What do the Sources Say? Exploring Heterogeneous Journalistic Data As a Graph

Ioana Manolescu
Inria Saclay-Île-de-France and Institut Polytechnique de Paris
Democratic societies crucially need the press

- To debate and express dissent
- To analyze, confirm or refute public statements
- To expose and explain society functioning

Fact-checking (Data) journalism

Socialist Romania, 1984
Projects and collaborations

**Google Award** (2015) with X. Tannier (U. Paris Sud)

**ANR ContentCheck** (2016-2020) with
X. Tannier (Sorbonne Université), S. Cazalens, P. Lamarre, J.-M. Petit, M. Plantevit (U. Lyon), F. Goasdoué (U. Rennes 1), Les Décodeurs (Le Monde) [https://contentcheck.inria.fr](https://contentcheck.inria.fr)

**Inria Associated Team WebClaimExplain** (2017-2019), with AIST Japan (Julien Leblay)

Collaboration with H. Galhardas (University of Lisbon), A. Anadiotis, O. Balalau (CEDAR), E. Pietriga (ILDA)
Project discussed with Les Décodeurs: Twitter monitoring for analyzing political communication

**Social media** (Twitter and FB): (semi)structured data (JSON)
- Author, date, content, retweet no., whether it’s a retweet etc.

**Décodeurs’** database of French public figures (Excel)
- First name, last name, Twitter ID, position, political party when known

**Wikipedia** (RDF)
- Information about some best-known public figures, parties

**Question**: How do public figures communicate on Twitter?
Twitter monitoring for analyzing political communication

Main challenge: data heterogeneity

Solution: Tatooine mediator [Bonaque et al., VLDB 2016]

- Handles heterogeneous data
- Integrates ML classifiers as UDFs
Twitter monitoring for analyzing political communication

Tag clouds of French political tweets, Nov 2015: left, right, green parties
Twitter monitoring for analyzing political communication

Tatooine mediator [Bonaque et al., VLDB 2016]

Tatooine solution analysis:
- ✔ Handles heterogeneous data
- ✔ Integrates ML classifiers as UDFs
- ✗ No query language; need to write dedicated sub-queries in each underlying lg. + mediator (combination) plan
- ✗ Too complicated to install and use (many sub-systems)
Another project discussed with Les Décodeurs: fake news detection and propagation on Twitter

**Online fact-checks**: (semi)structured data sources (JSON, XML) listing

- Link to claim (media, social network etc.), **claim author**
- **Fact-check**, containing: analysis (details), final assessment, fc author, date, institution

Among the first published: [https://www.lemonde.fr/webservice/decodex/updates](https://www.lemonde.fr/webservice/decodex/updates)

Years later: **ClaimReview** by Google and others ([https://www.claimreviewproject.com/](https://www.claimreviewproject.com/))
Another project discussed with Les Décodeurs: fake new detection and propagation on Twitter

**Online fact-checks:** (semi)structured data sources (JSON, XML) listing

- Link to claim (media, social network etc.), **claim author**
- **Fact-check**, containing: analysis (details), final assessment, fc author, date

**Décodeurs’** database of French public figures (Excel)

- First name, last name, Twitter ID, position, political party when known

**Question:** When does a fake news post first cross into a supposedly legitimate community (e.g. members of the Parliament)?

- Looking for tweets connected to a fake news author, and to a community member; both connections are arbitrary paths (chains of author/likes/retweets/inParty/...
Graph-based integration of heterogeneous data sources

- The sources are not RDF. They can be (semi)structured, or unstructured (text).
- The sources may be very dynamic (projects started and abandoned as per news cycle and data availability).
- There is no schema. Data producers often uncollaborative.
- For most journalists, databases do not come naturally, and IT support is limited. They know keyword-based search...

