INTÉGRATION HOLISTIQUE DES GRAPHES BASÉE SUR LA PROGRAMMATION LINÉAIRE POUR L'ENTREPOSAGE DES OPEN DATA

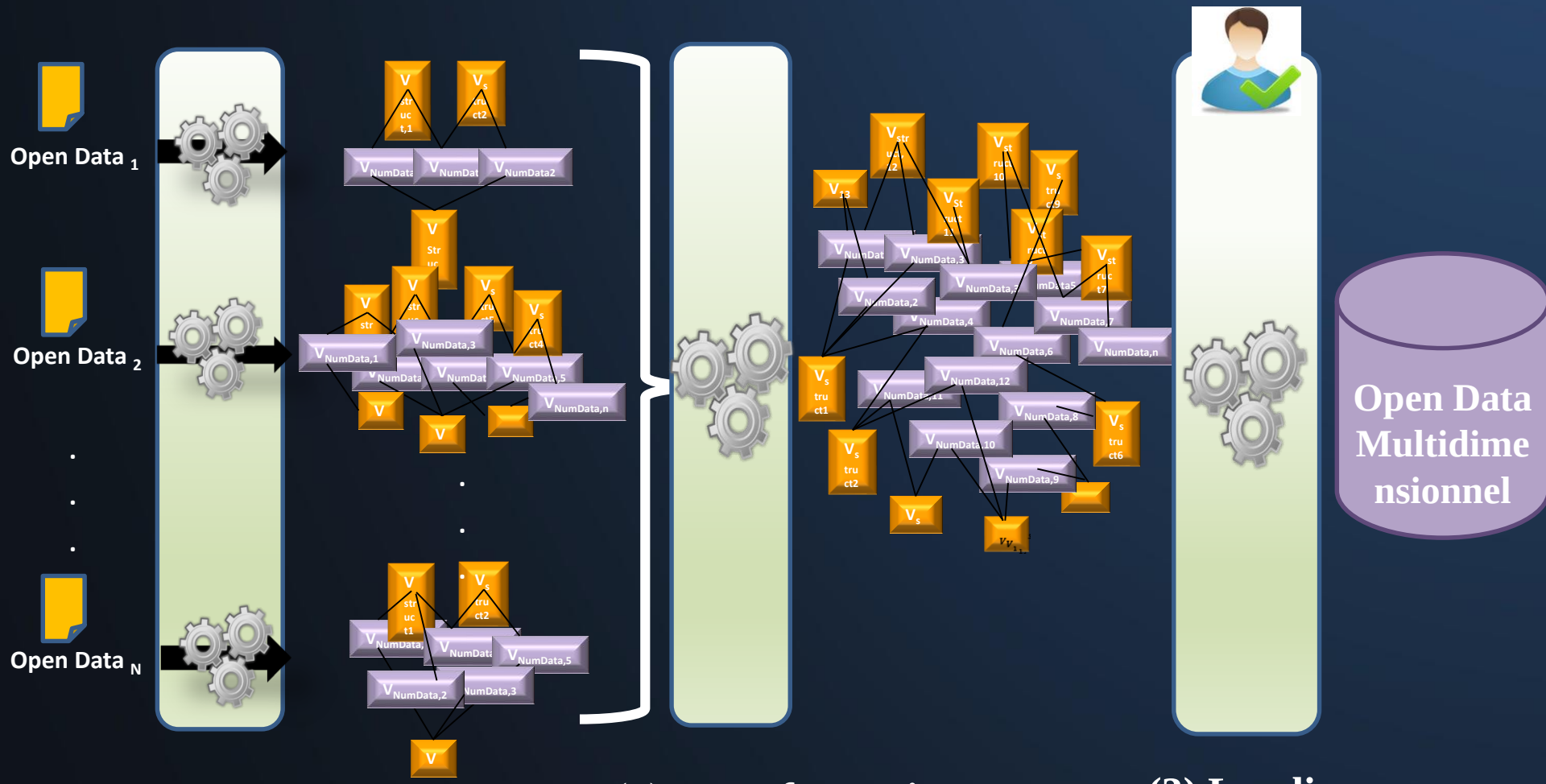
EDA' 2015

- **Présenté par : Imen Megdiche**
- **Directeur de thèse : Olivier Teste**
- **Co-directeur de thèse : Alain Berro**

