Holistic Statistical Open Data Integration Based on Integer Linear Programming

Alain Berro, Imen Megdiche, Olivier Teste

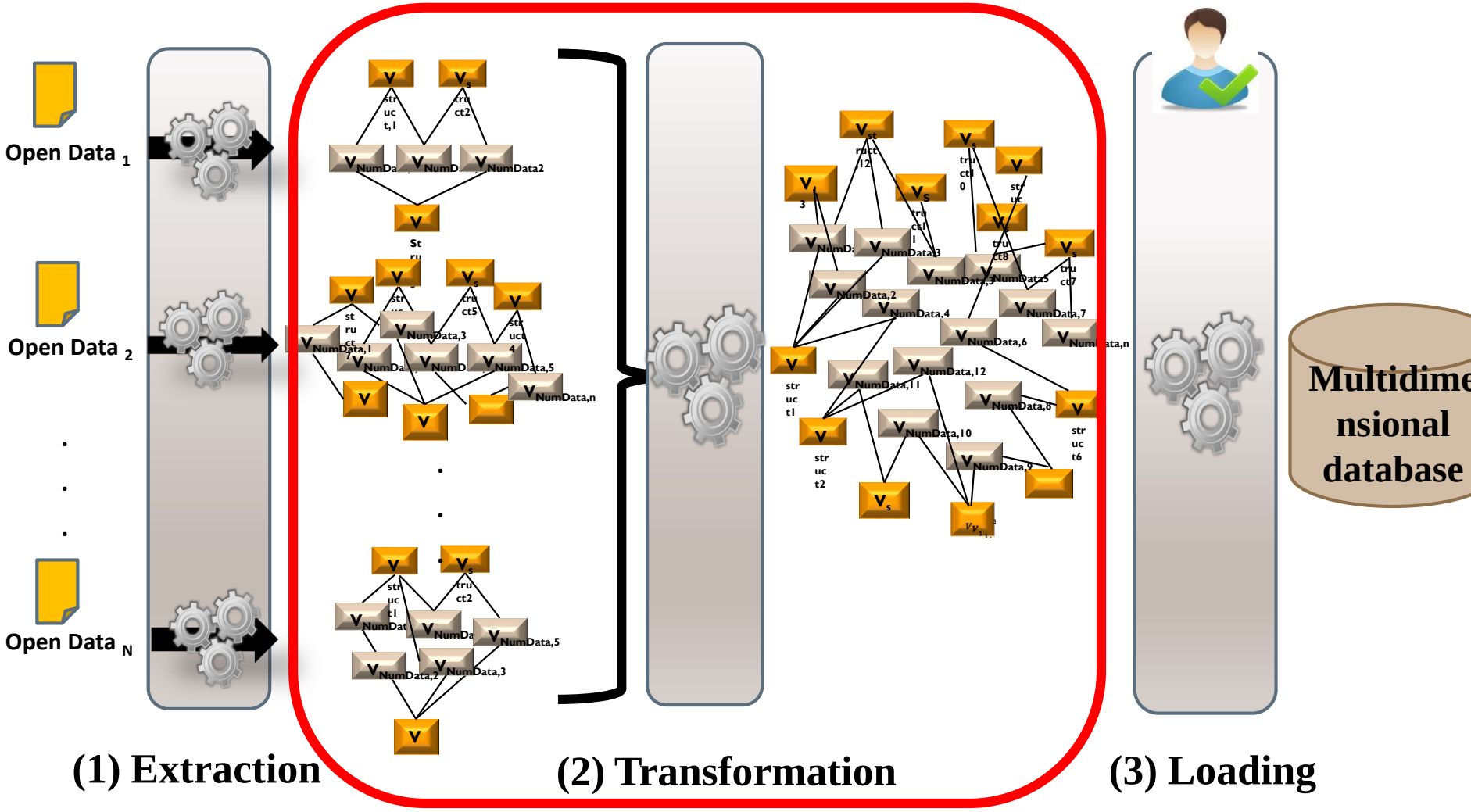
IRIT
CNRS - INPT - UPS - UT1 - UT2J

15/05/2015

Outline

- 1. Objective
- 2. State of the art
- 3. A linear Program For Holistic Matching
- 4. Evaluation
  - Matching quality
  - Matching performance
- 5. Conclusion and perspectives

1. Objective



(Berro et al., 2014)

1. Schema matching problem

SCHEMA 1

- Library
  - Item
    - Isbn
    - Author
    - Title
    - Year
  - Author
    - First name
    - Last name
  - Borrowed Items
    - Item
    - Borrower
  - Borrower
    - First name
    - Last name