Integrate heterogeneous sources within a graph, query w/ keywords
ConnectionLens: graph-based integration of heterogeneous data sources

https://team.inria.fr/cedar/connectionlens/

Joint work with: J. Leblay (AIST Japan),
H. Galhardas and C. Conceiçao (U. Portugal),
A. Anadiotis, O. Balalau, N. Barret, T. Bouganim, F. Chimienti,
M.-Y. Haddad, T. Merabti, P. Upadhyay (CEDAR) + past interns

S. Horel (Le Monde)

Ongoing work funded by ANR/DGA AI Chair SourcesSay
(https://sourcessay.inria.fr), DIM RFSI
**ConnectionLens principles** [Chanial et al., 2018]

Integrate **any kind of data** into a **graph**

**Extract entities** from any text node (regardless of the model of the data source where the text comes from)

- Same entity in two different text nodes = link among the text nodes (**densification** of the graph)

The graph is **heterogeneous** and **irregular**

Query it through **keywords**: find trees that connect 1 node matching each kwd

- Closely related to the Group Steiner Tree Problem (GSTP)
ConnectionLens principles [Chanial et al., 2018]

Integrate any kind of data into a graph

Extract entities from any text node (regardless of the model of the data source where the text comes from)

- Rest of the presentation based on state of the project
- 2+ years later:

Closer, related to the Group Steiner Tree Problem (GSTP)
ConnectionLens graph construction
The Balkany and their African connections
The Balkany and their African connections
The Balkanys and their African connections

Public officials transparency high authority (CSV)

<table>
<thead>
<tr>
<th>Name</th>
<th>Owner</th>
<th>Location</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dar Gyucy</td>
<td>P. Balkany</td>
<td>Marrakech</td>
<td>Real Estate</td>
</tr>
<tr>
<td>Moulin Cossy</td>
<td>I. Balkany</td>
<td>Giverny</td>
<td>Real Estate</td>
</tr>
</tbody>
</table>
The Balkanys and their African connections

<table>
<thead>
<tr>
<th>Name</th>
<th>Owner</th>
<th>Location</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dar Gyucy</td>
<td>P. Balkany</td>
<td>Marrakech</td>
<td>Real Estate</td>
</tr>
<tr>
<td>Moulin Cossy</td>
<td>I. Balkany</td>
<td>Giverny</td>
<td>Real Estate</td>
</tr>
</tbody>
</table>

National Directory of Elected Officials (JSON)

```json
[{
  name: "Levallois-Perret",
  mayor: "P. Balkany",
  city-council: [
    {name: "I. Balkany"},
    ...
  ],
}, ...
]```
The Balkanys and their African connections

<table>
<thead>
<tr>
<th>Name</th>
<th>Owner</th>
<th>Location</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dar Gyucy</td>
<td>P. Balkany</td>
<td>Marrakech</td>
<td>Real Estate</td>
</tr>
<tr>
<td>Moulin Cossy</td>
<td>I. Balkany</td>
<td>Giverny</td>
<td>Real Estate</td>
</tr>
</tbody>
</table>

National Directory of Elected Officials (JSON)

```json
[
  {
    name: "Levallois-Perret",
    mayor: "P. Balkany",
    city-council: [
      {name: "I. Balkany"},
      ...
    ]
  }, ...
]
```

dbpedia.org (RDF)

```json
{
  dbr:Marrakech
  {
    dbr:name "Marrakech"
    rdf:type dbo:City ;
    dbo:country dbr:Morrocco .
  }
  dbr:Morocco
  {
    dbr:name "Morocco"
    rdf:type dbo:Country
    dbo:locatedIn dbr:Africa .
  }
  dbr:CentralAfricanRepublic
  {
    dbr:name "Central African Republic"
    dbo:locatedIn dbr:Africa .
  }
}
The Balkanys and their African connections

Public officials transparency high authority (CSV)

<table>
<thead>
<tr>
<th>Name</th>
<th>Owner</th>
<th>Location</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dar Gyucy</td>
<td>P. Balkany</td>
<td>Marrakech</td>
<td>Real Estate</td>
</tr>
<tr>
<td>Moulin Cossy</td>
<td>I. Balkany</td>
<td>Giverny</td>
<td>Real Estate</td>
</tr>
</tbody>
</table>

