓ Given a set of path patterns and of examples
- Combine paths to build graph patterns

2. Build graph patterns

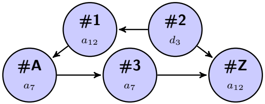
- ✓ Given a set of path patterns and of examples
- Combine paths to build graph patterns

Combination based on the instances

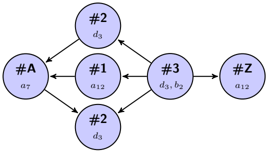


2. Build graph patterns

- ✓ Given a set of path patterns and of examples
- Combine paths to build graph patterns



Outcome a set of graph patterns

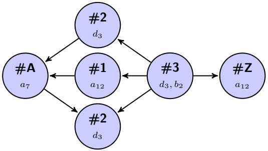


3. Select graph patterns

✓ Given a set of graph patterns and of examples

- Select a good cover

Outcome a small set of graph patterns best matching the examples



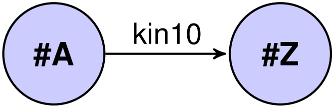
FpQm

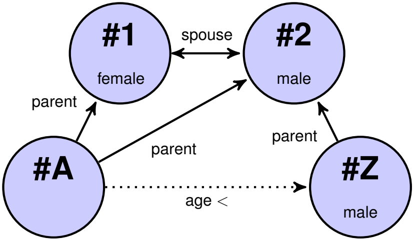
1. Enumerate connecting paths
   and mine frequent path patterns
2. Build graph patterns from path patterns
3. Select a subset of graph patterns

Alternating Scheme

initialize candidates
for each candidate **do**
    for each matching clause found with FpQm **do**
        if turns limit not reached and no equivalent clause **then**
            add to candidates
extract good pairs of adjacent clauses from the exploration tree

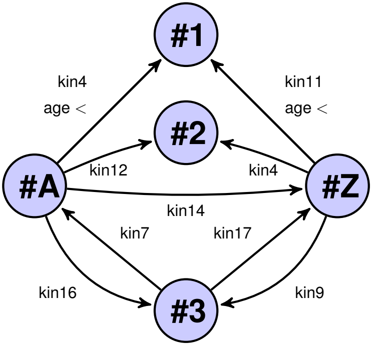
Examples from Kinship

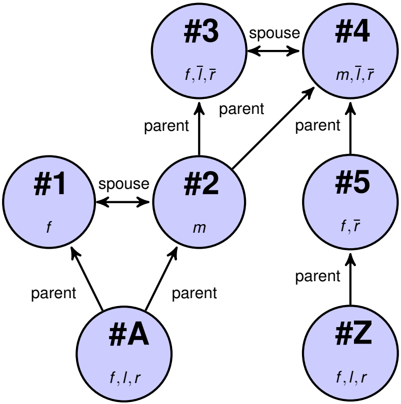




"Older brother"

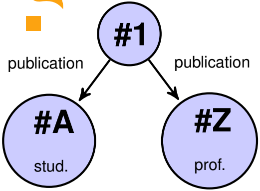
Examples from Kinship



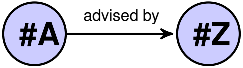


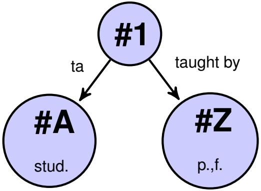
$$\left|E_{1,1}\right| = 23, \text{J} = 0.54$$

Examples from UW-CSE

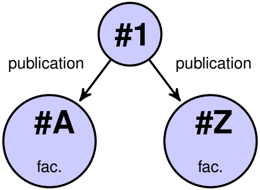


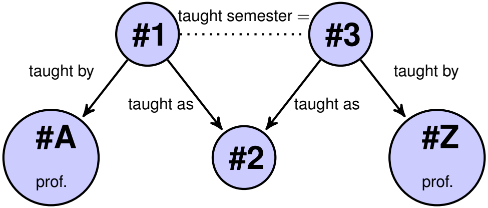
$$\left|E_{1,1}\right| = 41,\ \mathrm{J} = 0.24$$





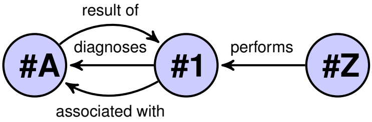
$$\left|E_{1,1}\right| = 21,\ \mathrm{J} = 0.10$$

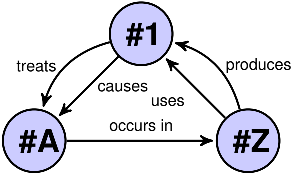




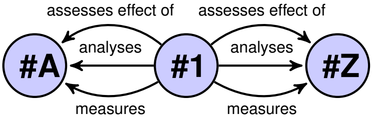
$$\left|E_{1,1}\right| = 24,\ \mathrm{J} = 0.14$$

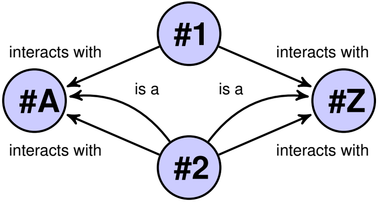
Examples from UMLS





$$\left|E_{1,1}\right| = 23,\ \mathrm{J} = 0.54$$





$$\left|E_{1,1}\right| = 506,\ \mathrm{J} = 0.62$$

But this is just ILP!?…

ILP tools *(from my uninitiated point of view)*

- general approach, encompassing various strategies
- progressive generalization / refinement of clauses
- heavy use of background knowledge, bias, types and co.

FpQM

- adapted to finding linked patterns
- purely data based, no additional knowledge
- relies on frequent paths

Experimental comparison: our approach out-performed c-armr on this task

How does it scale?

Dataset	$\|N\|$	$\|E\|$	#np.	#ep.	#cp.	$\|R\|$	$\|M\|$	Tot. T	T/clause max	T/clause avg
Kinship	381	24053	3	31	1	96	340	3h 36min	254s	38s
Umls	135	4181	–	46	–	15	81	13min 29s	79s	10s
Uwcse	1042	1674	6	7	5	8	25	39s	4s	2s

Strong impact on the running times:

- Network density
- Presence of symmetries

Relational Redescription Mining

- Find structurally different patterns covering (almost) the same pairs of nodes.
- An expressive tool for finding corresponding connections patterns in a network.

Relational Redescription Mining

- Find structurally different patterns covering (almost) the same pairs of nodes.
- An expressive tool for finding corresponding connections patterns in a network.

Thank you ...