

Integrating and exploring heterogeneous datasets

Nelly Barret

Postdoctoral researcher
Data Science group, DEIB, Politecnico di Milano

April 19, 2024



Outline

- 1 Motivation: data integration and exploration problems
- 2 Predihood: predicting neighbourhoods' environment
- 3 GeoAlign: spatial entity matching for Points of Interest
- 4 Abstra: first-sight overview of a dataset
- 5 Pathways: efficiently finding interesting paths
- 6 Systems developed
- 7 Conclusion

Outline

- 1 Motivation: data integration and exploration problems
- 2 Predihood: predicting neighbourhoods' environment
- 3 GeoAlign: spatial entity matching for Points of Interest
- 4 Abstra: first-sight overview of a dataset
- 5 Pathways: efficiently finding interesting paths
- 6 Systems developed
- 7 Conclusion

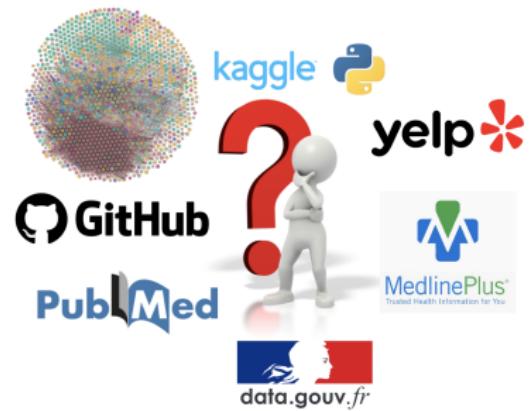
Data exploration and integration

Structured data models:

- **Relational** databases
- **Tables**

Semi-structured data models:

- **XML** documents
- **JSON** documents
- **RDF** graphs
- **Property** graphs



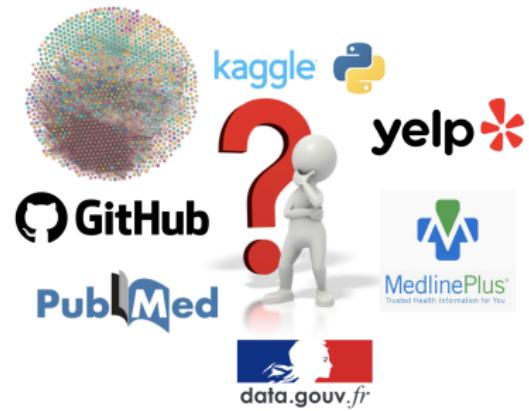
Data exploration and integration

Structured data models:

- **Relational** databases
- **Tables**

Semi-structured data models:

- **XML** documents
- **JSON** documents
- **RDF** graphs
- **Property** graphs



Dataset exploration and integration is hard: large, complex, irregular
Today's menu: focus on cartographic and semi-structured data

Outline

- 1 Motivation: data integration and exploration problems
- 2 Predihood: predicting neighbourhoods' environment
- 3 GeoAlign: spatial entity matching for Points of Interest
- 4 Abstra: first-sight overview of a dataset
- 5 Pathways: efficiently finding interesting paths
- 6 Systems developed
- 7 Conclusion

Motivation: heterogeneous data is everywhere

Name: Jane Doe

Job: French investigative journalist

Sex: F

Birth city: Paris

Residence city: Lyon



Wishes:

Learn Lyon neighbourhoods [BDF⁺21]

Aggregate city-level data

Visit Lyon's monuments [BDFM19]

Explore new datasets for her investigations [BMU24]

Reveal undeclared conflicts of interests [BGLM23a]

Skills:

Excel: ★★★★

Word: ★★★★

Rel. databases: ★

Semi-struct. data: N/A

Neighbourhood environment prediction

INSEE (French National Institute of Statistics)

- **IRIS**: small geo unit of 5K inhabitants (50K IRIS in FR)
- For each IRIS: 600 quantitative features
 - No high-level description of neighbourhoods' characteristics
 - Too many features for prediction



IRIS	Libellé de l'IRIS	Population en 2014 (princ)	Pop 0-2 ans en 2014 (princ)	Pop 3-5 ans en 2014 (princ)	Pop 6-10 ans en 2014 (princ)	Pop 11-17 ans en 2014 (princ)
IRIS	LIBIRIS	P14_POP	P14_POP0002	P14_POP0305	P14_POP0601	P14_POP1110
692640601	Belleruche	3736	301	211	302	445
692650000	Ville-sur-Jaméricourt (commune non inférée)	834	39	35	70	91
69266101	Charmelles	3567	168	103	181	177
69266102	Charles-Hemu	4908	218	169	220	337
69266103	Charpenne-Wilson	5916	174	195	245	352
69266201	Bois	2559	3	0	0	27
69266202	Onze-Novembre	2987	107	50	67	78
69266301	Tonkin-Sud	4356	242	199	261	274
69266302	Espace-Central	3181	188	126	175	191
69266401	Stalingrad	9	0	0	0	0
69266402	Tonkin-Ouest	2254	107	95	174	210
69266403	Tonkin-Nord	2309	102	83	93	151
69266501	Croix-Luzier-Ouest	3524	32	27	39	117
69266502	Croix-Luzier-Est	2382	78	38	44	116

Neighbourhood environment prediction

INSEE (French National Institute of Statistics)

- **IRIS**: small geo unit of 5K inhabitants (50K IRIS in FR)
- For each IRIS: 600 quantitative features
 - No high-level description of neighbourhoods' characteristics
 - Too many features for prediction



IRIS	Libellé de l'IRIS	Population en 2014 (princ)	Pop 0-2 ans en 2014 (princ)	Pop 3-5 ans en 2014 (princ)	Pop 6-10 ans en 2014 (princ)	Pop 11-17 ans en 2014 (princ)
IRIS	LIBIRIS	P14_POP	P14_POP0002	P14_POP0305	P14_POP0610	P14_POP1117
69264001	Belleruche	3736	301	211	302	445
69265001	Ville-sur-Jaméricourt (commune non inférée)	831	29	35	70	91
69266101	Charmelles	3567	168	103	181	177
69266102	Charles-Hemu	4908	218	169	220	337
69266103	Charpenne-Wilson	5916	174	195	245	352
69266201	Bois	2559	3	0	0	27
69266202	Onze-Novembre	2987	107	50	67	78
69266301	Tonkin-Sud	4356	242	199	261	274
69266302	Espace-Central	3181	188	126	175	191
69266401	SaintGrad	9	0	0	0	0
69266402	Tonkin-Ouest	2254	107	95	174	210
69266403	Tonkin-Nord	2309	102	83	93	151
69266501	Croix-Luizet-Ouest	3524	32	27	39	117
69266502	Croix-Luizet-Est	2382	78	38	44	116

Research contribution

Predict automatically the environment of any French neighbourhood, based on cartographic and city-level data

From raw features to environmental variables

Six environmental variables, defined with sociologists

- From hundreds of raw quantitative features, e.g., number of parks
- To few qualitative environmental variables, e.g., the landscape

Building type	Usage	Landscape	Social class	Morphology	Geography
Social housing	Housing	Urban	Lower	Central	Centre
Mixed	Shopping	Green areas	Low middle	Urban	North
Towers	Other	Forest	Middle	Peri-urban	North East
Subdivisions		Countryside	Up middle	Rural	East
Houses			Upper		...

Predict automatically any neighbourhood environment

- **Filter** the 600 features into lists of 30 features for each env. variable:
 - Remove descriptive, too precise, very correlated, useless features
- **Predict** the 6 environmental variables with the features lists
- With 7 **supervised** algorithms (manual annotation)



Predihood at work



A tool for visualizing IRIS

203 iris found for query lyon.

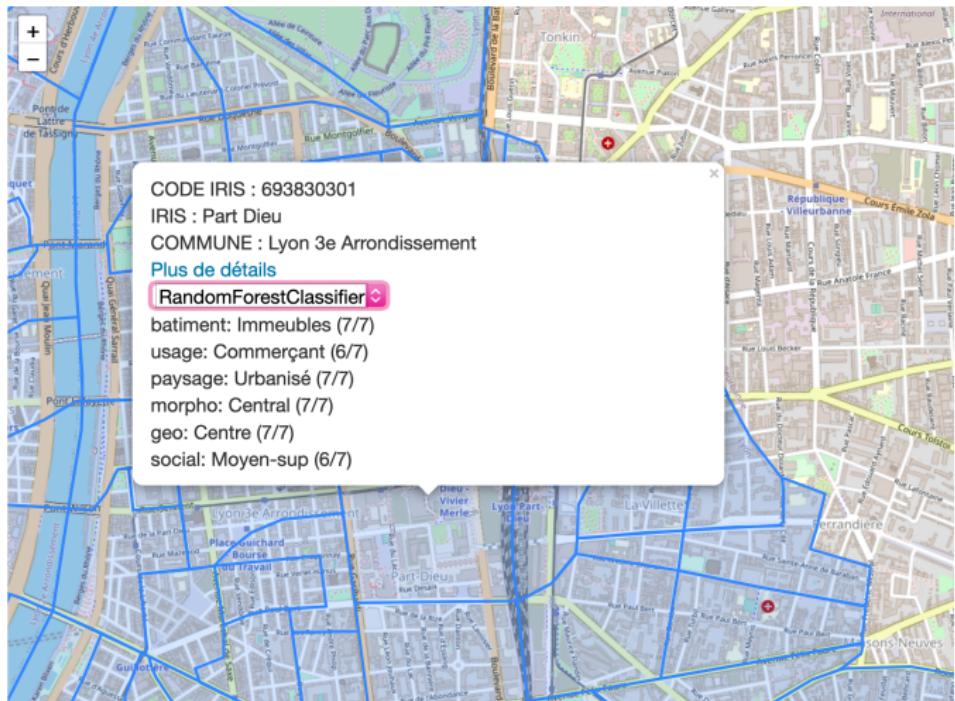
Minimal zoom level to display IRIS automatically

12 (actual zoom level = 15)

Search by IRIS code

Search by IRIS name or city

Clear



Outline

- 1 Motivation: data integration and exploration problems
- 2 Predihood: predicting neighbourhoods' environment
- 3 GeoAlign: spatial entity matching for Points of Interest
- 4 Abstra: first-sight overview of a dataset
- 5 Pathways: efficiently finding interesting paths
- 6 Systems developed
- 7 Conclusion

Motivation: heterogeneous data is everywhere

Name: Jane Doe

Job: French investigative journalist

Sex: F

Birth city: Paris

Residence city: Lyon



Wishes:

Learn Lyon neighbourhoods [BDF⁺21]

Visit Lyon's monuments [BDFM19]

Unify POIs across data providers

Explore new datasets for her investigations [BMU24]

Reveal undeclared conflicts of interests [BGLM23a]

Skills:

Excel: ★★★★

Word: ★★★★

Rel. databases: ★

Semi-struct. data: N/A

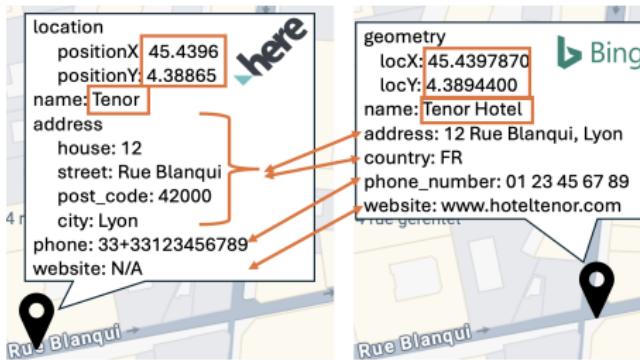
From cartographic entities to POIS

Cartographic data providers: Geonames, Bing, Here, OSM

→ No coordination between them

Point of Interest (POI): Duomo di Milano, restaurants, shops, ...

- Represented by one or several geographic entities (many providers)
- A set of attributes, with values (inconsistencies)



Research contribution

Find entities matching a unique real-world POI, with an adaptive formula

Adaptive formula for geographic entity matching

Given two entities e_1, e_2 , the **adaptive formula** relies on:

- The **similar degree** of e_1 and e_2 attributes
 - 13 measures: geo, text, type, ...
- The **weight/importance** of e_1 and e_2 attributes

$$f(e_1, e_2) = \sum_{i=1}^n \text{weight}_i * \text{sim}_i(\text{attribute}_i) > \theta$$

weight	sim. measure	attribute	
0.5	levenshtein	name	
0.4	distance	coordinates	
0.1	levenshtein	address	

Global threshold: 0.3

GeoAlign at work

GeoAlign

Search
Matching
Merging

Options
Help

Matching options

Global threshold: 0.2

Estimation of the quality

Quality of the correspondences

Threshold	TP	FP
0.1	4.5	7.0
0.2	4.0	6.5
0.3	4.5	6.0
0.4	4.0	6.5
0.5	4.5	6.0
0.6	4.0	6.5
0.7	4.5	5.5
0.8	4.0	5.0
0.9	4.5	5.0
1.0	4.0	5.0

Troyes

Name: Troyes
Coordinates: (48.298 ; 4.074)
Provider:
Type: places
Address: Rue Gabriel Grolez, Quartier de la Cité, Troyes, Aube, Grand Est, Metropolitan France, 10000, France
Phone: not specified
Website: not specified

Troyes, Aube, Grand Est

levenstein(name) = 1.000000
geobenchdistance(coordinates) = 0.000003

Outline

- 1 Motivation: data integration and exploration problems
- 2 Predihood: predicting neighbourhoods' environment
- 3 GeoAlign: spatial entity matching for Points of Interest
- 4 Abstra: first-sight overview of a dataset
- 5 Pathways: efficiently finding interesting paths
- 6 Systems developed
- 7 Conclusion

Motivation: heterogeneous data is everywhere

Name: Jane Doe

Job: French investigative journalist

Sex: F

Birth city: Paris

Residence city: Lyon



Wishes:

Learn Lyon neighbourhoods [BDF⁺21]

Visit Lyon's monuments [BDFM19]

Explore new datasets for her investigations [BMU24]

Reveal undeclared conflicts of interests [BGLM23a]

Skills:

Excel: ★★★★

Word: ★★★★

Rel. databases: ★

Semi-struct. data: N/A

Simple
descriptions

What does the dataset describe?



- Real-world objects and relationships between them

What does the dataset describe?



- Real-world objects and relationships between them
- Entity-Relationship models [RG03]

What does the dataset describe?



- Real-world objects and relationships between them
- Entity-Relationship models [RG03]
- Need to compute them from the dataset!

What does the dataset describe?



```
<person id="person1">
  <name>Alice</name>
  <address>
    <street>2, Second Street</street>
    <province>Georgia</province>
    <country>USA</country>
  </address>
  <mailbox>
    <mail from="person1@test.fr" to="person2@test.fr">
      <parlist>
        <listitem><text>Task 1</text></listitem>
        <listitem>
          <parlist>
            <listitem><text>Sub task 1</text></listitem>
            <listitem><text>Sub task 2</text></listitem>
            <listitem><text>Sub task 3</text></listitem>
          </parlist>
        </listitem>
      </parlist>
    </mail>
  </mailbox>
</person>
```

- Real-world objects and relationships between them
- Entity-Relationship models [RG03]
- Need to compute them from the dataset!
- What about semi-structured data models (nesting)?

What does the dataset describe?