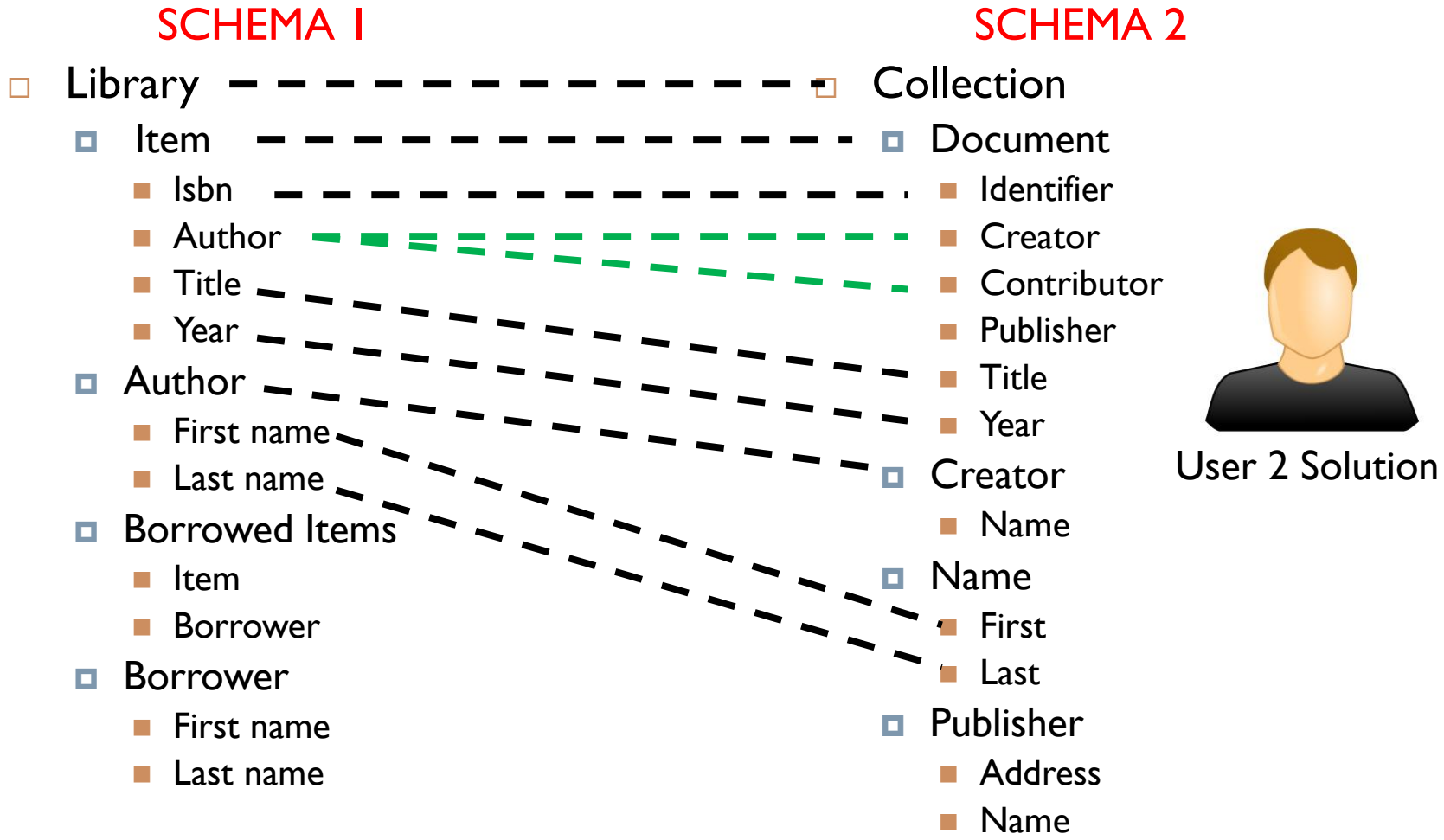
SCHEMA 2

- Collection
  - Document
    - Identifier
    - Creator
    - Contributor
    - Publisher
    - Title
    - Year
  - Creator
    - Name
  - Name
    - First
    - Last
  - Publisher
    - Address
    - Name

User 1 Solution

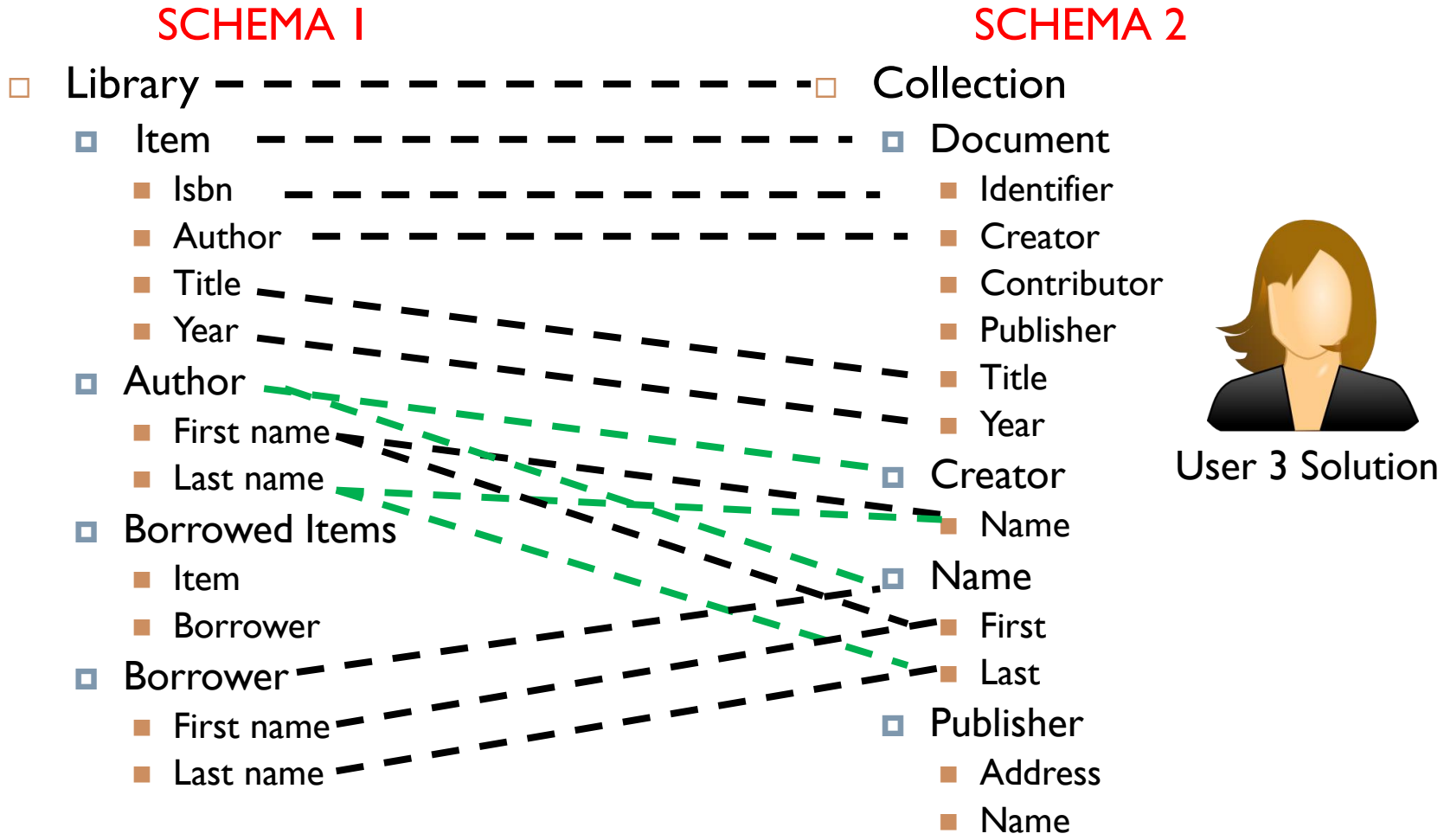
Task 3 from the benchmark http://infolab.stanford.edu/~melnik/mm/sfa/

