User-oriented exploration of semi-structured datasets

Nelly Barret

Inria Saclay and Institut Polytechnique de Paris
Supervised by Ioana Manolescu and Karen Bastien

March 15, 2024

Outline

Outline

Data exploration by non-technical users (NTUs)

Conflicts of Interest
in the biomedical domain
[ABB+21] w/ S. Horel

Data exploration by non-technical users (NTUs)

Conflicts of Interest
in the biomedical domain
[ABB+21] w/ S. Horel


```
<PubmedArticleSet>
  <PubmedArticle>
    <ArticleTitle>Characteristic Features of Nonalcoholic Fatty Liver Disease in
    Japan with a Focus on the Roles of Age, Sex and Body Mass Index.</ArticleTitle>
    <JournalTitle>Gut and liver</JournalTitle>
    <pubmedLink>https://pubmed.ncbi.nlm.nih.gov/31887811</pubmedLink>
    <Year>2020</Year>
    <DOI>10.5009/gnl19236</DOI>
    <KeywordList>
      <Keywords>Age</Keywords>
      <Keywords>Body mass index</Keywords>
      <Keywords>Lean NAFLD</Keywords>
    </KeywordList>
    <AuthorList>
      <Author>
        <Name>Maki Tobari</Name>
        <Affiliation>Department of Internal Medicine and Gastroenterology, Tokyo
        Women's Medical University Yachiyo Medical Center, Chiba, Japan.</
        Affiliation>
      </Author>
      <Author>
        <Name>Etsuko Hashimoto</Name>
        <Affiliation>Department of Internal Medicine and Gastroenterology, Tokyo
        Women's Medical University, Tokyo, Japan.</Affiliation>
      </Author>
    </AuthorList>
  </PubmedArticle>
  <PubmedArticle>
    <ArticleTitle>Efficacy of Current Traction Techniques for Endoscopic Submucosal
    Dissection.</ArticleTitle>
    <JournalTitle>Gut and liver</JournalTitle>
    <pubmedLink>https://pubmed.ncbi.nlm.nih.gov/31887810</pubmedLink>
    <Year>2020</Year>
    <DOI>10.5009/gnl19266</DOI>
    <AuthorList>
      <Author>
        <Name>Seiichiro Abe</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Shih Yea Sylvia Wu</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Mai Ego</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Hiroyuki Takamaru</Name>
```

Data exploration by non-technical users (NTUs)

Conflicts of Interest
in the biomedical domain
[ABB+21] w/ S. Horel


```xml
<PubmedArticleSet>
  <PubmedArticle>
    <ArticleTitle>Characteristic Features of Nonalcoholic Fatty Liver Disease in
    Japan with a Focus on the Roles of Age, Sex and Body Mass Index.</ArticleTitle>
    <JournalTitle>Gut and liver</JournalTitle>
    <pubmedLink>https://pubmed.ncbi.nlm.nih.gov/31887811</pubmedLink>
    <Year>2020</Year>
    <DOI>10.5009/gnl19236</DOI>
    <KeywordList>
      <Keywords>Age</Keywords>
      <Keywords>Body mass index</Keywords>
      <Keywords>Lean NAFLD</Keywords>
    </KeywordList>
    <AuthorList>
      <Author>
        <Name>Maki Tobari</Name>
        <Affiliation>Department of Internal Medicine and Gastroenterology, Tokyo
        Women's Medical University Yachiyo Medical Center, Chiba, Japan.</
        Affiliation>
      </Author>
      <Author>
        <Name>Etsuko Hashimoto</Name>
        <Affiliation>Department of Internal Medicine and Gastroenterology, Tokyo
        Women's Medical University, Tokyo, Japan.</Affiliation>
      </Author>
    </AuthorList>
  </PubmedArticle>
  <PubmedArticle>
    <ArticleTitle>Efficacy of Current Traction Techniques for Endoscopic Submucosal
    Dissection.</ArticleTitle>
    <JournalTitle>Gut and liver</JournalTitle>
    <pubmedLink>https://pubmed.ncbi.nlm.nih.gov/31887810</pubmedLink>
    <Year>2020</Year>
    <DOI>10.5009/gnl19266</DOI>
    <AuthorList>
      <Author>
        <Name>Seiichiro Abe</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Shih Yea Sylvia Wu</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Mai Ego</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Hiroyuki Takamaru</Name>
```


Is this dataset useful for the investigations?

Data exploration by non-technical users (NTUs)

Conflicts of Interest
in the biomedical domain
[ABB+21] w/ S. Horel

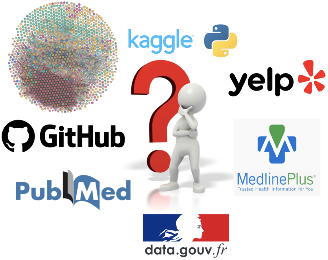

```xml
<PubmedArticleSet>
  <PubmedArticle>
    <ArticleTitle>Characteristic Features of Nonalcoholic Fatty Liver Disease in
    Japan with a Focus on the Roles of Age, Sex and Body Mass Index.</ArticleTitle>
    <JournalTitle>Gut and liver</JournalTitle>
    <pubmedLink>https://pubmed.ncbi.nlm.nih.gov/31887811</pubmedLink>
    <Year>2020</Year>
    <DOI>10.5009/gnl19236</DOI>
    <KeywordList>
      <Keywords>Age</Keywords>
      <Keywords>Body mass index</Keywords>
      <Keywords>Lean NAFLD</Keywords>
    </KeywordList>
    <AuthorList>
      <Author>
        <Name>Maki Tobari</Name>
        <Affiliation>Department of Internal Medicine and Gastroenterology, Tokyo
        Women's Medical University Yachiyo Medical Center, Chiba, Japan.</
        Affiliation>
      </Author>
      <Author>
        <Name>Etsuko Hashimoto</Name>
        <Affiliation>Department of Internal Medicine and Gastroenterology, Tokyo
        Women's Medical University, Tokyo, Japan.</Affiliation>
      </Author>
    </AuthorList>
  </PubmedArticle>
  <PubmedArticle>
    <ArticleTitle>Efficacy of Current Traction Techniques for Endoscopic Submucosal
    Dissection.</ArticleTitle>
    <JournalTitle>Gut and liver</JournalTitle>
    <pubmedLink>https://pubmed.ncbi.nlm.nih.gov/31887810</pubmedLink>
    <Year>2020</Year>
    <DOI>10.5009/gnl19266</DOI>
    <AuthorList>
      <Author>
        <Name>Seiichiro Abe</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Shih Yea Sylvia Wu</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Mai Ego</Name>
        <Affiliation>Endoscopy Division, National Cancer Center Hospital, Tokyo,
        Japan.</Affiliation>
      </Author>
      <Author>
        <Name>Hiroyuki Takamaru</Name>
```


How are authors connected to biomedical companies?

Semi-structured data exploration

Several semi-structured data models:

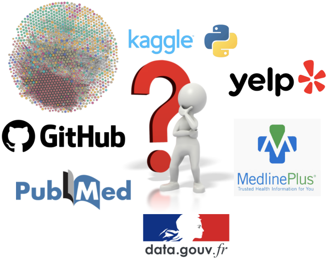
- **XML** documents
- **JSON** documents
- **RDF** graphs
- **Property** graphs



Semi-structured data exploration

Several semi-structured data models:

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



Semi-structured dataset exploration is hard: complex, irregular structure

Outline

Thesis approach

The problem

How to help users **explore unknown heterogeneous semi-structured datasets**?

Thesis approach

The problem

How to help users **explore unknown heterogeneous semi-structured datasets**?

Our approach

Automatically and efficiently compute from semi-structured datasets

Thesis approach

The problem

How to help users **explore unknown heterogeneous semi-structured datasets**?

Our approach

Automatically and efficiently compute from semi-structured datasets

1. A global, **easy-to-grasp overview** of the data

Thesis approach

The problem

How to help users **explore unknown heterogeneous semi-structured datasets**?