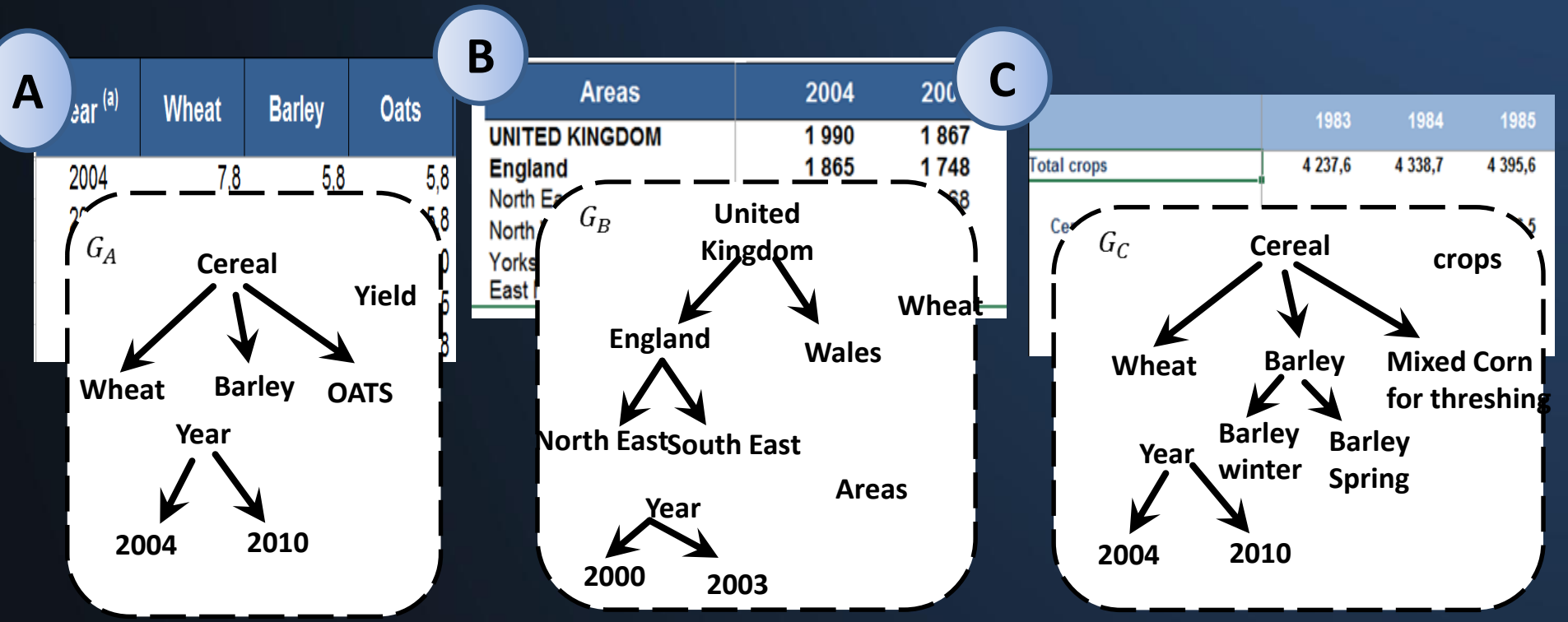
Plan

1. Contexte de recherche
2. Etat de l’art
3. Approche proposée : LP4HM
4. Validation expérimentale
5. Conclusion et perspectives