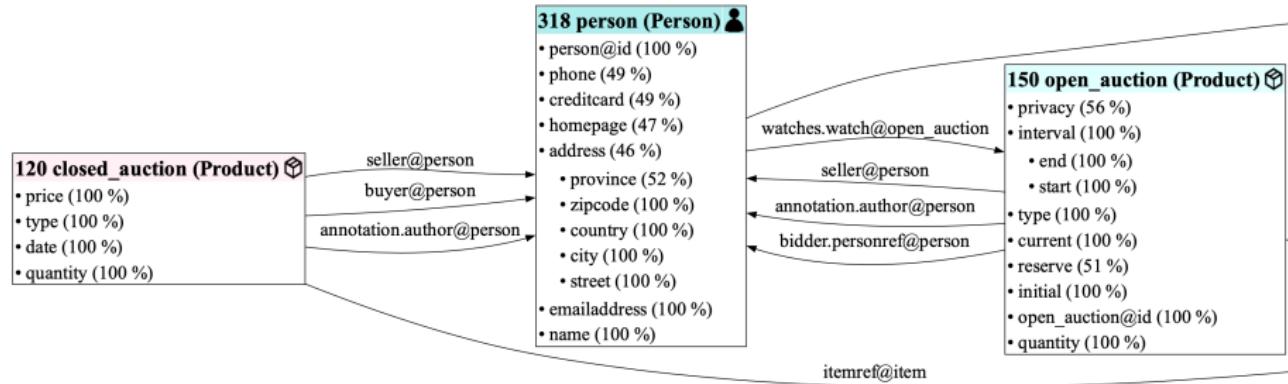
```
<person id="person1">
  <name>Alice</name>
  <address>
    <street>2, Second Street</street>
    <province>Georgia</province>
    <country>USA</country>
  </address>
  <mailbox>
    <mail from="person1@test.fr" to="person2@test.fr">
      <parlist>
        <listitem><text>Task 1</text></listitem>
        <listitem>
          <parlist>
            <listitem><text>Sub task 1</text></listitem>
            <listitem><text>Sub task 2</text></listitem>
            <listitem><text>Sub task 3</text></listitem>
          </parlist>
        </listitem>
      </parlist>
    </mail>
  </mailbox>
</person>
```

- Real-world objects and relationships between them
- Entity-Relationship models [RG03]
- Need to compute them from the dataset!
- What about semi-structured data models (nesting)?
- Keep it simple and of controllable size

Research contribution: data abstraction

Abstra: Lightweight Entity-Relationship diagrams [BMU22, BMU24]

- Automatically and efficiently from semi-structured data
- Compact yet meaningful data overviews
- Ideal for first-sight dataset discovery



The Abstra approach

- ① Integrate all data sources in a graph (ConnectionLens) [ABC⁺22]
- ② Summarize the graph
- ③ Among summary nodes, identify entities and their attributes
- ④ In the summary, identify relationships between the entities
- ⑤ Propose a simple category to each entity (best-effort)

Background: from heterogeneous data to data graphs

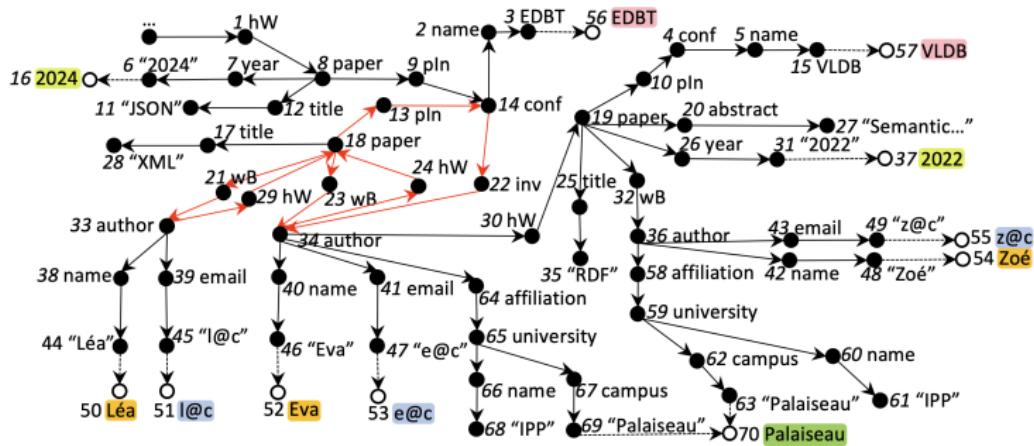
ConnectionLens [ABC⁺22]:

- ① Ingests any dataset into a **directed graph**
 - Generic, flexible, fine granularity

Background: from heterogeneous data to data graphs

ConnectionLens [ABC⁺22]:

- ① Ingests any dataset into a **directed graph**
 - Generic, flexible, fine granularity
- ② Extracts **Named Entities** (NEs) from all text nodes
 - date , email address , People , Place , Organization , ...



Data graph summarization

We need a **compact representation of large data graphs**

Data graph summarization

We need a **compact representation of large data graphs**

Challenges:

- Heterogeneous graphs originating from different data models
- Node and/or edge labels may be empty

Data graph summarization

We need a **compact representation of large data graphs**

Challenges:

- Heterogeneous graphs originating from different data models
- Node and/or edge labels may be empty

We aim for a **quotient graph summary**:

- Based on **equivalence** between nodes of the original graph
- We prefer **small summaries** (number of nodes)

Quotient summarization across data models

Each data model has its own syntax:

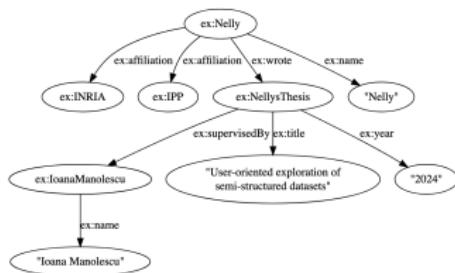
XML

```
<root>
  <student id="s1" thesisref="t1">
    <name>Nelly</name>
    <affiliation>Inria</affiliation>
    <affiliation>IPP</affiliation>
  </student>
  <researcher id="r1">
    <name>Ioana Manolescu</name>
  </researcher>
  <thesis id="t1" year="2024">
    <title>User-oriented exploration of
      semi-structured datasets</title>
    <supervisor supref="r1">
    </supervisor>
  </thesis>
</root>
```

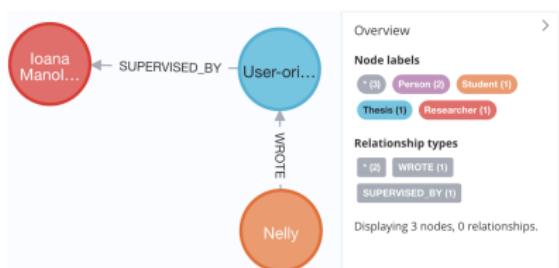
JSON

```
{
  "student": {
    "name": "Nelly",
    "affiliation": ["Inria", "IPP"],
    "thesis": {
      "year": "2024",
      "title": "User-oriented exploration of
        semi-structured datasets",
      "supervisor": {
        "name": "Ioana Manolescu"
      }
    }
  }
}
```

RDF



PG



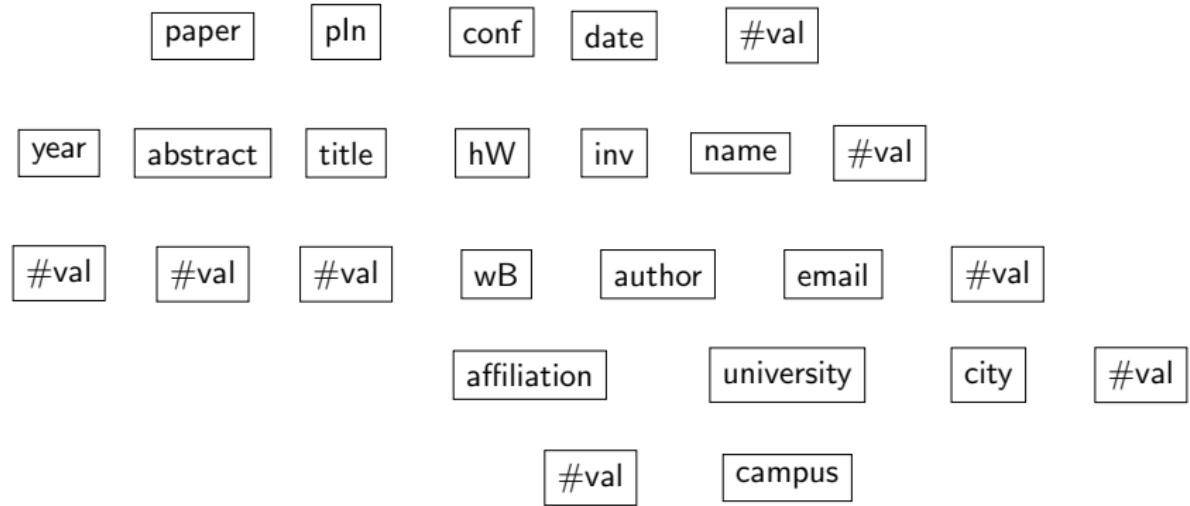
Summarization based on same-kind nodes

We identify **node kinds** in each model based on the respective best practices for data design:

- XML: elements with the same **label** (or type)
- JSON: nodes on the same **path from the root**
- RDF [GGM20]: depending on **node type(s)** or, if absent, **incoming and outgoing properties**
- PG: adaptation of the above [GGM20]

The summary (collection graph) \mathcal{G}

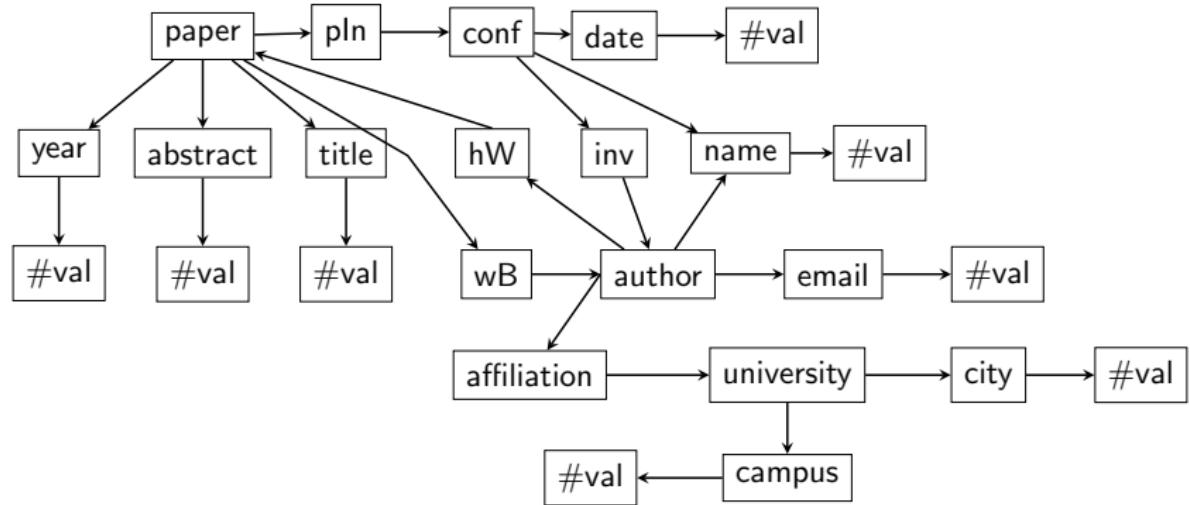
Collection node for each equivalence class



The summary (collection graph) \mathcal{G}

Collection node for each equivalence class

Collection edge $C_s \rightarrow C_t$ if a data edge exists

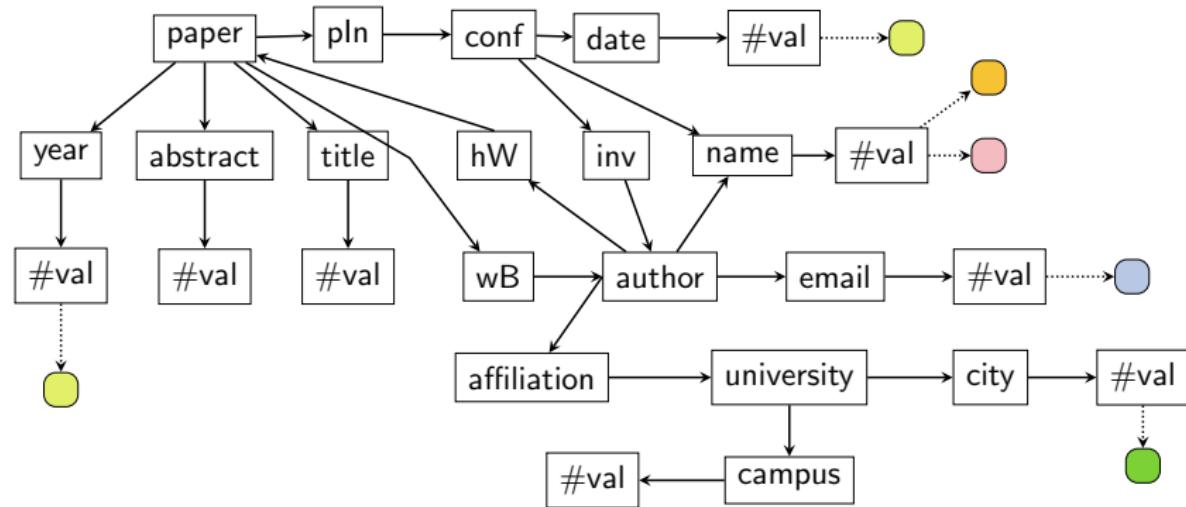


The summary (collection graph) \mathcal{G}

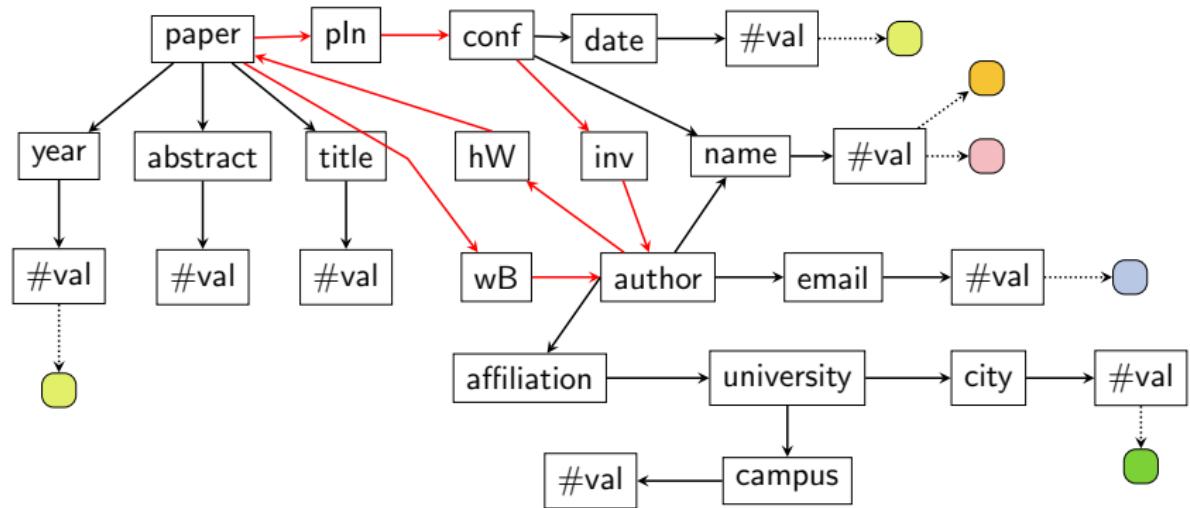
Collection node for each equivalence class

Collection edge $C_s \rightarrow C_t$ if a data edge exists

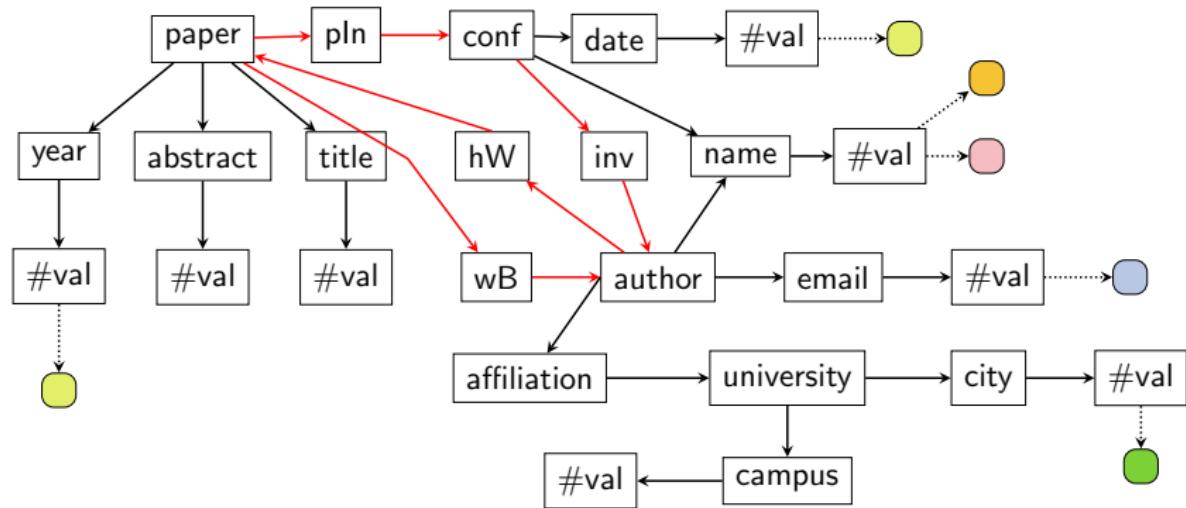
Entity profile for each **leaf collection node**: reflects NEs in the leaves



Identifying entities in the collection graph \mathcal{G}

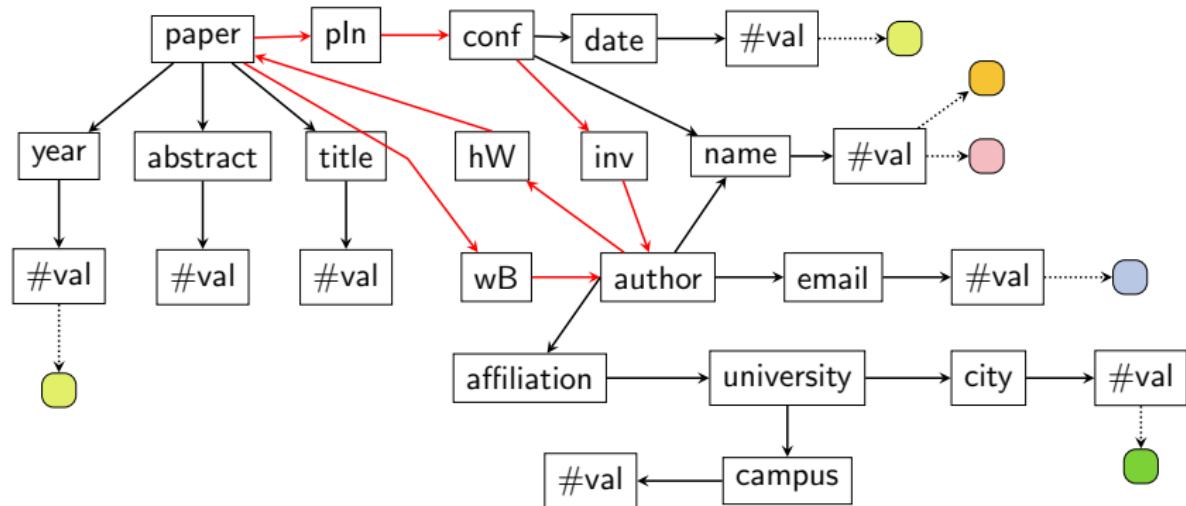


Identifying entities in the collection graph \mathcal{G}



Which collections represent **entities** in the E-R diagram?

