KNOWLEDGE GRAPH REFINEMENT
KEY DISCOVERY AND LINK INVALIDATION

FATIHA SAÏS

LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

Joint work with: N. Pernelle, L. Papaleo, J. Raad and D. Symeonidou

3ÈME JOURNÉE RI-IA SOUTENUE PAR L’AFIA ET ARIA, PARIS 2019
OUTLINE

• Introduction
  ▪ Linked Data
  ▪ Knowledge graphs
  ▪ Knowledge graph refinement

• Key discovery

• Link invalidation

• Conclusion
Linked Data - Datasets under an open access
- 1,139 datasets
- over 100B triples
- about 500M links
- several domains

Ex. DBPedia : 1.5 B triples

An RDF Graph is a set of triples.

- Its nodes are (labelled by) the subjects and objects appearing in the triples.
- Its edges are labelled by the properties

```plaintext
http://dbpedia.org/resource/CNRS

dbo:president

http://dbpedia.org/resource/Alain_Fuchs

dbo:birthPlace

http://dbpedia.org/resource/Chimie_Physique

http://dbpedia.org/resource/Lausane

dbo:domainOf

dbo:worksOn

dbo:created

xsd:date"1939-10-19"
```
NEED OF KNOWLEDGE
The role of knowledge in AI

[Artificial Intelligence 47 (1991)]

ON THE THRESHOLDS OF KNOWLEDGE

Douglas B. Lenat
MCC
3500 W. Balcones Center
Austin, TX 78759

Edward A. Feigenbaum
Computer Science Department
Stanford University
Stanford, CA 94305

Abstract

We articulate the three major findings of AI to date:
(1) The Knowledge Principle: if a program is to perform
complex task well, it must know a great deal about the
world in which it operates. (2) A plausible extension of
that principle, called the Breadth Hypothesis: there are two
additional abilities necessary for intelligent behavior in
unexpected situations: falling back on increasingly general
knowledge, and analogizing to specific but far-flung
knowledge. (3) AI as Empirical Inquiry: we must test our
ideas experimentally, on large problems. Each of these
three hypotheses proposes a particular threshold to cross,
which leads to a qualitative change in emergent intelligence.
Together, they determine a direction for future AI research.

The knowledge principle: “if a program is
to perform a complex task well, it must
know a great deal about the world in
which it operates.”
ONTOMETRY, A DEFINITION

“An ontology is an explicit, formal specification of a shared conceptualization.”

[Thomas R. Gruber, 1993]

Conceptualization: abstract model of domain related expressions
Specification: domain related
Explicit: semantics of all expressions is clear
Formal: machine-readable
Shared: consensus (different people have different perceptions)
SEMANTIC WEB: ONTOLOGIES

RDFS – Resource Description Framework Schema
- Lightweight ontologies

OWL – Web Ontology Language
- Expressive ontologies

Source: https://it.wikipedia.org/wiki/File:W3C-Semantic_Web_layerCake.png
OWL – WEB ONTOLOGY LANGUAGE

- **Classes**: concepts or collections of objects (individuals)
- **Properties**:
  - owl:DataTypeProperty (attribute)
  - owl:ObjectProperty (relation)
- **Individuals**: ground-level of the ontology (instances)

- **Axioms**
  - owl:subClassOf
  - owl:subPropertyOf
  - owl:inverseProperty
  - owl:FunctionalProperty
  - owl:minCardinality
  - …
ONTOLOGY LEVELS: KNOWLEDGE ENGINEERING VIEW

Conceptual level:
- classes, properties (relations)

Instance level:
- facts (individuals)
KNOWLEDGE
GRAPHS
WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007

- **DBpedia**

2007

- **yago**

2012

- **Google Knowledge Graph**

2013

- **Facebook Graph Search**

2015

- **LinkedIn Graph**

2016

- **Microsoft Graph**

**Academic side**

**Commercial side**
WEB SEARCH WITHOUT KNOWLEDGE GRAPHS
WEB SEARCH WITH KNOWLEDGE GRAPHS
QUESTION ANSWERING WITH KNOWLEDGE GRAPHS

barack obama mother

Ann Dunham

Ann Dunham - Wikipedia
https://en.wikipedia.org/wiki/Ann_Dunham
Stanley Ann Dunham, née le 29 novembre 1942 à Wichita et morte le 7 novembre 1995 à Honolulu, est une anthropologue américaine spécialisée dans l'anthropologie économique et le développement rural. Elle est la mère de Barack Obama, le 44e...