1. Schema matching problem



Task 3 from the benchmark http://infolab.stanford.edu/~melnik/mm/sfa/

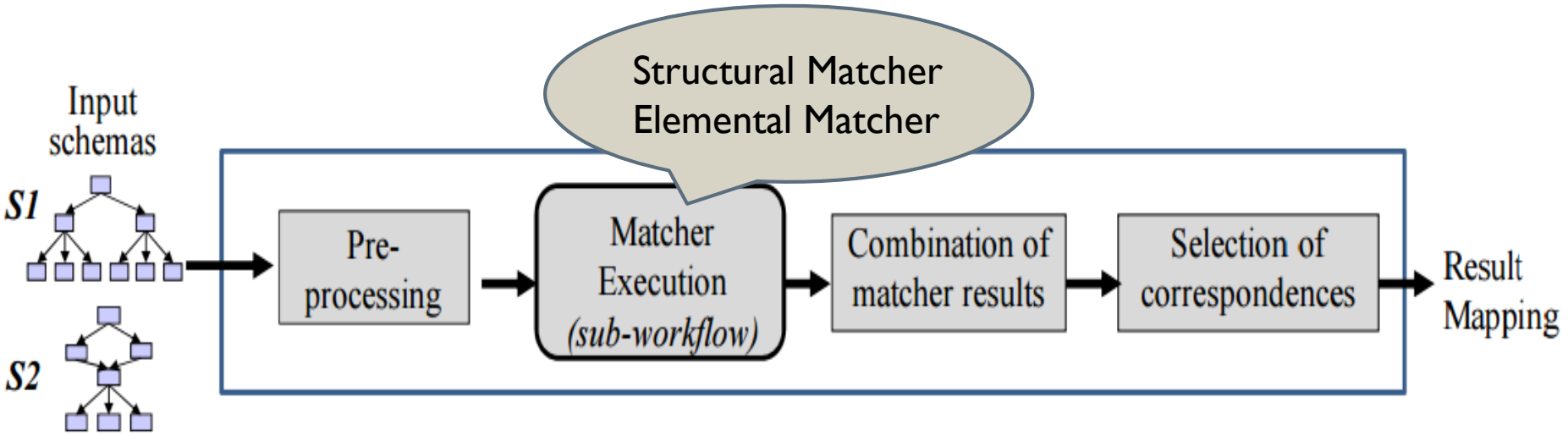
1. Schema matching problem



Task 3 from the benchmark http://infolab.stanford.edu/~melnik/mm/sfa/

2. State of the art

Schema Matching Problem



General workflow for matching
(Bellahsene et al. 2011)

