Exploitation and Merge of Information Sources for Public Procurement Improvement

Exploitation and Merge of Information Sources for Public Procurement Improvement

R. Nai, E. Sulis, P. Pasteris, M. Giunta, R.Meo

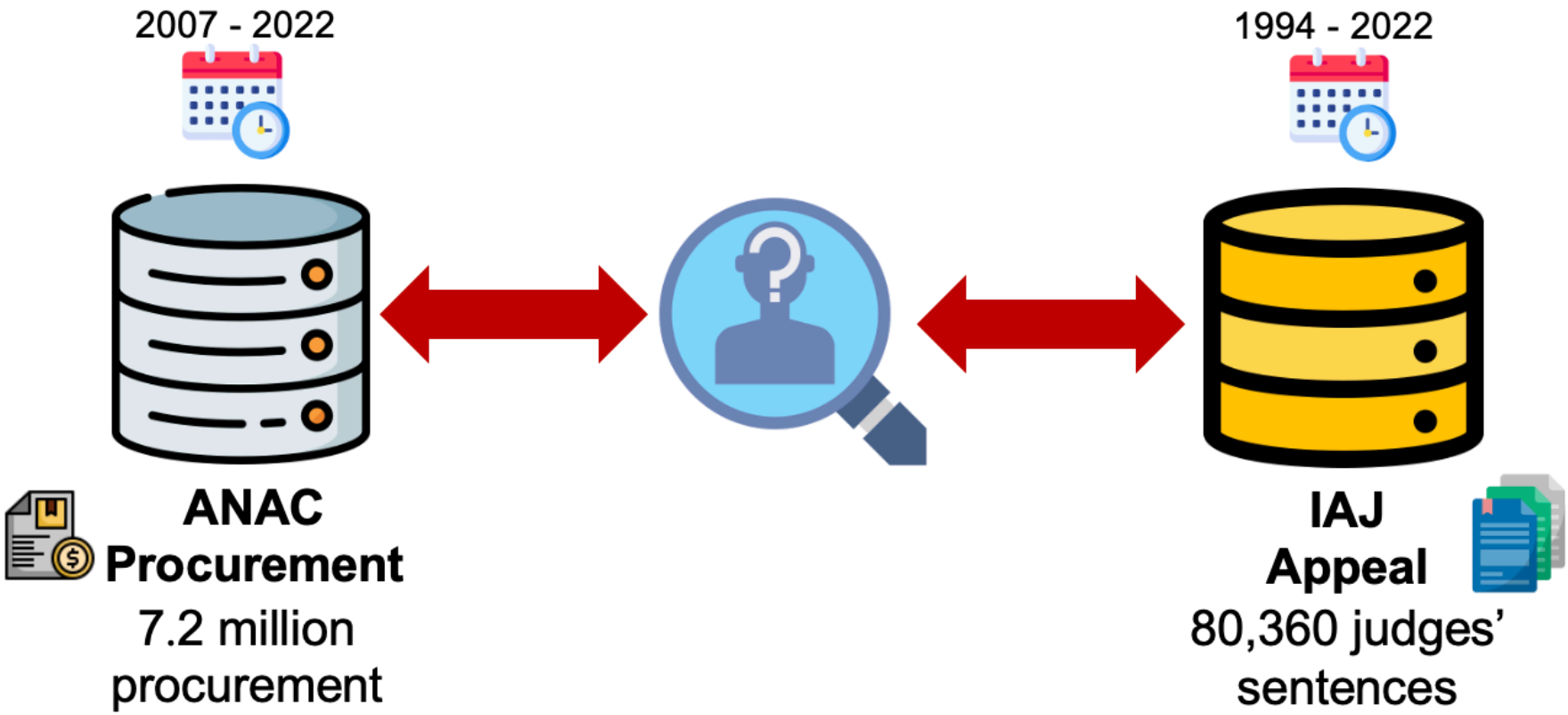
Computer Science Department, University of Turin

23th September, 2022

Introduction

- In the Internet age, we need on one side, search-based applications, on the other side we need to extract information from large corpora.
- This work investigates the automatic knowledge extraction from a set of public law archives.
- Two legal datasets:
  - public procurement from (Italian) National Anti-Corruption Authority (ANAC);
  - the archive of the Italian Administrative Justice (IAJ) appeals.