Family of Barack Obama - Wikipedia
The family of Barack Obama, the 44th President of the United States, and his wife Michelle Obama is made up of people of Kenyan (Luo), African-American, and Old Stock...
TOWARDS A KNOWLEDGE-POWERED DIGITAL ASSISTANT

- Natural access and storage of knowledge
- Chat bots
- Personalization
- Emotion

Cortana (Microsoft)
KNOWLEDGE GRAPH ADOPTION [2019]

**KNOWLEDGE GRAPH - DEFINITION**

**RDF Graphs**

- dbo:Museum
  - `rdf:type` http://dbpedia.org/ontology/Museum
  - dbo:museum-of http://fr.dbpedia.org/resource/Musée_du_Louvre
  - dbo:museumName http://fr.dbpedia.org/resource/Musée_du_Louvre

- dbo:Painting
  - `rdf:type` aw:Painting

**Ontology hierarchy**

- Thing
  - is-a dbo:Agent
  - is-a dbo:PopulatedPlace
  - is-a dbo:Person

- dbo:Museum
  - dbo:city
  - dbo:museum-of

- dbo:Painting
  - dbo:author

**Ontology axioms and rules**

- owl:equivalentClass(dbo:Municipality, dbo:Place)
- owl:equivalentClass(dbo:Place, dbo:Wikidata:Q532)
- owl:equivalentClass(dbo:Village, dbo:PopulatedPlace)
- owl:disjointClass(dbo:PopulatedPlace, dbo:Municipality)
- owl:disjointClass(dbo:PopulatedPlace, dbo:Artist)
- owl:FunctionalProperty(dbo:city)
- owl:InverseFunctionalProperty(dbo:museum-of)

- dbo:birthPlace(X, Y) => dbo:cityenOf(X, Y)
- dbo:parentOf(X, Y) => dbo:child(Y, X)

**Querying (SPARQL)**

PREFIX dbo: <http://dbpedia.org/ontology#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

**Reasoners:** (Pellet, Fact++, Hermit, etc.)

- KG saturation: infer whatever can be inferred from the KG.
- KG consistency checking: no contradictions
- KG repairing
- ...

**PREFIX**

- prefix dbo: <http://dbpedia.org/ontology#>
- prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
### Knowledge Graph Completeness?

<table>
<thead>
<tr>
<th>Name</th>
<th>Instances</th>
<th>Facts</th>
<th>Types</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBpedia (English)</td>
<td>4,806,150</td>
<td>176,043,129</td>
<td>735</td>
<td>2,813</td>
</tr>
<tr>
<td>YAGO</td>
<td>4,595,906</td>
<td>25,946,870</td>
<td>488,469</td>
<td>77</td>
</tr>
<tr>
<td>Freebase</td>
<td>49,947,845</td>
<td>3,041,722,635</td>
<td>26,507</td>
<td>37,781</td>
</tr>
<tr>
<td>Wikidata</td>
<td>15,602,060</td>
<td>65,993,797</td>
<td>23,157</td>
<td>1,673</td>
</tr>
<tr>
<td>NELL</td>
<td>2,006,896</td>
<td>432,845</td>
<td>285</td>
<td>425</td>
</tr>
<tr>
<td>OpenCyc</td>
<td>118,499</td>
<td>2,413,894</td>
<td>45,153</td>
<td>18,526</td>
</tr>
<tr>
<td>Google’s Knowledge Graph</td>
<td>570,000,000</td>
<td>18,000,000,000</td>
<td>1,500</td>
<td>35,000</td>
</tr>
<tr>
<td>Google’s Knowledge Vault</td>
<td>45,000,000</td>
<td>271,000,000</td>
<td>1,100</td>
<td>4,469</td>
</tr>
<tr>
<td>Yahoo! Knowledge Graph</td>
<td>3,443,743</td>
<td>1,391,054,990</td>
<td>250</td>
<td>800</td>
</tr>
</tbody>
</table>

KNOWLEDGE GRAPH CORRECTNESS?

About: Donald Trump

An Entity of Type: person, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Donald John Trump (born June 14, 1946) is an American businessman, author, television producer, politician, and the Republican Party nominee for President of the United States in the 2016 election. He is the chairman and president of The Trump Organization, which is the principal holding company for his real estate ventures and other business interests. During his career, Trump has built office towers, hotels, casinos, golf courses, an urban development project in Manhattan, and other branded facilities worldwide.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:birthName</td>
<td>Donald John Trump (en)</td>
</tr>
<tr>
<td>dbo:birthPlace</td>
<td>dbr:Queens</td>
</tr>
<tr>
<td></td>
<td>dbr:New_York_City</td>
</tr>
<tr>
<td>dbo:birthYear</td>
<td>1946-01-01 (xsd:date)</td>
</tr>
<tr>
<td>dbo:child</td>
<td>dbr:Donald_Trump_Jr.</td>
</tr>
<tr>
<td></td>
<td>dbr:Tiffany_Trump</td>
</tr>
<tr>
<td></td>
<td>dbr:Eric_Trump</td>
</tr>
<tr>
<td></td>
<td>dbr:Ivanka_Trump</td>
</tr>
<tr>
<td></td>
<td>dbr:Donald_Trump</td>
</tr>
</tbody>
</table>