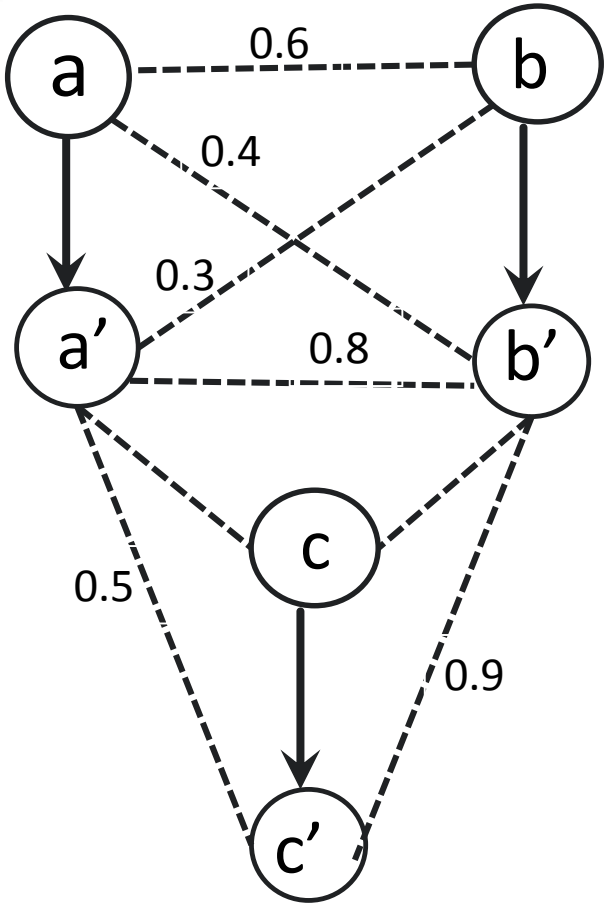
2. State of the art

Schema Matching Problem

Approach	Type	Internal Model	Combinatorial Optimisation Reduction			Dependant to threshold
			Problem	Complexity	Used in	
COMA++	Pairwise	Directed acyclic graph	-	-	-	Yes
Similarity Flooding	Pairwise	Labelled graphs	Stable Marriage	Polynomial	Selection	Yes
BMatch	Pairwise	Tree	-	-	-	Yes
CODI	Pairwise	Labelled graphs	Max-Sat	NP-Hard	Structural matcher & Selection	Yes
OLA	Pairwise	Labelled graphs	Maximum weighted graph matching	Polynomial	Selection	Yes
DCM	Holistic	List of attributes	-	-	-	Yes
PORSCHE	Holistic	Tree	-	-	-	No
PLASMA	Holistic	Tree	-	-	-	Yes
LP4HM	Holistic	Directed acyclic graph	Maximum weighted graph matching	Polynomial	Structural matcher & Selection	No

3. A Linear Program For Holistic Matching (LP4HM)

Global optimal solution



Schema Matching
Find the best set of correspondences having 1:1 cardinality

Structural constraints

3. A Linear Program For Holistic Matching (LP4HM)

$$\max \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} sim_{i_k,j_l}\ x_{i_k,j_l}$$

Objective function

$$s.t. \quad \sum_{l=1}^{n_j} x_{i_k,j_l} \leq 1, \quad \forall k \in [1, n_i] \quad \forall i \in [1, N-1]\ \forall j \in [i+1, N]$$

$$sim_{i_k,j_l}\ x_{i_k,j_l} \geq seuil\ x_{i_k,j_l} \quad \forall i \in [1, N-1]\ \forall j \in [i+1, N] \quad \forall k \in [1, n_i],\ \forall l \in [1, n_j]$$

$$x_{i_k,j_l} \leq x_{i_{pred(k)},j_{pred(l)}} \quad \forall i \in [1, N-1]\ \forall j \in [i+1, N] \quad \forall k \in [1, n_i],\ \forall l \in [1, n_j]$$

$$x_{i_k,j_l} + x_{i_{k'},j_{l'}} - (dir_{i_{k,k'}} dir_{j_{l,l'}}) \leq 1 \quad \forall i \in [1, N-1]\ \forall j \in [i+1, N] \quad \forall k, k' \in [1, n_i],\ \forall l, l' \in [1, n_j]$$

Linear constraints

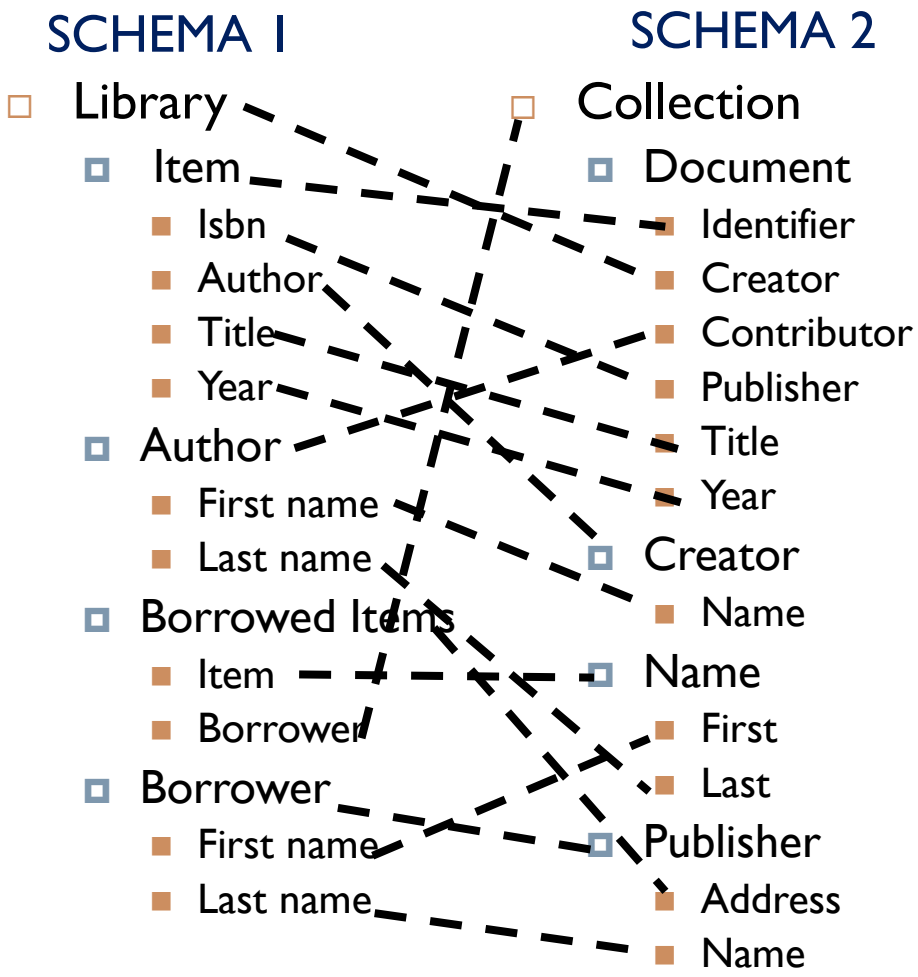
$$x_{i_k,j_l} \in \{0, 1\} \quad \forall i \in [1, N-1]\ \forall j \in [i+1, N] \quad \forall k \in [1, n_i],\ \forall l \in [1, n_j]$$

