

User-oriented exploration of semi-structured datasets

Nelly Barret

4th year PhD student

Supervised by Ioana Manolescu

Inria Saclay and Institut Polytechnique de Paris

January 25, 2024



Context: data is the new gold (1/3)





Context: data is the new gold (2/3)

Our digital world comes:

- In **various contexts**: science, health, political life
- At **various scales**: home, city, country, world
- By **different actors**: scientists, businesses, policy makers
- With **different needs**, constraints, abilities

We are **overwhelmed** by (raw) data, we need:

- Data-driven applications
- Data journalism
- Knowledge graphs
- Artificial “intelligence”
- ...



Context: data is the new gold (3/3)

Very **heterogeneous** data:

- Mainly RDF (1K datasets in the LODC)
- Also: XML, JSON, relational, Property Graph...

Detection of **entities** of interest:

- People, Place, email, ...



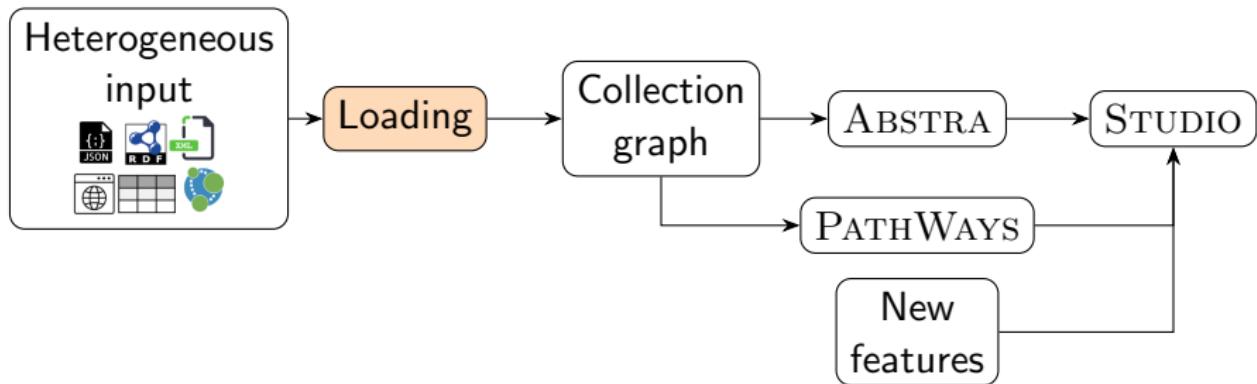
With **heterogeneous** data, users need:

- ① A **uniform** integration, view over the data
- ② **Efficient** algorithms and applications
- ③ A global **understanding**, description
- ④ Interesting **entity connections**

Create a unique data graph

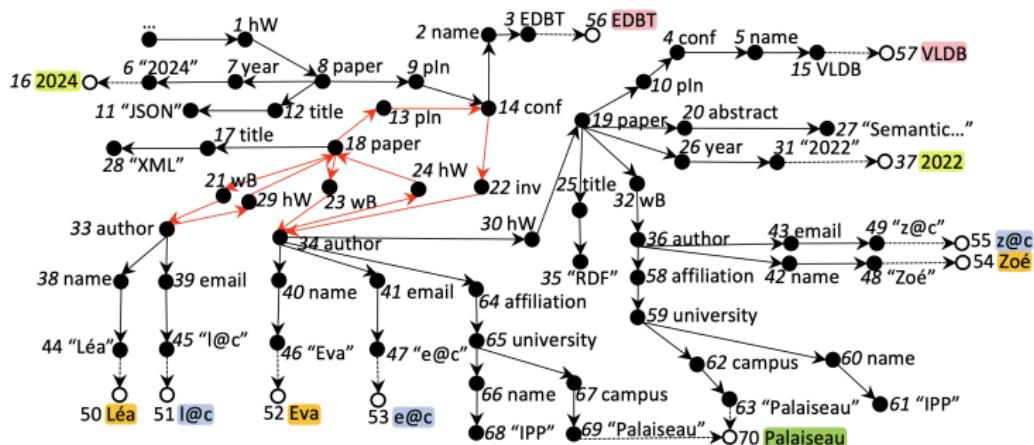
“A **uniform** integration, view”

Angelos Anadiotis Oana Balalau Ioana Manolescu et al...
IPP, EPFL Inria, IPP Inria, IPP INSEC, ...



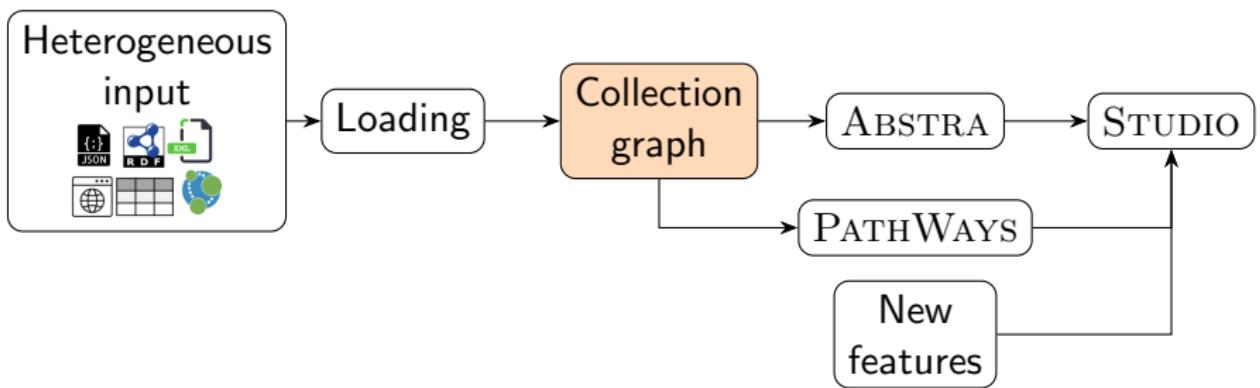
Graph construction

- Ingest any dataset into a **directed graph** (\bullet , \rightarrow)
- Extract **named entities**, NEs, from the graph values (\circ , $--\rightarrow$):
 - Temporal: **date**, time reference
 - Web: **URI**, **email address**, hashtag, Twitter citation
 - Complex entities: **People**, **Place**, **Organization**