Donald Trump is the child of himself!
KNOWLEDGE GRAPH REFINEMENT

Completeness  Correctness
** KNOWLEDGE GRAPH REFINEMENT**

- Completeness
- Correctness

- Key discovery
  - Data Linking
  - Data Fusion

- Link Invalidation
  - Contextual identity
OUTLINE

• Introduction
• Key discovery
• Link invalidation
• Conclusion
KEY DISCOVERY FOR DATA LINKING
DATA LINKING

Data linking or Identity link detection consists in detecting whether two descriptions of resources refer to the same real world entity (e.g. person, article, gene).
DATA LINKING

Data linking or Identity link detection consists in detecting whether two descriptions of resources refer to the same real world entity (e.g. person, article, gene).
DATA LINKING

Data linking or Identity link detection consists in detecting whether two descriptions of resources refer to the same real world entity (e.g. person, article, gene).
KEY DISCOVERY FOR DATA LINKING

Rule-based data linking approaches [Saïs et al. 2009, Al Bakri et al. 2015]: need for knowledge to be declared in an ontology language or other languages.

\[ \text{homepage}(X, Y) \land \text{homepage}(Z, Y) \Rightarrow \text{sameAs}(X, Z) \]

Then we may infer:

\begin{align*}
\text{sameAs}(\text{museum11}, \text{museum21}) \\
\text{sameAs}(\text{museum12}, \text{museum22}) \\
\text{sameAs}(\text{museum13}, \text{museum23})
\end{align*}

A key: is a set of properties that uniquely identifies every instance of a class

<table>
<thead>
<tr>
<th>museum11</th>
<th>...</th>
<th>homepage</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.louvre.com">www.louvre.com</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>museum12</th>
<th>...</th>
<th>homepage</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.musee-orsay.fr">www.musee-orsay.fr</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>museum13</th>
<th>...</th>
<th>homepage</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.quai-branly.fr">www.quai-branly.fr</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>museum14</th>
<th>...</th>
<th>homepage</th>
</tr>
</thead>
</table>
KEY DISCOVERY FOR DATA LINKING

Rule-based data linking approaches [Saïs et al. 2009, Al Bakri et al. 2015]: need for knowledge to be declared in an ontology language or other languages.

\[
\text{homepage}(X, Y) \land \text{homepage}(Z, Y) \Rightarrow \text{sameAs}(X, Z)
\]

Then we may infer:

\[
\text{sameAs}(\text{museum11, museum21})
\text{sameAs}(\text{museum12, museum22})
\text{sameAs}(\text{museum13, museum23})
\]

A key: is a set of properties that uniquely identifies every instance of a class

<table>
<thead>
<tr>
<th>museum</th>
<th>homepage</th>
<th>museum</th>
<th>homepage</th>
</tr>
</thead>
<tbody>
<tr>
<td>museum13</td>
<td><a href="http://www.quai-branly.fr">www.quai-branly.fr</a></td>
<td>museum23</td>
<td>...</td>
</tr>
<tr>
<td>museum14</td>
<td>...</td>
<td>museum24</td>
<td>...</td>
</tr>
</tbody>
</table>

How to automatically discover keys from KGs?
KEY VALIDITY: EXACT KEYS

A key is a set of properties that **uniquely identifies** every instance in the data.

<table>
<thead>
<tr>
<th></th>
<th>FirstName</th>
<th>LastName</th>
<th>Birthdate</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>15/02/88</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompson</td>
<td>02/09/75</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>15/02/85</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>25/01/72</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person4</td>
<td>Simon</td>
<td>Roche</td>
<td>06/12/90</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Jane</td>
<td>Ser</td>
<td>15/05/87</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Sara</td>
<td>Khan</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Theo</td>
<td>Martin</td>
<td>06/12/90</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Marc</td>
<td>Blanc</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

*Is [LastName] a key? ✗*

*Is [FirstName, LastName] a key? ✓*

**Exact keys**
A key is a set of properties that uniquely identifies every instance in the data.

<table>
<thead>
<tr>
<th>Person</th>
<th>FirstName</th>
<th>LastName</th>
<th>Birthdate</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>15/02/88</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompson</td>
<td>02/09/75</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>15/02/85</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>25/01/72</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person4</td>
<td>Simon</td>
<td>Roche</td>
<td>06/12/90</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Jane</td>
<td>Ser</td>
<td>15/05/87</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Sara</td>
<td>Khan</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Theo</td>
<td>Martin</td>
<td>06/12/90</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Marc</td>
<td>Blanc</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

Is [FirstName, LastName] a key? ✔

Is [Birthdate] a key? ✗

Is [Birthdate] a key with 2 exceptions? ✔
A key is a set of properties that **uniquely identifies** every instance in the data.