Introduction

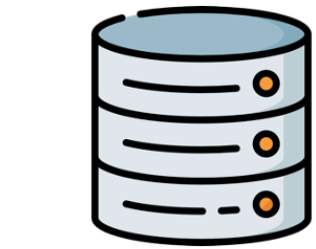


Introduction

Research Question

How can we automatically extract information from legal archives to identify the entities involved in a public procurement?

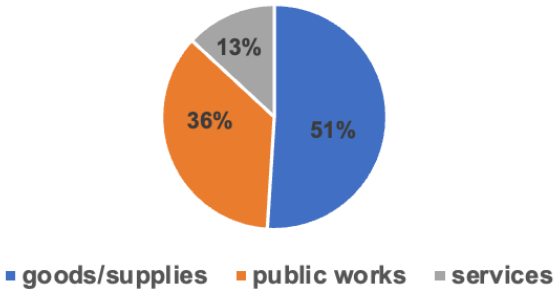
Case study



ANAC Procurement

7.2 million procurement

Identifier: `CIG` (shared key)



Contractors (PA): 42,393

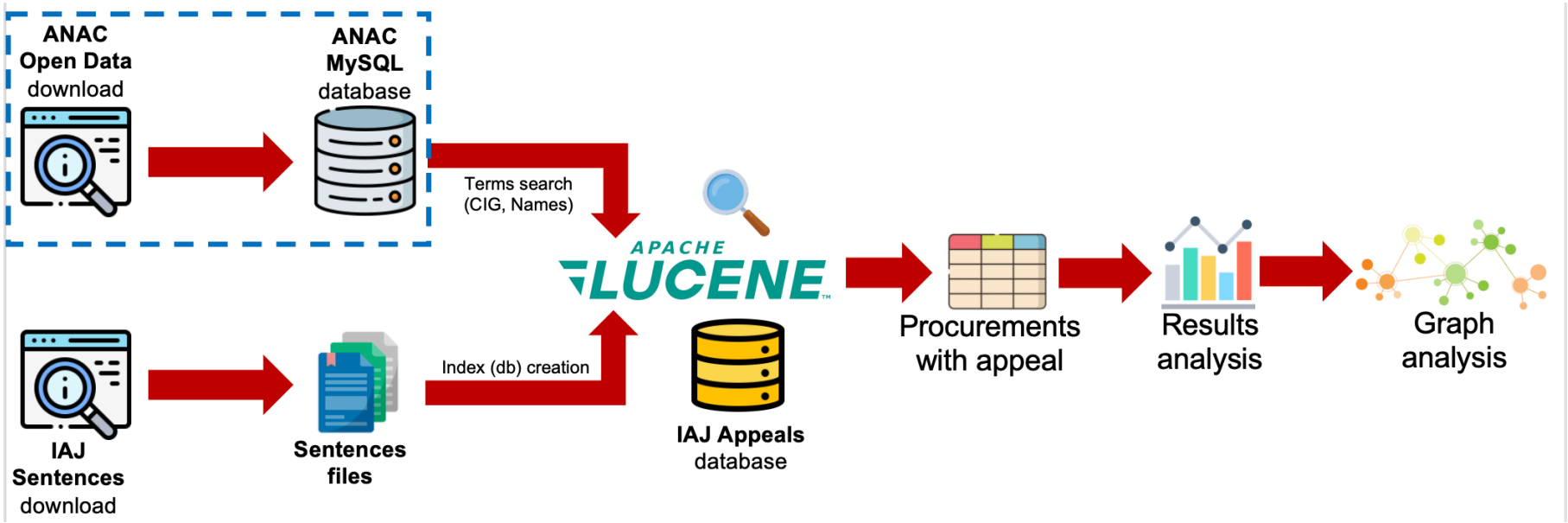
- Municipalities
- Universities
- Hospitals
- etc…