1. Un processus ETL basé sur les graphes



1. Exemples d'open data

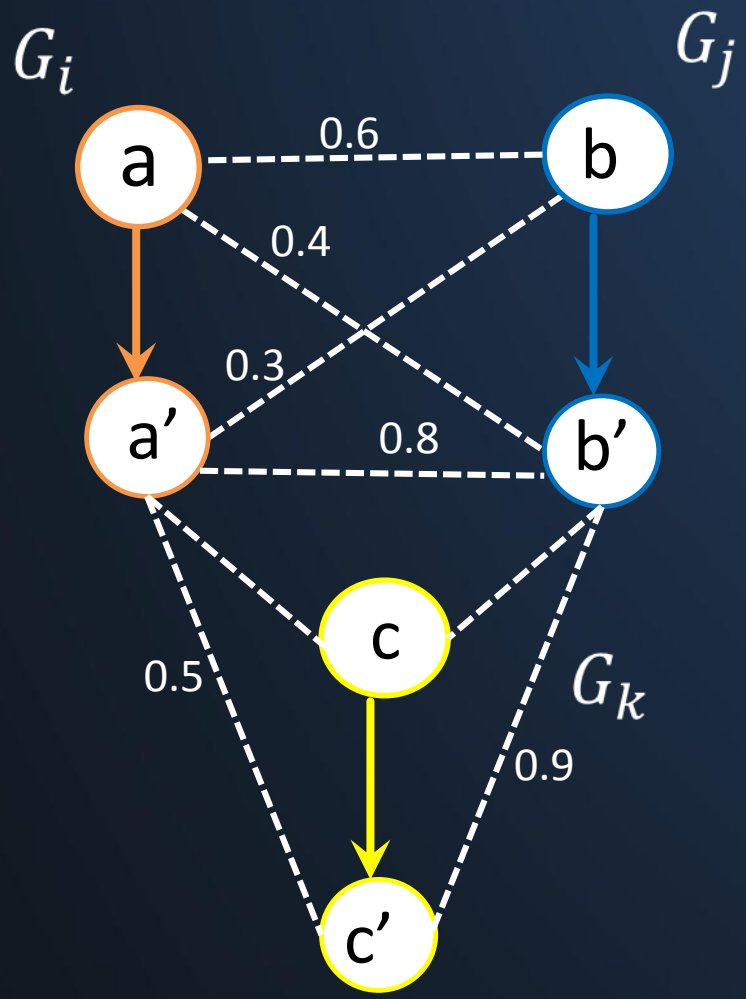


Comment intégrer plusieurs graphes hiérarchiques de concepts hétérogènes ?

2. Etat de l'art : matching de schémas

Approche	Type	Représentation interne	Type de matching
COMA++	Pairwise	Arbre	élément+ structure
Similarity Flooding	Pairwise	Graphes (RDF)	élément+ structure
BMatch	Pairwise	Arbre	élément+ structure
PSM	Holistique pré-matching	Liste d'attributs	élément
PORSCHE	Holistique pré-matching	Arbre	élément+ structure
PLASMA	Holistique pré-matching	Arbre	élément+ structure
LP4HM	Holistique	Arbre	élément+ structure

3. Principe de notre approche



Matching de schémas
Chercher le meilleur ensemble de correspondances de cardinalité 1:1

+

Contraintes de structure

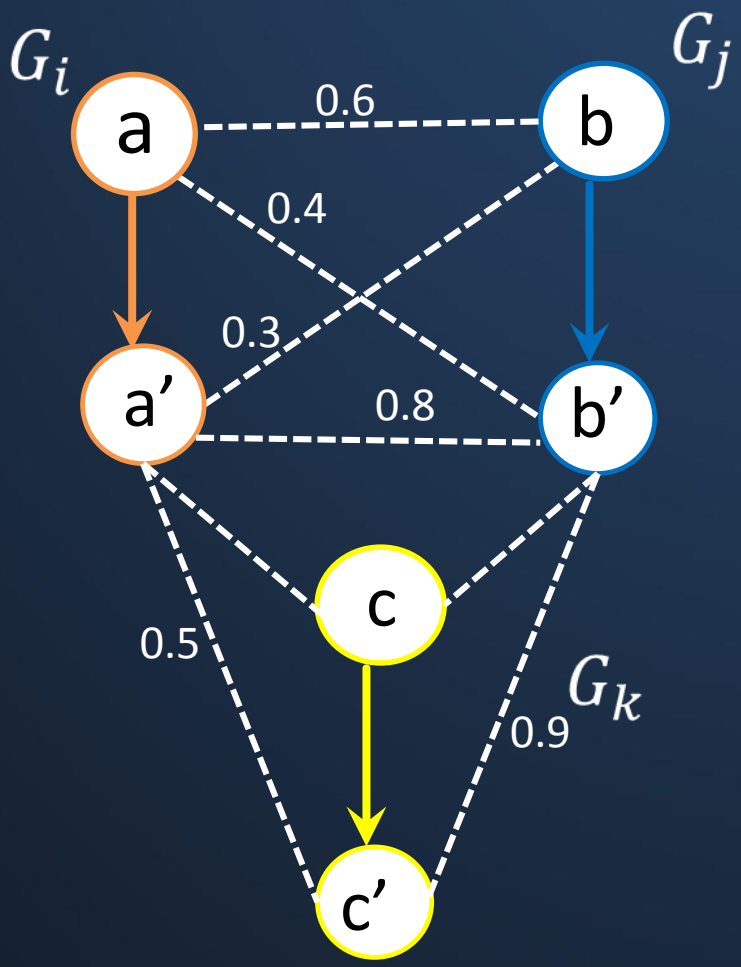
3. Principe de notre approche

Programme linéaire pour le matching holistique.