<table>
<thead>
<tr>
<th>Person</th>
<th>FirstName</th>
<th>LastName</th>
<th>Birthdate</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>15/02/88</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompson</td>
<td>02/09/75</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>15/02/85</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>25/01/72</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person4</td>
<td>Simon</td>
<td>Roche</td>
<td>06/12/90</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Jane</td>
<td>Ser</td>
<td>15/05/87</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Sara</td>
<td>Khan</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Theo</td>
<td>Martin</td>
<td>06/12/90</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Marc</td>
<td>Blanc</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

Is [FirstName,LastName] a key? **✔**

Is [Birthdate] a key with 2 exceptions? **✔**

Is [Birthdate and (Profession =“Actor”)] a key? **✔**
OWL2 Semantics

- **A Key for a class:** a combination of properties that uniquely identify each instance of a class:

  \[ \text{hasKey}(CE \ (OPE_1 \ ... \ OPE_m) \ (DPE_1 \ ... \ DPE_n)) \]

  \[
  \forall X, \forall Y, \forall Z_1, \ldots, Z_n, \forall T_1, \ldots, T_m \wedge \text{ce}(X) \wedge \text{ce}(Y) \bigwedge_{i=1}^{n} (\text{ope}_i(X, Z_i) \wedge \text{ope}_i(Y, Z_i)) \\
  \bigwedge_{i=1}^{m} (\text{dpe}_i(X, T_i) \wedge \text{dpe}_i(Y, T_i)) \Rightarrow X = Y
  \]

**owl:hasKey(Book(Author) (Title))** means:

\[
\text{Book}(x_1) \wedge \text{Book}(x_2) \wedge \\
\text{Author}(x_1, y) \wedge \text{Author}(x_2, y) \wedge \text{Title}(x_1, w) \wedge \text{Title}(x_2, w) \Rightarrow \text{sameAs}(x_1, x_2)
\]
Related Work

- Approaches in relational databases are not applicable
  - Closed world assumption
  - Do not consider multi-valued properties
  - No ontologies (semantics cannot be used)

Contributions

- **KD2R [ISSW 2011, JWS 2013]:** exact key discovery
  - Danai Symeonidou PhD, Qualinca ANR Project (2012-2016)
- **SAKey [ISWC 2014]:** n-almost key discovery
  - Danai Symeonidou PhD, Qualinca ANR Project (2012-2016)
- **VICKEY [ISWC 2017]:** conditional key discovery
  - Collaboration with INRA, Telecom ParisTech and Aalborg University (Danemark).
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least $2^n$ property combinations
  - need of efficient filtering and prunings
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least $2^n$ property combinations
  - need of efficient filtering and prunings
- For each combination scan all the instances
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least $2^n$ property combinations
  - need of efficient filtering and prunings
- For each combination scan all the instances
  - maximal non-keys derive minimal keys

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Phone</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>0169154259</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompson</td>
<td>0169154226</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>0425154012</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>0425154009</td>
</tr>
<tr>
<td>Person5</td>
<td>Simon</td>
<td>Roche</td>
<td>0321455823</td>
</tr>
<tr>
<td>Person6</td>
<td>Jane</td>
<td>Ser</td>
<td>0425462914</td>
</tr>
<tr>
<td>Person7</td>
<td>Sara</td>
<td>Khan</td>
<td>0425462915</td>
</tr>
<tr>
<td>Person8</td>
<td>Theo</td>
<td>Martin</td>
<td>0321455823</td>
</tr>
<tr>
<td>Person9</td>
<td>Marc</td>
<td>Blanc</td>
<td>0169154228</td>
</tr>
</tbody>
</table>

Is [LastName] a non-key? ➔ scan only a part of the data
SAKEY: N-ALMOST KEY DISCOVERY

- **SAKey** allows $n$ exceptions in the data
- **Exception set** $E_P$: set of instances that share values for the set of properties $P$
- **n-almost key**: a set of properties where $|E_P| \leq n$
- **n-non key**: a set of properties where $|E_P| \geq n+1$

n=4

<table>
<thead>
<tr>
<th>5-non keys</th>
<th>4-almost keys</th>
</tr>
</thead>
</table>

- All sets of properties that contain at least 5 exceptions
- All sets of properties that contain at most 4 exceptions
SAKEY: EVALUATION

Evaluation on 13 different datasets (OAEI, Qualinca project, Dbpedia, …)

Scalability
- Big classes (dbo:NaturalPlace more than 16 million triples and 243 properties): non-key discovery in 1min and key derivation 5min)

Quality
- Data linking with SAKey keys: obtains close or better results than expert keys
- Exceptions: important increase of recall and weak decrease of the precision.