Awards: 1,635,609 (~22.71% of Procurement)

Economic operators: 265,039

Data gathering

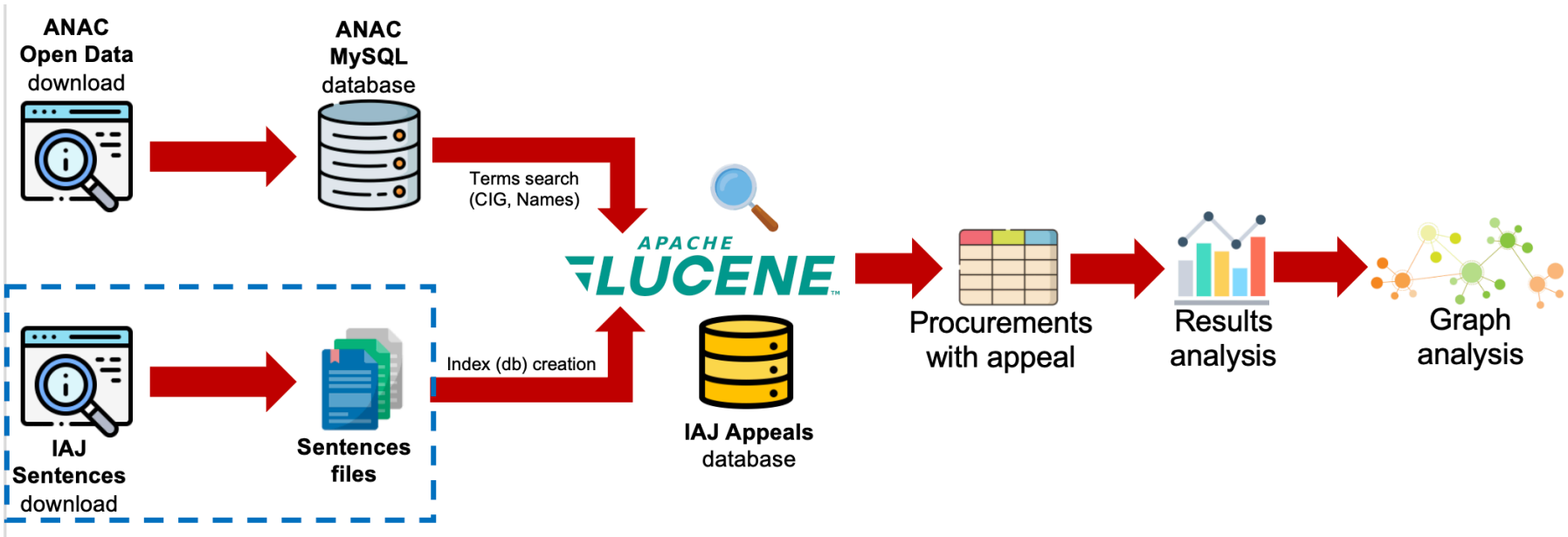
- We imported the Open Data of the procurement into an InnoDB table of a MySQL database (whose size is 5.5 GB).



Workflow of the research approach from data gathering to results analysis.