Our approach

Automatically and efficiently compute from semi-structured datasets

1. A global, **easy-to-grasp overview** of the data
2. The **interesting connections between Named Entities**

Research contributions

Abstra: data overviews [BMU22, BMU24]

- **Lightweight Entity-Relationship diagrams**
  - Compact yet meaningful data overviews
  - Ideal for first-sight dataset discovery

Research contributions

Abstra: data overviews [BMU22, BMU24]

- **Lightweight Entity-Relationship diagrams**
  - Compact yet meaningful data overviews
  - Ideal for first-sight dataset discovery

PathWays: interesting Named Entity connections [BGLM23b, BGLM23a, BGLM24]

- **Interesting entity paths** in and across datasets
  - Complete set of NE-to-NE interesting connections
  - Ideal for exploring connections within and across datasets

Outline

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?

```xml
<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>
```

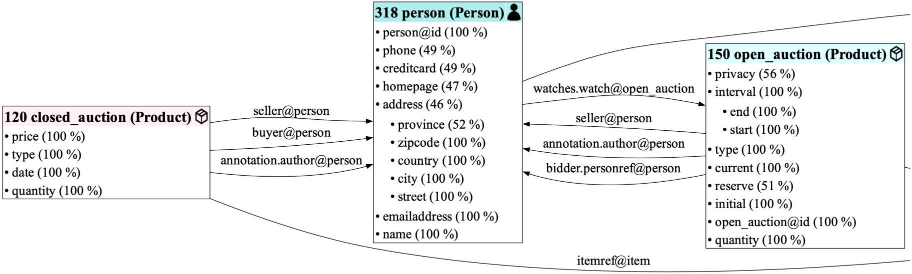
- Real-world objects and relationships between them
- Entity-Relationship models [RG03]
- Need to compute them from the dataset!
- What about semi-structured data models (nesting)?

What does the dataset describe?

```xml
<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>
```

- 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

What does the dataset describe?



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

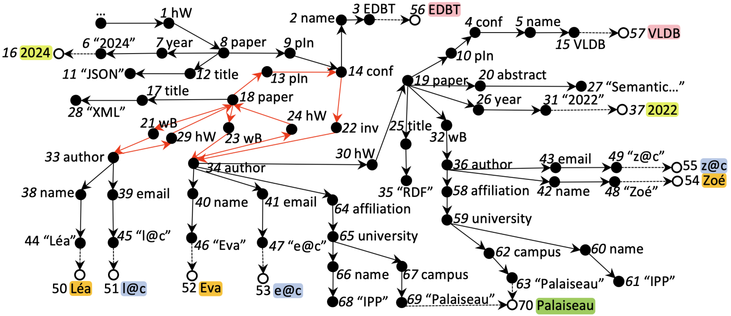
ConnectionLens [ABC+22]:

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:

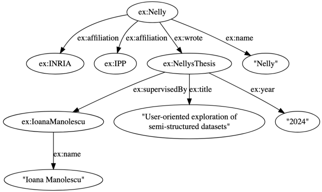
XML

```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">
  </thesis>
</root>
```

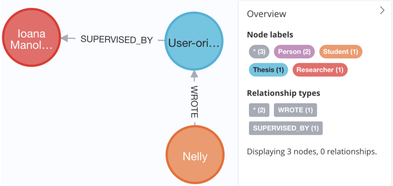
JSON

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

RDF



PG



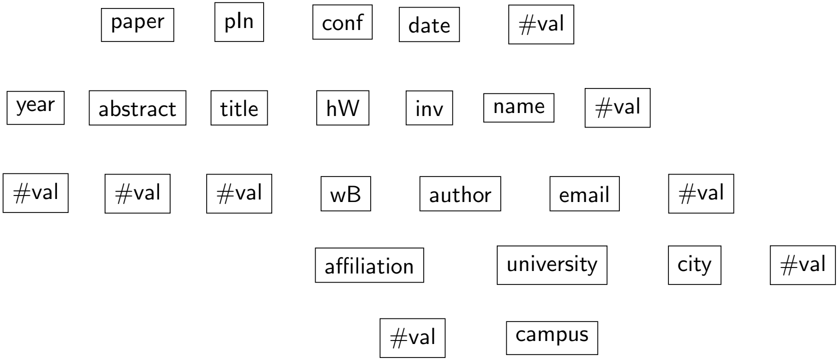
Summarization based on same-kind nodes

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) $\mathcal{G}$

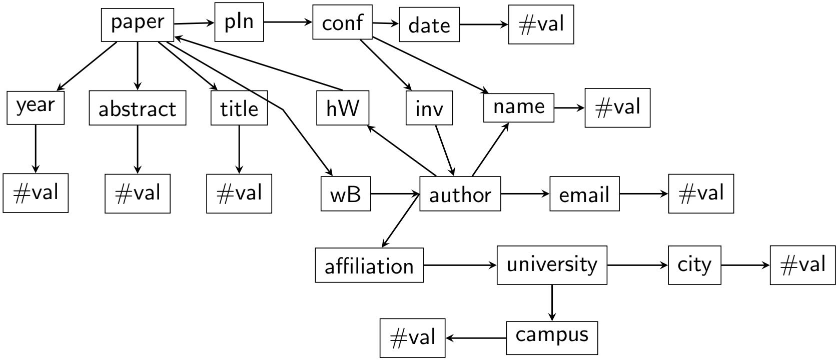
Collection node for each equivalence class



The summary (collection graph) $\mathcal{G}$

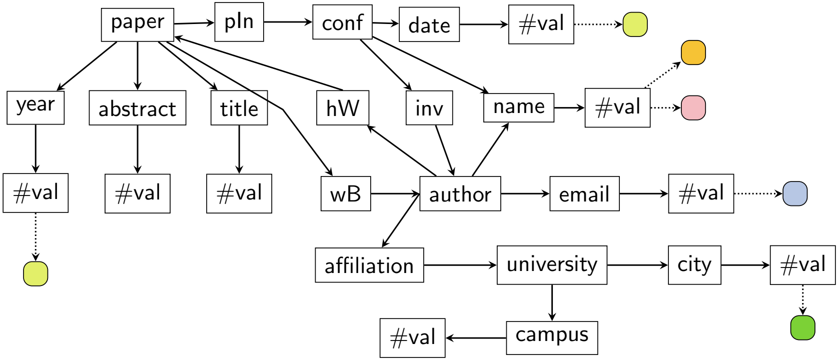
Collection node for each equivalence class

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

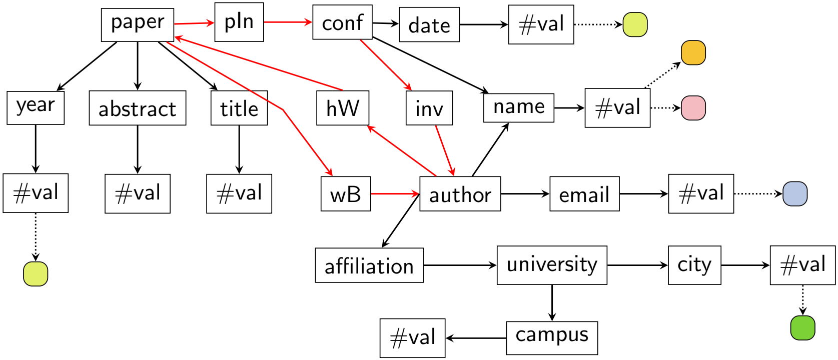


The summary (collection graph) $\mathcal{G}$

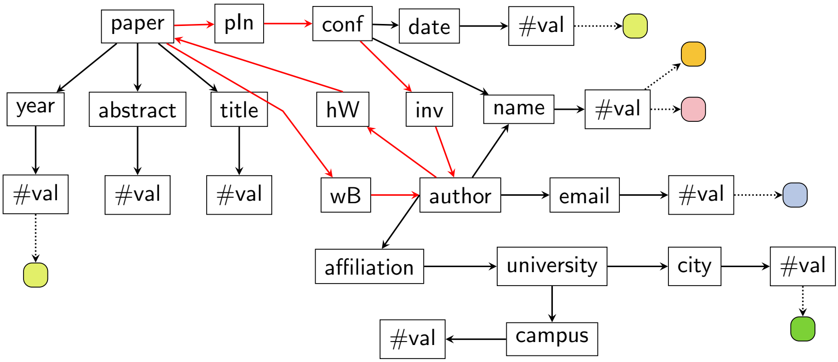
Collection node for each equivalence class
Collection edge $C_s \rightarrow C_t$ if a data edge exists
Entity profile for each **leaf collection node**: reflects NEs in the leaves