Identifying entities in the collection graph \mathcal{G}



Which collections represent **entities** in the E-R diagram?

Which collections represent **entity attributes**?

Requirements and algorithm

- We need an algorithm to identify entity roots and attributes for the E-R diagram
 - For complex, potentially cyclic, collection graphs

Requirements and algorithm

- We need an algorithm to identify entity roots and attributes for the E-R diagram
 - For complex, potentially cyclic, collection graphs

Greedy selection of few entities in \mathcal{G}

- ① Assign a **score** to each collection node
- ② While less than E_{max} entity roots, or data coverage $< cov_{min}$
 - ① Elect the next highest-scored eligible collection node as an entity root
 - ② Compute its **boundary**, i.e., attribute set
 - ③ **Update** the collection graph to reflect the selection of an entity
 - ④ Recompute the scores

How to score a collection node?

Reflect the **weight** of this node and its structure in the dataset

- ① w_{desc_k}, w_{leaf_k} : # descendants, leaf descendants, at depth k

How to score a collection node?

Reflect the **weight** of this node and its structure in the dataset

- ➊ w_{desc_k}, w_{leaf_k} : # descendants, leaf descendants, at depth k
- ✖ Not clear how to pick k

How to score a collection node?

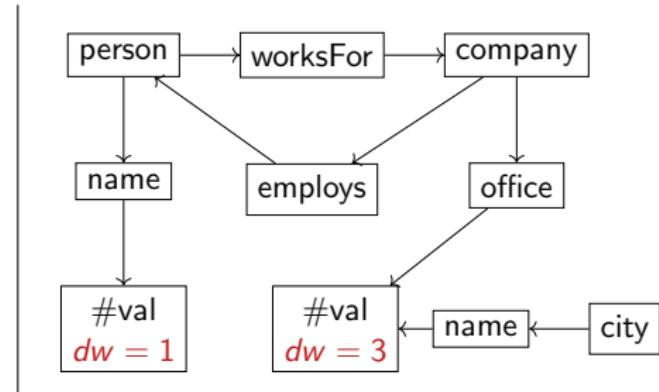
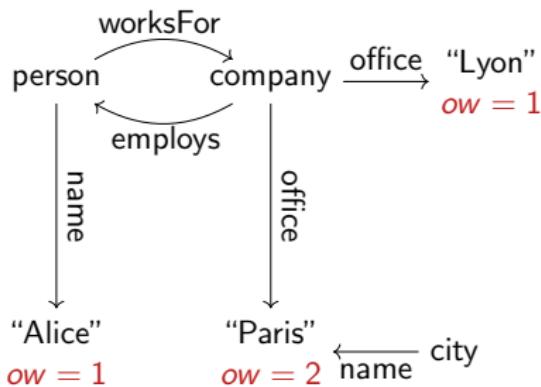
Reflect the **weight** of this node and its structure in the dataset

- ① w_{desc_k} , w_{leaf_k} : # descendants, leaf descendants, at depth k
- ② Directed Acyclic Graph (DAG) rooted in each node: w_{DAG}

Data weight

Own weight ow of a leaf node: its in-degree

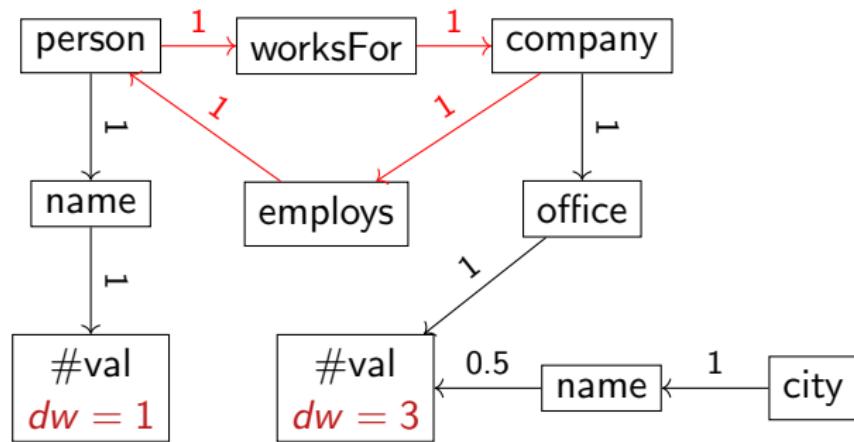
Data weight dw of a leaf collection node: the sum of its nodes' ow



Data weight DAG propagation

Leaf collection dw is propagated back to all ancestors which are not in a cycle

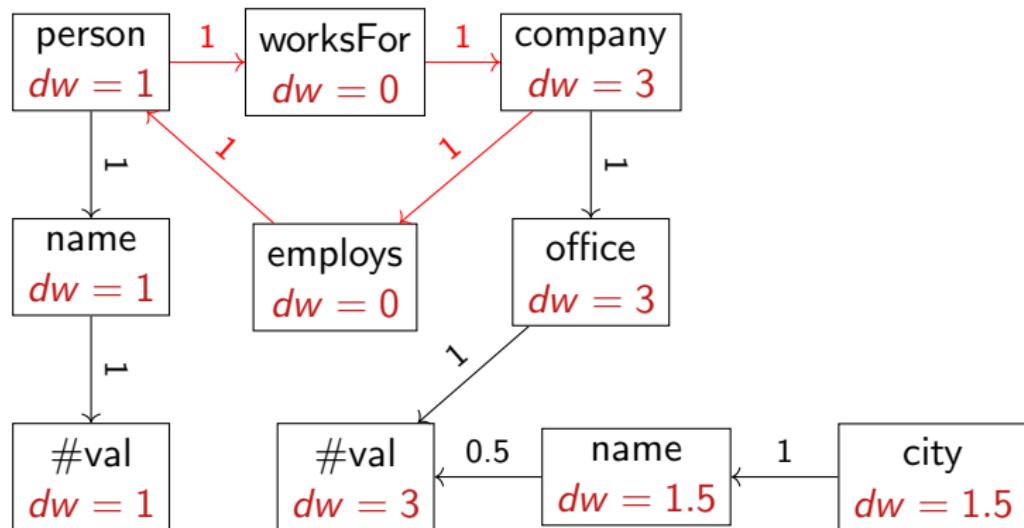
- **Edge transfer factor:** $\frac{|\text{nodes in } C_t \text{ having a parent in } C_s|}{|C_t|}$



Data weight DAG propagation

Leaf collection dw is propagated back to all ancestors which are not in a cycle

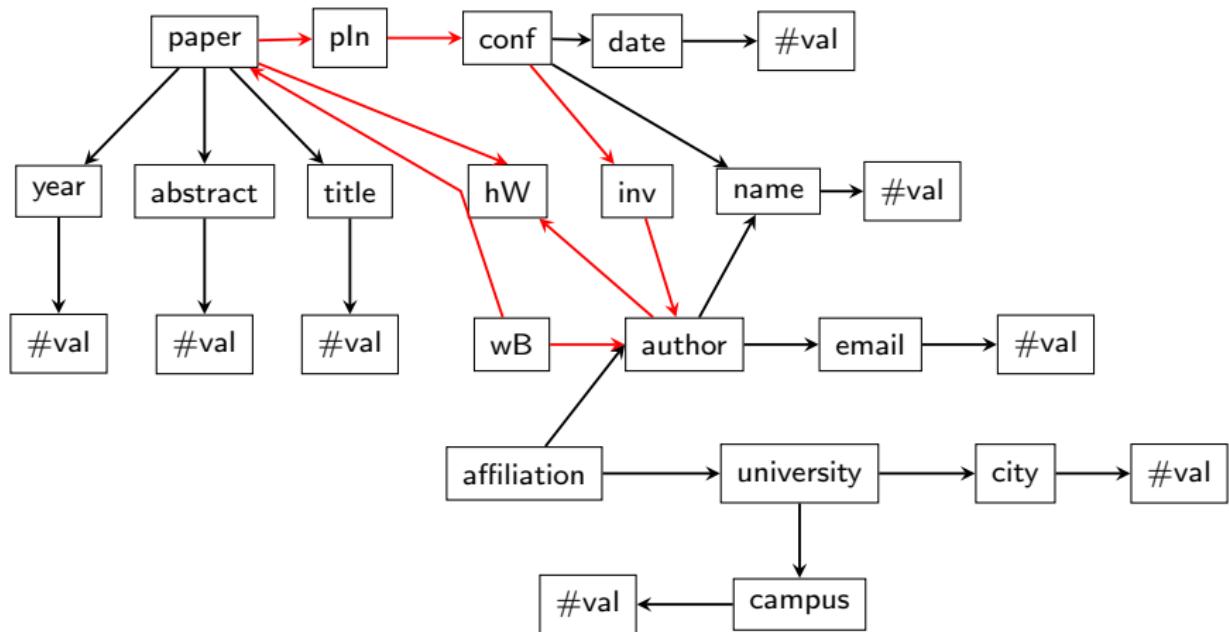
- **Edge transfer factor:**
$$\frac{\text{nodes in } C_t \text{ having a parent in } C_s}{|C_t|}$$



How to score a collection node?

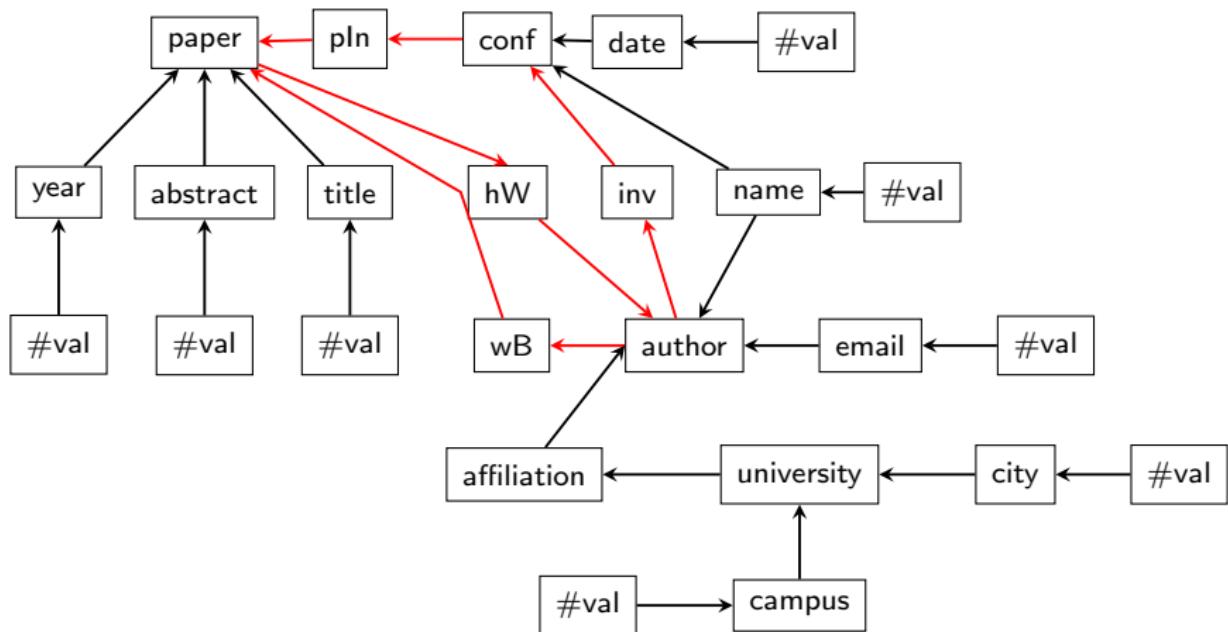
- ① w_{desc_k}, w_{leaf_k} : # descendants, leaf descendants, at depth k
- ② Directed Acyclic Graph (DAG) rooted in each node: w_{DAG}
- ③ $w_{PageRank}$: PageRank algorithm on \mathcal{G}