Data gathering

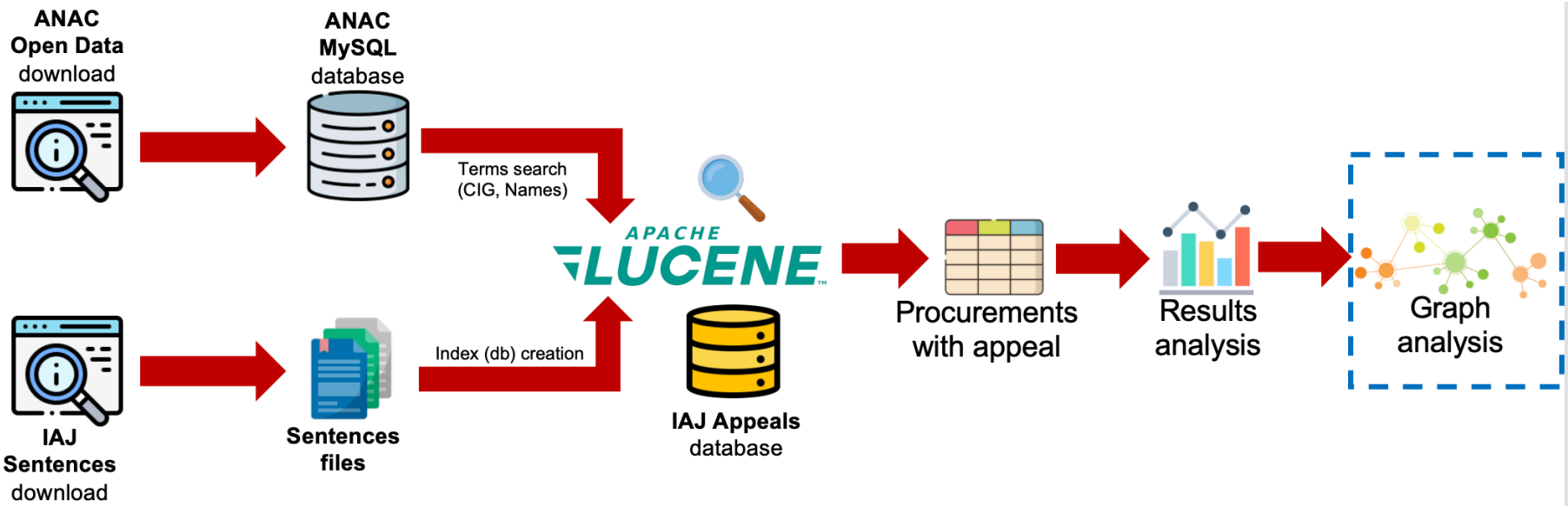
- We obtained the IAJ judgments via web scraping.
- They were indexed using tools specialized in Information Retrieval like Apache Lucene (`lucene.apache.org`).



Workflow of the research approach from data gathering to results analysis.