<u>Objectif</u> : maximiser la somme des similarités des correspondances.

<u>Contraintes</u> :

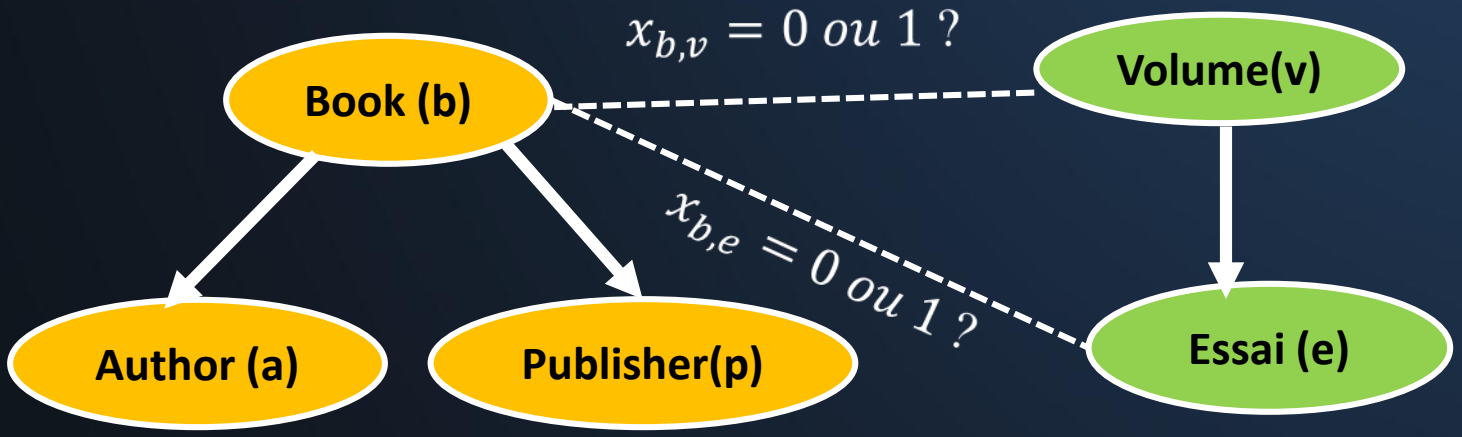
- **Sur les correspondances entre éléments**
  - **Cardinalité 1:1**
  - **Seuil de similarité**
- **Sur la structure des graphes**
  - **Cohérence**
  - **Hiérarchies strictes**



3. Formulation de LP4HM

- Variables de décision

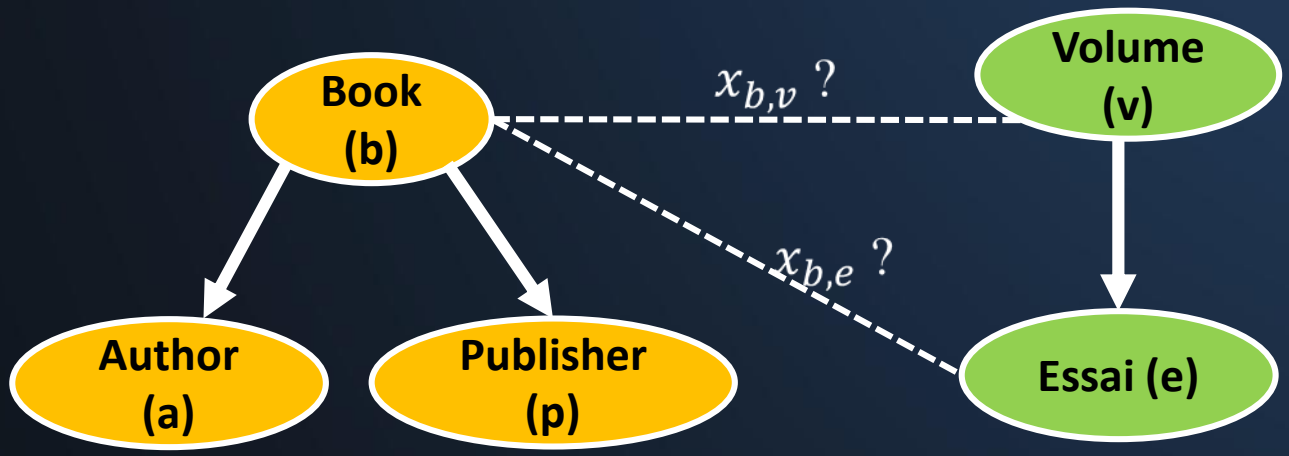
  Y-a-t-il une correspondance entre deux nœuds de deux graphes ? (Oui / Non)