PageRank score of a collection graph node



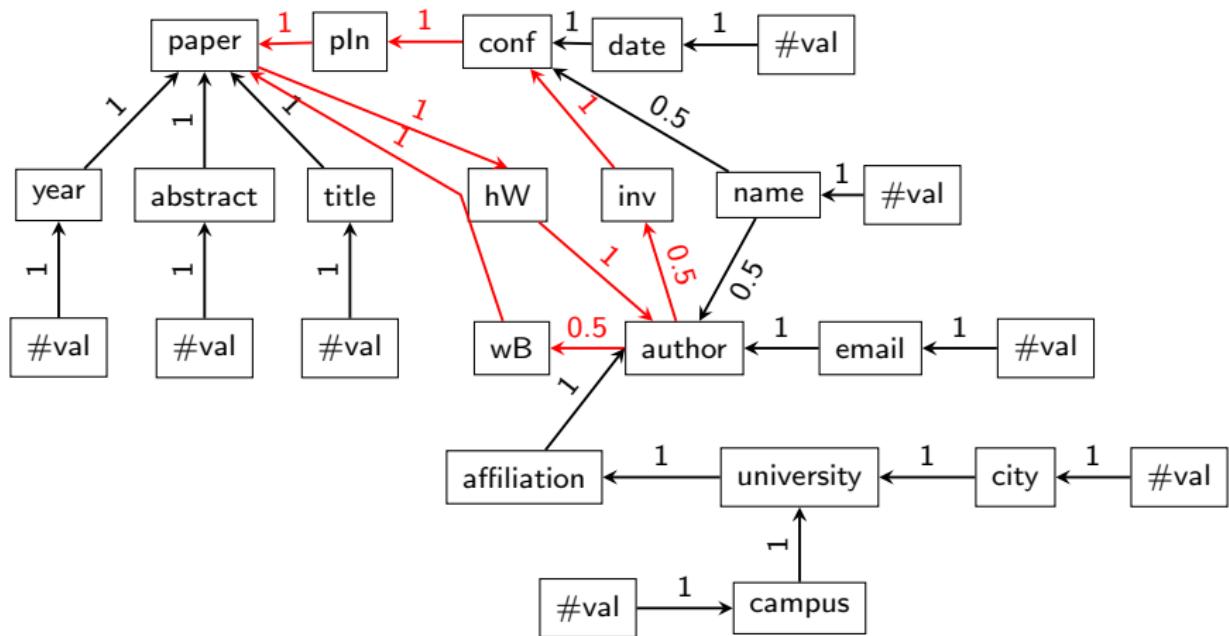
The collection graph \mathcal{G}

PageRank score of a collection graph node



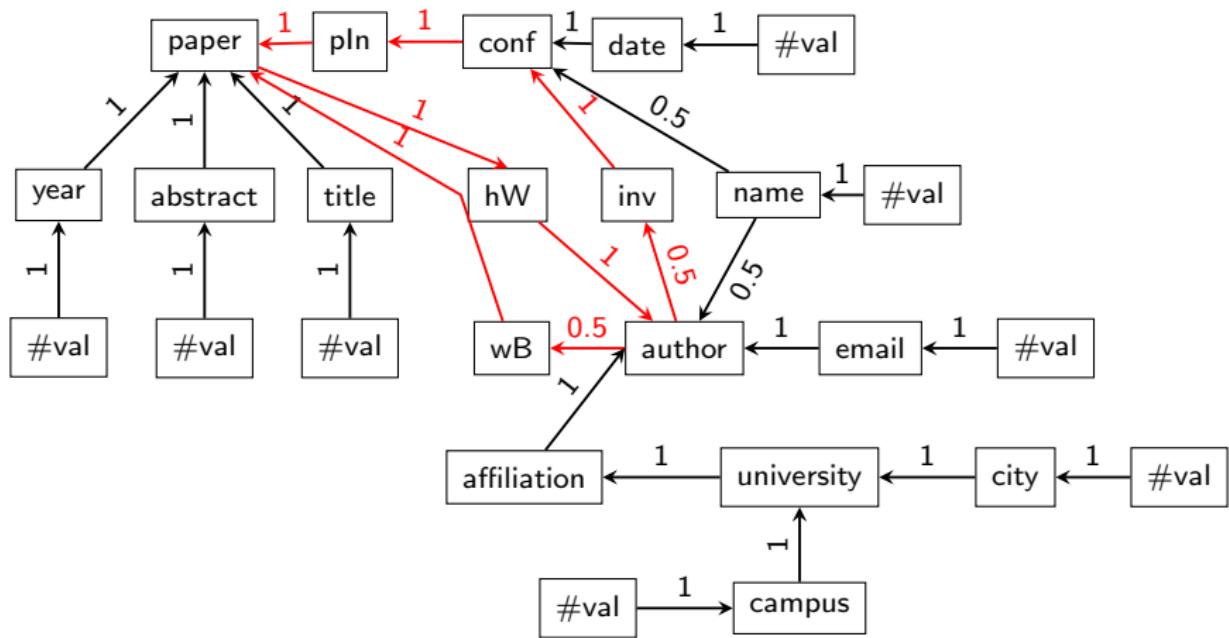
The reverse collection graph \mathcal{G}_R

PageRank score of a collection graph node



The reverse collection graph \mathcal{G}_R with PR edge weights

PageRank score of a collection graph node



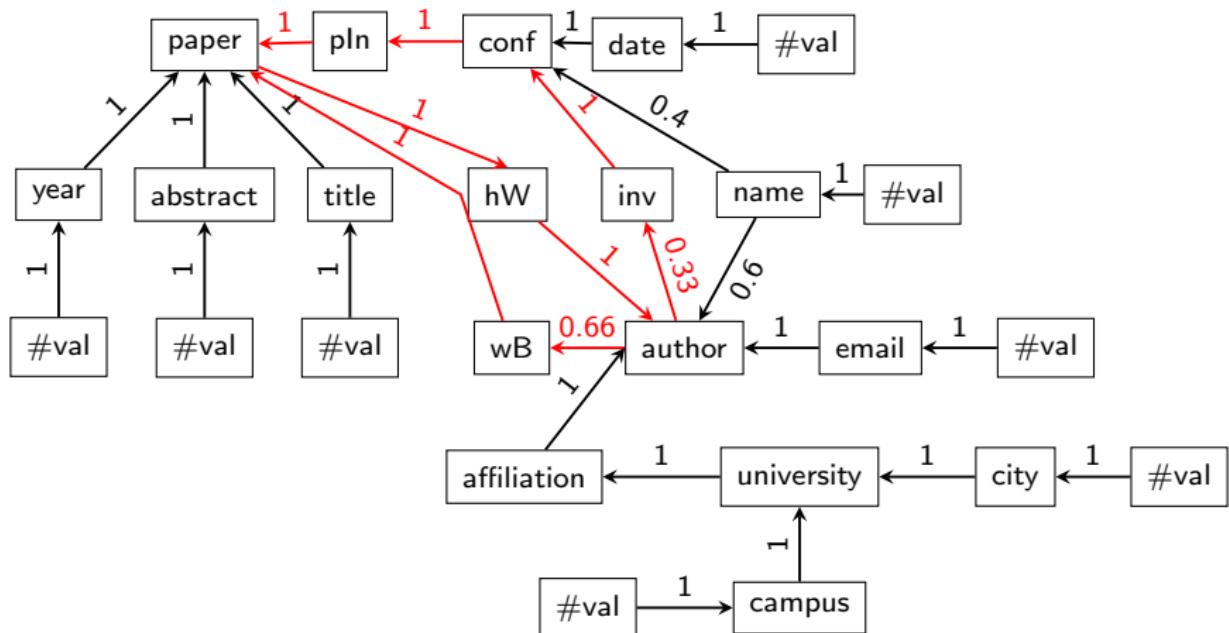
The reverse collection graph \mathcal{G}_R with PR edge weights

Collections distribute their score based solely on their connectivity

How to score a collection node?

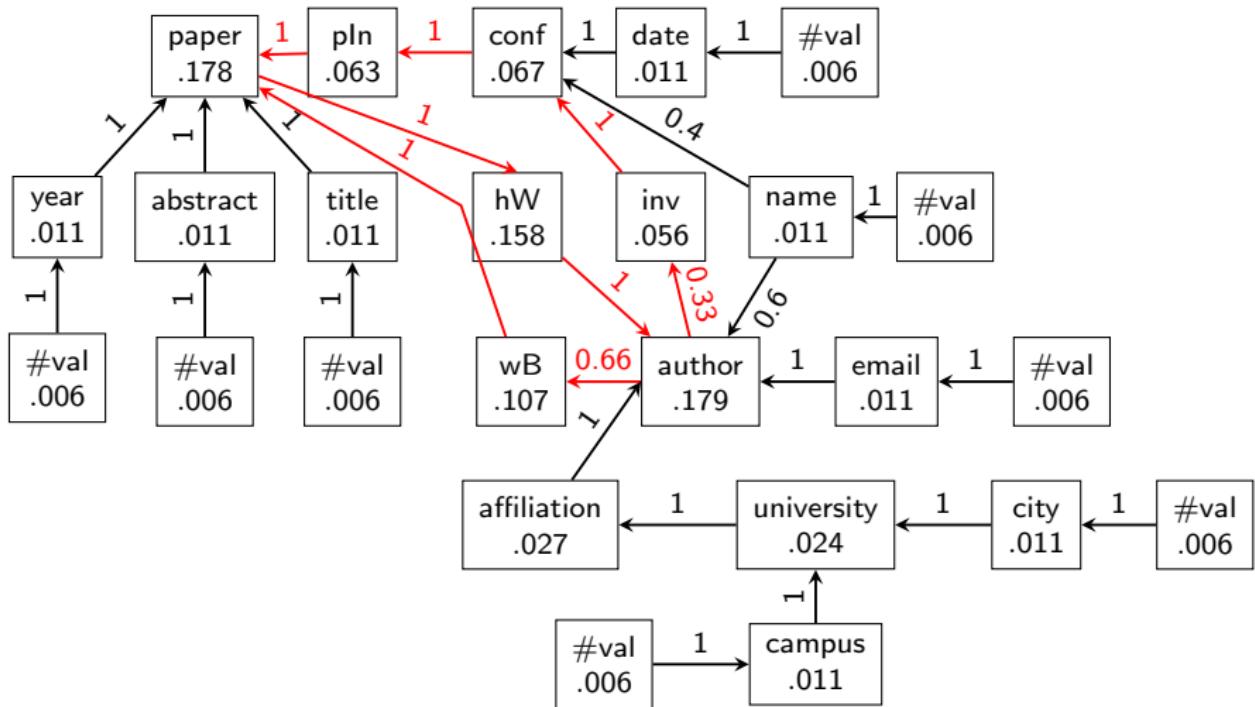
- ➊ w_{desc_k}, w_{leaf_k} : # descendants, leaf descendants, at depth k
 - ➋ w_{DAG} : dw bottom-up propagation on \mathcal{G} (outside cycles)
 - ➌ $w_{PageRank}$: PageRank algorithm on \mathcal{G}
 - ➍ $w_{dwPageRank}$: PageRank algorithm on \mathcal{G} with dw -tuned PR edge weights
-  Reflects both the topology and where actual data is

The data-weighted PageRank score

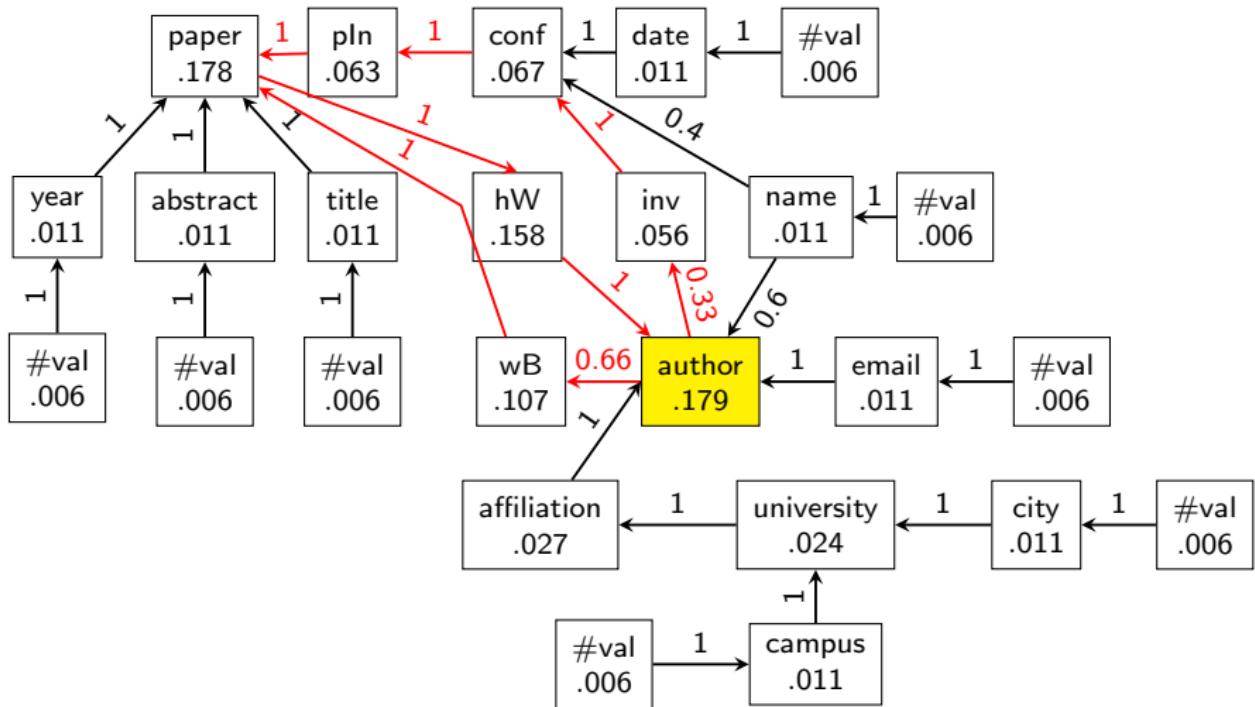


The reverse collection graph \mathcal{G}_R with *dw*-tuned PR edge weights

The data-weighted PageRank score

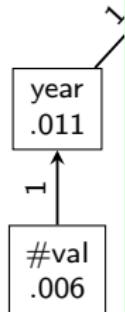


The data-weighted PageRank score



The data-weighted PageRank score

[paper | 1 | pln | 1 | conf | 1 | date | 1 | #val]

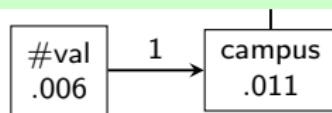


Propagates scores across the collection graph

Works on cyclic collection graphs

The score reflects the topology and where the data is

A collection node distributes its weight



#val
006

How to compute an entity boundary?

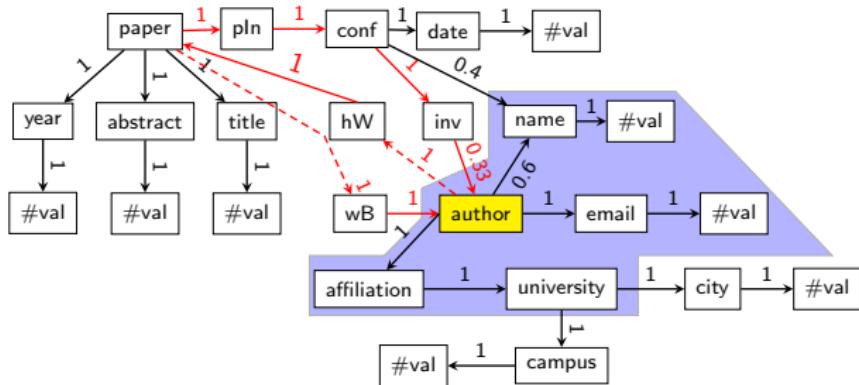
Collections in \mathcal{G} representing attributes of this entity

How to compute an entity boundary?

Collections in \mathcal{G} representing attributes of this entity

“Those that contribute to the entity’s weight”

- The boundary may go far (for deep-structure entities)
- Easy to define for w_{desc_k} , w_{leaf_k} , w_{DAG} . Example for w_{desc_2}

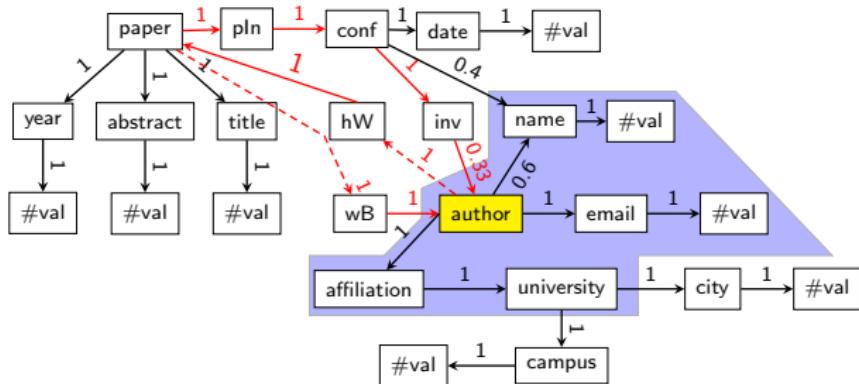


How to compute an entity boundary?