Identifying entities in the collection graph $\mathcal{G}$

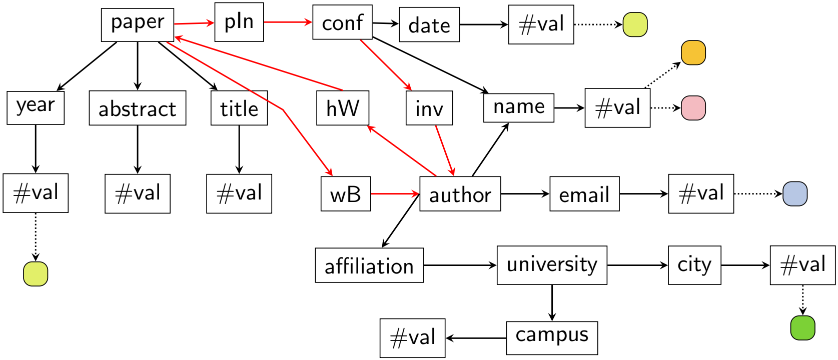


Identifying entities in the collection graph $\mathcal{G}$



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 $\boxed{\text{score}}$ 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 $\boxed{\text{boundary}}$, i.e., attribute set
   3. $\boxed{\text{Update}}$ 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

1. w_{desc_k}, w_{leaf_k}: # descendants, leaf descendants, at depth k
   - ⊗ Not clear how to pick k

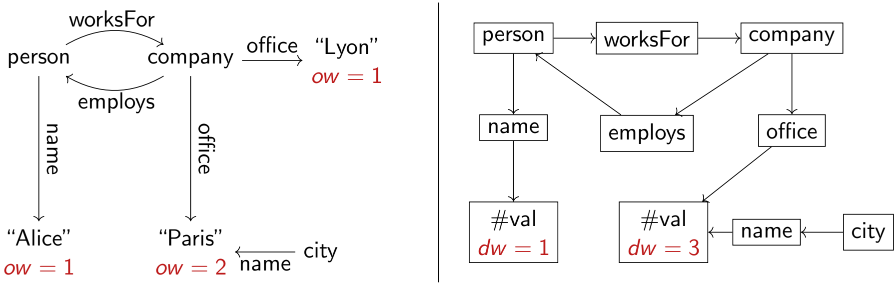
How to score a collection node?

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

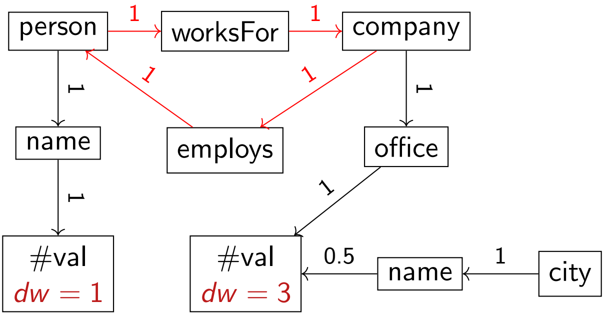
Own weight *ow* of a leaf node: its in-degree
Data weight *dw* of a leaf collection node: the sum of its nodes' *ow*



Data weight DAG propagation

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

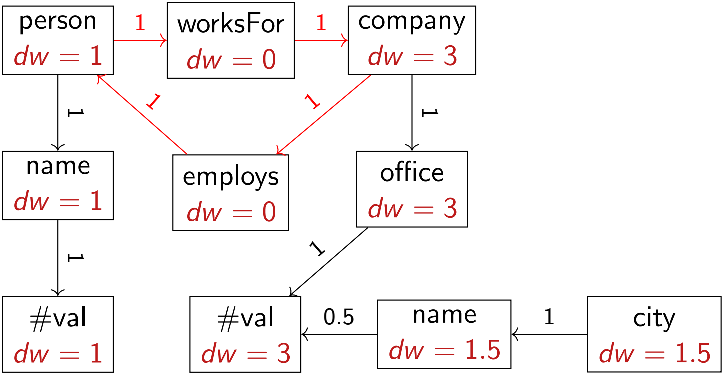
- **Edge transfer factor:** $\frac{|\text{nodes in } C_t \text{ having a parent in } C_s|}{|C_t|}$



Data weight DAG propagation

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

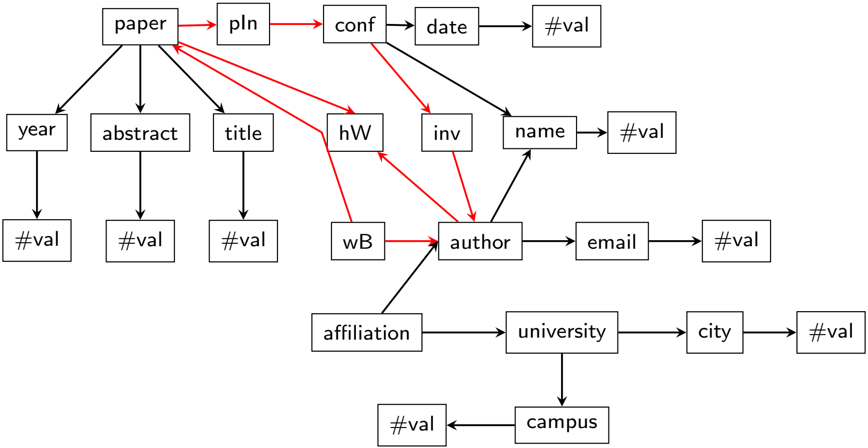
- **Edge transfer factor:** $\frac{|\text{nodes in } C_t \text{ having a parent in } C_s|}{|C_t|}$



How to score a collection node?

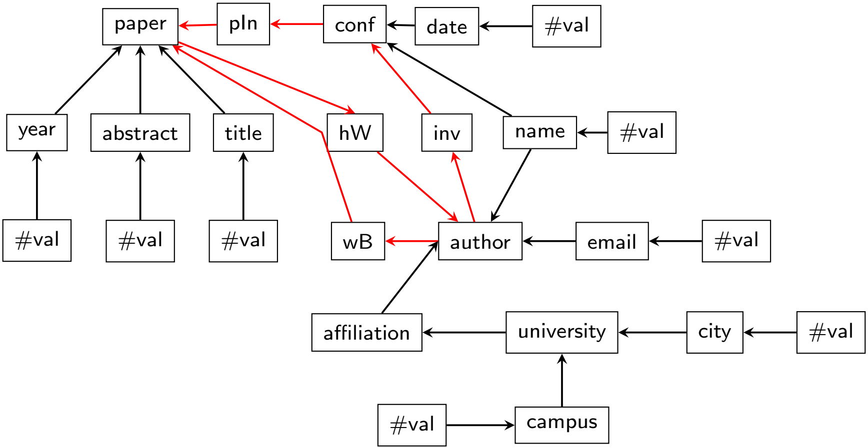
1. w_{desc_k}, w_{leaf_k}: # descendants, leaf descendants, at depth k
2. Directed Acyclic Graph (DAG) rooted in each node: w_{DAG}
3. $w_{PageRank}$: PageRank algorithm on $\mathcal{G}$

PageRank score of a collection graph node



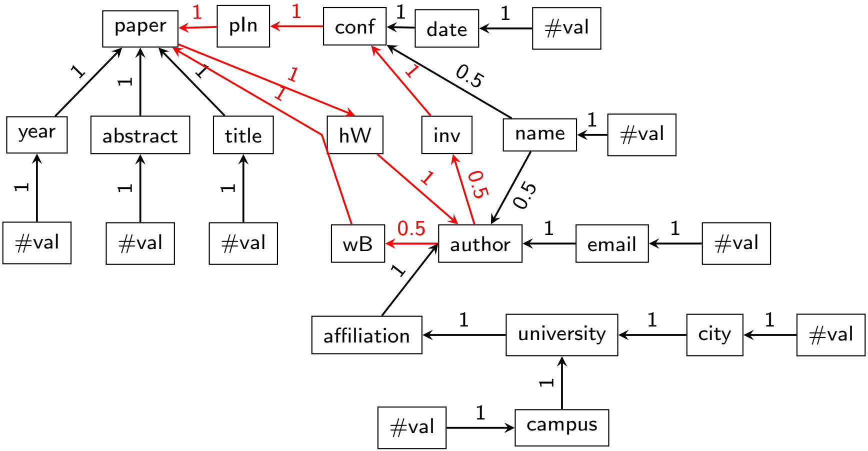
The collection graph $\mathcal{G}$

PageRank score of a collection graph node



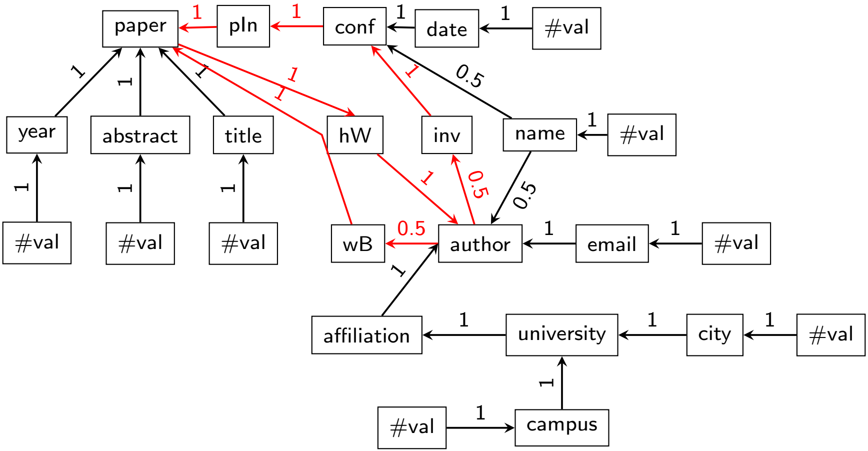
The reverse collection graph $\mathcal{G}_R$