3. Formulation de LP4HM

❑ Contrainte de cardinalité 1:1

Chaque nœud d'un graphe G_i doit correspondre à au plus un nœud d'un graphe G_j

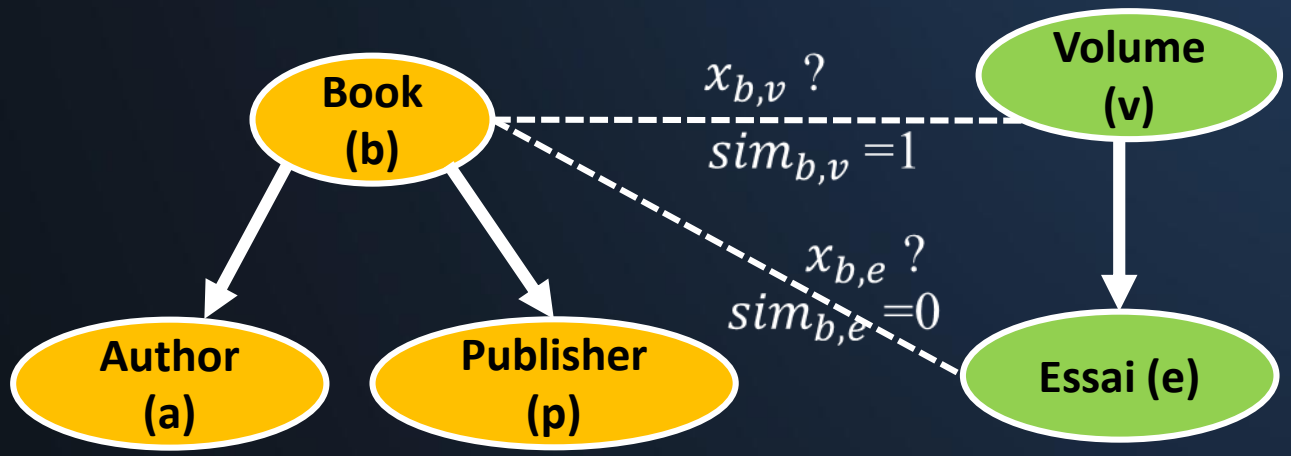


$$x_{b,v} + x_{b,e} <= 1$$

3. Formulation de LP4HM

❑ Contrainte de seuil

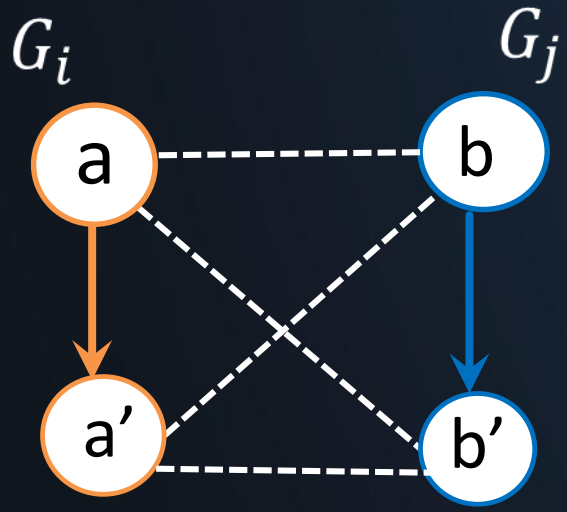
Pour un seuil donné, les similarités des correspondances de la solution optimale doivent être supérieures à ce seuil.