National Directory of Elected Officials (JSON)

```json
[{
  name: "Levallois-Perret",
  mayor: "P. Balkany",
  city-council: [{
    name: "I. Balkany",
    ...]
  }, ...]
```

dbpedia.org (RDF)

```rdf
{ dbr:Marrakech
dbr:name "Marrakech"
rdf:type dbo:City ;
dbo:country dbr:Morrocco .
dbr:Morocco
dbr:name "Morocco"
rdf:type dbo:Country
dbo:locatedIn dbr:Africa .
dbr:CentralAfricanRepublic
dbr:name "Central African Republic"
dbo:locatedIn dbr:Africa .
}
```

Libération – Nov. 13, 2014 (Text)

**Balkany mineur de fonds**

L’élu de Levallois-Perret est soupçonné d’avoir touché 5 millions de dollars de commission en 2009 grâce à son rôle d’intermédiaire entre Areva et la Centrafrique dans le dossier Uramin. [...]
How are the Balkany connected to Africa and "real estate"?

<table>
<thead>
<tr>
<th>Name</th>
<th>Owner</th>
<th>Location</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dar Gyucy</td>
<td>P. Balkany</td>
<td>Marrakech</td>
<td>Real Estate</td>
</tr>
<tr>
<td>Moulin Cossy</td>
<td>I. Balkany</td>
<td>Giverny</td>
<td>Real Estate</td>
</tr>
</tbody>
</table>

National Directory of Elected Officials (JSON)
```json
[{
  name: "Levallois-Perret",
  mayor: "P. Balkany",
  city-council: [
    {name: "I. Balkany"},
    ...
  ]
}, ...]
```

Libération – Nov. 13, 2014 (Text)

**Balkany mineur de fonds**

L’élue de Levallois-Perret est soupçonnée d’avoir touché 5 millions de dollars de commission en 2009 grâce à son rôle d’intermédiaire entre Areva et la Centrafrique dans le dossier Uramin. […]

---

Ioana Manolescu, Inria and Institut Polytechnique de Paris  
IRISA, March 2021
Idea: integrate all data sources into a heterogeneous graph
Graph construction stages

1. **Primary node and edge construction**
   - Direct for XML, JSON, RDF, HTML
   - 1 relational tuple = 1 node; primary keys-foreign keys as links
   - Convert information from PDF into: (a) JSON, and (b) RDF describing tables
Graph construction stages

1. **Primary node and edge construction**
   - Direct for XML, JSON, RDF, HTML
   - 1 relational tuple = 1 node; PK-FKs as links
   - [Optional] segment text documents
   - Extract information from PDF into: (a) JSON, and (b) RDF describing tables

<table>
<thead>
<tr>
<th></th>
<th>Ile-de-France</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Paris</td>
</tr>
<tr>
<td>Députés</td>
<td>Essonne</td>
</tr>
<tr>
<td>Sénateurs</td>
<td>Yvelines</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Députés</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paris</td>
<td>Alice, Bob</td>
</tr>
<tr>
<td>Carole</td>
<td>David</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Sénateurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erin</td>
<td>Fred</td>
</tr>
<tr>
<td>Fred</td>
<td>George</td>
</tr>
</tbody>
</table>
Graph construction stages

2. **Entity extraction**

- From all text nodes of all the sources: entity node child of text node
- [VLDB2018]: based on Stanford NER
- [BDA2020] Developed and trained new entity extractor from French, based on Flair framework
Graph construction stages

2. **Entity extraction**
   - From all text nodes of all the sources: entity node child of text node
   - [VLDB2018]: based on Stanford NER
   - [BDA2020] Developed and trained new entity extractor from French, based on Flair framework

3. **Entity disambiguation**
   - For each recognized entity, e.g., "Hollande" the place or the person?
   - Built novel disambiguation pipeline for French, based on Ambiverse framework
     - Based on knowledge bases (WikiData, YAGO) and Wikipedia
     - Helpful on well-known entities
Graph construction stages

2. Entity extraction

- From all text nodes of all the sources: entity node child of text node
- [VLDB2018]: based on Stanford NER
- [BDA2020] Developed and trained new entity extractor from French, based on Flair framework