<table>
<thead>
<tr>
<th># exceptions</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 1</td>
<td>25.6%</td>
<td>100%</td>
<td>41%</td>
</tr>
<tr>
<td>2, 3</td>
<td>47.6%</td>
<td>98.1%</td>
<td>64.2%</td>
</tr>
<tr>
<td>4, 5</td>
<td>47.9%</td>
<td>96.3%</td>
<td>63.9%</td>
</tr>
<tr>
<td>6, ..., 16</td>
<td>48.1%</td>
<td>96.3%</td>
<td>64.1%</td>
</tr>
<tr>
<td>17</td>
<td>49.3%</td>
<td>82.8%</td>
<td>61.8%</td>
</tr>
</tbody>
</table>

Tool available at: https://www.lri.fr/sakey
To discover even more keys in a dataset
VICKEY: CONDITIONAL-KEY DISCOVERY

To discover even more keys in a dataset

**Conditional key:** a key, valid for instances of a class satisfying a specific condition

<table>
<thead>
<tr>
<th>Instance</th>
<th>FirstName</th>
<th>LastName</th>
<th>Gender</th>
<th>Lab</th>
<th>Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>instance1</td>
<td>Claude</td>
<td>Dupont</td>
<td>Female</td>
<td>Paris-Sud</td>
<td>France</td>
</tr>
<tr>
<td>instance2</td>
<td>Claude</td>
<td>Dupont</td>
<td>Male</td>
<td>Paris-Sud</td>
<td>Belgium</td>
</tr>
<tr>
<td>instance3</td>
<td>Juan</td>
<td>Rodríguez</td>
<td>Male</td>
<td>INRA</td>
<td>Spain, Italy</td>
</tr>
<tr>
<td>instance4</td>
<td>Juan</td>
<td>Salvez</td>
<td>Male</td>
<td>INRA</td>
<td>Spain</td>
</tr>
<tr>
<td>instance5</td>
<td>Anna</td>
<td>Georgiou</td>
<td>Female</td>
<td>INRA</td>
<td>Greece, France</td>
</tr>
<tr>
<td>instance6</td>
<td>Pavlos</td>
<td>Markou</td>
<td>Male</td>
<td>Paris-Sud</td>
<td>Greece</td>
</tr>
<tr>
<td>instance7</td>
<td>Marie</td>
<td>Legendre</td>
<td>Female</td>
<td>INRA</td>
<td>France</td>
</tr>
</tbody>
</table>
To discover even more keys in a dataset

**Conditional key:** a key, valid for instances of a class satisfying a specific condition

<table>
<thead>
<tr>
<th>Instance</th>
<th>FirstName</th>
<th>LastName</th>
<th>Gender</th>
<th>Lab</th>
<th>Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>instance1</td>
<td>Claude</td>
<td>Dupont</td>
<td>Female</td>
<td>Paris-Sud</td>
<td>France</td>
</tr>
<tr>
<td>instance2</td>
<td>Claude</td>
<td>Dupont</td>
<td>Male</td>
<td>Paris-Sud</td>
<td>Belgium</td>
</tr>
<tr>
<td>instance3</td>
<td>Juan</td>
<td>Rodríguez</td>
<td>Male</td>
<td>INRA</td>
<td>Spain, Italy</td>
</tr>
<tr>
<td>instance4</td>
<td>Juan</td>
<td>Salvez</td>
<td>Male</td>
<td>INRA</td>
<td>Spain</td>
</tr>
<tr>
<td>instance5</td>
<td>Anna</td>
<td>Georgiou</td>
<td>Female</td>
<td>INRA</td>
<td>Greece, France</td>
</tr>
<tr>
<td>instance6</td>
<td>Pavlos</td>
<td>Markou</td>
<td>Male</td>
<td>Paris-Sud</td>
<td>Greece</td>
</tr>
<tr>
<td>instance7</td>
<td>Marie</td>
<td>Legendre</td>
<td>Female</td>
<td>INRA</td>
<td>France</td>
</tr>
</tbody>
</table>

\{LastName\} is a key under the condition \{Lab=INRA\}

Algorithm: discovers minimal conditional keys from maximal non-keys (SAKey)
VICKEY: EVALUATION

Goal: evaluate the quality of data linking using:

- Classical keys discovered by SAKey
- Conditional keys discovered by VICKEY
- Both classical keys and conditional keys