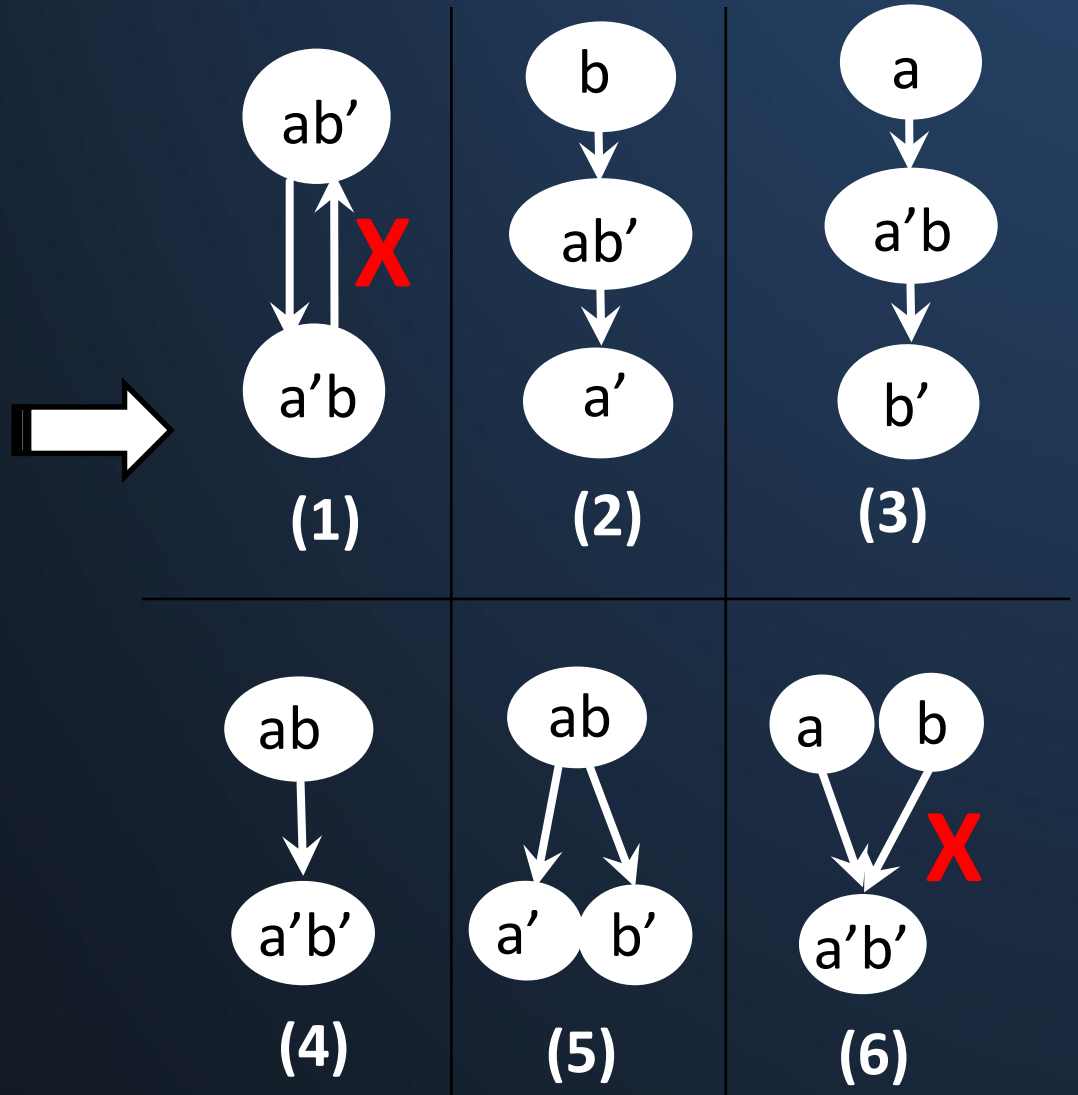


$$(sim_{b,v} - \text{seuil}) \cdot x_{b,v} \geq 0$$

3. Formulation de LP4HM

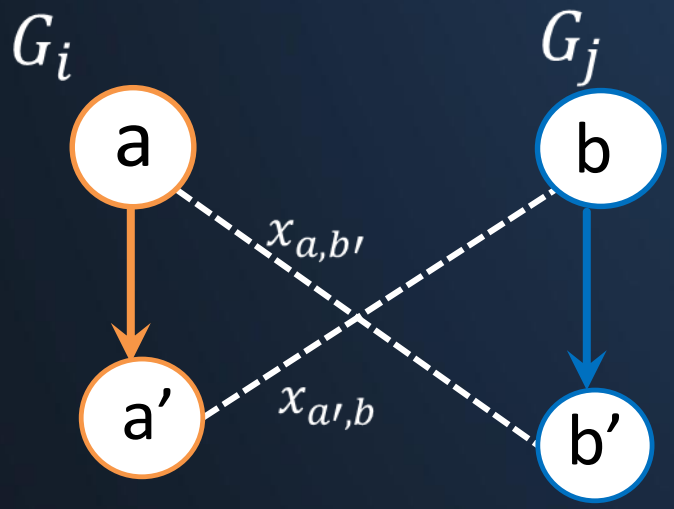
- Contraintes de structures





3. Formulation de LP4HM

- Contrainte de cohérence

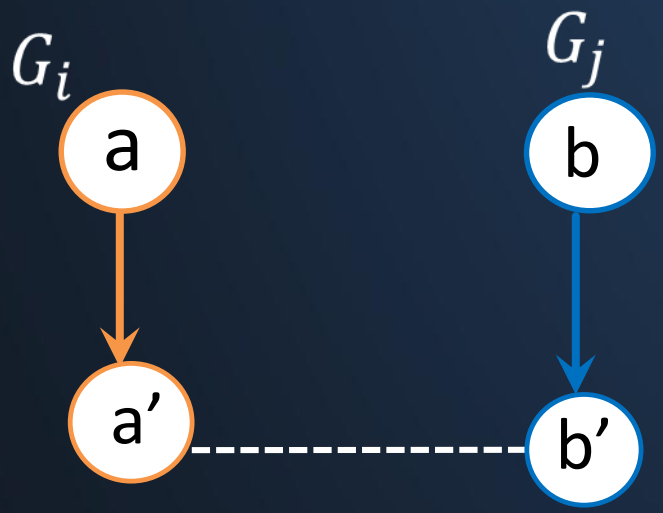


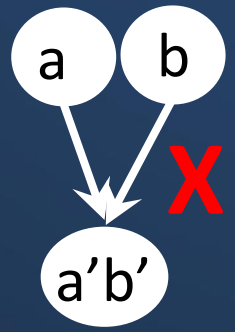