3. Entity disambiguation

- For each recognized entity, e.g., "Hollande" the place or the person?
- Built novel disambiguation pipeline for French, based on Ambiverse framework
  - Based on knowledge bases (WikiData, YAGO) and Wikipedia
  - Helpful on well-known entities
Graph construction stages

4. Node matching

- To create sameAs edges:
  - Strong sameAs edges: equivalent nodes 1.
  - Weak sameAs edges: similar nodes .85

- Appropriate distance functions

- New: more normalization → better matching

- Remains quadratic at the core ☹, so...

Node factorization (heuristic): create only one node per label per document (or per graph)
Thanks S. Horel (Le Monde), E. Pietriga (Inria & U. Paris Saclay), J. Feitz (X17)

Ioana Manolescu, Inria and Institut Polytechnique de Paris
IRISA, March 2021
ConnectionLens graph querying
Querying problem statement

- Given the graph $G = (N, E)$ built out of the datasets $D$ and a query $Q = \{w_1, ..., w_m\}$, return the **$k$ highest-score minimal answer trees**.
- An answer tree is a set of edges which (i) form a tree (ii) contain at least one node whose label matches each keyword $w_i$.
- We are interested in **minimal answer trees**, that is:
  - Removing an edge from the tree should make it lack some keyword(s).
  - If a keyword matches more than one nodes in the answer tree, then all these matching nodes must be equivalent.
Search space and complexity

- The problem is related to the (Group) Steiner Tree Problem
- Given weighted graph G, and nodes n1, . . . , nm, the Steiner Tree Problem (STP) consists of finding the smallest-cost tree in G that connects all the nodes.
  - Known NP-hard problem in |G|
- Differences with our problem:
  - Each edge can be taken in any direction: exponential increase in search space size
  - We need the k smallest-cost trees, not just one.
  - Each keyword may match several nodes, not just one.
- Our approach: enumerate solutions until time-out or max number of solutions reached.
  - Return best k solutions found
GAM (Grow and Aggressive Merge) Algorithm

- Builds trees “backward” from the keyword matches
- **GROW** adds an edge to the root of a tree, **MERGE** merges trees with the same root
- Exploration (GROW) order:
  1. Favor trees matching the largest number of query keywords
  2. To break ties, favor smaller trees
  3. To break second tie between (t1, e1), (t2, e2), we prefer the pair with the higher specificity edge.

The specificity of $e = n_1 \xrightarrow{l} n_2$ is: $s(e) = \frac{2}{N_{n_1}^l + N_{n_2}^l}$
Implementation and performance
Implementation

- Java (228 classes/45K LOC), Python (25 classes/2700 LOC), JS + CSS
- Available in Inria Gitlab: https://gitlab.inria.fr/cedar/connectionlens

ConnectionLens graph construction

Nodes+edges

Relational DB

Parallel, in-mem keyword search

In-memory graph

Keyword search algorithm
Implementation

- Graph creation time mostly **linear in the size of the data**
- Costliest operations involve ML (disambiguation, extraction)
  - Nov 2020: **batch extraction**, 20x speed-up on GPU, 2x speed-up on regular server
  - Oct 2020: **extraction policies** replace or avoid extraction in some parts of the data

![Graph construction time comparison chart](image-url)

**Figure 6:** Graph construction time (seconds).
Graph creation performance: batch extraction
https://arxiv.org/abs/2012.08830

Figure 7: YAGO loading time (minutes) using Flair.