Decision variables

3. LP4HM : cardinality “1-1”

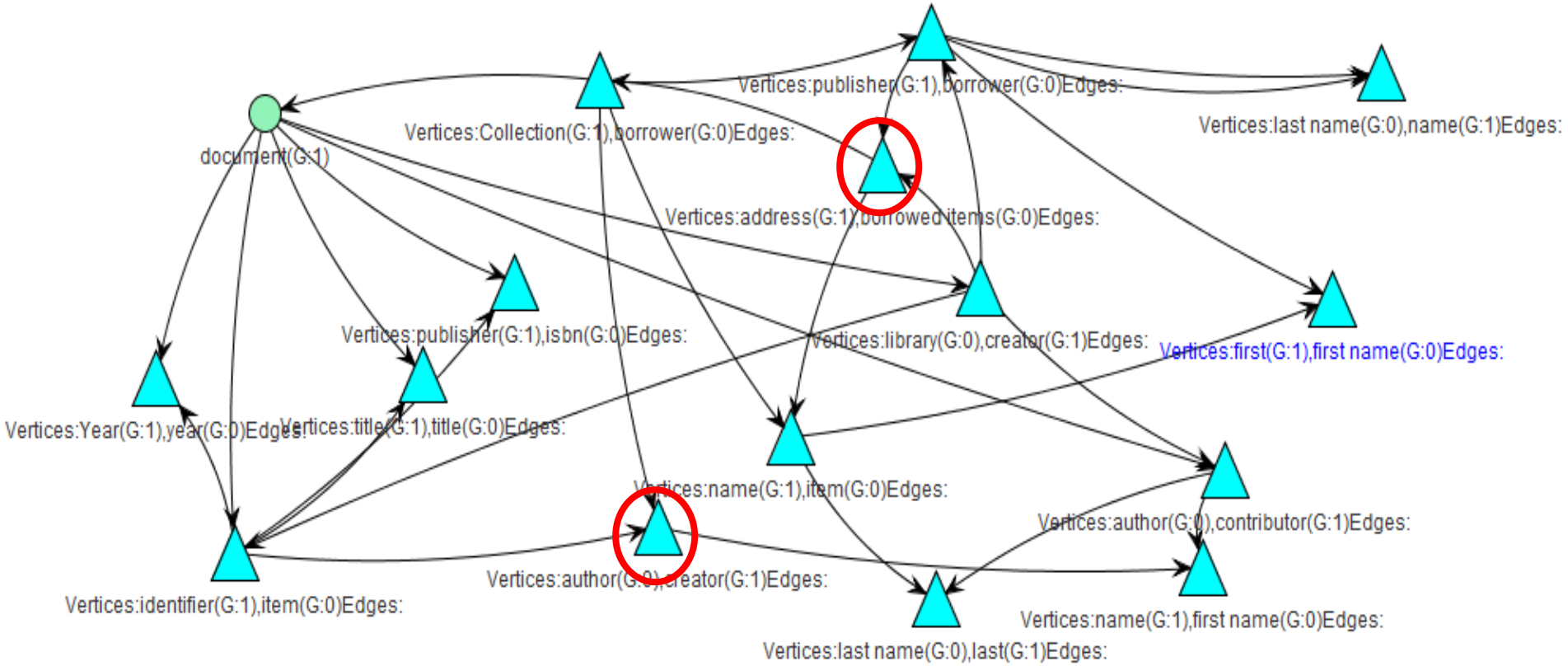
Each node in a graph G_i could match with at most one node in another graph G_j

$$\sum_{l=1}^{n_j} x_{i_k,j_l} \leq 1,$$



Task 3 from the benchmark http://infolab.stanford.edu/~melnik/mm/sfa/

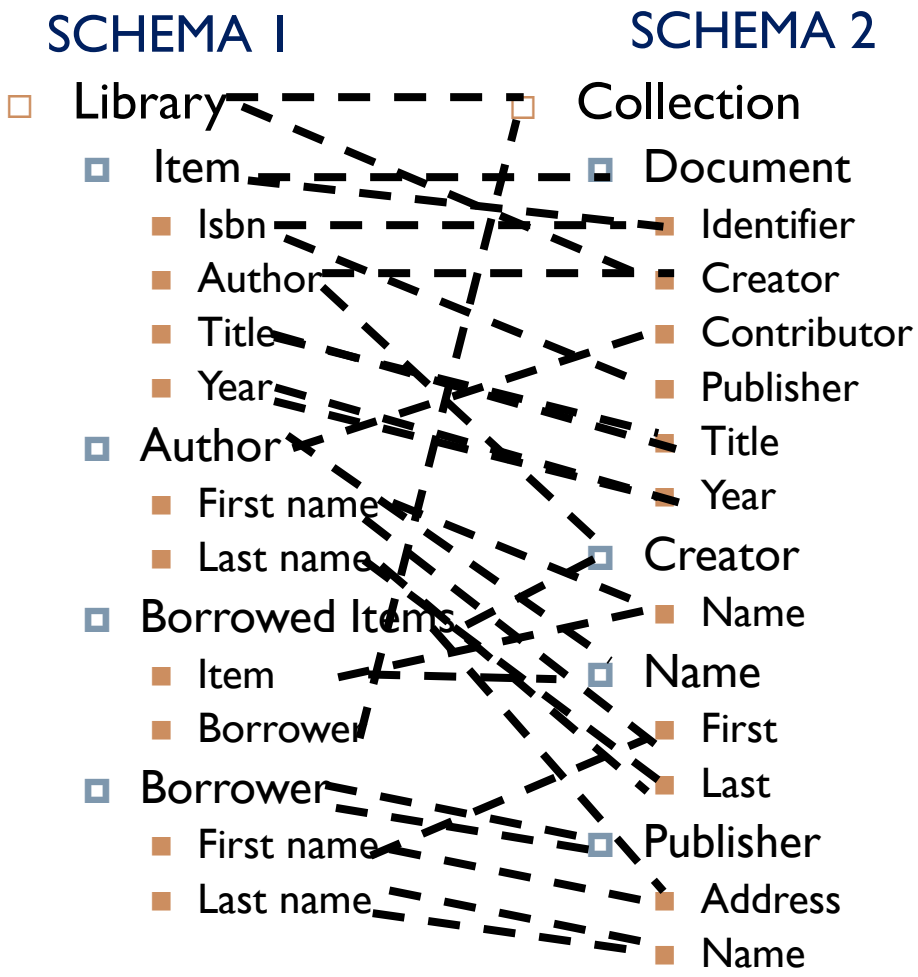
3. LP4HM : cardinality “1-1”