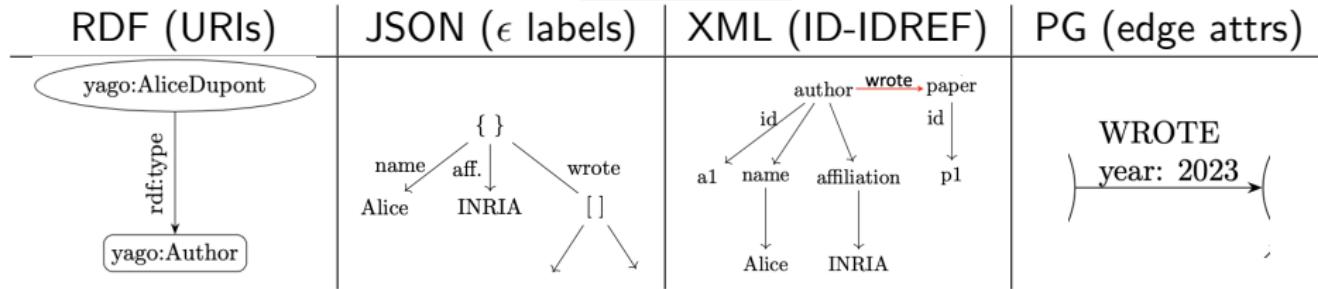
Create a compact representation of the data graph

“Efficient algorithms and applications”



A uniform view of data formats

Each data format has its own specificities:



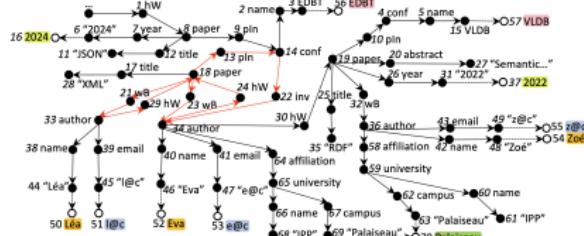
But, we **encode** the same logic:

- **Record**: piece of data, an object
- **Value**: record with no children
- **Same-kind records**: schema or “intuitive” order
- **Relationship**: how records relate

Compact representation (summarization)

Three equivalence relations:

- Per **label** for XML
 - Per **path** for JSON and relational data
 - Per **type or edge neighbourhood** for RDF and PG [GGM20]



$$\left\{ \begin{array}{ll} EC_1 & \{N_8, N_{18}, N_{19}\} \\ EC_2 & \{N_{33}, N_{34}, N_{36}\} \\ EC_3 & \{N_4, N_5\} \\ EC_4 & \{N_2, N_5, N_{38}, N_{40}, N_{42}\} \\ EC_i & \dots \end{array} \right\}$$

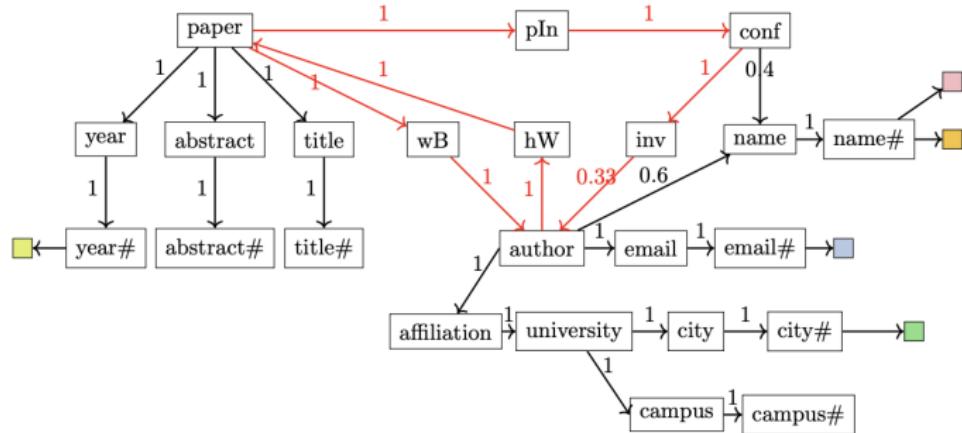
The collection graph

One **collection node** for each equivalence class

One **collection edge** $C_s \rightarrow C_t$:

- Between two collection nodes if a data edge exists
- **Edge transfer factor:** $\frac{|C_t \rightarrow C_s|}{|C_t|}$
- **At-most-one:** 1:1 cardinality

An **entity profile** for each **leaf collection node**: presence of entities



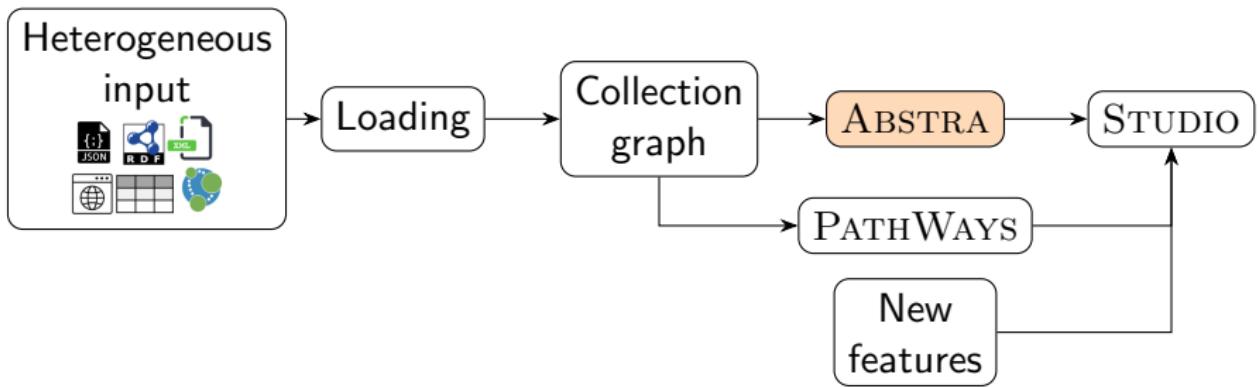
Build an Entity-Relationship model

“A global **understanding, description**”

Nelly Barret
Inria, IPP

Prajna Upadhyay
Inria

Ioana Manolescu
Inria, IPP



ABSTRA: get an overview of the data

Problem statement

How to produce a **compact** and **expressive** description out of **any** dataset?