Collaboration with Stéphane Horel (Le Monde)

Data: XML, PDF → JSON, HTML

<table>
<thead>
<tr>
<th></th>
<th>(N)</th>
<th>(E)</th>
<th>(N)</th>
<th>(N_P)</th>
<th>(N_O)</th>
<th>(N_L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML</td>
<td>32,028,429</td>
<td>19,851,904</td>
<td>1,483,631</td>
<td>584,734</td>
<td>126,629</td>
<td></td>
</tr>
<tr>
<td>JSON</td>
<td>1,025,307</td>
<td>432,303</td>
<td>75,297</td>
<td>7,320</td>
<td>4,139</td>
<td></td>
</tr>
<tr>
<td>HTML</td>
<td>246,636</td>
<td>185,479</td>
<td>3,726</td>
<td>7,227</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>33,300,372</td>
<td>20,469,686</td>
<td>1,562,654</td>
<td>665,167</td>
<td>131,088</td>
<td></td>
</tr>
</tbody>
</table>

*Table 3: Statistics on Conflict of Interest application graph.*

Collaboration with Stéphane Horel (Le Monde)

Data: XML, PDF → JSON, HTML

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>E</td>
<td>N</td>
<td>N_P</td>
<td>N_O</td>
<td>N_L</td>
</tr>
<tr>
<td>XML</td>
<td>32,028,429</td>
<td>19,851,904</td>
<td>1,483,631</td>
<td>584,734</td>
<td>126,629</td>
</tr>
<tr>
<td>JSON</td>
<td>1,025,307</td>
<td>432,303</td>
<td>75,297</td>
<td>7,320</td>
<td>4,139</td>
</tr>
<tr>
<td>HTML</td>
<td>246,636</td>
<td>185,479</td>
<td>3,726</td>
<td>7,227</td>
<td>320</td>
</tr>
<tr>
<td>Total</td>
<td>33,300,372</td>
<td>20,469,686</td>
<td>1,562,654</td>
<td>665,167</td>
<td>131,088</td>
</tr>
</tbody>
</table>

Table 3: Statistics on Conflict of Interest application graph.

Collaboration with Stéphane Horel (Le Monde)
Graph quality experiments

- PDF extraction accuracy: 63%
- F1 score for entity extraction from French:
  - Flair stacked forward and backward embeddings with French fastText embeddings: 73%
  - Spacy: 63%
  - StanfordNER: 45%
- F1 score of disambiguation: 86%
ConnectionLens in the scientific landscape
ConnectionLens in the scientific landscape (1)

**Data integration** for structured, semistructured and unstructured data

- “Ad-hoc” (combinations of sources to be unioned, joined, or chained)
- No schema, ontologies, queries known in advance
- Mediator approaches inappropriate due to complexity, lack of structure, and performance
- Now: graph warehouse
- Lack of structure forces reachability queries instead of join: price to pay for powerful integration

**Data cleaning** aspects: Similarity links require value or entity matching

- Avoid constructing structured objects (“clean tuples”): don’t seem necessary
ConnectionLens in the scientific landscape (2)

Graph construction
- Users of entity extraction modules, trained a model for French
- May require re-training, especially if specialized language

Keyword search on structured data
- Previously studied for relational, graph, or XML databases
  - Typically assume structure/regularity in the graph
  - Exploit favorable properties of the score function
  - First work to solve such queries across heterogeneous data sources, w/o assumptions on cost function, sub-optimal structure prop., w/ bidirectional search
  - In-memory graph store and parallel query processor (200x speed-up)
Ongoing work

- Extending and improving the in-memory query processor (A. Anadiotis, F. Chimienti, IM)
- Relationship extraction (O. Balalau, M. Mohanty, IM)
- Natural language querying of the graph (O. Balalau, P. Upadhyay, IM + PhD in fall 2021)
- Improving the quality of graph linking (T. Bouganim, H. Galhardas, IM)
- Abstracting CL graphs (N. Barret, H. Galhardas, P. Upadhyay, IM)

- Applications:
  - Conflicts of Interest database (w/ S. Horel and G. Fooks, Aston U., UK)
  - Mediacités (w/ WeDoData)
Why data journalism?

Because I grew up in a dictatorship, and I value free press

Because journalists are threatened and killed still today in Europe

Because the press' economic model is threatened by IT giants

Because this industry is currently underserved by IT – and we could really make an impact!
Useful links

**ConnectionLens**: https://team.inria.fr/cedar/connectionlens/

To continue within SourcesSay (ANR + DGA AI Chair, 2020-2024): https://sourcessay.inria.fr