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



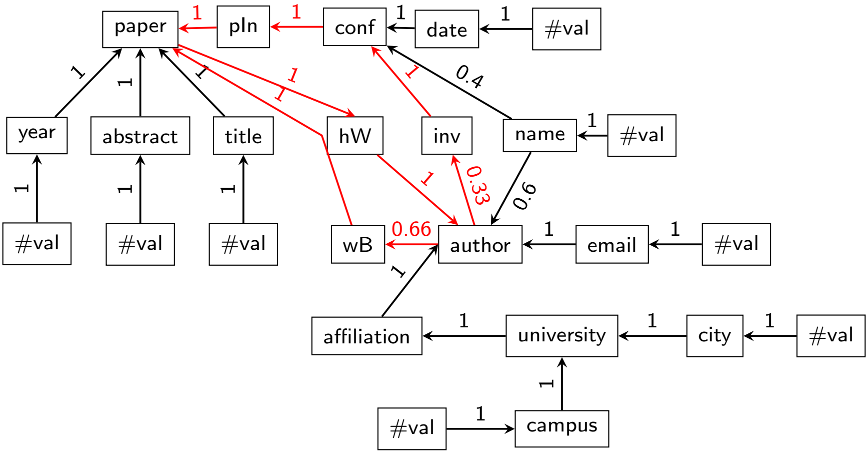
The reverse collection graph $\mathcal{G}_R$ with PR edge weights

Collections distribute their score based solely on their connectivity

How to score a collection node?

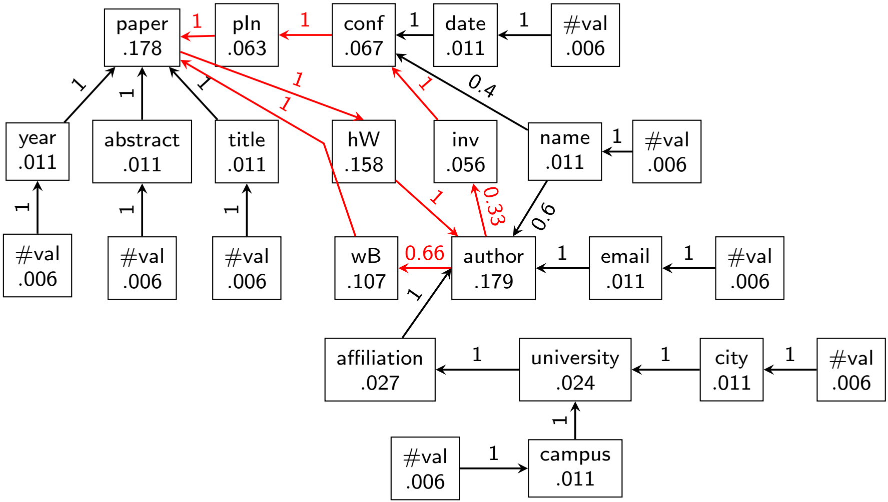
1. w_{desc_k}, w_{leaf_k}: # descendants, leaf descendants, at depth k
2. w_{DAG}: dw bottom-up propagation on $\mathcal{G}$ (outside cycles)
3. $w_{PageRank}$: PageRank algorithm on $\mathcal{G}$
4. $w_{dwPageRank}$: PageRank algorithm on $\mathcal{G}$ with dw-tuned PR edge weights
   - ✓ Reflects both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

The data-weighted PageRank score

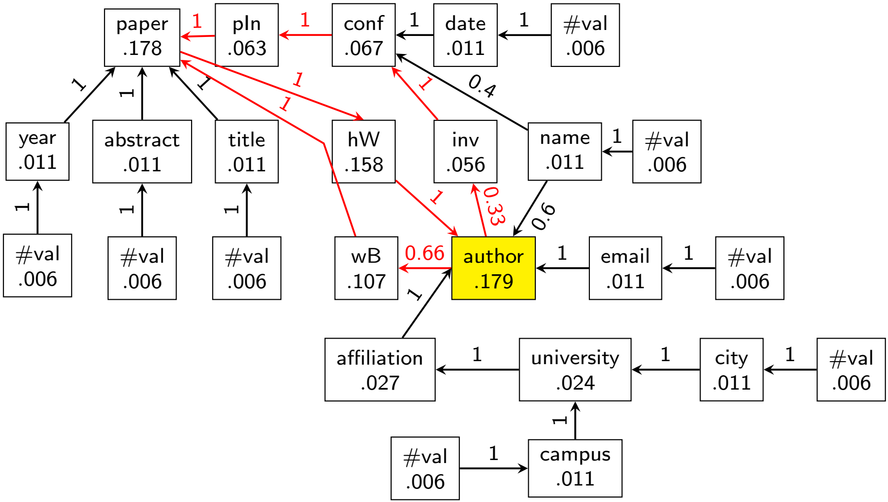


The reverse collection graph $\mathcal{G}_R$ with *dw*-tuned PR edge weights

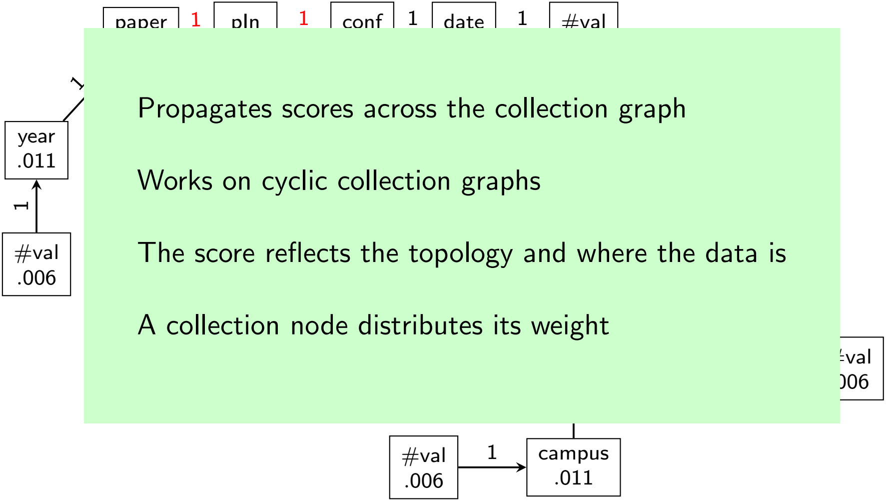
The data-weighted PageRank score



The data-weighted PageRank score



The data-weighted PageRank score



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

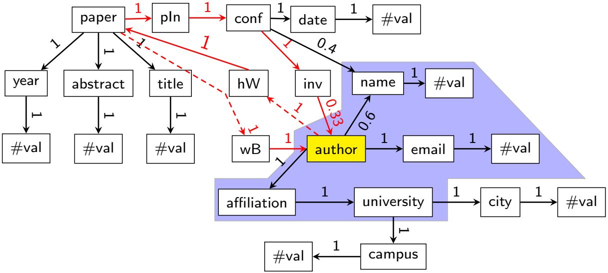
How to compute an entity boundary?