Collections in \mathcal{G} representing attributes of this entity

“Those that contribute to the entity’s weight”

- The boundary may go far (for deep-structure entities)
- Easy to define for w_{desc_k} , w_{leaf_k} , w_{DAG} . Example for w_{desc_2}



Does not apply for PageRank-based scores

Data-acyclic flooding boundary $bound_{dfI-ac}$

Idea: the collection nodes

- **Reachable** from the entity root
- **Mainly** part of **this entity**
- The path between the entity root and this collection's nodes is **not data cyclic**

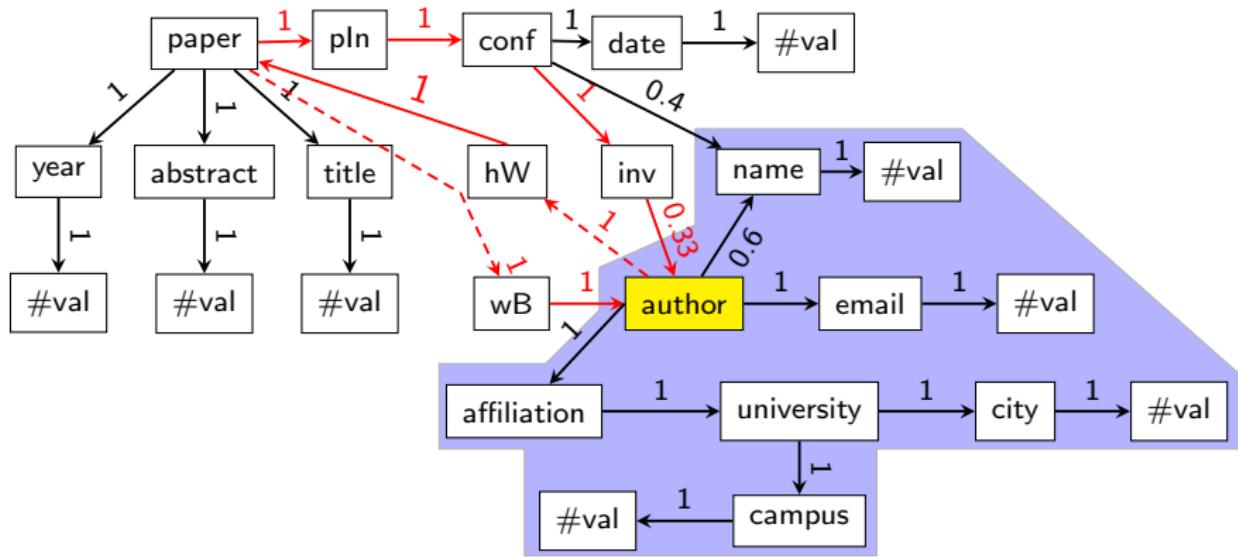
Data-acyclic flooding boundary $bound_{dfI-ac}$

Idea: the collection nodes

- **Reachable** from the entity root
- **Mainly** part of **this entity**
 - Edge transfer factor $\geq f_{min}$
 - At-most-one: each C_s node has at most one child in C_t
- The path between the entity root and this collection's nodes is **not data cyclic**
 - If the path in the collection graph has no in-cycle edges
 - Or, the collection graph path has in-cycle edges, but they are not in the data

Data-acyclic flooding boundary $bound_{dfI-ac}$

- **Reachable** from the entity root
- **Mainly** part of **this entity**
- The path is **not data cyclic**



How to update the collection graph after selecting an entity?

Reflect the allocation of data nodes and edges to one entity

How to update the collection graph after selecting an entity?

Reflect the allocation of data nodes and edges to one entity

① *updateboolean*

- Collection nodes and edges in the boundary of the entity
 - Very efficient
 - Sufficient for w_{desc_k} , w_{leaf_k} , $WDAG$

How to update the collection graph after selecting an entity?

Reflect the allocation of data nodes and edges to one entity

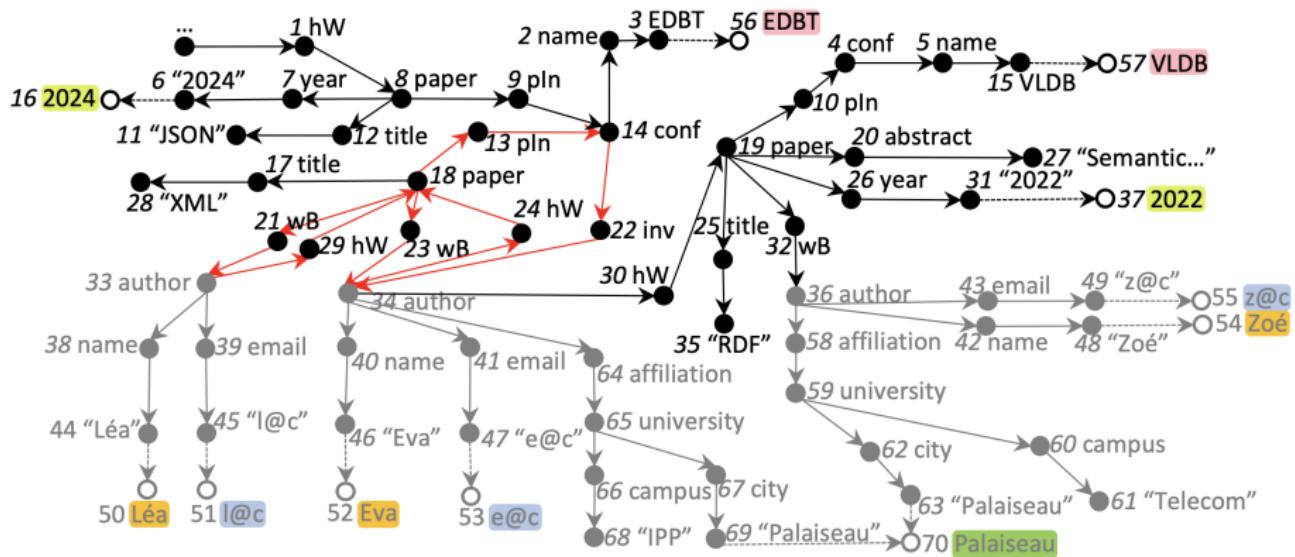
① $update_{boolean}$

- Collection nodes and edges in the boundary of the entity
 - Very efficient
 - Sufficient for w_{desc_k} , w_{leaf_k} , $WDAG$

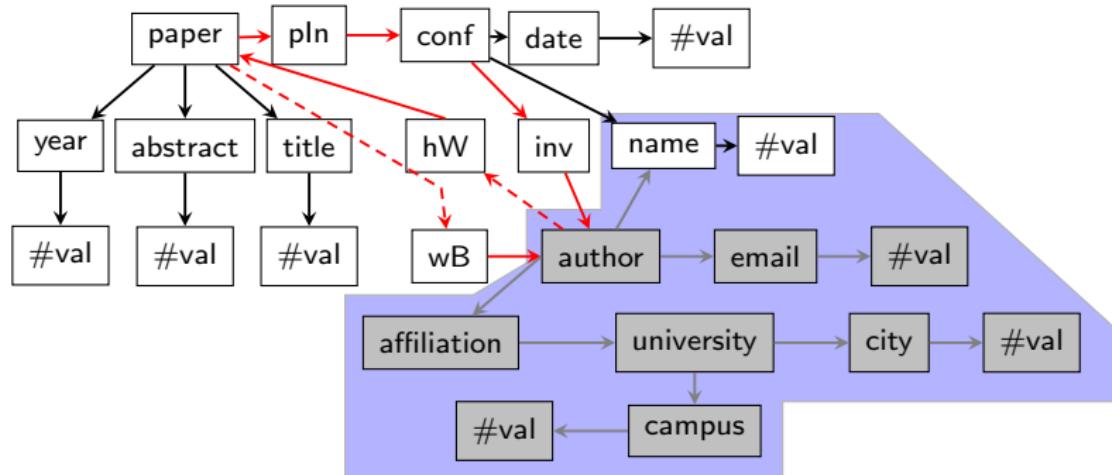
② $update_{exact}$

- Graph nodes and edges
 - Much more costly
 - Required for $W_{PageRank}$, $W_{dwPageRank}$

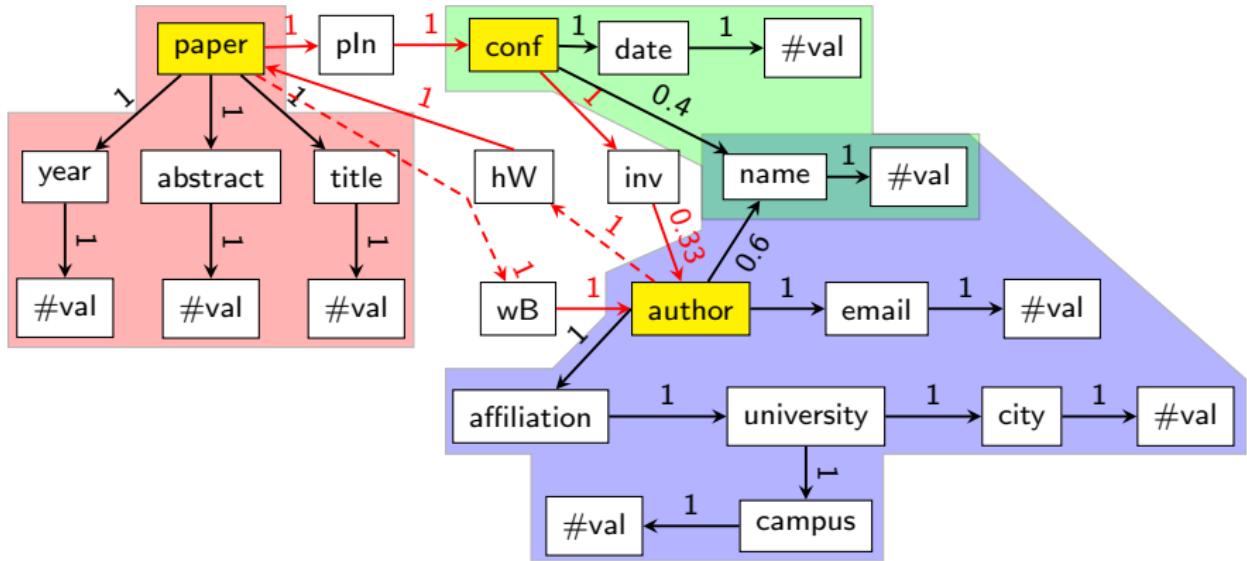
Exact graph update



Exact graph update

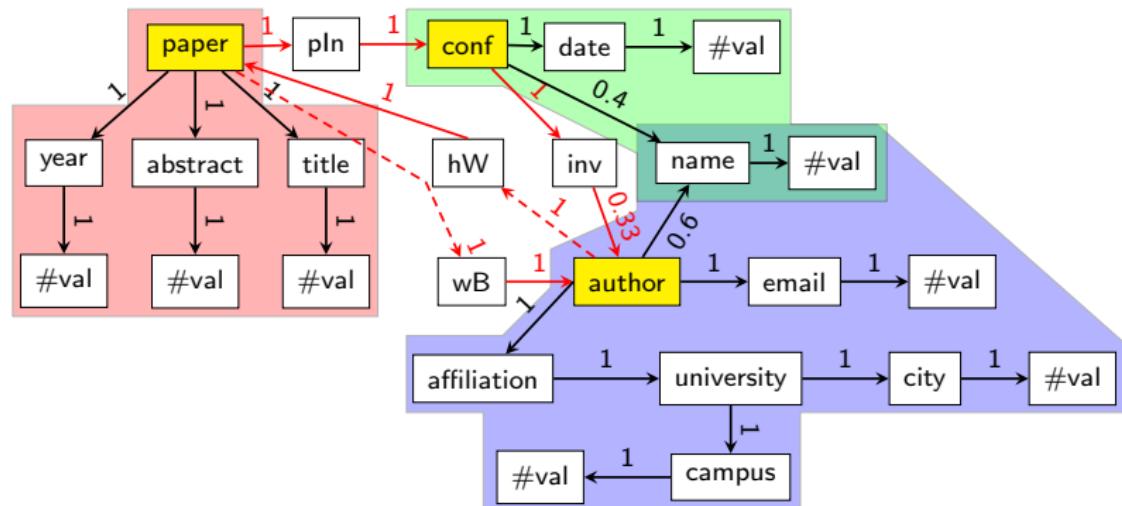


Selected entities and their boundaries



Finding relationships between entities

Relationship: a path from an entity to another



- paper → wB → author
- paper → pln → conf
- author → hW → paper
- conf → inv → author

Entity classification

Assign a semantic category to each entity

Input: an entity E , categories \mathcal{K} , semantic properties \mathcal{P}

- \mathcal{K} : Person, ScientificPaper, Event, Website, Mountain, ...
- \mathcal{P} : {label:"address", domain:[Pers., Org.], range:[Place]}, ...

Output: a category for E

Entity classification

Assign a semantic category to each entity

Input: an entity E , categories \mathcal{K} , semantic properties \mathcal{P}

- \mathcal{K} : Person, ScientificPaper, Event, Website, Mountain, ...
- \mathcal{P} : {label:"address", domain:[Pers., Org.], range:[Place]}, ...

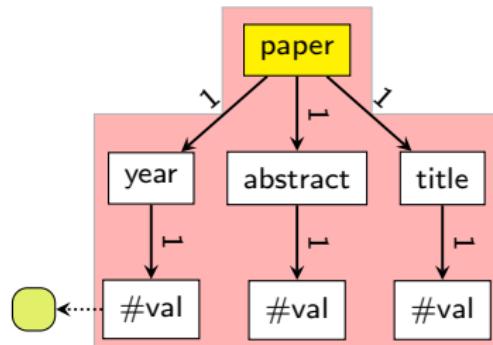
Output: a category for E

Algorithm:

- Compare:
 - The common name of all nodes in the entity root (if it exists) with $k \in \mathcal{K}$ (*conf, paper, author*)
 - Its attribute names with $p \in \mathcal{P}$ (*affiliation, email, ...*)
 - Its entity profiles with $p.\text{range} \in \mathcal{P}$ (, , , ...)
- Each good match votes for one or few categories

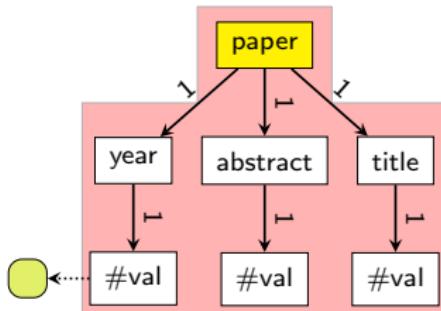
Entity classification

Name	Similar to	Votes for
paper	ResearchPublication (0.85) News (0.63)	ResearchPublication News



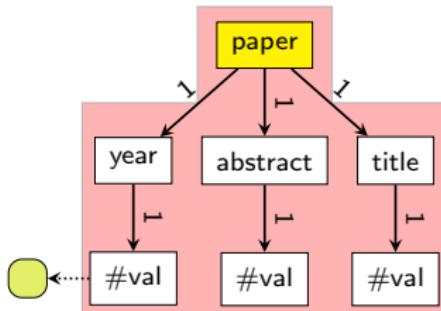
Entity classification

Attribute	Similar to	Votes for
abstract	abstract (1.0) summary (0.92) preface (0.47)	ResearchPublication Book
title	title (1.0) honorific title (0.87)	ResearchPublication Movie Person
year	year publication (0.85 + █)	Event Book ResearchPublication, ...



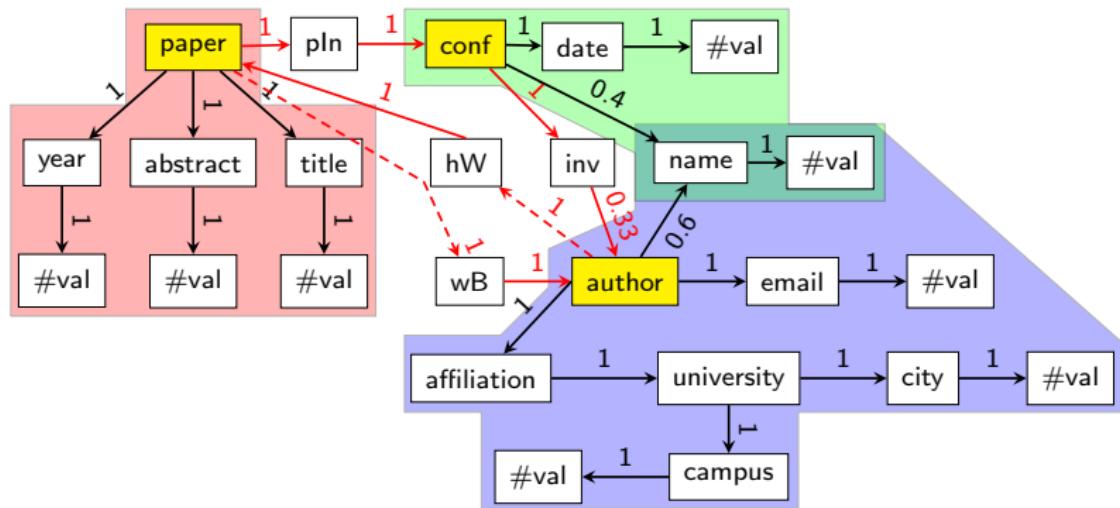
Entity classification

Attribute	Similar to	Votes for
abstract	abstract (1.0) summary (0.92) preface (0.47)	ResearchPublication Book
title	title (1.0) honorific title (0.87)	ResearchPublication Movie Person
year	year publication (0.85 + █)	Event Book ResearchPublication, ...