- ① A **high-level, global description**, easy to grasp for **NTUs**
- ② Focus on the data **meaning** more than the **syntax**

⇒ Retrieve / build **the Entity-Relationship model** behind any dataset

ABSTRA: get an overview of the data

Problem statement

How to produce a **compact** and **expressive** description out of **any** dataset?

- ① A **high-level, global description**, easy to grasp for **NTUs**
- ② Focus on the data **meaning** more than the **syntax**

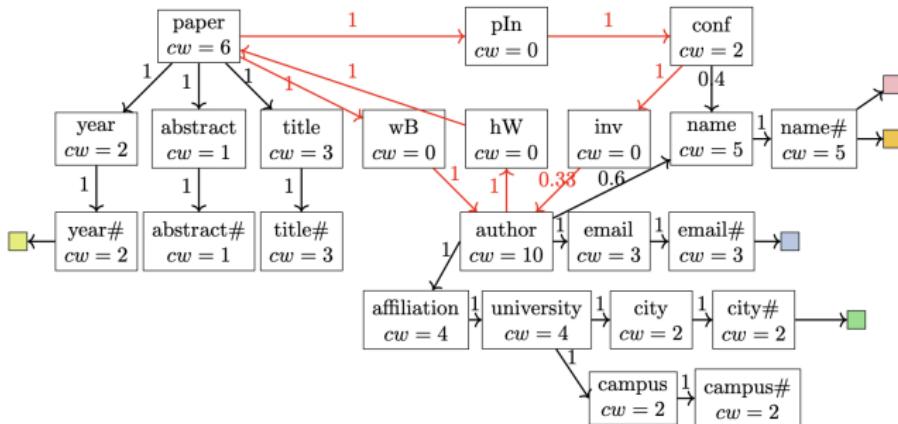
⇒ Retrieve / build **the Entity-Relationship model** behind any dataset

	Data Summarization	Schema inference	Abstra
Several data formats	✗	~	✓
Content and structure	~	~	✓
No syntactic detail	✓	✗	✓
First-sight discovery	~	✗	✓

Main collections selection

Election of few main collections, representing mostly the dataset

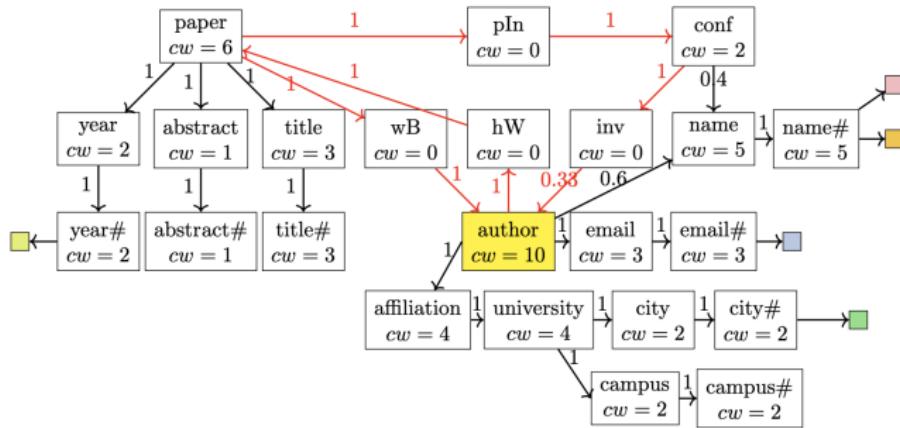
- ① Assign a **weight** to each collection
- ② While less than E_{max} main collections or data coverage $< cov_{min}$
 - ① Pick C^* , the next heaviest collection
 - ② Compute the **boundary** of C^*
 - ③ **Update** the collection graph to reflect the selection of C^*
 - ④ Recompute the weights



Collections weights, boundaries and graph updates

Collection weight

- W_{desc_k}
- W_{leaf_k}
- W_{DAG}
- $W_{PageRank}$
- $W_{dwPageRank}$



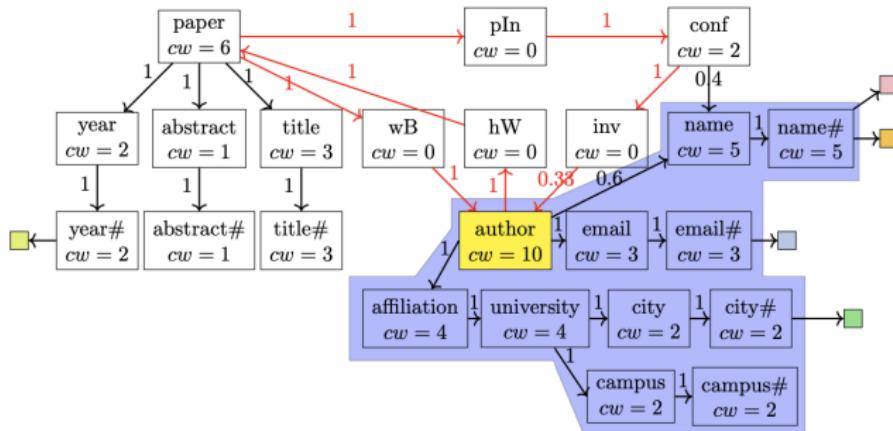
Collections weights, boundaries and graph updates

Collection weight

- w_{desc_k}
- w_{leaf_k}
- w_{DAG}
- $w_{PageRank}$
- $w_{dwPageRank}$

Boundary

- $bound_{desc}$
- $bound_{leaf}$
- $bound_{DAG}$
- $bound_{flood}$
- $bound_{acyclic-flood}$



Collections weights, boundaries and graph updates

Collection weight

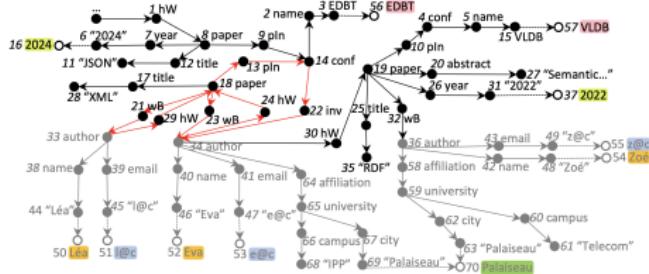
- w_{desc_k}
- w_{leaf_k}
- w_{DAG}
- $w_{PageRank}$
- $w_{dwPageRank}$