Use of **Yago** and **Dbpedia** datasets (**9 classes**): Actor, Album, Book, Film, Mountain, Museum, Organization, Scientist, University
VICKEY: EVALUATION

Goal: evaluate the quality of data linking using:

- Classical keys discovered by SAKey
- Conditional keys discovered by VICKEY
- Both classical keys and conditional keys

Use of Yago and Dbpedia datasets (9 classes): Actor, Album, Book, Film, Mountain, Museum, Organization, Scientist, University

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actor</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAKey Keys</td>
<td>0.27</td>
<td>0.99</td>
<td>0.43</td>
</tr>
<tr>
<td>Conditional keys</td>
<td>0.57</td>
<td>0.99</td>
<td>0.73</td>
</tr>
<tr>
<td>SAKey Keys + Conditional keys</td>
<td>0.6</td>
<td>0.99</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Album</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAKey Keys</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>Conditional keys</td>
<td>0.15</td>
<td>0.99</td>
<td>0.26</td>
</tr>
<tr>
<td>SAKey Keys + Conditional keys</td>
<td>0.15</td>
<td>0.99</td>
<td>0.26</td>
</tr>
<tr>
<td><strong>Film</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAKey Keys</td>
<td>0.04</td>
<td>0.99</td>
<td>0.08</td>
</tr>
<tr>
<td>Conditional keys</td>
<td>0.38</td>
<td>0.96</td>
<td>0.54</td>
</tr>
<tr>
<td>SAKey Keys + Conditional keys</td>
<td>0.39</td>
<td>0.98</td>
<td>0.55</td>
</tr>
</tbody>
</table>
KEY DISCOVERY: SUMMARY

- **Different methods** (KD2R, SAKey, VICKEY, Linkkey [Atencia et al. 2014], Rocker [Soru et al. 2015]) that discover three different kinds of keys

- **Relevance** of exact-keys, n-almost and conditional keys for data linking

- Relying on the strategy of **non-key search first** prevents the use of well-known quality metrics to prune the search space (e.g., support)

Possible improvements

- **More expressive keys** such as key graphs or referring expressions may be discovered

- **Different key semantics** can co-exist: how to choose the good key semantics using the data characteristics (e.g. completeness)
OUTLINE

• Introduction
• Key discovery
• Link invalidation
• Conclusion
IDENTITY PROBLEM

- [Halpin et al. 2010] showed that 37% of owl:sameAs links randomly selected among 250 identity links between books were incorrect.

- In [Jaffri et al., 2008], the authors discuss how erroneous use of owl:sameAs in the interlinking of the DBpedia and DBLP datasets has resulted in publications becoming incorrectly assigned to different authors.

- Automatic data linking tools do not guarantee 100% precision, because of:
  - Errors, missing information, data freshness, etc.
**OWL:sameAs Predicate**

- `owl:sameAs`, indicates that two different descriptions refer to the same entity.
- A strict semantics,
  1) Reflexive,
  2) Symmetric,
  3) Transitive and
  4) Fulfils property sharing:

\[
\forall X \forall Y \text{owl:sameAs}(X, Y) \land p(X, Z) \Rightarrow p(Y, Z)
\]
DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

owl:sameAs(b1, b2)?
DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

owl:sameAs(b1, b2)?
Which kind of information to use for detecting erroneous Identity links?
DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

A Semantic Web Primer

A Semantic Web Primer

2007

2007

208

208

Paul Grauth

Grigoris Antoniou

P. Grauth

G. Antoniou

UNA (S1) > UNA (S2) > UNA (S3)

DETECTION OF ERRONEOUS IDENTITIES
DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

Ontology Axioms:
- \( \text{Func}(\text{nbPages}) \)
- \( \text{LC}(\text{author}) \)
- \( \text{Func}(\text{title}) \)
- \( \text{Disj}(\text{Science-fiction, Memoir}), \ldots \)
DETECTION OF ERRONEOUS IDENTITY LINKS

Inconsistency-based

Content-Based

UNA

Trustworthiness

Ontology axioms

Network-Based

Community Detection

Network Metrics

Papaleo et al. 2014

Paulheim 2014

Raad et al. 2018

Guéret et al. 2016

[ de Melo, 2013 ]

[ Valdestilhas et al., 2017 ]
DETECTION OF ERRONEOUS
IDENTITY LINKS

Inconsistency-based

UNA

Trustworthiness

Ontology axioms

Content-Based

Network-Based

Community Detection

Network Metrics

Cudré-Mauroux et al. 2009

Papaleo et al. 2014

Paulheim 2014

Raad et al. 2018

Guéret et al. 2016

[ de Melo, 2013 ]

[ Valdestilhas et al., 2017 ]
AXIOM-BASED APPROACHES
**ONTOLOGY AXIOM VIOLATION**

**Principle:** use of ontology axioms (functionality, local completeness, asymmetry, etc.) to detect inconsistencies or error candidates in the linked resources descriptions.