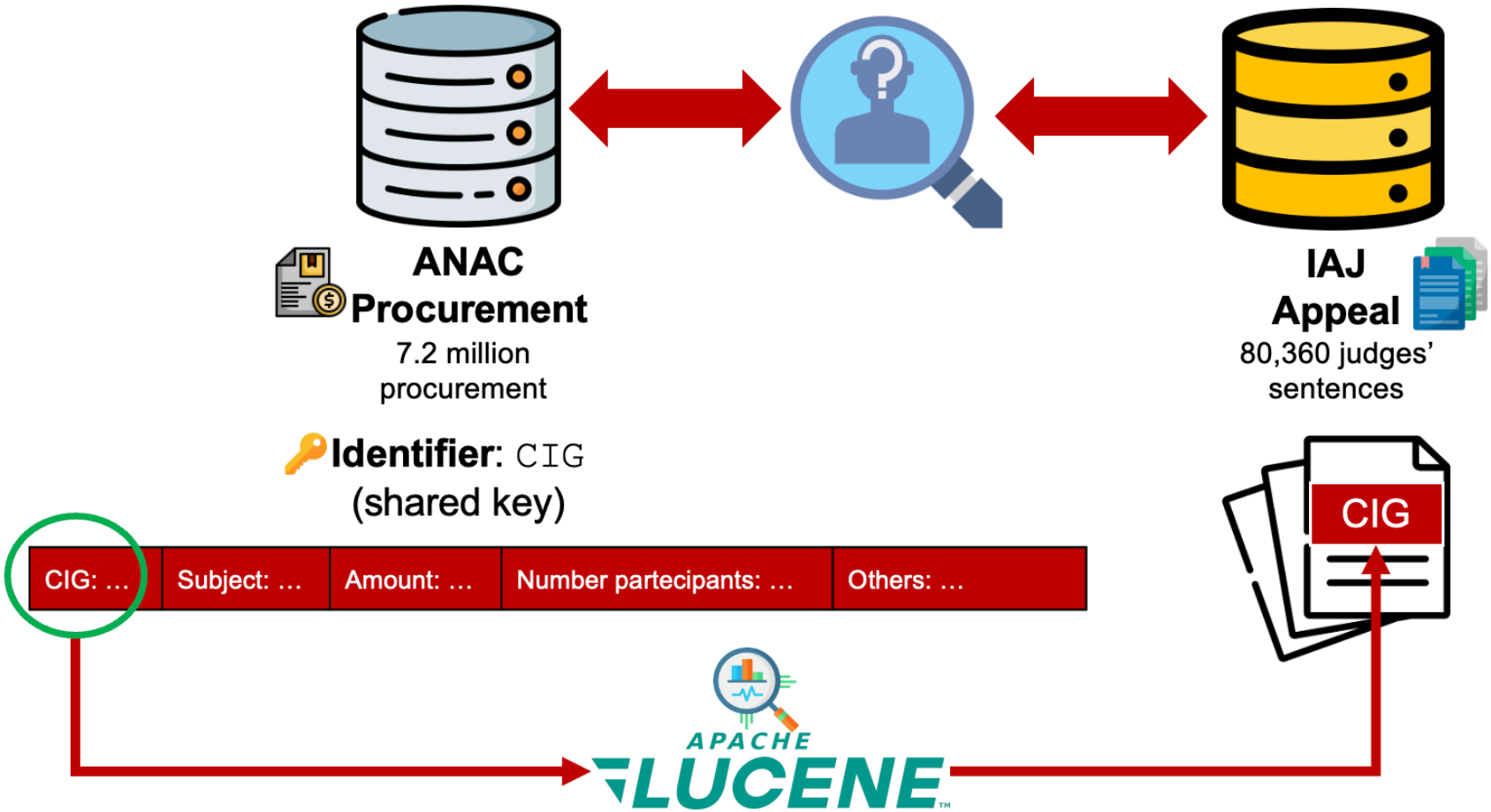
Data gathering

- We imported the ANAC Open Data on the procurement between a public entity and a private company into the Neo4j (`neo4j.com`) graph NoSQL database.



Workflow of the research approach from data gathering to results analysis.

Search by procurement ID (CIG)