Boundary

- $bound_{desc}$
- $bound_{leaf}$
- $bound_{DAG}$
- $bound_{flood}$
- $bound_{acyclic-flood}$

Graph update

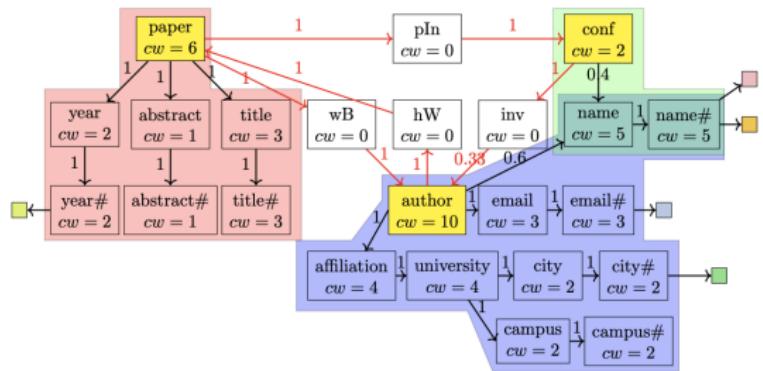
- $update_{boolean}$
- $update_{exact}$



Find relationships between main collections

Possible relationships

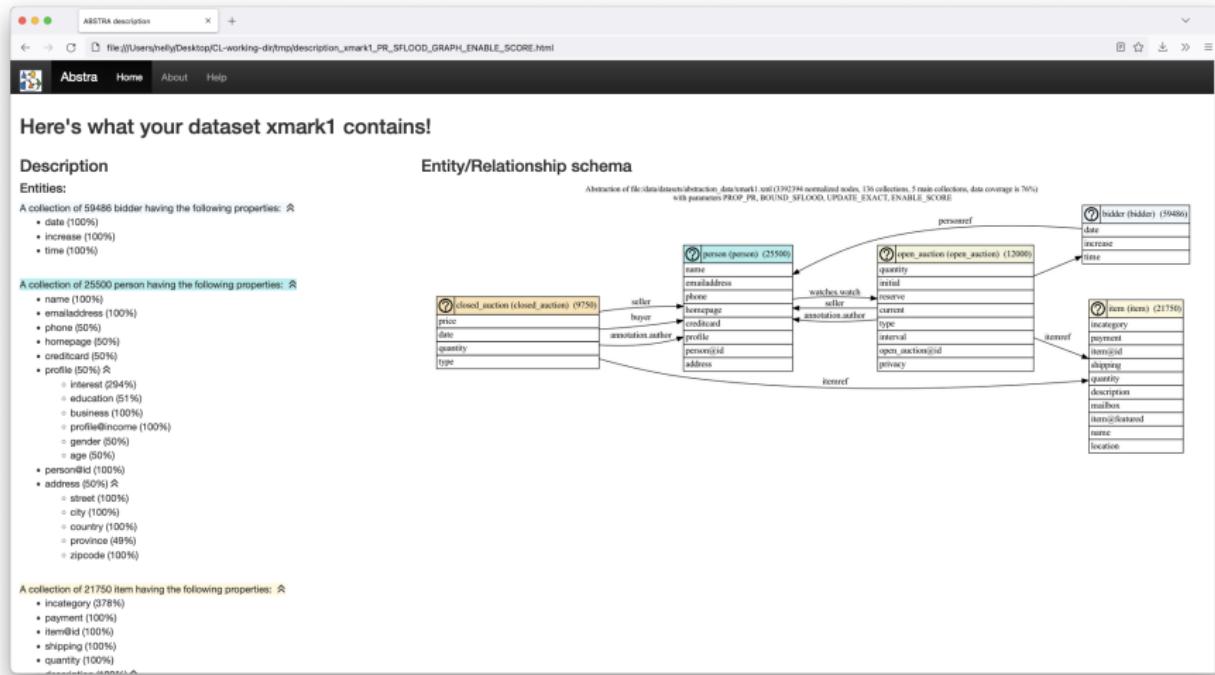
The **set of relationships** connecting a pair of collections is the set of their paths.



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

The final output in ABSTRA

<https://team.inria.fr/cedar/projects/abstra/>



Enumerate entity paths

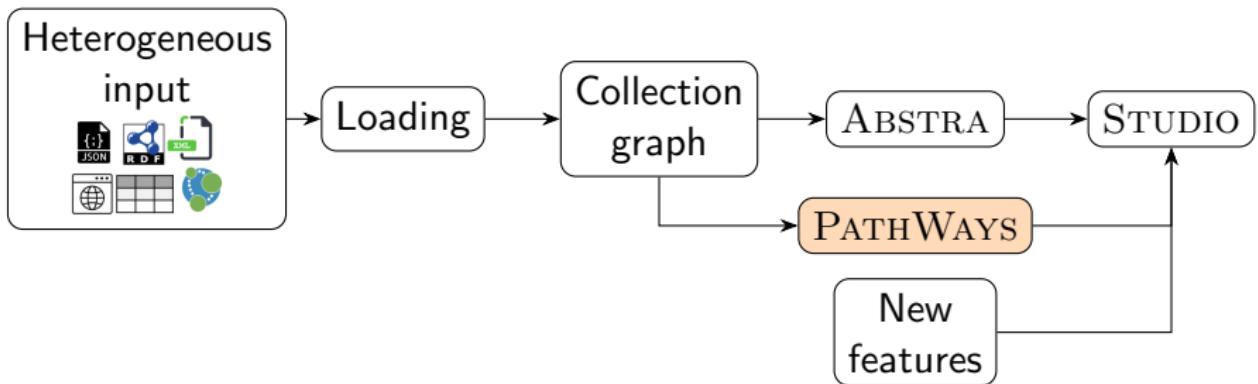
“Interesting **entity connections**”

Nelly Barret
Inria, IPP

Antoine Gauquier
IMT

Jia Jean Law
IPP

Ioana Manolescu
Inria, IPP



PATHWAYS: find interesting connections in the data

Problem statement

How to **interactively** explore **entity connections** in **heterogeneous datasets**?

- ① No query writing, nor prior knowledge
- ② A **tabular, high-level output**, easy to grasp for **NTUs**
- ③ Do it **efficiently** even if the data graph is large

⇒ Connect **named entities** (People, Places, ...) **in and across** datasets.