3. LP4HM : strict hierarchies

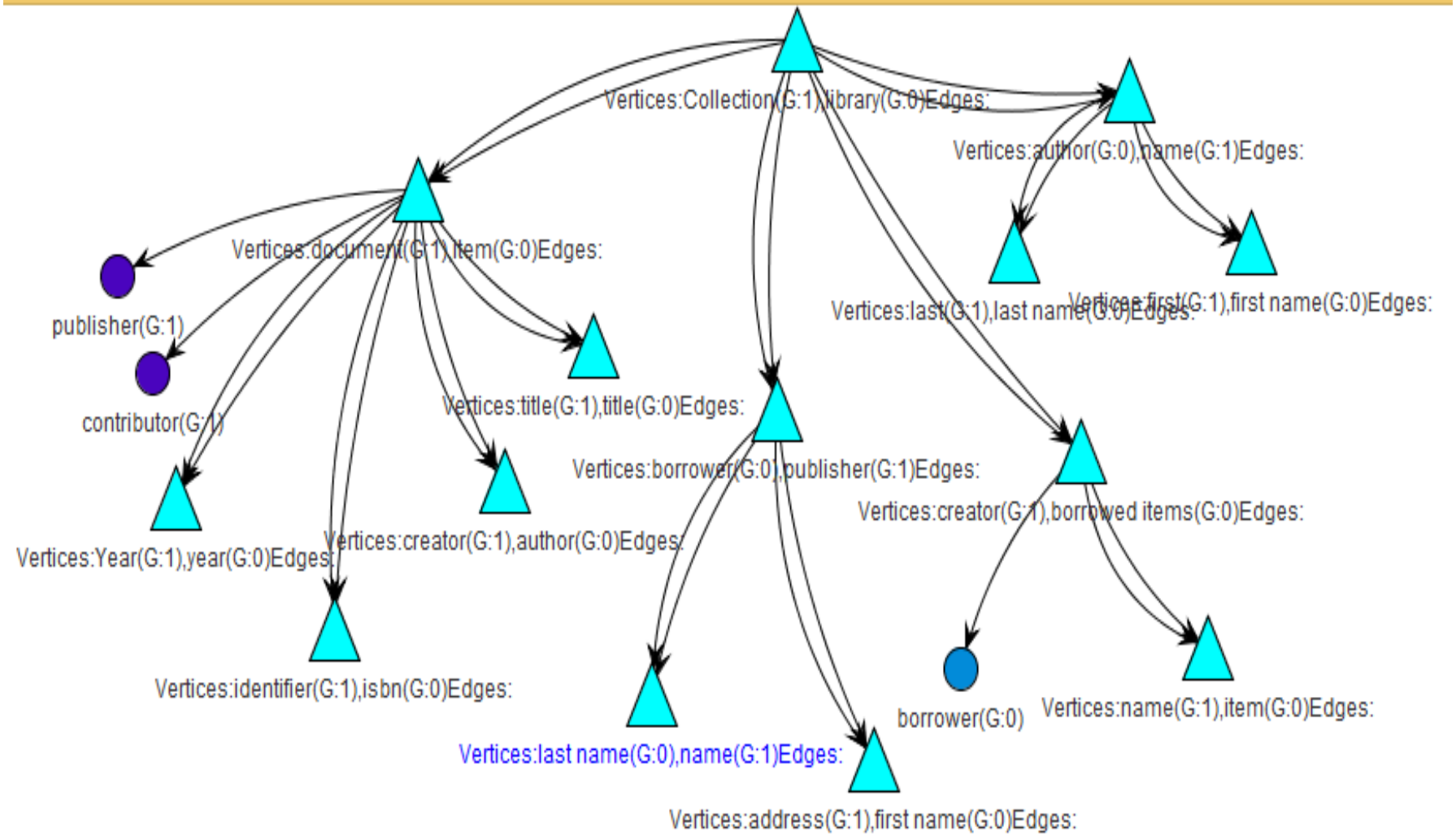
Each node should have at most one parent :
if children nodes match so their predecessor have to match too