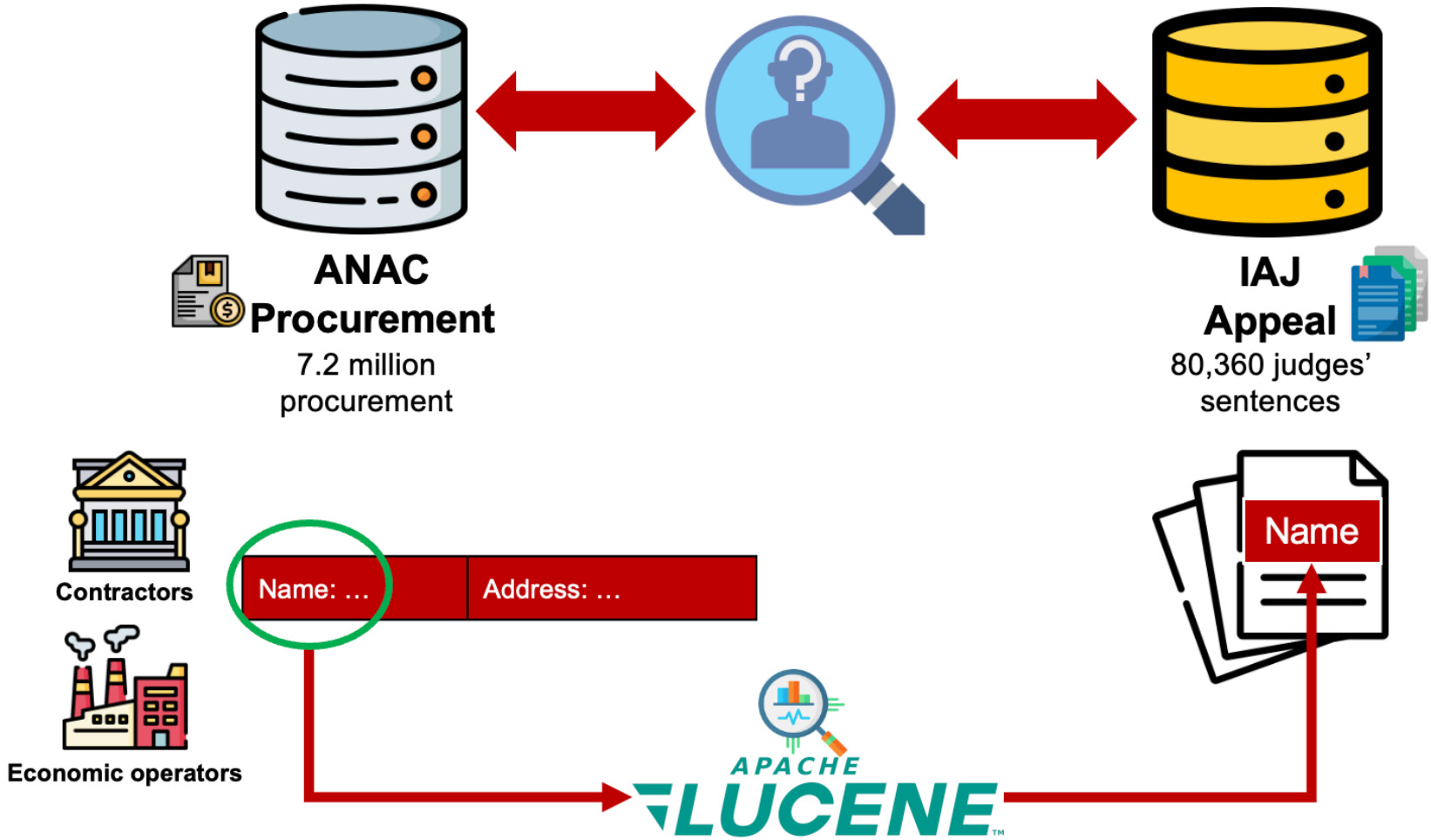
Search by procurement ID (CIG)

- The total number of CIG found is **8,062**: this means that the probability that a sentence in our archive refers to a CIG is about **10%**.

Topic	Value
Procurement type	services: 61.634% public works: 22.163% goods/supplies: 16.248%

Quantitative description of ANAC procurement by application of Lucene on the CIG inside sentences.

Search by contract authority and economic operators' denomination



Search by contract authority and economic operators' denomination

Type	**Names found** (total)	**Names found** (percent)	**Time**
contracting authorities	37,890	6.164%	23 min
economic operators	152,934	24.880%	2h:15m

Search result of the contracting authorities and economic operators in sentences by Lucene.

Definition of the litigation measure to estimate economic operator aggressiveness

known by APACHE LUCENE

$$litigation\_tendency = \frac{n\_recourses}{total\_participations}$$ for a single economic operator

unknown

Partecipations to procurements for an **economic operator**

Definition of the litigation measure to estimate economic operator aggressiveness

known by APACHE LUCENE

$$litigation\_tendency = \frac{n\_recourses}{total\_participations}$$ for a single economic operator

unknown

Partecipations to procurements for an **economic operator**

known (**Awards** table)

$$p\_award = \frac{n\_awards}{total\_participations}$$ for a single economic operator

unknown

Definition of the litigation measure to estimate economic operator aggressiveness

Partecipations to procurements for an **economic operator**

known by APACHE LUCENE

$$litigation\_tendency = \frac{n\_recourses}{total\_participations}$$ for a single economic operator

unknown

known (**Awards** table)

$$p\_award = \frac{n\_awards}{total\_participations}$$ for a single economic operator