PATHWAYS: find interesting connections in the data

Problem statement

How to **interactively** explore **entity connections** in **heterogeneous datasets**?

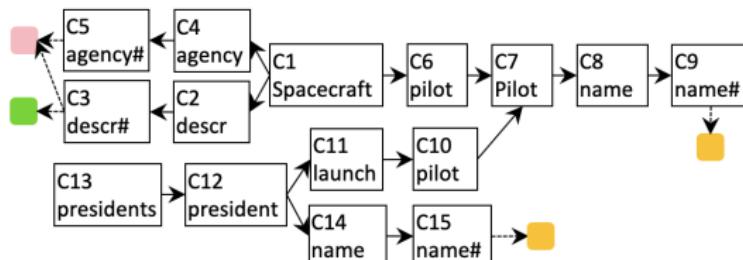
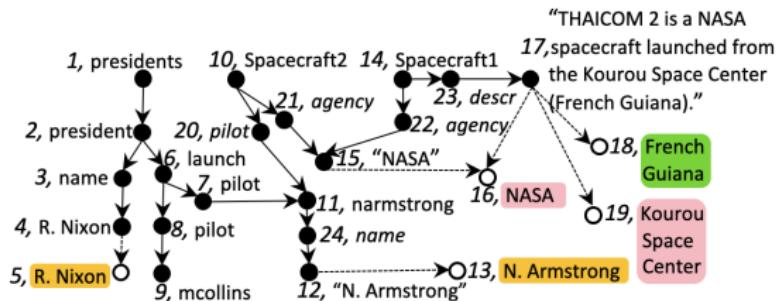
- ① No query writing, nor prior knowledge
- ② A **tabular, high-level output**, easy to grasp for **NTUs**
- ③ Do it **efficiently** even if the data graph is large

⇒ Connect **named entities** (People, Places, ...) **in and across** datasets.

	Keyword search	Graph query	Reachability query	PATHWAYS
No query writing	✓	✗	✗	✓
Tabular output	~	~	✗	✓
Efficient	✗	✓	✓	✓

Scenario and terminology

- A **data (entity) path** is a path in the data graph
- A **collection (entity) path** is a path in the collection graph
- The evaluation of a collection path leads to a set of data paths