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"

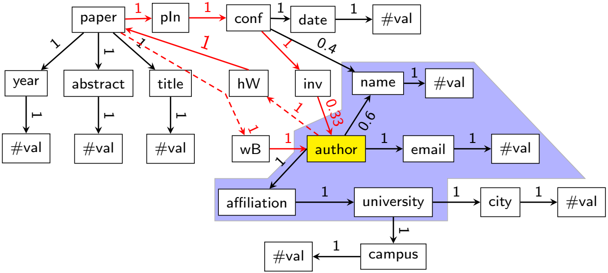
- The boundary may go far (for deep-structure entities)
- Easy to define for w_{desc_k}, w_{leaf_k}, w_{DAG}. Example for w_{desc_2}



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 $bound_{dfl-ac}$

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**

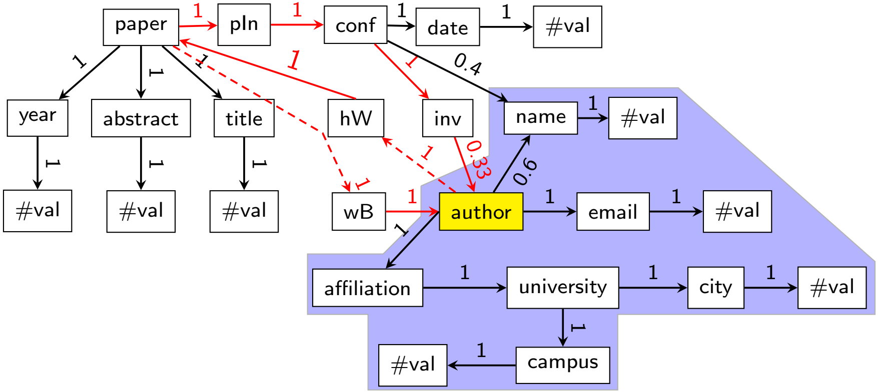
Data-acyclic flooding boundary $bound_{dfl-ac}$

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 $bound_{dfl-ac}$

- **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

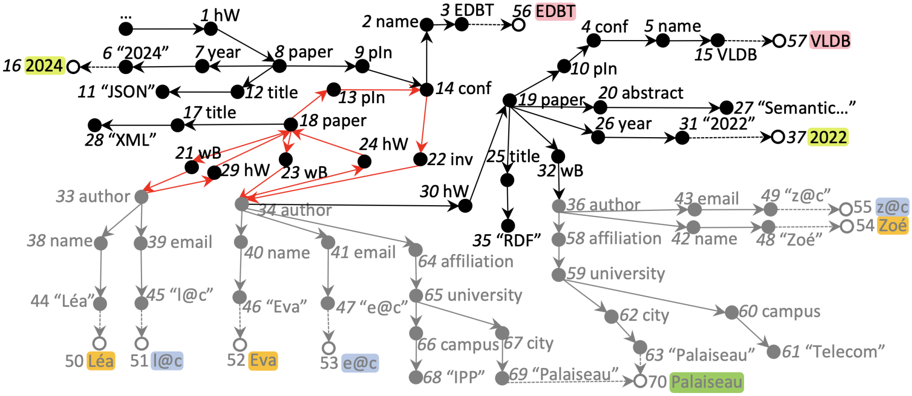
1. $update_{boolean}$
   - Collection nodes and edges in the boundary of the entity
     - Very efficient
     - Sufficient for w_{desc_k}, w_{leaf_k}, w_{DAG}

How to update the collection graph after selecting an entity?

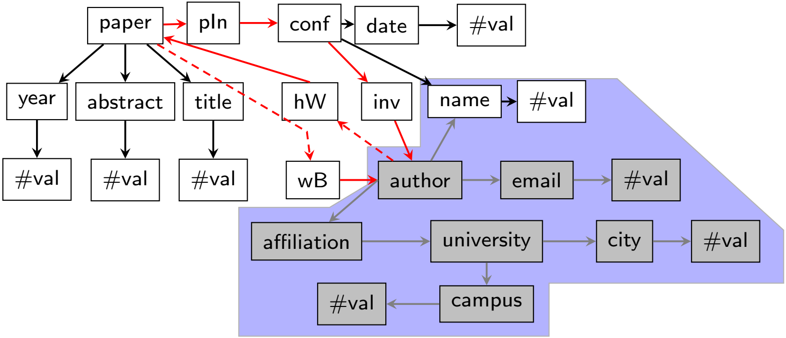
Reflect the allocation of data nodes and edges to one entity

1. $update_{boolean}$
   - Collection nodes and edges in the boundary of the entity
     - Very efficient
     - Sufficient for w_{desc_k}, w_{leaf_k}, w_{DAG}
2. $update_{exact}$
   - Graph nodes and edges
     - Much more costly
     - Required for $w_{PageRank}$, $w_{dwPageRank}$

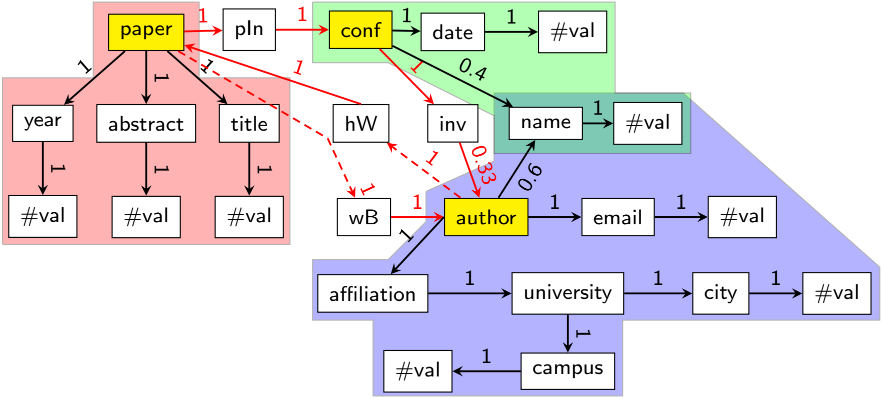
Exact graph update



Exact graph update

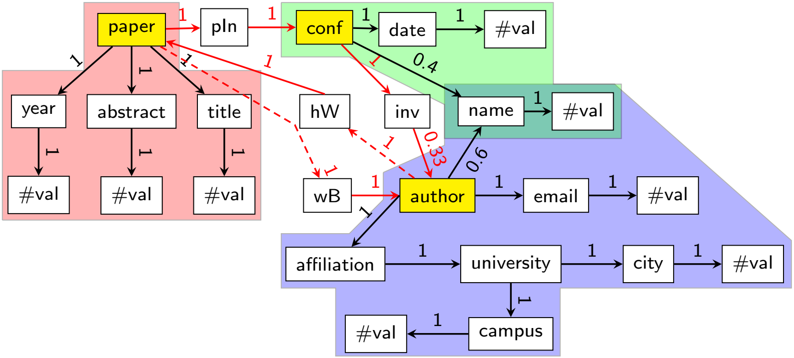


Selected entities and their boundaries



Finding relationships between entities

Relationship: a path from an entity to another



- paper → wB → author
- paper → pIn → 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]}`, ...

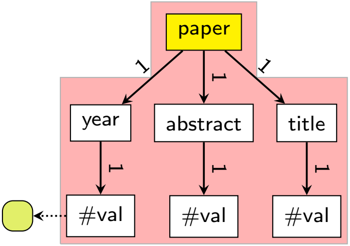
Output: a category for E

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 $p.range \in \mathcal{P}$ (, , , ...)
- Each good match votes for one or few categories

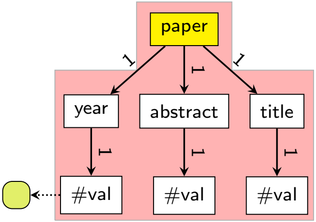
Entity classification

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



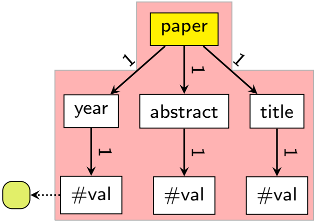
Entity classification

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, ...



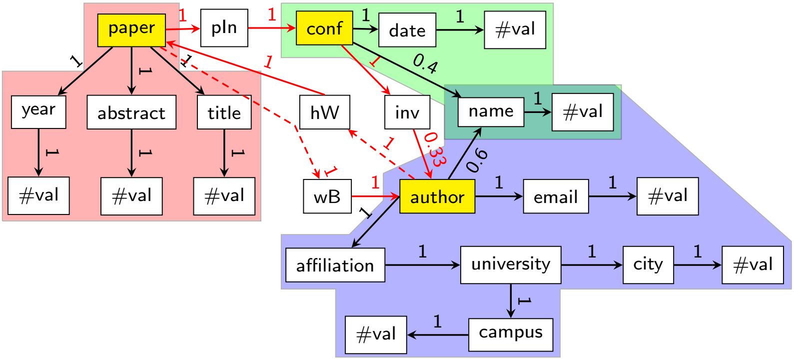
Entity classification

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, ...