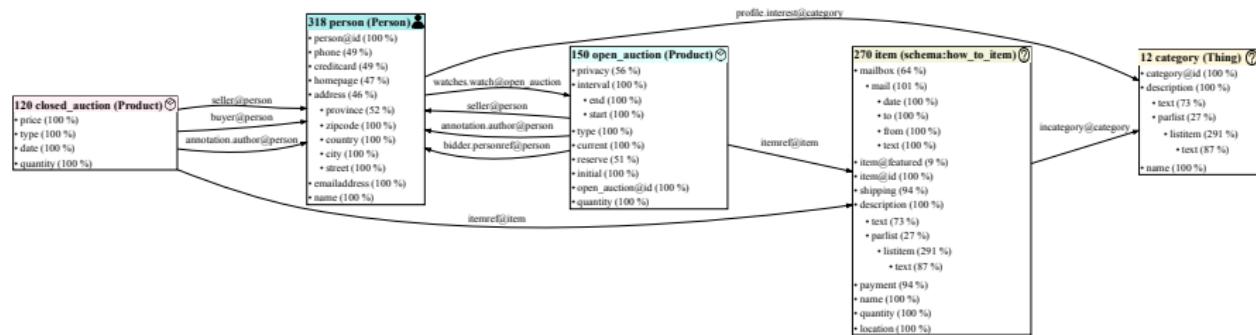


Entity classification

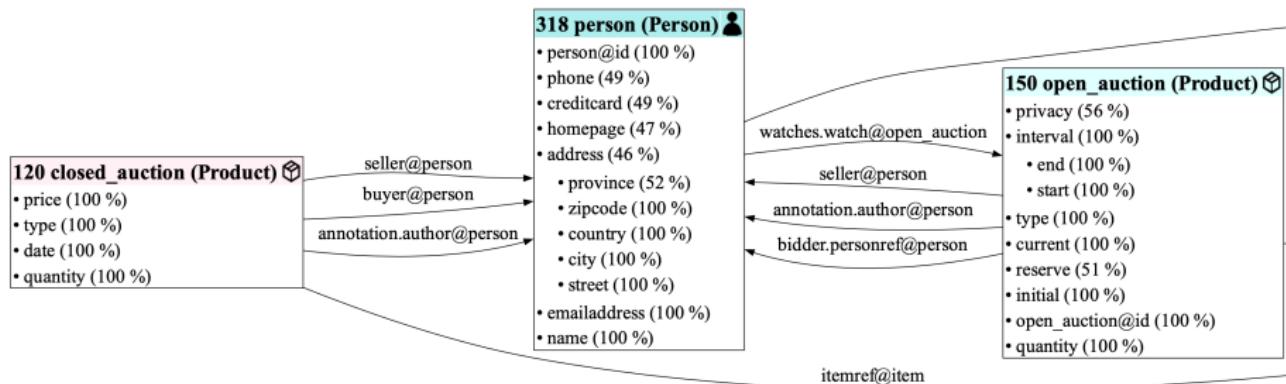
- paper nodes classified as **ResearchPublication**
- author nodes classified as **Researcher**
- conference nodes classified as **Event**



Abstra output: a lightweight Entity-Relationship diagram



Abstra output: a lightweight Entity-Relationship diagram



Experimental evaluation

On main **semi-structured** data models: 8 JSON, 7 RDF, 5 XML, 3 PG

- 10 synthetic, 13 real-world
- 5M to 14M nodes
- Collection graphs:
 - 26 to 4.8K collections
 - 14/23 have cycles

Experimental evaluation

On main **semi-structured** data models: 8 JSON, 7 RDF, 5 XML, 3 PG

- 10 synthetic, 13 real-world
- 5M to 14M nodes
- Collection graphs:
 - 26 to 4.8K collections
 - 14/23 have cycles

Graphs stored in PostgreSQL, algorithms in SQL and Java

Experimental evaluation

On main **semi-structured** data models: 8 JSON, 7 RDF, 5 XML, 3 PG

- 10 synthetic, 13 real-world
- 5M to 14M nodes
- Collection graphs:
 - 26 to 4.8K collections
 - 14/23 have cycles

Graphs stored in PostgreSQL, algorithms in SQL and Java

We evaluate:

- ① Entity selection quality
- ② Scalability

Entity selection quality with ($w_{dwPageRank}$, $bound_{fl-ac}$)

Dataset name	$ C $	$ \mathcal{ME} $	$ \mathcal{MR} $	cov	\mathcal{ME}	d_{max}	$ \mathcal{ME}_i $
Mondial 	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Entity selection quality with ($w_{dwPageRank}$, $bound_{fl-ac}$)

Dataset name	$ C $	$ \mathcal{ME} $	$ \mathcal{MR} $	cov	\mathcal{ME}	d_{max}	$ \mathcal{ME}_i $
Mondial 	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Entity selection quality with ($w_{dwPageRank}$, $bound_{fl-ac}$)

Dataset name	$ C $	$ \mathcal{ME} $	$ \mathcal{MR} $	cov	\mathcal{ME}	d_{max}	$ \mathcal{ME}_i $
Mondial 	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Entity selection quality with ($w_{dwPageRank}$, $bound_{fl-ac}$)

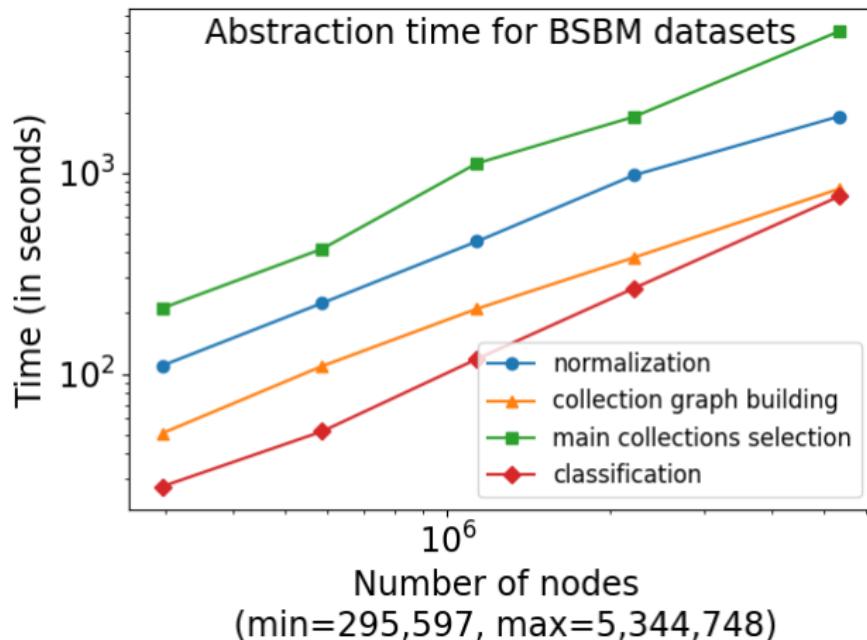
Dataset name	$ C $	$ \mathcal{ME} $	$ \mathcal{MR} $	cov	\mathcal{ME}	d_{max}	$ \mathcal{ME}_i $
Mondial ⚙	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 ⚙	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 ⚙	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Entity selection quality with ($w_{dwPageRank}$, $bound_{fl-ac}$)

Dataset name	$ C $	$ \mathcal{ME} $	$ \mathcal{MR} $	cov	\mathcal{ME}	d_{max}	$ \mathcal{ME}_i $
Mondial ⚙	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 ⚙	136	5	10	0.91	Person	4	25,500
					Item	7	21,750
					Open_Auction	8	12,000
					Closed_Auction	8	9,750
					Category	2	1,000
XMark4 ⚙	136	5	10	0.90	Person	4	102,000
					Item	7	87,000
					Open_Auction	8	48,000
					Closed_Auction	8	39,000
					Category	2	4,000
Wikimedia	59	2	0	1.0	Page	4	54,750
					Namespace	3	32

Abstra selects frequent, coherent and semantically central entities

Experimental evaluation: scalability



Our abstraction method scales up linearly in the data size

Related work

Data summarization

- Structural
 - Quotient [GGM20, KC10, MS99]
(the one we adopt to build \mathcal{G})
 - Non-quotient [GW97]
- Pattern mining [ZLVK16]
- Statistical [HS12]
- Hybrid [RGSB17]

Schema inference

- XML [CGS11]
- JSON [BCGS19]
- RDF [GLSW22]
- PG [LBH21]

- Data summarization and schema inference are tied to one data model
- Schemas are often not suited to NTUs

A JSON schema from social network data using [BCGS19]

```
▼ __Content:  
  ▼ _id:  
    ▼ __Content:  
      ▼ $oid:  
        __Kind: "StrType"  
        __Kind: "Record"  
  ▼ code:  
    __Kind: "NumType"  
  ▼ event:  
    ▼ __Content:  
      ▼ @:  
        ▼ __Content:  
          ▼ action:  
            __Kind: "StrType"  
          ▼ attachments:  
            ▼ __Content:  
              ▼ __Content:  
                ▼ @:  
                  ▼ __Content:  
                    ▼ audio:  
                      ▼ __Content:  
                        ▼ @:  
                          ▼ __Content:  
                            ▼ album_id:  
                              __Kind: "NumType"  
                            __Kind: "StrType"  
                            __Kind: "Record"  
                          ▼ artist:  
                            __Kind: "StrType"  
                            __Kind: "Record"  
                          ▼ content_restricted:  
                            __Kind: "NumType"  
                            __Kind: "Record"  
                          ▼ date:  
                            __Kind: "NumType"  
                            __Kind: "Record"  
                          ▼ duration:  
                            __Kind: "NumType"  
                            __Kind: "Record"  
                          ▼ genre_id:  
                            __Kind: "NumType"  
                            __Kind: "Record"  
                          ▼ id:  
                            __Kind: "NumType"  
                            __Kind: "Record"  
                          ▼ lyrics_id:  
                            __Kind: "NumType"  
                            __Kind: "Record"  
                          ▼ owner_id:  
                            __Kind: "NumType"  
                            __Kind: "Record"
```

Outline

- 1 Motivation: data integration and exploration problems
- 2 Predihood: predicting neighbourhoods' environment
- 3 GeoAlign: spatial entity matching for Points of Interest
- 4 Abstra: first-sight overview of a dataset
- 5 Pathways: efficiently finding interesting paths
- 6 Systems developed
- 7 Conclusion

Motivation: heterogeneous data is everywhere

Name: Jane Doe

Job: French investigative journalist

Sex: F

Birth city: Paris

Residence city: Lyon



Wishes:

Learn Lyon neighbourhoods [BDF⁺21]

Visit Lyon's monuments [BDFM19]

Explore new datasets for her investigations [BMU24]

Reveal undeclared conflicts of interests [BGLM23a]

Skills:

Excel: ★★★★

Word: ★★★★

Rel. databases: ★

Semi-struct. data: N/A

Entity-to-entity
paths

Research contribution

PathWays: interesting Named Entities connections

[BGLM23b, BGLM23a, BGLM24]

- Automatically and efficiently from semi-structured datasets
- Complete set of NE-to-NE interesting connections
- Ideal for exploring connections within and across datasets

COLUMNS	FILTERS	DENSITY	EXPORT	ETENDRE LE TEXTE
#val	agency	Spacecraft	description	#val
Algeria	http://purl.org/net/schemas/space/agency	http://data.kasabi.com/dataset/nasa/spacecraft/2002-054A	http://purl.org/dc/elements/1.1/description	Alsat
Argentina	http://purl.org/net/schemas/space/agency	http://data.kasabi.com/dataset/nasa/spacecraft/1997-002B	http://purl.org/dc/elements/1.1/description	Aerospatiale
Argentina	http://purl.org/net/schemas/space/agency	http://data.kasabi.com/dataset/nasa/spacecraft/1998-069B	http://purl.org/dc/elements/1.1/description	Argentinean National Commission of Space Activities
Australia	http://purl.org/net/schemas/space/agency	http://data.kasabi.com/dataset/nasa/spacecraft/1967-118A	http://purl.org/dc/elements/1.1/description	Sparta
Australia	http://purl.org/net/schemas/space/agency	http://data.kasabi.com/dataset/nasa/spacecraft/1967-118A	http://purl.org/dc/elements/1.1/description	Weapons Research Establishment

Rows per page: 10 ▾ 1–10 of 3903 < >

How are Named Entities connected?

Enumerate paths between (value) nodes in which NEs have been detected

- On the **data graph** (expensive)
- On the **collection graph** (much faster)
- Regardless of the edge direction

How are Named Entities connected?

Enumerate paths between (value) nodes in which NEs have been detected

- On the **data graph** (expensive)
- On the **collection graph** (much faster)
- Regardless of the edge direction

Each collection graph path, evaluated on the data graph, turns into a relation (set of data paths)

How are Named Entities connected?

Enumerate paths between (value) nodes in which NEs have been detected

- On the **data graph** (expensive)
- On the **collection graph** (much faster)
- Regardless of the edge direction

Each collection graph path, evaluated on the data graph, turns into a relation (set of data paths)

Challenges:

- Finding only **interesting** paths (to be seen)
- **Efficiently** evaluating the paths over the data graph: multi-query optimization [BGLM24]

What makes a NE-to-NE path interesting?

Some paths connecting Person NEs (■) to Organization NEs (■)

- ■ ← #val ← Name ← Author → Affiliation → #val → ■

What makes a NE-to-NE path interesting?

Some paths connecting Person NEs (■) to Organization NEs (■)

- ■ ← #val ← Name ← Author → Affiliation → #val → ■
- ■ ← #val ← Name ← Author ← Authors ← Article → Journal → #val → ■

What makes a NE-to-NE path interesting?

Some paths connecting Person NEs (■) to Organization NEs (■)

- ■ ← #val ← Name ← Author → Affiliation → #val → ■
- ■ ← #val ← Name ← Author ← Authors ← Article → Journal → #val → ■
- ■ ← #val ← COI ← Article → Journal → #val → ■ ← #val → ■

What makes a NE-to-NE path interesting?

Some paths connecting Person NEs (■) to Organization NEs (■)

- ■ ← #val ← Name ← Author → Affiliation → #val → ■
- ■ ← #val ← Name ← Author ← Authors ← Article → Journal → #val → ■
- ■ ← #val ← COI ← Article → Journal → #val → ■ ← #val → ■

Which paths are most interesting and deserve to be evaluated?

What makes a NE-to-NE path interesting?

Some paths are **unreliable**: we face entity extraction errors

- E.g., “John Hopkins University Hospital”
person
- False positives, or wrong entity type attribution, e.g., “THC”
org.

What makes a NE-to-NE path interesting?

Some paths are **unreliable**: we face entity extraction errors

- E.g., John Hopkins University Hospital
 - person
- False positives, or wrong entity type attribution, e.g., THC
 - org.

Some paths are **structurally weak**: we face information dilution

- E.g., a paper has 50 authors

What makes a NE-to-NE path interesting?

Some paths are **unreliable**: we face entity extraction errors

- E.g., John Hopkins University Hospital
 - person
- False positives, or wrong entity type attribution, e.g., THC
 - org.

Some paths are **structurally weak**: we face information dilution

- E.g., a paper has 50 authors

Path interestingness: based on **edge reliability** and **edge force**

What makes a NE-to-NE path interesting?

① Reliability $r(C_i \rightarrow \blacksquare)$ of an extraction collection edge

- The ratio of NEs having the type \blacksquare , and extracted from C_i
- Path reliability: minimum extraction edge reliability

What makes a NE-to-NE path interesting?

① Reliability $r(C_i \rightarrow \blacksquare)$ of an extraction collection edge

- The ratio of NEs having the type \blacksquare , and extracted from C_i
- Path reliability: minimum extraction edge reliability

② Force $f(C_i \rightarrow C_j)$ of a structural collection edge

- The inverse of the maximal source node out-degree among data edges represented by $C_i \rightarrow C_j$
- Path force: product of edge forces

What makes a NE-to-NE path interesting?

- ① **Reliability** $r(C_i \rightarrow \blacksquare)$ of an extraction collection edge
 - The ratio of NEs having the type \blacksquare , and extracted from C_i
 - Path reliability: minimum extraction edge reliability
- ② **Force** $f(C_i \rightarrow C_j)$ of a structural collection edge
 - The inverse of the maximal source node out-degree among data edges represented by $C_i \rightarrow C_j$
 - Path force: product of edge forces
- ③ Rank paths on their **reliability**, then their **force**

What makes a NE-to-NE path interesting?