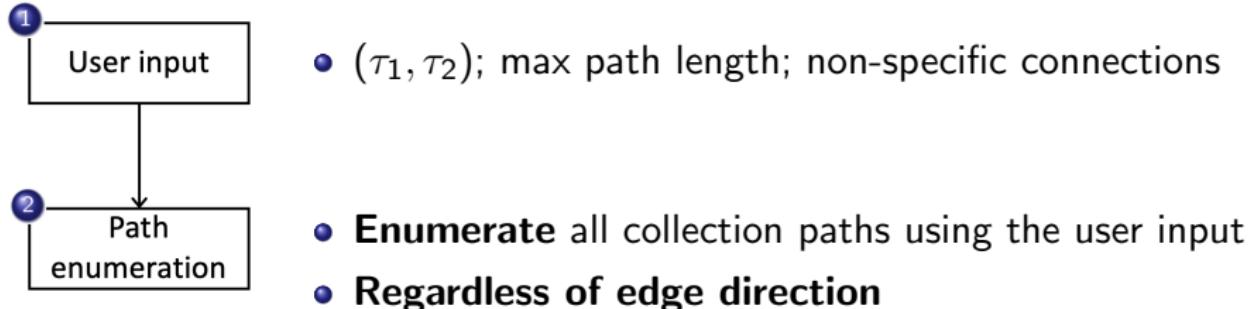
Collection (entity) path enumeration

1

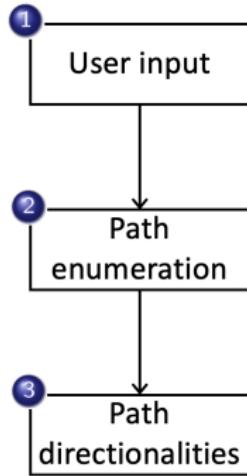
User input

- (τ_1, τ_2) ; max path length; non-specific connections

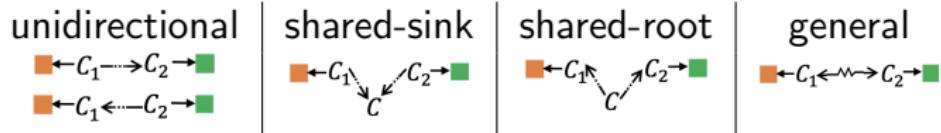
Collection (entity) path enumeration



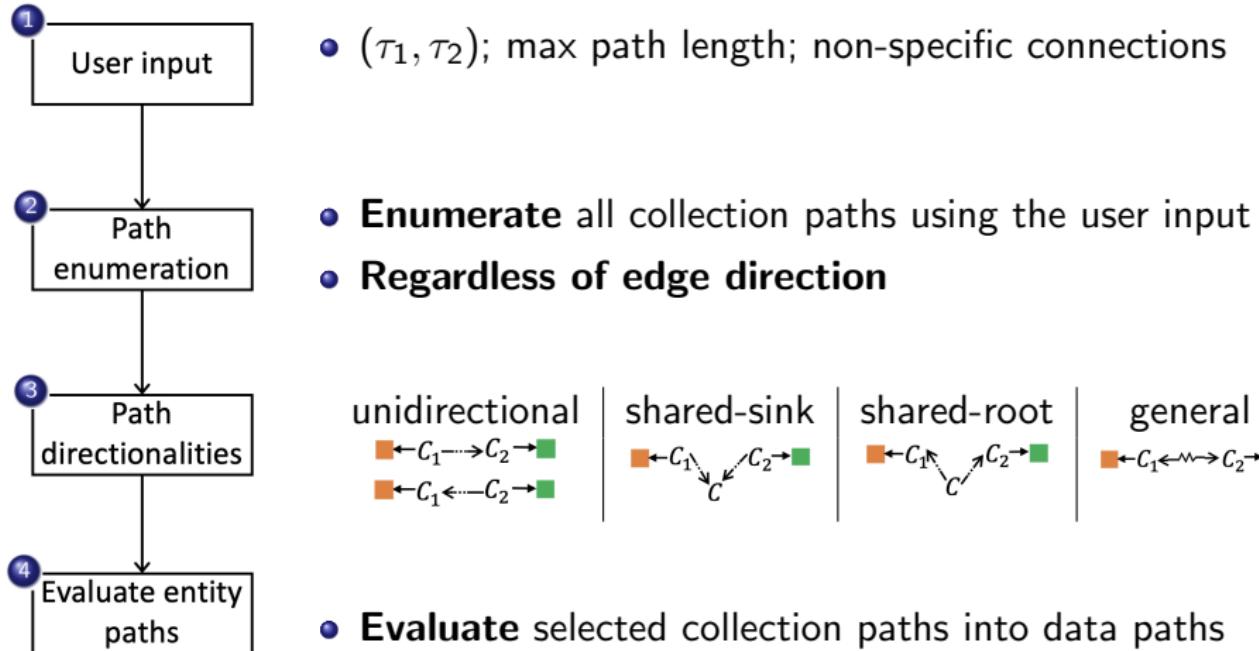
Collection (entity) path enumeration



- (τ_1, τ_2) ; max path length; non-specific connections
- **Enumerate** all collection paths using the user input
- **Regardless of edge direction**

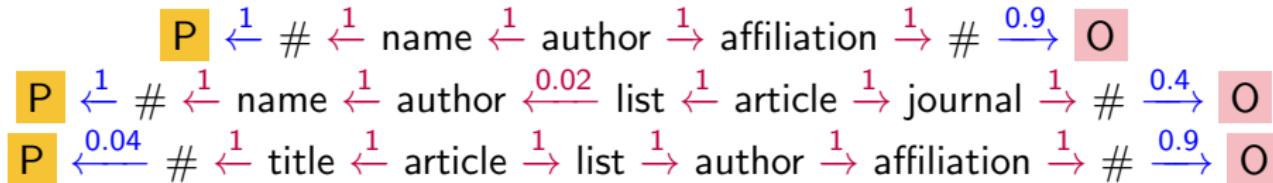


Collection (entity) path enumeration



Collection paths interestingness

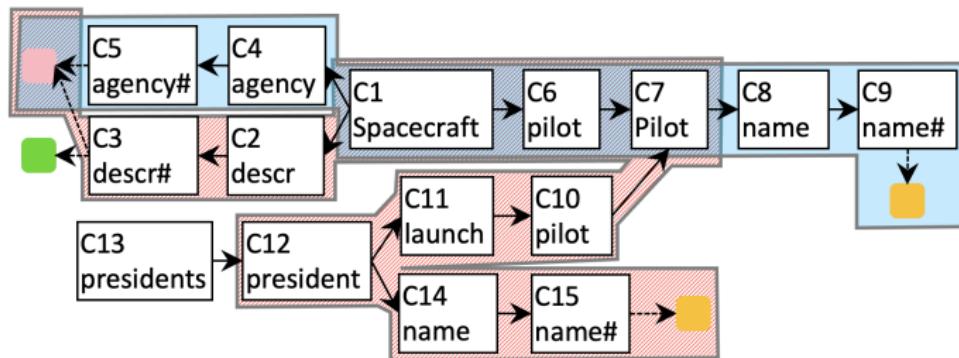
- Many enumerated paths are non-interesting
- Humans can “see/feel it”
- How to quantify the **interestingness** of a path?
 - ① Compute the **reliability** r of its extracted NEs
 - How many NEs of a given type are extracted from a leaf collection?
 - ② Compute the **force** f of each structural collection edge
 - What is the maximal data edge cardinality behind a collection edge?
 - ③ Rank paths on their reliability, then their force
 - ④ Take the top- k or those having $r \geq \theta$



Optimized data paths computation (1/2)

Assumption: enumerated collection paths (largely) overlap

- There exist common sub-paths between collection paths
- Common sub-paths should be evaluated only once as views
 - Saves computation time
- Collection paths are rewritten using views
 - Reduces the number of joins



Optimized data paths computation (2/2)

Greedily select the most profitable views to materialize

Input: collection paths \mathcal{P} , candidate views \mathcal{V}

Output: a set of views, a set of rewritings

① While there are some $v \in \mathcal{V}$:

- ① For each pair (p, v) , compute $ben(p, v) \leftarrow costEval(p) - costEval(p, v)$
- ② Store v_{max} , the view maximizing $ben(v) \leftarrow \sum ben(p, v) - costMat(v)$
- ③ For each path p , rewrite it, if possible, using v_{max}

■ \leftarrow agency# \leftarrow agency \leftarrow Spacecraft \rightarrow pilot \rightarrow Pilot \rightarrow name \rightarrow name# \rightarrow ■
■ \leftarrow agency# \leftarrow agency \leftarrow v \rightarrow name \rightarrow name# \rightarrow ■

```
SELECT le.label, C5.label, C4.label, v.C1label, v.C6label, v.C7label, C8.label, C9.label, re.label  
FROM nEntities le, nodes C5, edges C4, view v, edges C8, nodes C9, nEntities re  
WHERE le.leafId=C5.id AND C4.t=C5.id AND C4.s=v.C1id AND C8.s=v.C7id AND C8.t=C9.id AND re.leafId=C9.id  
AND le.type = ■ AND re.type = ■;
```

Data path results in PathWAYS

<https://team.inria.fr/cedar/projects/pathways/>

PathWays +

localhost:8080/gui/pathways

PathWays Home About Help

Load a PathWays result from database

pathways_pubmedcoi

- (PERSON, ORGANIZATION), max 100 paths of max size 20
- (PERSON, LOCATION), max 100 paths of max size 20

Run PathWays on a dataset

Enter a database name:

Left entity type:

Right entity type:

Run Pathways

Result

Sort queries by length Sort queries by number of associated data paths Hide/show queries without associated data paths

▶ Name#val — Name — Author — AuthorList — PubmedArticle — CoiStatement — CoiStatement#val (860 data paths)

ID	Name#val ▲ ▼	Name ▲ ▼	Author ▲ ▼	AuthorList ▲ ▼	PubmedArticle ▲ ▼	CoiStatement ▲ ▼	CoiStatement#val ▲ ▼
2901	Giampiero Mazzaglia	Name	Author	AuthorList	PubmedArticle	CoiStatement	... Bayer ...
2931	Giampiero Mazzaglia	Name	Author	AuthorList	PubmedArticle	CoiStatement	... Pfizer ...
5531	Paolo Angelo Cortesi	Name	Author	AuthorList	PubmedArticle	CoiStatement	... Bayer ...
5561	Paolo Angelo Cortesi	Name	Author	AuthorList	PubmedArticle	CoiStatement	... Pfizer ...