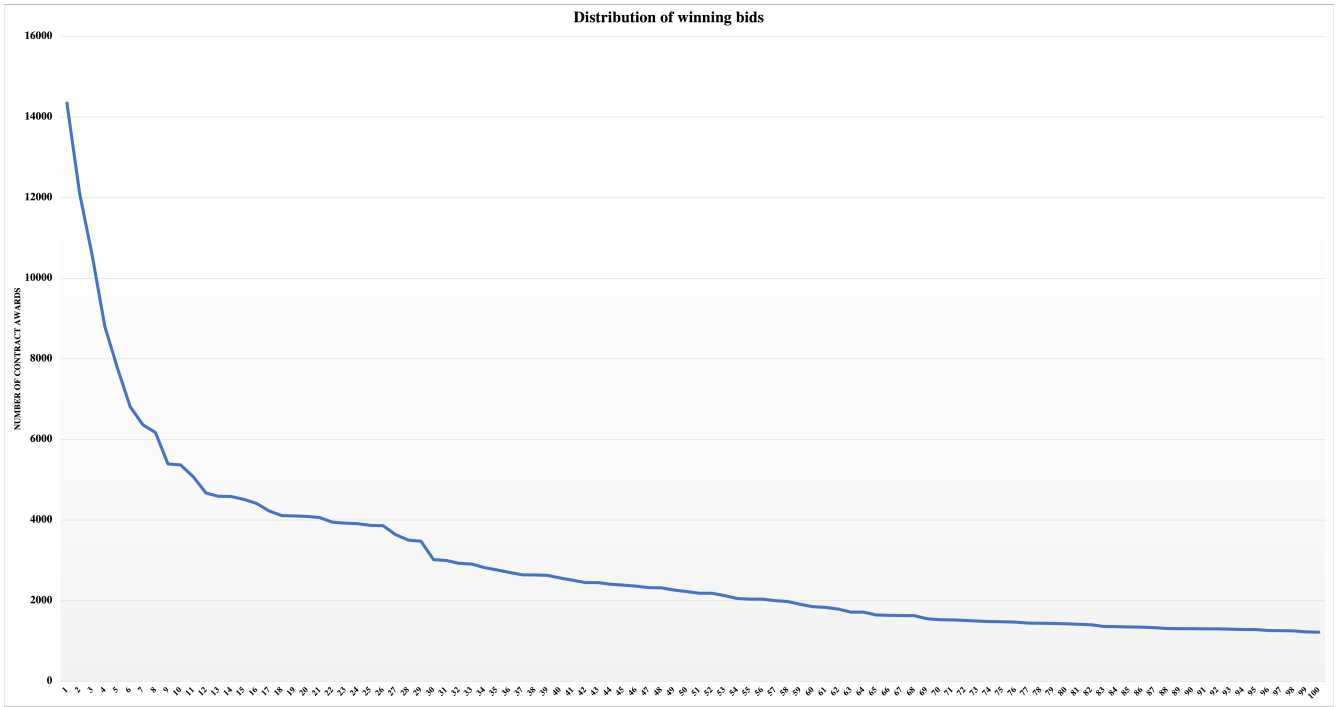
unknown

Estimation:

$$\hat{p\_award} = \frac{n\_tenders\_awarded}{total\_participants} \rightarrow \hat{total\_partitipations} = \frac{n\_awards}{\hat{p\_award}}$$