- ① **Reliability** $r(C_i \rightarrow \blacksquare)$ of an extraction collection edge
 - The ratio of NEs having the type \blacksquare , and extracted from C_i
 - Path reliability: minimum extraction edge reliability
- ② **Force** $f(C_i \rightarrow C_j)$ of a structural collection edge
 - The inverse of the maximal source node out-degree among data edges represented by $C_i \rightarrow C_j$
 - Path force: product of edge forces
- ③ Rank paths on their **reliability**, then their **force**
- ④ Take a top- k or those having $r \geq \theta$

What makes a NE-to-NE path interesting?

Some paths connecting Person NEs (■) to Organization NEs (■)

- ```
graph LR; P1[Person] -- "#val 1.0" --> N1[Name]; N1 -- "#val 1.0" --> A1[Author]; A1 -- "#val 1.0" --> AF1[Affiliation]; AF1 -- "#val 0.91" --> O1[Organization]
```

  - Reliable; strong
- ```
graph LR; P2[Person] -- "#val 1.0" --> N2[Name]; N2 -- "#val 1.0" --> A2[Author]; A2 -- "0.02" --> A2_2[Authors]; A2_2 -- "1.0" --> AR2[Article]; AR2 -- "1.0" --> J2[Journal]; J2 -- "#val 1.0" --> O2[Organization]
```

 - Reliable; weak
- ```
graph LR; P3[Person] -- "#val 0.09" --> COI3[COI]; COI3 -- "1.0" --> AR3[Article]; AR3 -- "1.0" --> J3[Journal]; J3 -- "#val 0.05" --> P4[Person]; P4 -- "#val 0.09" --> COI4[COI]; COI4 -- "0.04" --> O3[Organization]
```

  - Not reliable; strong

# PathWays output: data paths as tables

Connect  to  Maximum depth of a path

Sort by

 #val agency Spacecraft description  #val (3903 paths)

 #val (175 paths)

 #val agency Spacecraft name  #val (133 paths)

 #val agency Spacecraft missionProfile  #val (71 paths)

# PathWays output: data paths as tables

| COLUMNS   | FILTERS                                  | DENSITY                                                  | EXPORT                                      | ETENDRE LE TEXTE                                    |
|-----------|------------------------------------------|----------------------------------------------------------|---------------------------------------------|-----------------------------------------------------|
| #val      | agency                                   | Spacecraft                                               | description                                 | #val                                                |
| Algeria   | http://purl.org/net/schemas/space/agency | http://data.kasabi.com/dataset/nasa/spacecraft/2002-054A | http://purl.org/dc/elements/1.1/description | Alsat                                               |
| Argentina | http://purl.org/net/schemas/space/agency | http://data.kasabi.com/dataset/nasa/spacecraft/1997-002B | http://purl.org/dc/elements/1.1/description | Aerospatiale                                        |
| Argentina | http://purl.org/net/schemas/space/agency | http://data.kasabi.com/dataset/nasa/spacecraft/1998-069B | http://purl.org/dc/elements/1.1/description | Argentinean National Commission of Space Activities |
| Australia | http://purl.org/net/schemas/space/agency | http://data.kasabi.com/dataset/nasa/spacecraft/1967-118A | http://purl.org/dc/elements/1.1/description | Sparta                                              |
| Australia | http://purl.org/net/schemas/space/agency | http://data.kasabi.com/dataset/nasa/spacecraft/1967-118A | http://purl.org/dc/elements/1.1/description | Weapons Research Establishment                      |
| Australia | http://purl.org/net/schemas/space/agency | http://data.kasabi.com/dataset/nasa/spacecraft/1985-076B | http://purl.org/dc/elements/1.1/description | Hughes                                              |
| Australia | http://purl.org/net/schemas/space/agency | http://data.kasabi.com/dataset/nasa/spacecraft/1987-078A | http://purl.org/dc/elements/1.1/description | Aussat                                              |

Rows per page: 10 ▾ 1–10 of 3903 < >

# Experimental evaluation

On 3 **semi-structured** datasets: Yelp (JSON), PubMed (XML), Nasa (RDF):

- Real-world datasets
- 57K to 230K nodes
- 300 to 6K NEs of a given type

# Experimental evaluation

On 3 **semi-structured** datasets: Yelp (JSON), PubMed (XML), Nasa (RDF):

- Real-world datasets
- 57K to 230K nodes
- 300 to 6K NEs of a given type

We **evaluate** path interestingness

# Experimental evaluation: path interestingness

|        | $(\tau_1, \tau_2)$           | min $p_{\text{rel}}$ | max $p_{\text{rel}}$ | $p_{\text{rel}}^{20}$ | $ \mathcal{P} $ | $ \mathcal{P}' $ | $R = \frac{ \mathcal{P}' }{ \mathcal{P} }$ |
|--------|------------------------------|----------------------|----------------------|-----------------------|-----------------|------------------|--------------------------------------------|
| PubMed | (Person, Organization)       | 0.0150               | 0.9142               | 0.0409                | 52              | 20               | 38.45%                                     |
|        | (Person, Location)           | 0.0150               | 0.9107               | 0.0150                | 30              | 20               | 66.66%                                     |
|        | (Location, Organization)     | 0.0150               | 0.9107               | 0.0232                | 34              | 20               | 58.82%                                     |
|        | (Person, Person)             | 0.0150               | 0.9774               | 0.0150                | 24              | 20               | 83.33%                                     |
|        | (Organization, Organization) | 0.0150               | 0.4158               | 0.0232                | 31              | 20               | 64.51%                                     |
|        | (Location, Location)         | 0.0150               | 0.0954               | 0.0150                | 20              | 20               | 100.00%                                    |
| Nasa   | (Person, Organization)       | 0.0014               | 0.0645               | 0.0178                | 191             | 20               | 10.47%                                     |
|        | (Person, Location)           | 0.0014               | 0.0645               | 0.0077                | 142             | 20               | 14.08%                                     |
|        | (Location, Organization)     | 0.0014               | 0.1016               | 0.0077                | 115             | 20               | 17.39%                                     |
|        | (Person, Person)             | 0.0014               | 0.0232               | 0.0077                | 110             | 20               | 18.18%                                     |
|        | (Organization, Organization) | 0.0014               | 0.0581               | 0.0077                | 92              | 20               | 21.73%                                     |
|        | (Location, Location)         | 0.0014               | 0.3790               | 0.0077                | 67              | 20               | 29.85%                                     |
| Yelp   | (Location, Organization)     | 0.0002               | 0.9997               | 0.0002                | 8               | 8                | 100.00%                                    |
|        | (Location, Location)         | 0.0002               | 1.0000               | 0.0002                | 11              | 11               | 100.00%                                    |

# Experimental evaluation: path interestingness

|        | $(\tau_1, \tau_2)$           | $\min p_{\text{rel}}$ | $\max p_{\text{rel}}$ | $p_{\text{rel}}^{20}$ | $ \mathcal{P} $ | $ \mathcal{P}' $ | $R = \frac{ \mathcal{P}' }{ \mathcal{P} }$ |
|--------|------------------------------|-----------------------|-----------------------|-----------------------|-----------------|------------------|--------------------------------------------|
| PubMed | (Person, Organization)       | 0.0150                | 0.9142                | 0.0409                | 52              | 20               | 38.45%                                     |
|        | (Person, Location)           | 0.0150                | 0.9107                | 0.0150                | 30              | 20               | 66.66%                                     |
|        | (Location, Organization)     | 0.0150                | 0.9107                | 0.0232                | 34              | 20               | 58.82%                                     |
|        | (Person, Person)             | 0.0150                | 0.9774                | 0.0150                | 24              | 20               | 83.33%                                     |
|        | (Organization, Organization) | 0.0150                | 0.4158                | 0.0232                | 31              | 20               | 64.51%                                     |
|        | (Location, Location)         | 0.0150                | 0.0954                | 0.0150                | 20              | 20               | 100.00%                                    |
| Nasa   | (Person, Organization)       | 0.0014                | 0.0645                | 0.0178                | 191             | 20               | 10.47%                                     |
|        | (Person, Location)           | 0.0014                | 0.0645                | 0.0077                | 142             | 20               | 14.08%                                     |
|        | (Location, Organization)     | 0.0014                | 0.1016                | 0.0077                | 115             | 20               | 17.39%                                     |
|        | (Person, Person)             | 0.0014                | 0.0232                | 0.0077                | 110             | 20               | 18.18%                                     |
|        | (Organization, Organization) | 0.0014                | 0.0581                | 0.0077                | 92              | 20               | 21.73%                                     |
|        | (Location, Location)         | 0.0014                | 0.3790                | 0.0077                | 67              | 20               | 29.85%                                     |
| Yelp   | (Location, Organization)     | 0.0002                | 0.9997                | 0.0002                | 8               | 8                | 100.00%                                    |
|        | (Location, Location)         | 0.0002                | 1.0000                | 0.0002                | 11              | 11               | 100.00%                                    |

# Experimental evaluation: path interestingness

|        | $(\tau_1, \tau_2)$           | $\min p_{\text{rel}}$ | $\max p_{\text{rel}}$ | $p_{\text{rel}}^{20}$ | $ \mathcal{P} $ | $ \mathcal{P}' $ | $R = \frac{ \mathcal{P}' }{ \mathcal{P} }$ |
|--------|------------------------------|-----------------------|-----------------------|-----------------------|-----------------|------------------|--------------------------------------------|
| PubMed | (Person, Organization)       | 0.0150                | 0.9142                | 0.0409                | 52              | 20               | 38.45%                                     |
|        | (Person, Location)           | 0.0150                | 0.9107                | 0.0150                | 30              | 20               | 66.66%                                     |
|        | (Location, Organization)     | 0.0150                | 0.9107                | 0.0232                | 34              | 20               | 58.82%                                     |
|        | (Person, Person)             | 0.0150                | 0.9774                | 0.0150                | 24              | 20               | 83.33%                                     |
|        | (Organization, Organization) | 0.0150                | 0.4158                | 0.0232                | 31              | 20               | 64.51%                                     |
|        | (Location, Location)         | 0.0150                | 0.0954                | 0.0150                | 20              | 20               | 100.00%                                    |
| Nasa   | (Person, Organization)       | 0.0014                | 0.0645                | 0.0178                | 191             | 20               | 10.47%                                     |
|        | (Person, Location)           | 0.0014                | 0.0645                | 0.0077                | 142             | 20               | 14.08%                                     |
|        | (Location, Organization)     | 0.0014                | 0.1016                | 0.0077                | 115             | 20               | 17.39%                                     |
|        | (Person, Person)             | 0.0014                | 0.0232                | 0.0077                | 110             | 20               | 18.18%                                     |
|        | (Organization, Organization) | 0.0014                | 0.0581                | 0.0077                | 92              | 20               | 21.73%                                     |
|        | (Location, Location)         | 0.0014                | 0.3790                | 0.0077                | 67              | 20               | 29.85%                                     |
| Yelp   | (Location, Organization)     | 0.0002                | 0.9997                | 0.0002                | 8               | 8                | 100.00%                                    |
|        | (Location, Location)         | 0.0002                | 1.0000                | 0.0002                | 11              | 11               | 100.00%                                    |

Both reliability and force downgrade meaningless paths (NE errors or structurally weak)

# Related work

## Structured querying

- SQL, SPARQL, GQL  
[DFG<sup>+</sup>22]

## Assisted struct. querying

- Interactive queries [DAB16]
- Guided query writing  
[ERAAL18, KKBS10]
- NL2SQL [KSHL20]

## Keyword-based search

- Unidirectional  
[ABC<sup>+</sup>02, LOF<sup>+</sup>08]
- Bi-directional [ABC<sup>+</sup>22]

## Path search in struct. queries

- SPARQL extensions:  
[ASMH18, AMSH18,  
AMM23]
- For PGs: [DFG<sup>+</sup>22]

- Pathways users need no knowledge of the graph structure or values
- Less intimidating for NTUs

# Outline

- 1 Motivation: data integration and exploration problems
- 2 Predihood: predicting neighbourhoods' environment
- 3 GeoAlign: spatial entity matching for Points of Interest
- 4 Abstra: first-sight overview of a dataset
- 5 Pathways: efficiently finding interesting paths
- 6 Systems developed
- 7 Conclusion

# Systems developed

## Predihood

City environment prediction



17 Python core classes

DATA 2020 [BDF<sup>+</sup>21]

## GeoAlign

Entity matching for POIs



41 PHP core classes

SIGSPATIAL 2019 [BDFM19]

## Abstra

Abstractions as E-R diagrams



65 Java core classes

CIKM 2022 [BMU22]

## PathWays

Interesting NE-to-NE paths



18 Java core classes

ESWC 2023 [BGLM23b]

# Outline

- 1 Motivation: data integration and exploration problems
- 2 Predihood: predicting neighbourhoods' environment
- 3 GeoAlign: spatial entity matching for Points of Interest
- 4 Abstra: first-sight overview of a dataset
- 5 Pathways: efficiently finding interesting paths
- 6 Systems developed
- 7 Conclusion

# Lessons learned

**Data integration and exploration** are difficult:

- Lack of schema or schema heterogeneity
- Data quality: wrong, null, missing values, ...
- Large amounts of data
- Bring out insights and knowledge from raw data

From the **user point of view**:

- ① User-friendly interfaces
- ② No technical detail
- ③ High-level representation



# Thanks

Nelly BARRET

✉ nelly.barret@polimi.it

🌐 <https://nelly-barret.github.io/>

🏡 Data Science group

DEIB, Politecnico di Milano

Milano



POLITECNICO  
MILANO 1863



LYON

PALAISEAU

LYON?

# References I



B Aditya, Gaurav Bhalotia, Soumen Chakrabarti, Arvind Hulgeri, Charuta Nakhe, S Sudarshanxe, et al.

BANKS: browsing and keyword searching in relational databases.

In VLDB'02: *Proceedings of the 28th International Conference on Very Large Databases*, pages 1083–1086. Elsevier, 2002.



Angelos Anadiotis, Oana Balalau, Catarina Conceicao, et al.

Graph integration of structured, semistructured and unstructured data for data journalism.

*Inf. Systems*, 104, 2022.



Angelos Christos Anadiotis, Ioana Manolescu, and Madhulika Mohanty.

Integrating connection search in graph queries.

In ICDE, April 2023.



Christian Aebeloe, Gabriela Montoya, Vinay Setty, and Katja Hose.

Discovering diversified paths in knowledge bases.

*Proc. VLDB Endow.*, 11(12):2002–2005, 2018.

Code available at: <http://qweb.cs.aau.dk/jedi/>.



Christian Aebeloe, Vinay Setty, Gabriela Montoya, and Katja Hose.

Top-k diversification for path queries in knowledge graphs.

In Marieke van Erp, Medha Atre, Vanessa López, Kavitha Srinivas, and Carolina Fortuna, editors, *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018*, volume 2180 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.



Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani.

Parametric schema inference for massive JSON datasets.

*VLDB J.*, 28(4), 2019.

# References II



Nelly Barret, Fabien Duchateau, Franck Favetta, Aurélien Gentil, and Loïc Bonneval.

An environmental study of french neighbourhoods.

In *Data Management Technologies and Applications: 9th International Conference, DATA 2020*, pages 267–292. Springer, 2021.



Nelly Barret, Fabien Duchateau, Franck Favetta, and Ludovic Moncla.

Spatial entity matching with geoalign (demo paper).

In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 580–583, 2019.



Nelly Barret, Antoine Gauquier, Jia Jean Law, and Ioana Manolescu.

Exploring heterogeneous data graphs through their entity paths.

In *Advances in Databases and Information Systems*, volume 13985 of *Lecture Notes in Computer Science*, pages 163–179. Springer, 2023.



Nelly Barret, Antoine Gauquier, Jia Jean Law, and Ioana Manolescu.

PATHWAYS: entity-focused exploration of heterogeneous data graphs (demonstration).

In *ESWC*, 2023.



Nelly Barret, Antoine Gauquier, Jia Jean Law, and Ioana Manolescu.

Exploring heterogeneous data graphs through their entity paths.

*Inf. Systems SUBM*, 2024.



Nelly Barret, Ioana Manolescu, and Prajna Upadhyay.

ABSTRA: toward generic abstractions for data of any model (demonstration).

In *CIKM*, 2022.