▶ CoiStatement#val — CoiStatement — PubmedArticle — AuthorList — Author — Affiliation — Affiliation#val (480 data paths)

▶ Name#val — Name — Author — AuthorList — PubmedArticle — ArticleTitle — ArticleTitle#val (8 data paths)

▶ Name#val — Name — Author — Affiliation — Affiliation#val (71 data paths)

▶ CoiStatement#val — CoiStatement — PubmedArticle — ArticleTitle — ArticleTitle#val (12 data paths)

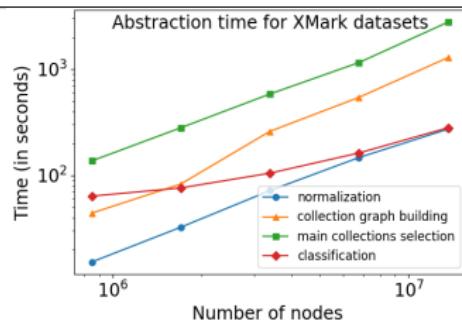
Authors: Nelly Barret @ Inria, Antoine Gauquier @ IMT Nord Europe, Jean Law @ Ecole Polytechnique, Ioana Manolescu @ Inria

Main contact: nelly.barret@inria.fr

Quick overview of experiments

On widely-used **open data formats**: JSON, RDF, XML and PG.

ABSTRA



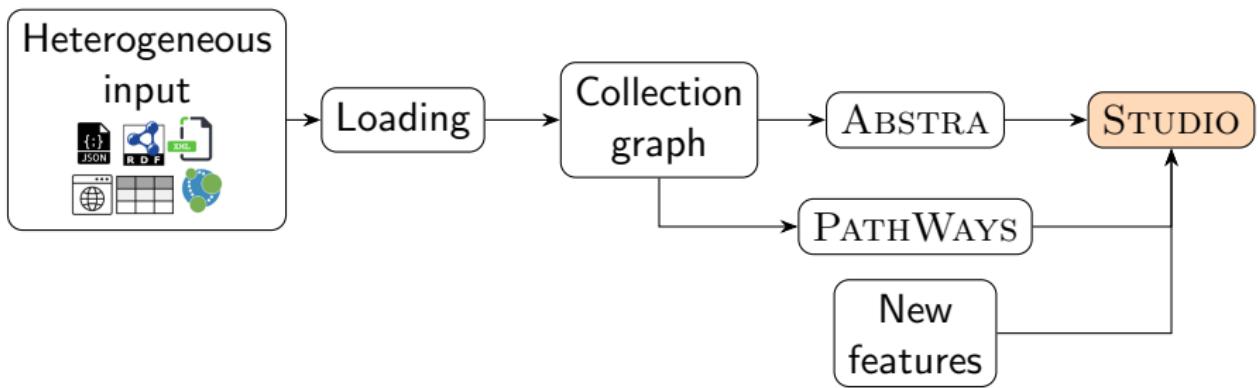
PATHWAYS

(τ_1, τ_2)	T_0	$T = T_R + T_{Q_{NV}}$	$s = T_0/T$
(τ_P, τ_O)	250.36	4.10	61x
(τ_P, τ_L)	37.29	19.38	2x
(τ_L, τ_O)	151.29	20.47	7x
(τ_P, τ_P)	152.59	44.27	3x
(τ_L, τ_L)	169.64	71.63	2x
(τ_O, τ_O)	317.92	23.24	13x

- User study
- Comparison to schemas

- # paths: 0 to very high
- Filter spurious paths
(path interestingness)

Future work, takeaways and open questions



Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

- French media are interested (DataJournos, CFI)

Connection Studio Projects

The screenshot shows a user interface for managing projects in Connection Studio. At the top, there's a blue header bar with the title "Connection Studio Projects". On the left, there's a dropdown menu labeled "Sort by Project name". On the right, there are two buttons: "END SESSION" and "CREATE A PROJECT". Below the header, there's a grid of seven project cards, each with a close button ("X") in the top right corner and a "MANAGE" button at the bottom.

Project Name	Description	Created On	Latest File Addition
Project Cac	1 files Created on: 2023-07-13 11:32:13 Latest file addition: 2023-07-13 11:32:13		
Project Cac40	1 files Created on: 2023-07-05 16:12:38 Latest file addition: 2023-07-05 16:12:38		
Project Hatvp Cac	2 files Created on: 2023-07-11 16:03:48 Latest file addition: 2023-07-11 16:39:39		
Project Hatvp Cac40	2 files Created on: 2023-07-05 15:46:07 Latest file addition: 2023-07-05 16:25:52		
Project Hatvpsmall	No files uploaded yet, add one!		
Project Pubmed	1 files Created on: 2023-07-05 09:46:07 Latest file addition: 2023-07-05 09:46:07		
Project Recac40	1 files Created on: 2023-07-12 23:24:56 Latest file addition: 2023-07-12 23:24:56		

Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

The screenshot shows the Connection Studio interface. At the top, there's a blue header bar with the title "Connection Studio" in large white font, and "Uploaded files" to its right. On the far left of the header is a "Explore" link. To the right of the title, there's a small British flag icon followed by the text "Project: Hatvp Cac". Below the header is a sub-header "Uploaded files". On the left side of the main area, there's a button labeled "DISPLAY ADVANCED OPTIONS" with a gear icon. The main area is a table with two rows of data. The columns are labeled "ID", "File", "Path", "Type", and "Creation date". The first row contains file "hatvp-cleaned.xml" (Type: XML, Creation date: 2023-07-11 16:03:48+02), and the second row contains file "Cac40.csv" (Type: CSV, Creation date: 2023-07-11 16:39:39+02). Each row has a small circular icon and a square icon to its right.