*nbPages is a Functional Property*
A logical **ontology-based method** to detect invalid sameAs statements

- Builds a contextual graph «around» each one of the two resources involved in the sameAs by exploiting ontology axioms on:
  - **functionality and inverse functionality** of properties and
  - **local completeness** of some properties, e.g., the author list of a book.

- Exploit the descriptions provided in these contextual graphs to eventually detect inconsistencies or high dissimilarities.
F is the set of RDF facts enriched by a set of \( \neg \text{synVals} \) facts in the form

\[ \neg \text{synVals}(w_1, w_2) \]

where \( w_1 \) and \( w_2 \), being literals and different.

EXAMPLES:
- \( \text{notSynVals}(’231’,’100’) \)
  for a functional property \( nbPages \)
- \( \text{notSynVals}(’New York’, ‘Paris’) \)
  for a functional property \( cityName \)

... knowledge from expert or extracted.
R the set of rules

(inverse) functional properties

- $R_{1_{FDP}} : \text{sameAs}(x, y) \land p_i(x, w_1) \land p_i(y, w_2) \rightarrow \text{synVals}(w_1)$
- $R_{2_{FOP}} : \text{sameAs}(x, y) \land p_j(x, w_1) \land p_j(y, w_2) \rightarrow \text{sameAs}(w_1)$
- $R_{3_{FDP}} : \text{sameAs}(x, y) \land p_k(x, w_1) \land p_k(y, w_2) \rightarrow \text{sameAs}(w_2)$

local complete properties

- $R_{4_{LC}} : \text{sameAs}(x, y) \land p(x, w_1) \rightarrow p(y, w_1)$
ONTOLOGY AXIOM VIOLATION

- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools [1], [2] and [3].


2-degree contextual graph
phone_number, hasAddress & city
(possible synvals computation)
ONTОLOGY AXIOM VIOLATION

- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools [1], [2] and [3].

<table>
<thead>
<tr>
<th>LM</th>
<th>LM Precision</th>
<th>linkInv precision</th>
<th>LM+linkInv precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>95.55%</td>
<td>37%</td>
<td>98.85%</td>
</tr>
<tr>
<td>1</td>
<td>69.71%</td>
<td>88.4%</td>
<td>95.19%</td>
</tr>
<tr>
<td>3</td>
<td>90.17%</td>
<td>42.30%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Limits:**
- Scalability problems
- Need of uniform vocabulary in datasets

[Papaleo et al. 2014]
IDENTITY PROBLEM AT LOD SCALE

> Several domains
> 558 M. identity links
> 28 B. RDF triples
> 48 K. equiv. classes

[Beek et al., 2018]

http://sameas.cc/explicit/img
The largest identity set contains 177,794 terms:

Different countries
Different cities
Albert Enstein

→ quality problems
IDENTITY LINK INVALIDATION

Inconsistency-based
- UNA
- Trustworthiness
- Ontology axioms
  - Cudré-Mauroux et al. 2009
  - Papaleo et al. 2014
  - Paulheim 2014

Content-Based

Network-Based
- Community Detection
  - Raad et al. 2018
- Network Metrics
  - Guéret et al. 2012

[ de Melo, 2013 ]
[ Valdestilhas et al., 2017 ]
Principle

- The quality of a link can be determined based on **how connected a node** is within the **network** in which it appears.
- Use of **network metrics and structures** can help to detect erroneous links.

[Guéret et al., 2012]

[Raad et al., ISWC 2018]
NETWORK BASED

- Considers the **identity network** build from the **explicit identity network** of sameAs links: removing of symmetric and reflexive links.

- Uses of Louvain **community detection** algorithm to detect subgraphs in the **identity network** that are highly connected.

- Defines a **ranking score** for each (intra-community and inter-community) identity link based on the **density of the community**.

[Raad et al., ISWC 2018]
INTRA-COMMUNITY ERRONEOUSNESS DEGREE

\[ \text{err}(e_C) = \frac{1}{w(e_C)} \times \left(1 - \frac{W_C}{|C| \times (|C| - 1)}\right) \]

INTER-COMMUNITY ERRONEOUSNESS DEGREE

\[ \text{err}(e_{C_{ij}}) = \frac{1}{w(e_{C_{ij}})} \times \left(1 - \frac{W_{C_{ij}}}{2 \times |C_i| \times |C_j|}\right) \]
Dataset

- LOD-a-lot dataset [Fernandez et al. 2017]: a compressed data file of 28B triples from LOD 2015 crawl
- An explicit identity network of 558.9M edges (links) and 179M nodes (resources)