# References III

-  Nelly Barret, Ioana Manolescu, and Prajna Upadhyay.  
Computing generic abstractions from application datasets.  
In *EDBT*, 2024.
-  Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani.  
Schemas for safe and efficient XML processing.  
In *ICDE*. IEEE Computer Society, 2011.
-  Gonzalo Diaz, Marcelo Arenas, and Michael Benedikt.  
SPARQLByE: querying rdf data by example.  
*Proceedings of the VLDB Endowment*, 9(13):1533–1536, 2016.
-  Alin Deutsch, Nadime Francis, Alastair Green, Keith Hare, Bei Li, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Wim Martens, Jan Michels, Filip Murlak, Stefan Plantikow, Petra Selmer, Oskar van Rest, Hannes Voigt, Domagoj Vrgoc, Mingxi Wu, and Fred Zemke.  
Graph pattern matching in GQL and SQL/PGQ.  
In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, pages 2246–2258, 2022.
-  Ahmed El-Roby, Khaled Ammar, Ashraf Aboulnaga, and Jimmy Lin.  
Sapphire: querying rdf data made simple.  
*arXiv preprint arXiv:1805.11728*, 2018.
-  François Goasdoué, Paweł Guzewicz, and Ioana Manolescu.  
RDF graph summarization for first-sight structure discovery.  
*The VLDB Journal*, 29(5), April 2020.

# References IV

-  Benoît Groz, Aurélien Lemay, Sławek Staworko, and Piotr Wieczorek.  
Inference of shape graphs for graph databases.  
In *ICDT*, volume 220, 2022.
-  Roy Goldman and Jennifer Widom.  
DataGuides: enabling query formulation and optimization in semistructured databases.  
In *VLDB*, 1997.
-  Katja Hose and Ralf Schenkel.  
Towards benefit-based RDF source selection for SPARQL queries.  
In *Proceedings of the 4th International Workshop on Semantic Web Information Management*, pages 1–8, 2012.
-  Shahin Khatchadourian and Mariano P Consens.  
ExpLOD: summary-based exploration of interlinking and RDF usage in the Linked Open Data Cloud.  
In *Extended semantic web conference*, pages 272–287. Springer, 2010.
-  Nodira Khoussainova, YongChul Kwon, Magdalena Balazinska, and Dan Suciu.  
SnipSuggest: context-aware autocomplete for SQL.  
*Proceedings of the VLDB Endowment*, 4(1):22–33, 2010.
-  Hyonji Kim, Byeong-Hoon So, Wook-Shin Han, and Hongrae Lee.  
Natural language to SQL: Where are we today?  
*Proceedings of the VLDB Endowment*, 13(10):1737–1750, 2020.
-  Hanâ Lbath, Angela Bonifati, and Russ Harmer.  
Schema inference for property graphs.  
In *EDBT*, 2021.

# References V



Guoliang Li, Beng Chin Ooi, Jianhua Feng, Jianyong Wang, and Lizhu Zhou.

EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data.

In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 903–914, 2008.



Tova Milo and Dan Suciu.

Index structures for path expressions.

In *International Conference on Database Theory*, pages 277–295. Springer, 1999.



Raghu Ramakrishnan and Johannes Gehrke.

*Database Management Systems (3rd edition)*.

McGraw-Hill, 2003.



Matteo Riondato, David García-Soriano, and Francesco Bonchi.

Graph summarization with quality guarantees.

*Data mining and knowledge discovery*, 31:314–349, 2017.

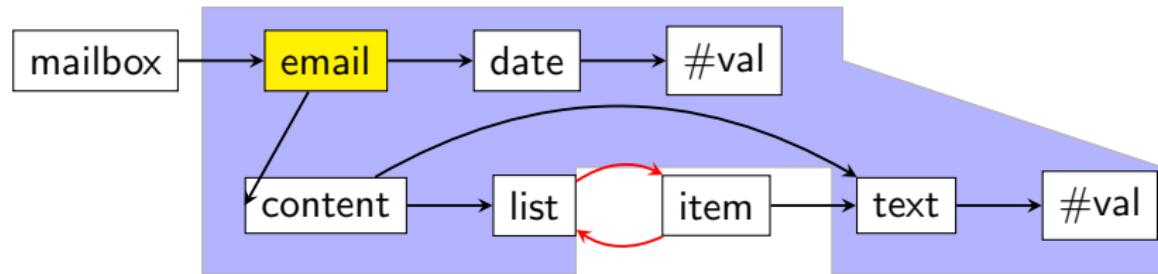


Mussab Zneika, Claudio Lucchese, Dan Vodislav, and Dimitris Kotzinos.

Summarizing linked data RDF graphs using approximate graph pattern mining.

In *19th International Conference on Extending Database Technology*, 2016.

# Data-acyclic flooding boundary



The boundary is truncated due to cyclic collection edges

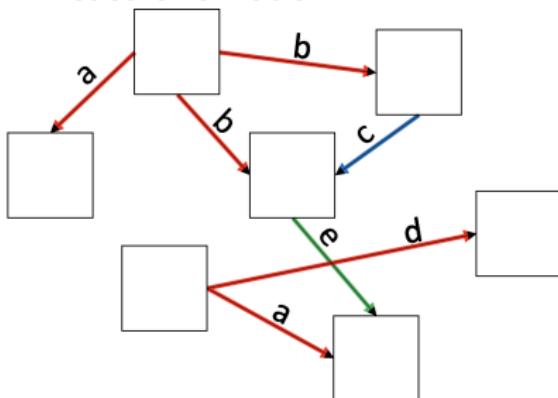
# Entity classification time

The **classification time** is composed of:

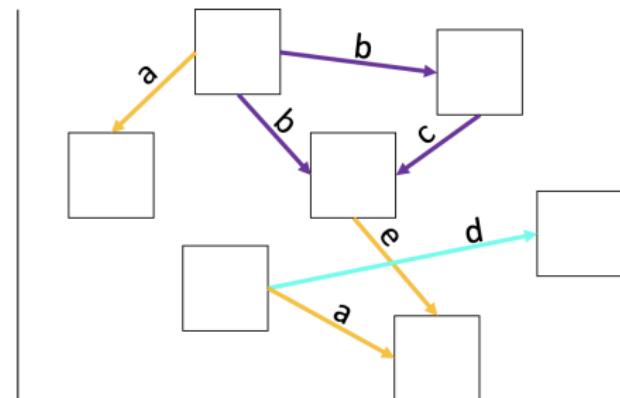
- Loading the Word2Vec semantic model
  - Constant, 4-8 seconds
- Comparing entity attributes with semantic properties
  - Varies with the number of entities and their number of attributes
  - May vary in a generated dataset of different sizes (different entity roots)
- Computing entity profiles
  - Linear in the input size

# RDF quotient graph summarization [GGM20]

- **Source clique**: set of outgoing properties co-occurring together on at least one node
- **Target clique**: set of incoming properties co-occurring together on at least one node



Properties “a”, “b”, “d” are in the same source clique



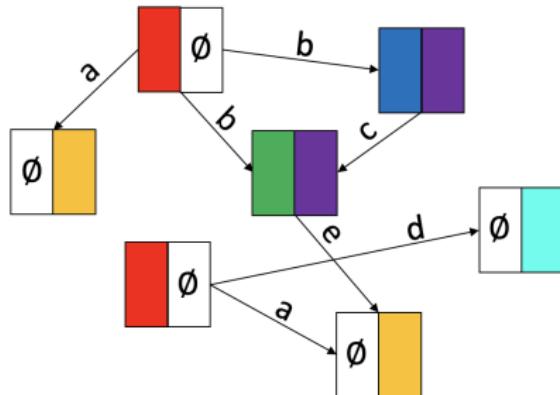
Properties “a” and “e” are in the same target clique

(c) Paweł Guzewic

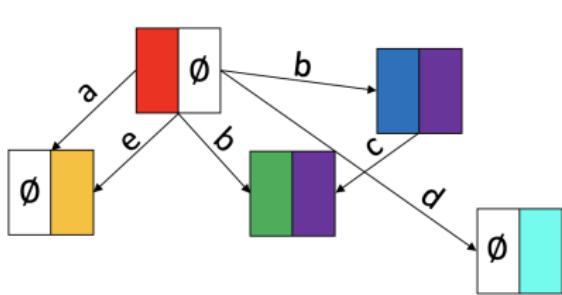
# Strong summary [GGM20]

## Strong S summary:

- Two nodes are **S equivalent** iff they have **both** the same source and target cliques



Source and target cliques for each node



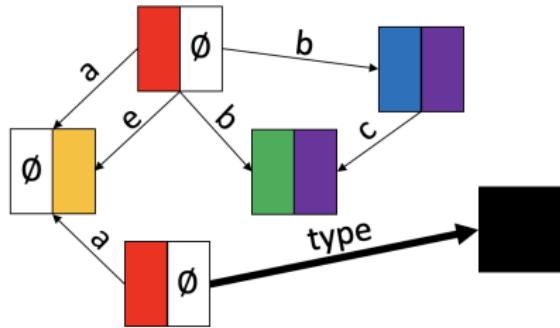
Strong summary

(c) Paweł Guzewic

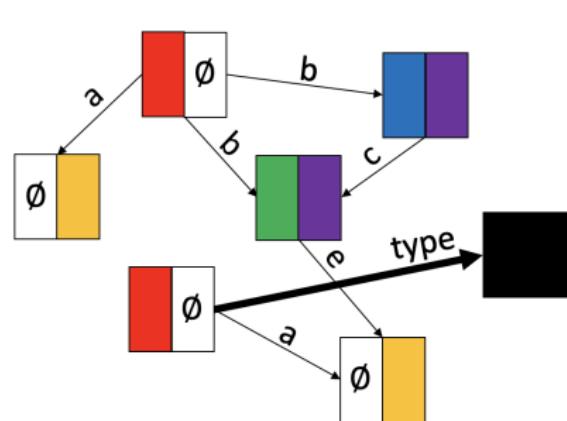
# Typed-strong summary [GGM20]

## Typed-strong TS summary:

- Two **typed** nodes are **TS equivalent** iff they have the same type set
- Two **untyped** nodes are **TS equivalent** iff they have **both** the same source and target cliques



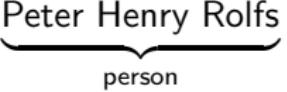
Source and target cliques for each node + an RDF type

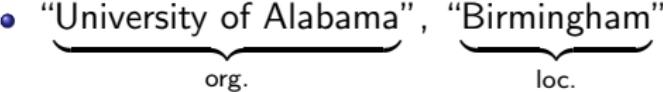
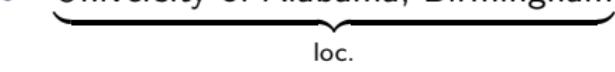


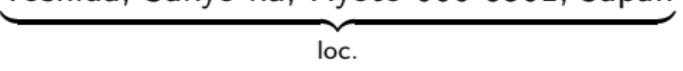
Typed-strong summary

(c) Paweł Guzewic

# Disagreement between Flair and ChatGPT

- False Flair positives:
  - Flair identifies “Av. Peter Henry Rolfs 36570-900 Vicos”  

- Flair mislead by capitalization:
  - Flair identifies “Claudin-7b” (but not ChatGPT)  

- Different token allocation:
  - “University of Alabama”, “Birmingham”  

  - “University of Alabama, Birmingham”  

- Missed non-English spelling/names:
  - ChatGPT finds “Antonio González”  

  - ChatGPT finds “Yoshida, Sakyō-ku, Kyoto 606-8501, Japan”  


# A comprehensive data exploration tool for NTUs

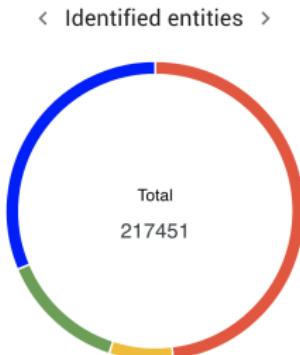
Explore

## Connection Studio Statistics



Project: Hatvp Cac

### Entities distribution by type



- Number of dates
- Number of Persons
- Number of Places
- Number of Organizations
- Number of hashtags

### Entity cloud

SVG    PNG

|                          |                          |                         |             |
|--------------------------|--------------------------|-------------------------|-------------|
| Retraitee                | Communauté               | Conseil de surveillance |             |
| VICE PRESIDENTE          | Conseil de Surveillance  | 22/06/2022              |             |
| Conseil d'Administration | Conseil d'administration | SEM                     |             |
| CONSEIL                  | Conseil Régional         | 03/20                   | PRESIDENTE  |
| SARL                     | SDIS                     | 06/07/2020              | 06/07/2020  |
| SCEA                     | Conseil départemental    | 07/20                   | 2018        |
| 11/07/2020               | 2015                     | Conseil                 | Vice GFA    |
|                          |                          | 05/20                   | 01/07/2021  |
|                          |                          | néant                   | 19/06/2022  |
|                          |                          | Retraité                | 15/07/2020  |
| Membre CA                | 02/20 AG                 | 04/20                   |             |
| CCAS                     | 2026                     | 01/20                   |             |
| 1901                     | 10/07/2020               | 12/20                   |             |
| Comité                   | 06/20                    | 01/20                   |             |
| Comité syndical          | 09/20                    | 12/20                   |             |
| Régional                 | CA                       | sci                     | PRESIDENT   |
| 2020                     | 08/20                    | 2019                    | Département |
|                          |                          | SCI                     | Membre      |
|                          |                          | France                  | 2021        |
|                          |                          | 2021                    | 16/07/2020  |
| 09/07/2020               | 11/20                    | M                       | Maire       |
| 27/09/2020               | 27/06/2021               | 2017                    | Député      |
|                          |                          | neant                   | Sénateur    |
| Education Nationale      | NEANT                    | 03/07/2020              |             |
| 15/03/2020               | Métropole                | 28/06/2020              | 30/06/2020  |
|                          | Bureau                   | 04/07/2020              |             |
|                          | Education nationale      | 08/07/2020              |             |
|                          | VICE PRESIDENT           |                         |             |
| MEMBRE CA                | CONSEIL D'ADMINISTRATION | 24/09/2017              |             |
| 07/07/2020               | 02/07/2021               | Communauté de communes  |             |
|                          |                          | 17/07/2020              |             |

# A comprehensive data exploration tool for NTUs

|                                                    |                              |                                |                                                                                  |                                     |
|----------------------------------------------------|------------------------------|--------------------------------|----------------------------------------------------------------------------------|-------------------------------------|
| Path 1<br>declaration.general.declarer.name#val    | Starting variable<br>decla   | Ending variable<br>deputyName  | <input checked="" type="button"/> EVALUATE THE QUERY                             | <input type="button"/> SAVE CHANGES |
| Path 2<br>declaration.financialInterest.items.item | Starting variable<br>decla   | Ending variable<br>item        | Join<br><input checked="" type="radio"/> Required <input type="radio"/> Optional | <input type="button"/>              |
| Path 3<br>item.company#val.extract:o               | Starting variable<br>item    | Ending variable<br>companyName | Join<br><input checked="" type="radio"/> Required <input type="radio"/> Optional | <input type="button"/>              |
| Path 4<br>item.nbShares#val                        | Starting variable<br>item    | Ending variable<br>nbShares    | Join<br><input type="radio"/> Required <input checked="" type="radio"/> Optional | <input type="button"/>              |
| Path 5<br>row.company_name.#val.extract:o          | Starting variable<br>csvline | Ending variable<br>companyName | Join<br><input checked="" type="radio"/> Required <input type="radio"/> Optional | <input type="button"/>              |

| COLUMNS | FILTERS                    | DENSITY | EXPORT      |
|---------|----------------------------|---------|-------------|
| decla   | deputyname                 | item    | companyname |
| 2660    | alain pierre marie rousset | 2743    | sanofi      |
| 1470    | edouard courtial           | 1511    | lvmh        |
| 1470    | edouard courtial           | 1543    | michelin    |

# Experimental evaluation: Flair VS ChatGPT NE extractors

|                    | GPT Person  | GPT Location | GPT Organization | GPT no entity |
|--------------------|-------------|--------------|------------------|---------------|
| Flair Person       | <b>5913</b> | 6            | 11               | 98            |
| Flair Location     | 25          | <b>1088</b>  | 507              | 905           |
| Flair Organization | 36          | 141          | <b>2988</b>      | <u>1797</u>   |
| Flair no entity    | 101         | <u>1335</u>  | <u>1233</u>      | —             |

Flair and ChatGPT mostly agree  
ChatGPT extraction has better quality