Si $x_{a,b'} = 1$ et $x_{a',b} = 1$ alors $dir_{a,a'} \,.\, dir_{b',b} = 1$

$$x_{a,b'} + x_{a',b} - dir_{a,a'} \,.\, dir_{b',b} <= 1$$

3. Formulation de LP4HM

- Contrainte de hiérarchies strictes





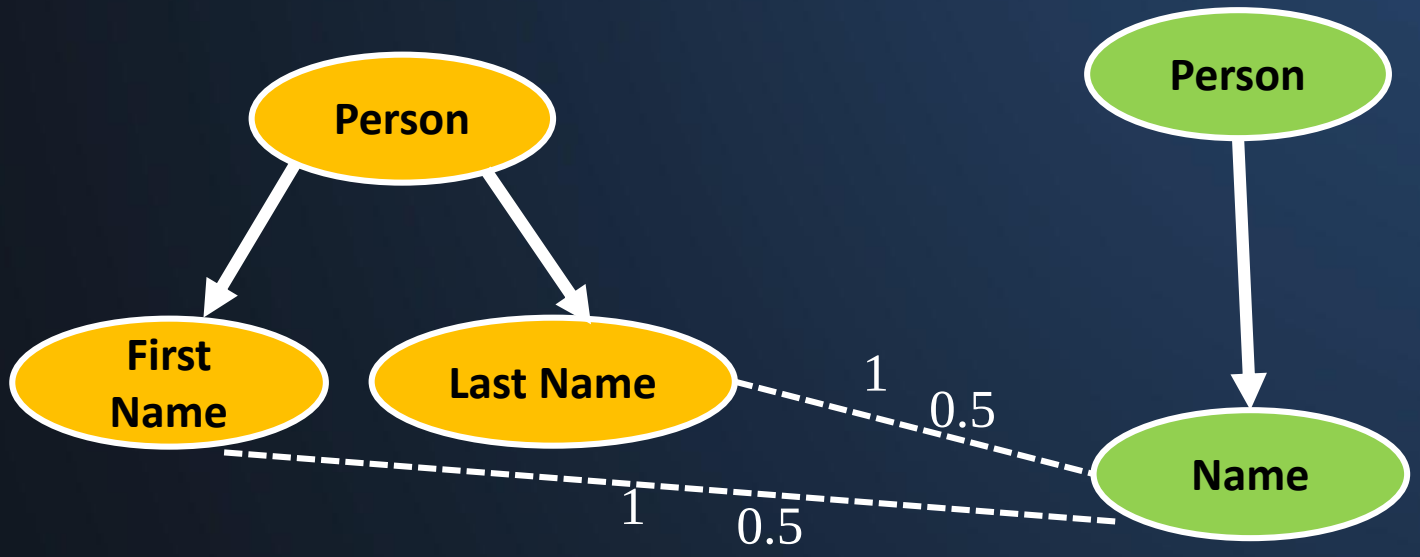
Si $x_{a\prime,b\prime} = 1$ alors $x_{pred(a'),pred(b')} = 1$