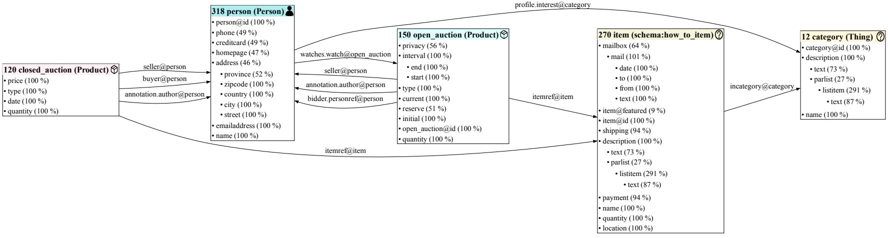


Entity classification

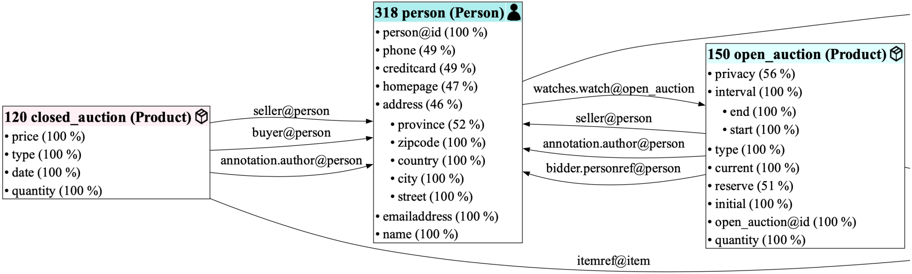
- paper nodes classified as `ResearchPublication`
- author nodes classified as `Researcher`
- conference nodes classified as `Event`



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	$\lvert C \rvert$	$\lvert \mathcal{ME} \rvert$	$\lvert \mathcal{MR} \rvert$	*cov*	$\mathcal{ME}$	d_{max}	$\lvert \mathcal{ME}_i \rvert$
Mondial $\circlearrowright$	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 $\circlearrowright$	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 $\circlearrowright$	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 $\circlearrowright$	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 $\circlearrowright$	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 $\circlearrowright$	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 $(w_{dwPageRank}, bound_{fl-ac})$