ID	File	Path	Type	Creation date
1	hatvp-cleaned.xml	file:///Users/nelly/Documents/boulot/theseNelly/connection-lens/..//connection-studio/demo-CFI	XML	2023-07-11 16:03:48+02
2	Cac40.csv	file:///Users/nelly/Documents/boulot/theseNelly/connection-lens/..//connection-studio/demo-CFI	CSV	2023-07-11 16:39:39+02

Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

Explore

Connection Studio Statistics

Project: Hatvp Cac

Entities distribution by type

Total 217451

Identified entities

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

Entity cloud

GFA [Données non publiées] SCEA [Données non publiées]
SARL [Données non publiées] EDUCATION NATIONALE
22/06/2022 Conseil d'Administration Sénateur
Conseil de surveillance VICE PRESIDENT France
02/07/2021 Comité syndical Membre CA
Communauté de communes 2020 16/07/2020
2008 Vice CA 2020 09/20 AG SCINEANT CONSEIL
17/07/2020 28/06/2020
15/03/2020 15/07/2020 Comité 01/07/2021 Maire 07/07/2020
06/20 10/07/2020 02/20 Communauté
RÉGIONAL M 2026 Conseil départemental
Régional 11/20 Métropole 27/06/2021 27/09/2020
Bureau Membre 10/20/08/20 2016 Education Nationale
Conseil d'administration 01/20 03/20 04/07/2020 08/07/2020
MEMBRE CA 2019 2017 Mme 11/07/2020
Retraitee 12/20 2022 Paris 05/20 03/07/2020 SEM
Education nationale 07/20 néant Conseil 06/07/2020
Retraitee 2021 2014 2012 PRESIDENTE
SCI [Données non publiées] 2015 09/07/2020
SDIS Conseillère Départementale
Conseil Régional Député 19/06/2022 sci [Données non publiées]
CONSEIL D'ADMINISTRATION CCAS
24/09/2017 Conseil de Surveillance
VICE PRESIDENTE 30/06/2020

Nelly Barret (Inria)

Semi-structured Data Exploration

January 25, 2024

35 / 40

Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

The screenshot shows the STUDIO interface with five defined data paths:

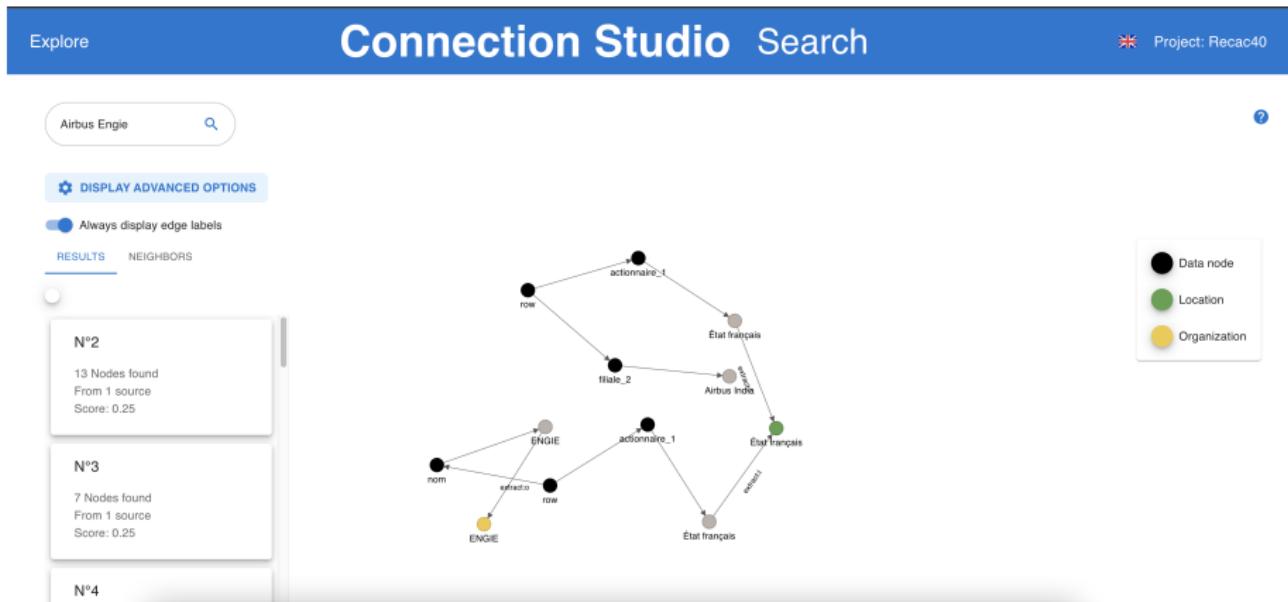
- Path 1: declaration.general.declarer.name#val. Starting variable: decla. Ending variable: deputyName. Join type: Required.
- Path 2: declaration.financialInterest.items.item. Starting variable: decla. Ending variable: item. Join type: Optional.
- Path 3: item.company#val.extract:o. Starting variable: item. Ending variable: companyName. Join type: Required.
- Path 4: item.nbShares#val. Starting variable: item. Ending variable: nbShares. Join type: Optional.
- Path 5: row.company_name.#val.extract:o. Starting variable: csvline. Ending variable: companyName. Join type: Required.

Below the paths is a table with the following data:

COLUMNS					
decla	deputynname	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

Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data



Future work: STUDIO

STUDIO: a data lake for ingesting, querying, cleaning and understanding heterogeneous data

The screenshot shows the STUDIO interface with the following configuration:

- Extraction model: Stanford Extractor
- File language: English
- Extraction policy: declaration.general.declarant.nom#val Person, declarations.declaration.origine#val NoExtract
- Split long texts: False

A blue button labeled "SAVE PARAMETERS" is visible on the right.

The results table displays the following data:

ID	Text
2998	als thom
2998	alsthom
2998	alsthom atlantique
2998	alstom

The cell containing "als thom" is highlighted with a blue border.

Takeaways and open questions

- ABSTRA: a dataset abstraction system for heterogeneous data
- PATHWAYS: an entity-focused exploration system
- STUDIO: a user-oriented data lake for data exploration

ABSTRA	PATHWAYS	STUDIO
EDBT 2024	ADBIS 2023	CoopIS 2023
		

Further opportunities

Nelly BARRET

✉ nelly.barret@inria.fr

💻 <https://pages.saclay.inria.fr/nelly.barret/>

🏡 Inria Saclay & Institut Polytechnique de Paris
Palaiseau



POLITECNICO
MILANO 1863