$$x_{a\prime,b\prime} <= x_{a,b}$$

3. LP4HM

$$
\begin{cases}
\max \quad \displaystyle\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} sim_{i_k,j_l}\ x_{i_k,j_l} \\
\\
s.t. \quad \displaystyle\sum_{l=1}^{n_j} x_{i_k,j_l} \leq 1, \ \forall k \in [1, n_i] \\
\qquad\qquad \forall i \in [1, N-1]\ \forall j \in [i+1, N] \\
\\
sim_{i_k,j_l}\ x_{i_k,j_l} \geq seuil\ x_{i_k,j_l} \\
\qquad\qquad \forall i \in [1, N-1]\ \forall j \in [i+1, N] \\
\qquad\qquad \forall k \in [1, n_i],\ \forall l \in [1, n_j] \\
\\
x_{i_k,j_l} \leq x_{i_{pred(k)},j_{pred(l)}} \\
\qquad\qquad \forall i \in [1, N-1]\ \forall j \in [i+1, N] \\
\qquad\qquad \forall k \in [1, n_i],\ \forall l \in [1, n_j] \\
\\
x_{i_k,j_l} + x_{i_{k'},j_{l'}} - (dir_{i_{k,k'}} dir_{j_{l,l'}}) \leq 1 \\
\qquad\qquad \forall i \in [1, N-1]\ \forall j \in [i+1, N] \\
\qquad\qquad \forall k, k' \in [1, n_i],\ \forall l, l' \in [1, n_j] \\
\\
x_{i_k,j_l} \in \{0, 1\} \ \forall i \in [1, N-1]\ \forall j \in [i+1, N] \\
\qquad\qquad \forall k \in [1, n_i],\ \forall l \in [1, n_j]
\end{cases}
$$

3. LP4HM relaxé



Relaxation des variables de décision binaires en variables fractionnaire dans l'intervalle [0,1].

4. Validation expérimentale

- Expérimentation avec le benchmark de Stanford

  Benchmark orienté utilisateurs pour évaluer le « pairwise matching » , 7 utilisateurs et 9 tâches.

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

5. Conclusion et perspectives

- Intégration holistique des open data
  - Un programme linéaire qui se focalise sur les structures hiérarchiques des graphes d'open data. Il génère une solution optimale globale.
    - Deux stratégies (non relaxé et relaxé) pour résoudre deux types de cardinalités (simple et complexe)
- Une qualité de matching compétitive par rapport à d'autres approches sans l'utilisation du seuil de similarité.
- Futurs travaux :
  - étudier l'extension de notre modèle sur des graphes labellisés tel que les ontologies

Références

- Aumueller, D., H.-H. Do, S. Massmann, et E. Rahm (2005). Schema and ontology matching with coma++. In Proceedings of the ACM International Conference on Management of Data, SIGMOD '05, pp. 906–908.
- Berro, A., Megdiche, I., Teste, O. (2014). A content-Driven ETL Processes for Open Data. In New Trends in Database and Information Systems II, Advances in Intelligent Systems and Computing. East-European Conference on Advances in Databases and Information Systems ADBIS'2014, pp. 19-40, Ohrid, Macedonia.
- Duchateau, F., Z. Bellahsene, et M. Roche (2007). Bmatch : a semantically context-based tool enhanced by an indexing structure to accelerate schema matching. In 23èmes Journées Bases de Données Avancées, BDA' 2007.
- Melnik, S., H. Garcia-Molina, et E. Rahm (2002). Similarity flooding : A versatile graph matching algorithm and its application to schema matching. In Proceedings of the 18th International Conference on Data Engineering, ICDE' 2002.
- PLASMA Benharkat, A., R. Rifaieh, S. Sellami, M. Boukhebouze, et Y. Amghar (2007). PLASMA : A Platform for schema matching and management. IBIS 5, 9–20.
- Saleem, K., Z. Bellahsene, et E. Hunt (2007). Performance oriented schema matching. In 18th International Conference Database and Expert Systems Applications, DEXA' 2007, pp. 844–853.
- Su, W., J. Wang, et F. Lochovsky (2006). Holistic schema matching for web query interfaces. In 10th International Conference on Extending Database Technology, EDBT' 2006, pp. 77–94.

Merci pour votre attention

Questions ?