$$x_{i_k,j_l} \leq x_{i_{pred(k)},j_{pred(l)}}$$



Task 3 from the benchmark http://infolab.stanford.edu/~melnik/mm/sfa/

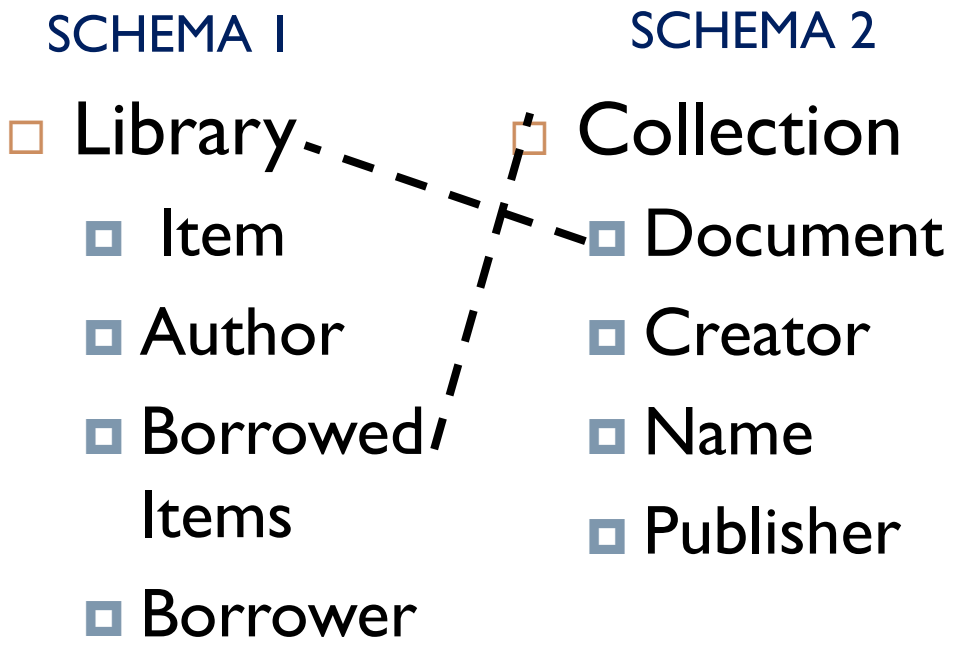
3. LP4HM : strict hierarchies



3. LP4HM : coherent structure

Prevent the generation of conflictual edges resulting from the matching of parents with childs' and childs with parents'

$$x_{i_k,j_l} + x_{i_{k'},j_{l'}} - (dir_{i_{k,k'}} dir_{j_{l,l'}}) \leq 1$$



Task 3 from the benchmark http://infolab.stanford.edu/~melnik/mm/sfa/

3. LP4HM : threshold

$$\max \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} sim_{i_k,j_l}$$

$$s.t. \quad \sum_{l=1}^{n_j} x_{i_k,j_l} \leq 1, \quad \forall k \in [1,$$

$$\forall i \in [1,$$

$$sim_{i_k,j_l}\ x_{i_k,j_l} \geq seuil\ x_{i_k,j_l}$$
$$\forall i \in [1, N-1]\ \forall j \in [i+1, N]$$
$$\forall k \in [1, n_i],\ \forall l \in [1, n_j]$$

$$x_{i_k,j_l} \leq x_{i_{pred(k)},j_{pred(l)}}$$
$$\forall i \in [1, N-1]\ \forall j \in [i+1, N]$$
$$\forall k \in [1, n_i],\ \forall l \in [1, n_j]$$

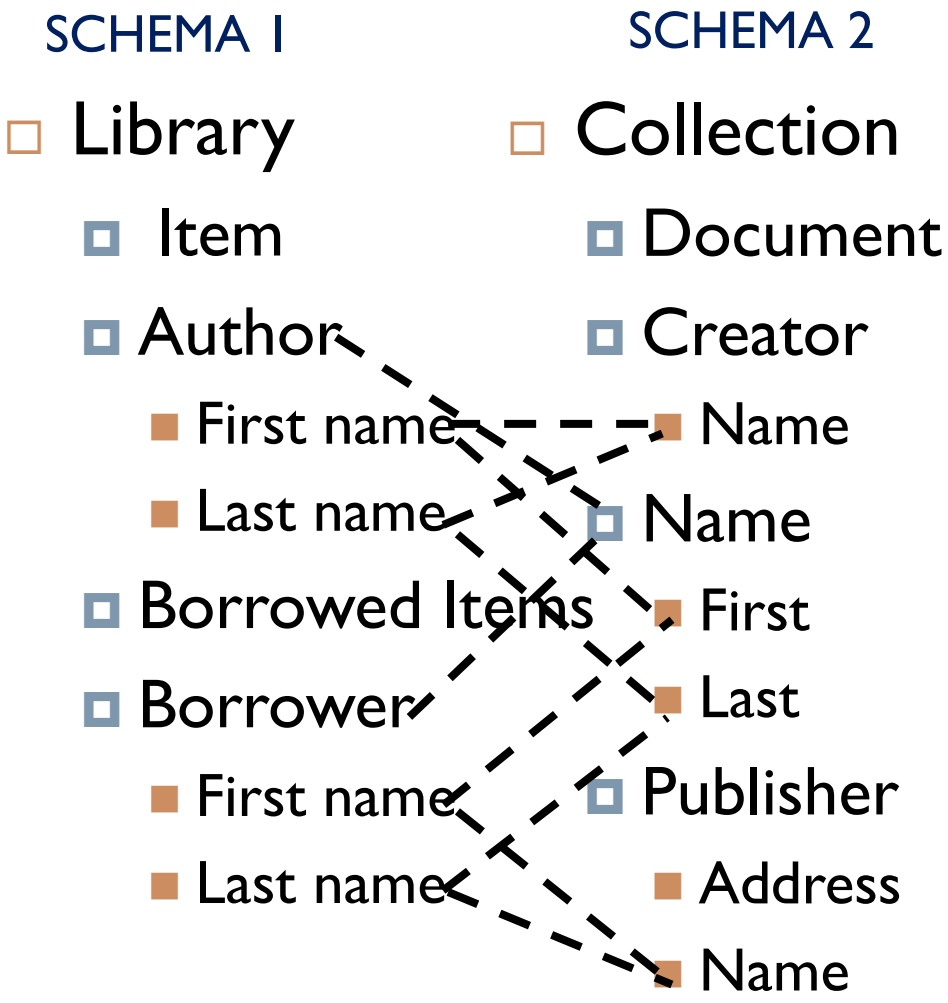
$$x_{i_k,j_l} + x_{i_{k'},j_{l'}} - (dir_{i_{k,k'}} dir_{j_{l,l'}}) \leq 1$$
$$\forall i \in [1, N-1]\ \forall j \in [i+1, N]$$
$$\forall k, k' \in [1, n_i],\ \forall l, l' \in [1, n_j]$$

$$x_{i_k,j_l} \in \{0, 1\}\ \forall i \in [1, N-1]\ \forall j \in [i+1, N]$$
$$\forall k \in [1, n_i],\ \forall l \in [1, n_j]$$