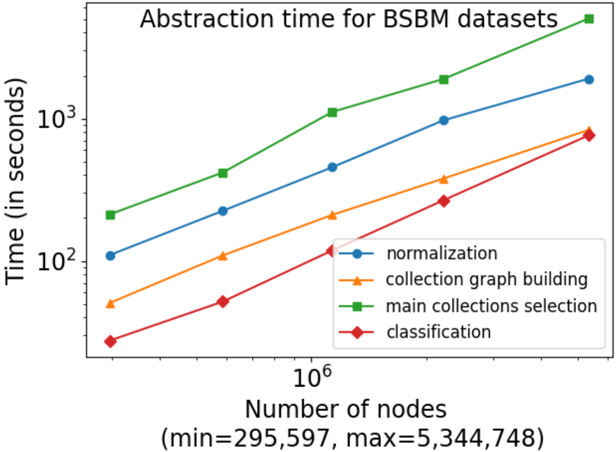
Dataset name	$\lvert C \rvert$	$\lvert \mathcal{ME} \rvert$	$\lvert \mathcal{MR} \rvert$	cov	$\mathcal{ME}$	d_{max}	$\lvert \mathcal{ME}_i \rvert$
Mondial $\circlearrowright$	168	5	8	0.85	City	3	3,152
					Province	3	1,455
					Country	4	231
					Organization	4	168
					River	4	135
PubMed	26	1	0	1.0	PubMedArticle	5	957
XMark1 $\circlearrowright$	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 $\circlearrowright$	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:
      ▾ 0:
        ▾ __Content:
          ▾ action:
              __Kind:                      "StrType"
          ▾ attachments:
            ▾ __Content:
              ▾ __Content:
                ▾ 0:
                  ▾ __Content:
                    ▾ audio:
                      ▾ __Content:
                        ▾ 0:
                          ▾ __Content:
                            ▾ album_id:
                                __Kind:    "NumType"
                            ▾ artist:
                                __Kind:    "StrType"
                            ▾ content_restricted:
                                __Kind:    "NumType"
                            ▾ date:
                                __Kind:    "NumType"
                            ▾ duration:
                                __Kind:    "NumType"
                            ▾ genre_id:
                                __Kind:    "NumType"
                            ▾ id:
                                __Kind:    "NumType"
                            ▾ lyrics_id:
                                __Kind:    "NumType"
                            ▾ owner_id:
                                __Kind:    "NumType"
```

Outline

Data is often used to find connections

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

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 (■)

- ■ $\leftarrow$ **#val** $\leftarrow$ **Name** $\leftarrow$ **Author** $\rightarrow$ **Affiliation** $\rightarrow$ **#val** $\rightarrow$ ■
- ■ $\leftarrow$ **#val** $\leftarrow$ **Name** $\leftarrow$ **Author** $\leftarrow$ **Authors** $\leftarrow$ **Article** $\rightarrow$ **Journal** $\rightarrow$ **#val** $\rightarrow$ ■

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., "$\underbrace{\text{John Hopkins}}_{\text{person}}$ University Hospital"
- False positives, or wrong entity type attribution, e.g., "$\underbrace{\text{THC}}_{\text{org.}}$"

What makes a NE-to-NE path interesting?

Some paths are unreliable: we face entity extraction errors

- E.g., "$\underbrace{\text{John Hopkins}}_{\text{person}}$ University Hospital"
- False positives, or wrong entity type attribution, e.g., "$\underbrace{\text{THC}}_{\text{org.}}$"

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., "$\underbrace{\text{John Hopkins}}_{\text{person}}$ University Hospital"
- False positives, or wrong entity type attribution, e.g., "$\underbrace{\text{THC}}_{\text{org.}}$"

Some paths are structurally weak: we face information dilution

- E.g., a paper has 50 authors

$\boxed{\text{Path interestingness}}$: based on edge reliability and edge force

What makes a NE-to-NE path interesting?

1. **Reliability** $r(C_i \dashrightarrow \blacksquare)$ 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** $r(C_i \dashrightarrow \blacksquare)$ 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** $r(C_i \dashrightarrow \blacksquare)$ 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?

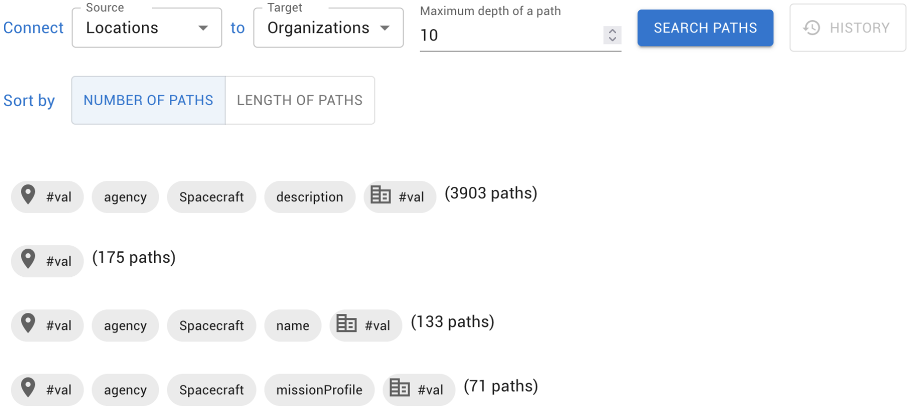
1. **Reliability** $r(C_i \dashrightarrow \blacksquare)$ of an extraction collection edge
   - The ratio of NEs having the type $\blacksquare$, and extracted from C_i
   - Path reliability: minimum extraction edge reliability
2. **Force** $f(C_i \rightarrow C_j)$ of a structural collection edge
   - 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 (■)

- ■ $\xleftarrow{1.0}$ **#val** $\xleftarrow{1.0}$ **Name** $\xleftarrow{1.0}$ **Author** $\xrightarrow{1.0}$ **Affiliation** $\xrightarrow{1.0}$ **#val** $\xrightarrow{0.91}$ ■
  - Reliable; strong
- ■ $\xleftarrow{1.0}$ **#val** $\xleftarrow{1.0}$ **Name** $\xleftarrow{1.0}$ **Author** $\xleftarrow{0.02}$ **Authors** $\xleftarrow{1.0}$ **Article** $\xrightarrow{1.0}$ **Journal** $\xrightarrow{1.0}$ **#val** $\xrightarrow{0.41}$ ■
  - Reliable; weak
- ■ $\xleftarrow{0.09}$ **#val** $\xleftarrow{1.0}$ **COI** $\xleftarrow{1.0}$ **Article** $\xrightarrow{1.0}$ **Journal** $\xrightarrow{1.0}$ **#val** $\xrightarrow{0.05}$ ■ $\xleftarrow{0.09}$ **#val** $\xrightarrow{0.04}$ ■
  - Not reliable; strong

PathWays output: data paths as tables



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

	(τ_1, τ_2)	min p_{rel}	max p_{rel}	p_{rel}^{20}	$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{P}'\rvert$	$R = \frac{\lvert\mathcal{P}'\rvert}{\lvert\mathcal{P}\rvert}$
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

	(τ_1, τ_2)	min p_{rel}	max p_{rel}	p_{rel}^{20}	$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{P}'\rvert$	$R = \frac{\lvert\mathcal{P}'\rvert}{\lvert\mathcal{P}\rvert}$
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

	(τ_1, τ_2)	min p_{rel}	max p_{rel}	p^{20}_{rel}	$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{P}'\rvert$	$R = \frac{\lvert\mathcal{P}'\rvert}{\lvert\mathcal{P}\rvert}$
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+22]

Assisted struct. querying

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

Keyword-based search

- Unidirectional [ABC+02, LOF+08]
- Bi-directional [ABC+22]

Path search in struct. queries

- SPARQL extensions: [ASMH18, AMSH18, AMM23]
- For PGs: [DFG+22]

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

Outline

Systems developed

Abstra for data abstraction:

- https://team.inria.fr/cedar/projects/abstra/
    - 65 Java core classes and 10K LOC
    - Demonstrated at CIKM 2022 [BMU22] (also BDA 2022)

PathWays for NE-to-NE paths:

- https://team.inria.fr/cedar/projects/pathways/
    - 18 Java core classes and 4K LOC
    - Demonstrated at ESWC 2023 [BGLM23b] (also BDA 2023)

Systems developed

Abstra for data abstraction:

- https://team.inria.fr/cedar/projects/abstra/
  - 65 Java core classes and 10K LOC
  - Demonstrated at CIKM 2022 [BMU22] (also BDA 2022)

PathWays for NE-to-NE paths:

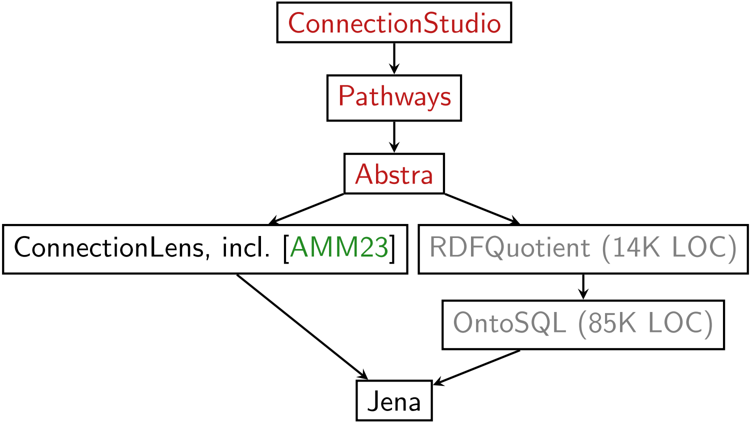
- https://team.inria.fr/cedar/projects/pathways/
  - 18 Java core classes and 4K LOC
  - Demonstrated at ESWC 2023 [BGLM23b] (also BDA 2023)

ConnectionStudio for NTU data exploration:

- https://connectionstudio.inria.fr/
  - 4K Java LOC and 21K JavaScript LOC (w/ T. Galizzi, S. Ebel, M. Mohanty)
  - Demonstrated at CoopIS 2023 [BEG+23] (also BDA 2023)

ConnectionStudio software pile

All deployed using Maven, hundreds of unit tests, etc.
Help from T. Galizzi, M. Mohanty
Several rounds of re-engineering (ML model memory consumption, etc.)



A comprehensive data exploration tool for NTUs

ConnectionStudio: a data lake for ingesting, exploring and querying heterogeneous data

1. Data abstractions as E-R diagrams (Abstra)
2. NE-to-NE paths as tables (PathWays)