Example: The B. Obama equality set that contain 440 nodes
Barack Obama’s Equality Set

DBpedia IRIs referring to the person Obama in different languages

IRIs referring to the presidency and the Obama administration

DBpedia IRIs referring to the person Obama in different functions

DBpedia IRIs referring to the person Obama, his senator career
Barack Obama’s Equality Set

Low err(e) for the links of this community

These two links have err(e) = 1

Most of the links have err(e) = 0.9

NETWORK BASED

[Raad et al., ISWC 2018]
• **Scales** to a graph of **28 billion** triples: **11 hours for the 4 steps**

No **benchmark** for qualitative evaluation

**Precision**: manual evaluation of 200 links

- The higher the error degree is the most likely the link will be erroneous: 100% of owl:sameAs with an **error degree <0.4** are correct

- Can theoretically **invalidate a large set of owl:sameAs links** on the LOD: 1% (1.26M owl:sameAs) have an **error degree** in [0.99, 1]

**Recall**: 780 incorrect links between 40 distinct resources have been introduced in the explicit identity graph. **Recall = 93 %**
Identity invalidation

- Different kinds of information can be used for link invalidation: axioms, resource descriptions and graph topology

- The efficiency of the proposed approaches depends on the characteristics of the knowledge graphs: volume, heterogeneity, ontology
Identity invalidation

- Different kinds of information can be used for link invalidation: axioms, resource descriptions and graph topology
- The efficiency of the proposed approaches depends on the characteristics of the knowledge graphs: volume, heterogeneity, ontology

Possible improvements

- Need for hybrid approaches for link invalidation
- Need for well-formalized weak-identity: contextual identity, similarity, …
- Need for approaches for difference links detection: useful for inconsistency checking
CONCLUSION

- Semantic Web standards, data and many applications are there

- Promising applications are emerging for which reasoning on data is central:
  - Web search, recommendation systems, chat-bots, …

- Many challenges remain to handle at large scale the incomplete, uncretain and evolving knowledge graphs
  - Combining numerical and symplcic AI is challenging but worthwhile to investigate more deeply.
KNOWLEDGE GRAPH REFINEMENT
KEY DISCOVERY AND LINK INVALIDATION

FATIHA SAÏS

LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

Joint work with: N. Pernelle, L. Papaleo, J. Raad and D. Symeonidou

3ÈME JOURNÉE RI-IA SOUTENUE PAR L’AFIA ET ARIA, PARIS 2019
REFERENCES (1)

[Atencia et al. 2014] Manuel Atencia, Jérôme David, Jérôme Euzenat:
Data interlinking through robust linkkey extraction. ECAI 2014: 15-20

[Al-Balkri et al. 2015] Mustafa Al-Bakri, Manuel Atencia, Steffen Lalande, Marie-Christine Rousset:
Inferring Same-As Facts from Linked Data: An Iterative Import-by-Query Approach. AAAI 2015: 9-15


[Beek et al., 2016 ] A contextualised semantics for owl: sameas.
W. Beek, S. Schlobach, and F. van Harmelen. In ESWC 2016

[CudreMauroux et al., 2009 ] idmesh: graph-based disambiguation of linked data.

[de Melo, 2013] Not quite the same: Identity constraints for the web of linked data.

REFERENCES (2)

Assessing linked data mappings using network measures. In ESWC 2012

When owl:sameAs isn’t the same: An analysis of identity in Linked Data. In ISWC 2010.

Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. In JWS 2012.

[Jaffri et al., 2008] URI disambiguation in the context of linked data.
A. Jaffri, H. Glaser, and I. Millard. In LDOW@WWW 2008.


[Papaleo et al., 2014] Logical detection of invalid sameas statements in rdf data.

An Automatic Key Discovery Approach for Data Linking. In Journal of Web Semantics


REFERENCES (3)

[Raad et al., 2017] Detection of contextual identity links in a knowledge base.

[Raad et al., 2018] Detecting Erroneous Identity Links on the Web using Network Metrics. J. Raad, W. Beek, F. van Harmelen, N. Pernelle and F. Saïs. ISWC 2018


[Saïs et al.09] Combining a Logical and a Numerical Method for Data Reconciliation.
Fatiha Saïs., Nathalie Pernelle and Marie-Christine Rousset.
In Journal of Data Semantics 2009.

[Soru et al. 2015] Tommaso Soru, Edgard Marx, Axel-Cyrille Ngonga Ngomo:
ROCKER: A Refinement Operator for Key Discovery. WWW 2015: 1025-1033


[Valdestilhas et al., 2017 ] Cedal: time-efficient detection of erroneous links in large-scale link repositories.