For a given threshold, all correspondences should be above this threshold

3. LP4HM Relaxed

From binary to fractional decision variables (in [0,1])



Task 3 from the benchmark http://infolab.stanford.edu/~melnik/mm/sfa/

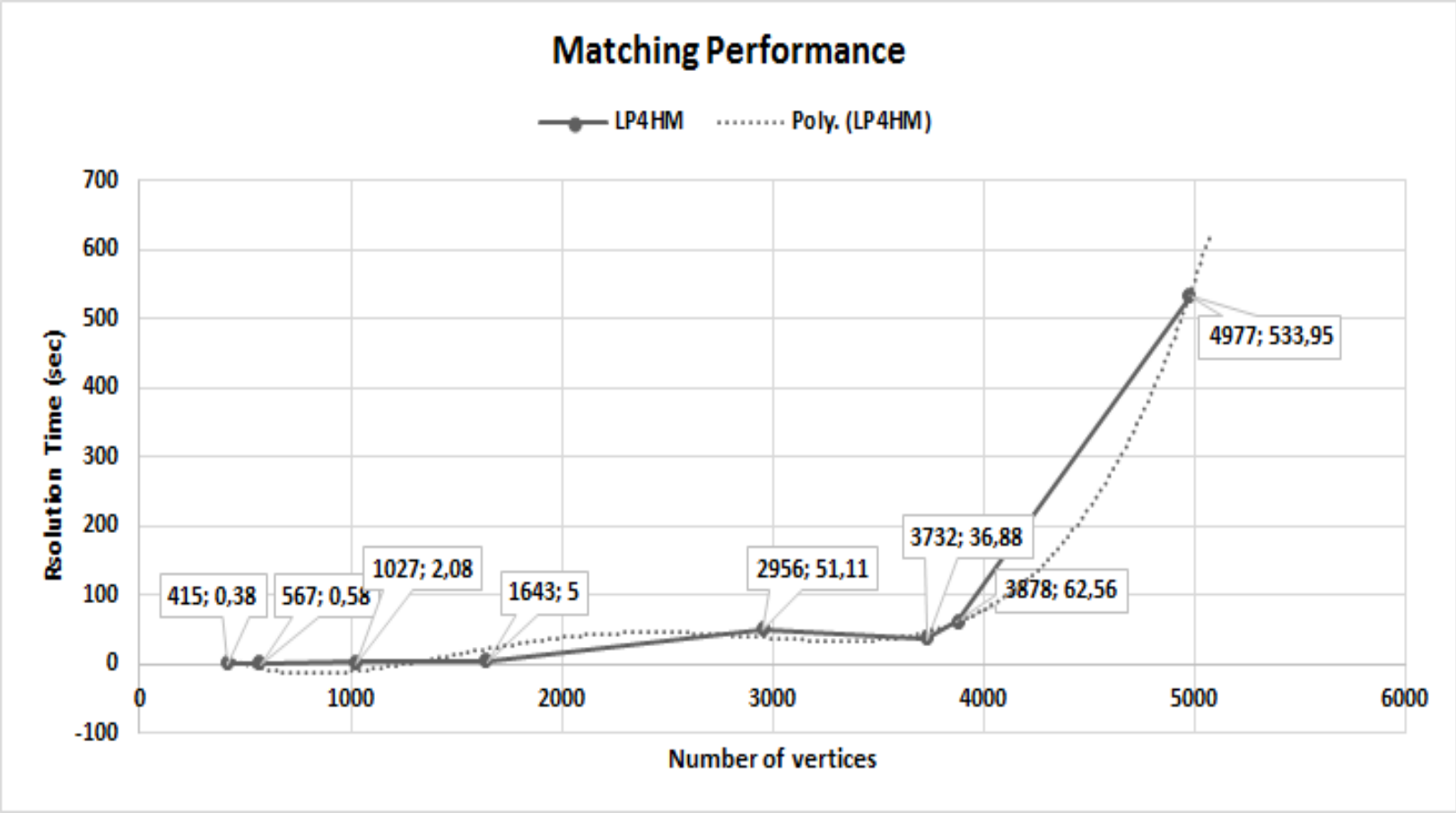
4. Evaluation
Matching quality

- Results on Similarity Flooding benchmark (7 users, 9 tasks)

	Precision	Recall	F-Measure	Accuracy	HSR
LP4HM	67%	58%	62%	30%	**81%**
LP4HM(Relaxed)	58%	**66%**	60%	23%	**81%**
COMA++	72%	50%	58%	32%	76%
Bmatch	22%	47%	28%	0%	69%
Similarity Flooding	**81%**	55%	**65%**	**43%**	80%

the benchmark http://infolab.stanford.edu/~melnik/mm/sfa/

4. Evaluation
Matching performance



5. Conclusion and perspectives

- An approach to holistically integrate open data
  - A linear program focused on hierarchical structural constraints
  - Two strategies to resolve two types of cardinality problem
- Evaluation based on matching quality and performance
  - Competitive results without similarity threshold
  - A polynomial trendline for several input open data
  - A generalised approach for N >= 2 graphs (holistic)
- Perspectives
  - Extend this model to handle labelled graphs
  - Experiments on larger datasets

Thank you for you attention
Questions ?

{berro,megdiche, teste}@irit.fr

IRIT, Toulouse, France