3. “Gentle introduction” to the data lake (w/ journalist input)

A comprehensive data exploration tool for NTUs

ConnectionStudio: a data lake for ingesting, exploring and querying heterogeneous data

1. Data abstractions as E-R diagrams (Abstra)
2. NE-to-NE paths as tables (PathWays)
3. “Gentle introduction” to the data lake (w/ journalist input)

Demonstrated to journalists at **DataJournos** (40) and **CFI** (60)

A comprehensive data exploration tool for NTUs

ConnectionStudio: a data lake for ingesting, exploring and querying heterogeneous data

1. Data abstractions as E-R diagrams (Abstra)
2. NE-to-NE paths as tables (PathWays)
3. "Gentle introduction" to the data lake (w/ journalist input)

Demonstrated to journalists at **DataJournos** (40) and **CFI** (60)

ConnectionStudio interesting for a first look at the data.
Still maturing...

Outline

Takeaways and next steps

We introduced:

1. A unified view over heterogeneous semi-structured data models
2. Abstra: a dataset abstraction system for semi-structured data
3. PathWays: an entity-focused exploration system
4. ConnectionStudio: a comprehensive data lake exploration tool

Takeaways and next steps

We introduced:

1. A unified view over heterogeneous semi-structured data models
2. Abstra: a dataset abstraction system for semi-structured data
3. PathWays: an entity-focused exploration system
4. ConnectionStudio: a comprehensive data lake exploration tool

Next steps:

- Generate PG schemas from abstractions [BEMM24]
- Migrate data graphs into PG graphs
- Enrich extracted NEs with RDF knowledge bases

Publications (1/2)

Abstra: **N. Barret**, T. Enache, N. Dobričic, S. Ebel, T. Galizzi, I. Manolescu, P. Upadhyay, M. Mohanty

1. Finding the PG schema of any (semi)structured dataset: a tale of graphs and abstraction, *SEAGraph'24*
2. Computing generic abstractions from application datasets, *EDBT'24*
3. Abstra: toward generic abstractions for data of any model, *CIKM'22*
4. Toward Generic Abstractions for Data of Any Model, *BDA'21*
5. Facilitating Heterogeneous Dataset Understanding, *BDA'21*

Publications (2/2)

PathWays: **N. Barret**, A. Gauquier, J. J. Law, I. Manolescu

1. Exploring heterogeneous data graphs through their entity paths, *INFSYS'24 – submitted*
2. Exploring heterogeneous data graphs through their entity paths, *ADBIS'23*
3. PathWays: entity-focused exploration of heterogeneous data graphs, *ESWC'23*

ConnectionStudio: **N. Barret**, S. Ebel, T. Galizzi, I. Manolescu, M. Mohanty

1. User-friendly exploration of highly heterogeneous data lakes, *EGC'24*
2. User-friendly exploration of highly heterogeneous data lakes, *CoopIS'23*

Thanks

- My PhD advisor: Ioana Manolescu
- Interns I co-supervised
- The CEDAR team
- My family

The CEDAR team at Saint-Rémy-lès-Chevreuse in 2023

References I

Angelos-Christos G. Anadiotis, Oana Balalau, Theo Bouganim, Francesco Chimienti, Helena Galhardas, Mhd Yamen Haddad, Stephane Horel, Ioana Manolescu, and Youssr Youssef.
Empowering investigative journalism with graph-based heterogeneous data management.
IEEE Data Eng. Bull., 44(3):12–26, 2021.

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/.

References II

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.

Nelly Barret, Simon Ebel, Théo Galizzi, Ioana Manolescu, and Madhulika Mohanty.
User-friendly exploration of highly heterogeneous data lakes.
In Mohamed Sellami, Maria-Esther Vidal, Boudewijn F. van Dongen, Walid Gaaloul, and Hervé Panetto, editors, *Cooperative Information Systems - 29th International Conference, CoopIS 2023, Groningen, The Netherlands, October 30 - November 3, 2023, Proceedings*, volume 14353 of *Lecture Notes in Computer Science*, pages 488–496. Springer, 2023.

Nelly Barret, Tudor Enache, Ioana Manolescu, and Madhulika Mohanty.
Finding the PG schema of any (semi)structured dataset: a tale of graphs and abstraction.
In *SEAGraph workshop*, 2024.

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.

References III

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.

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.

References IV

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, Pawel Guzewicz, and Ioana Manolescu.
RDF graph summarization for first-sight structure discovery.
The VLDB Journal, 29(5), April 2020.

Benoît Groz, Aurélien Lemay, Slawek 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.

References V

Shahan 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 autocompletion for SQL.
Proceedings of the VLDB Endowment, 4(1):22–33, 2010.

Hyeonji 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.

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.

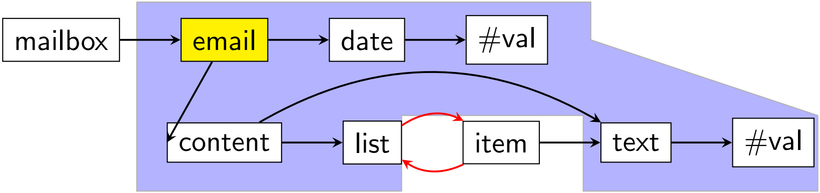
Raghu Ramakhrishnan and Johannes Gehrke.
Database Management Systems (3rd edition).
McGraw-Hill, 2003.

References VI

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

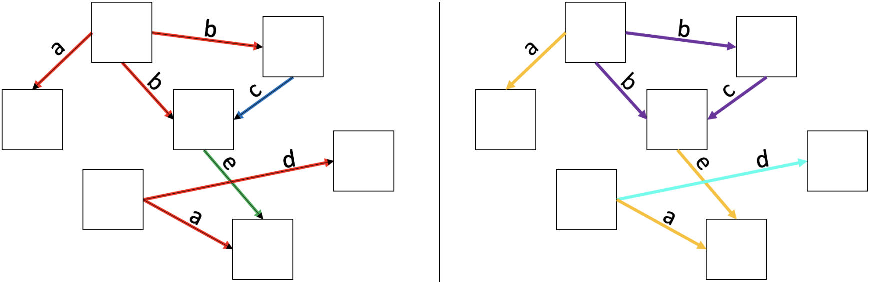
Entity classification time

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-occuring together on at least one node
- Target clique: set of incoming properties co-occuring together on at least one node



Properties "a", "b", "d" are in the same source clique

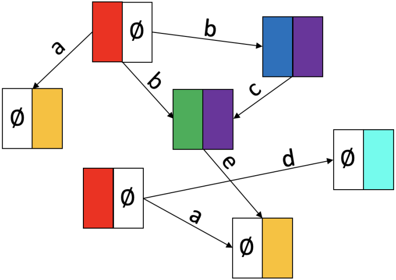
Properties "a" and "e" are in the same target clique

(c) Pawel Guzewic

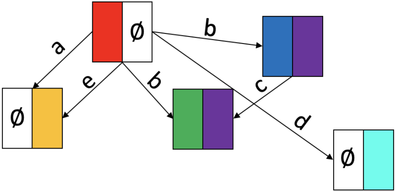
Strong summary [GGM20]

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



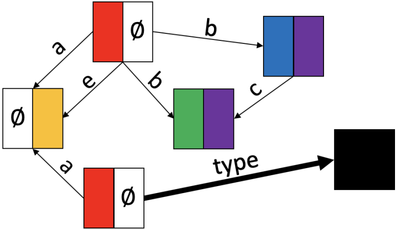
Strong summary

(c) Pawel Guzewic

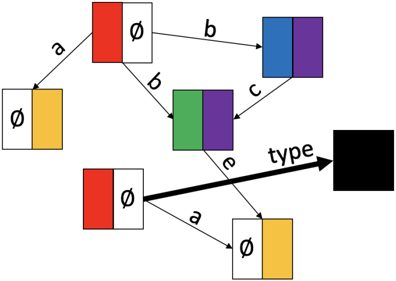
Typed-strong summary [GGM20]

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



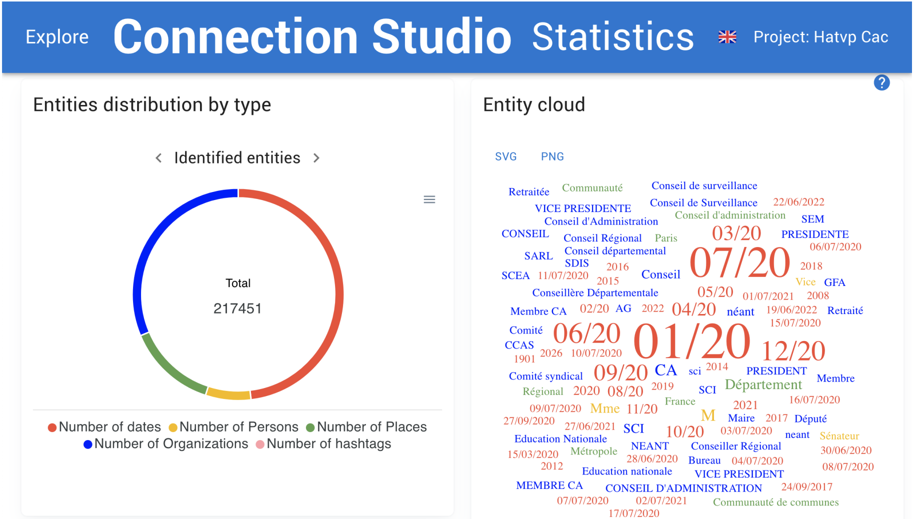
Typed-strong summary

(c) Pawel Guzewic

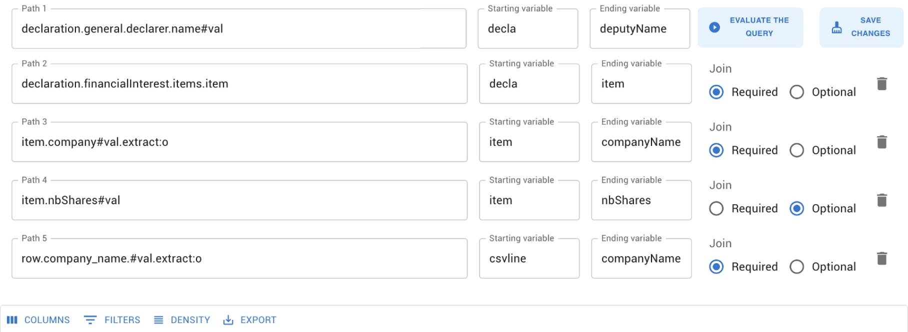
Disagreement between Flair and ChatGPT

- False Flair positives:
  - Flair identifies "Av. $\underbrace{\text{Peter Henry Rolfs}}_{\text{person}}$ 36570-900 Vicosa"
- Flair mislead by capitalization:
  - Flair identifies "$\underbrace{\text{Claudin-7b}}_{\text{person}}$" (but not ChatGPT)
- Different token allocation:
  - "$\underbrace{\text{University of Alabama}}_{\text{org.}}$", "$\underbrace{\text{Birmingham}}_{\text{loc.}}$"
  - "$\underbrace{\text{University of Alabama, Birmingham}}_{\text{loc.}}$"
- Missed non-English spelling/names:
  - ChatGPT finds "$\underbrace{\text{Antonio González}}_{\text{person}}$"
  - ChatGPT finds "$\underbrace{\text{Yoshida, Sakyo-ku, Kyoto 606-8501, Japan}}_{\text{loc.}}$"

A comprehensive data exploration tool for NTUs



A comprehensive data exploration tool for NTUs



decla	deputyname	item	companyname	nbshares	csvline
2660	alain pierre marie rousset	2743	sanofi	1200	352
1470	edouard courtial	1511	lvmh	29013	248
1470	edouard courtial	1543	michelin	162179	261

Experimental evaluation: Flair VS ChatGPT NE extractors

	GPT Person	GPT Location	GPT Organization	GPT no entity
Flair Person	**5913**	6	11	98
Flair Location	25	**1088**	507	905
Flair Organization	36	141	**2988**	1797
Flair no entity	101	1335	1233	–

Flair and ChatGPT mostly agree
ChatGPT extraction has better quality