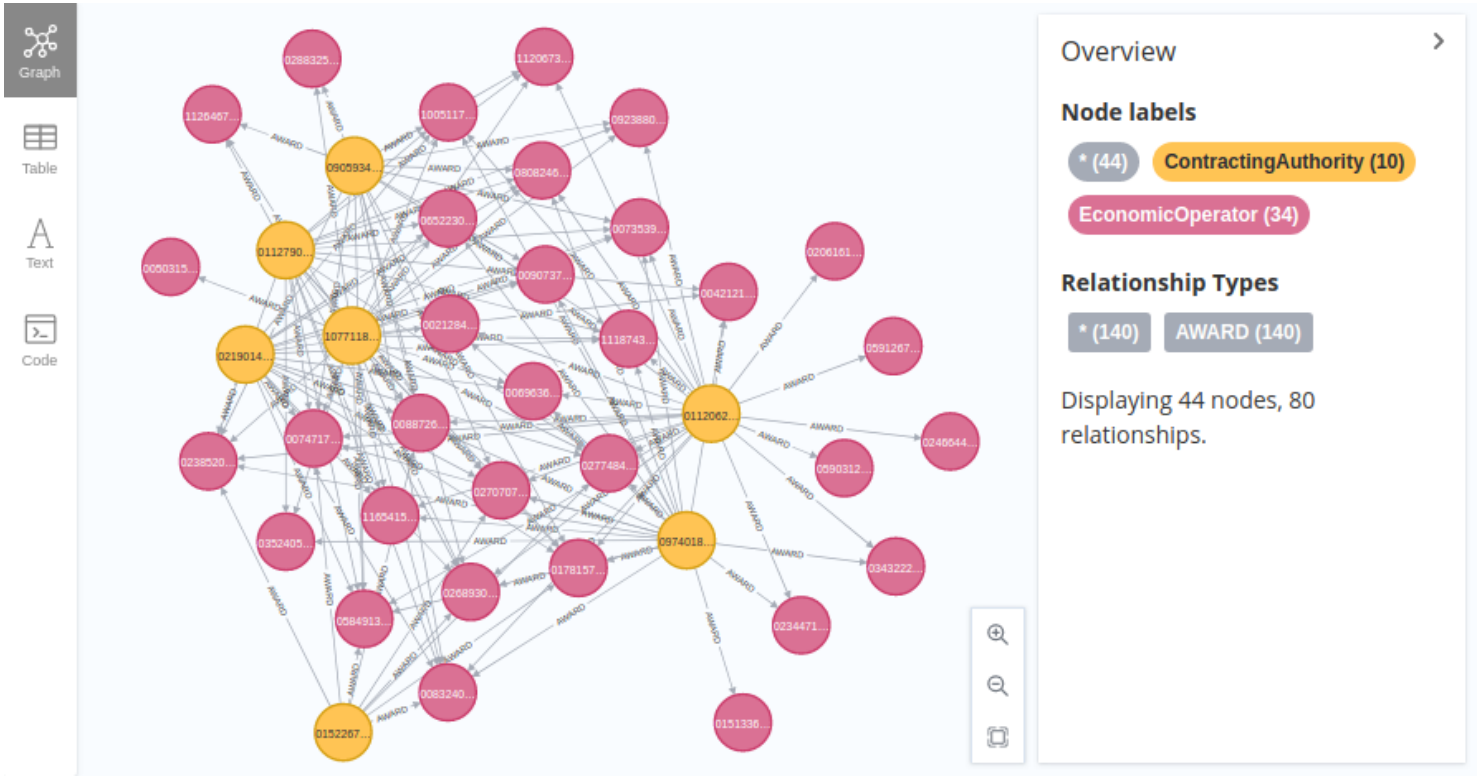
to every procurement (**Awards** table)

Definition of the litigation measure with estimation of participation in public tenders



The power-law distribution of the litigation measure computed by equation *litigation_tendency*.
We fitted it by the Maximum Likelihood Estimation $p(x) = 1.38 * x^{-2.4}$ with a RMSE $= 0.3$.

Analysis on the Graph



Part of the graph obtained in Neo4j containing the relationship `Awards` (edges) between the contracting authorities (yellow nodes) and the economic operators (pink nodes).

Analysis on the Graph

- We exploited the Neo4j library "Graph Data Science" (GDS);
  - algorithms of interest are built into the tool, avoiding the need of external applications.
- Two main graph algorithms: *community* and *betweenness centrality* detection;
  - for the *community detection*, Neo4j GDS library contains the Louvain method;
  - for the *betweenness centrality* Brandes' approximate algorithm.

Analysis on the Graph

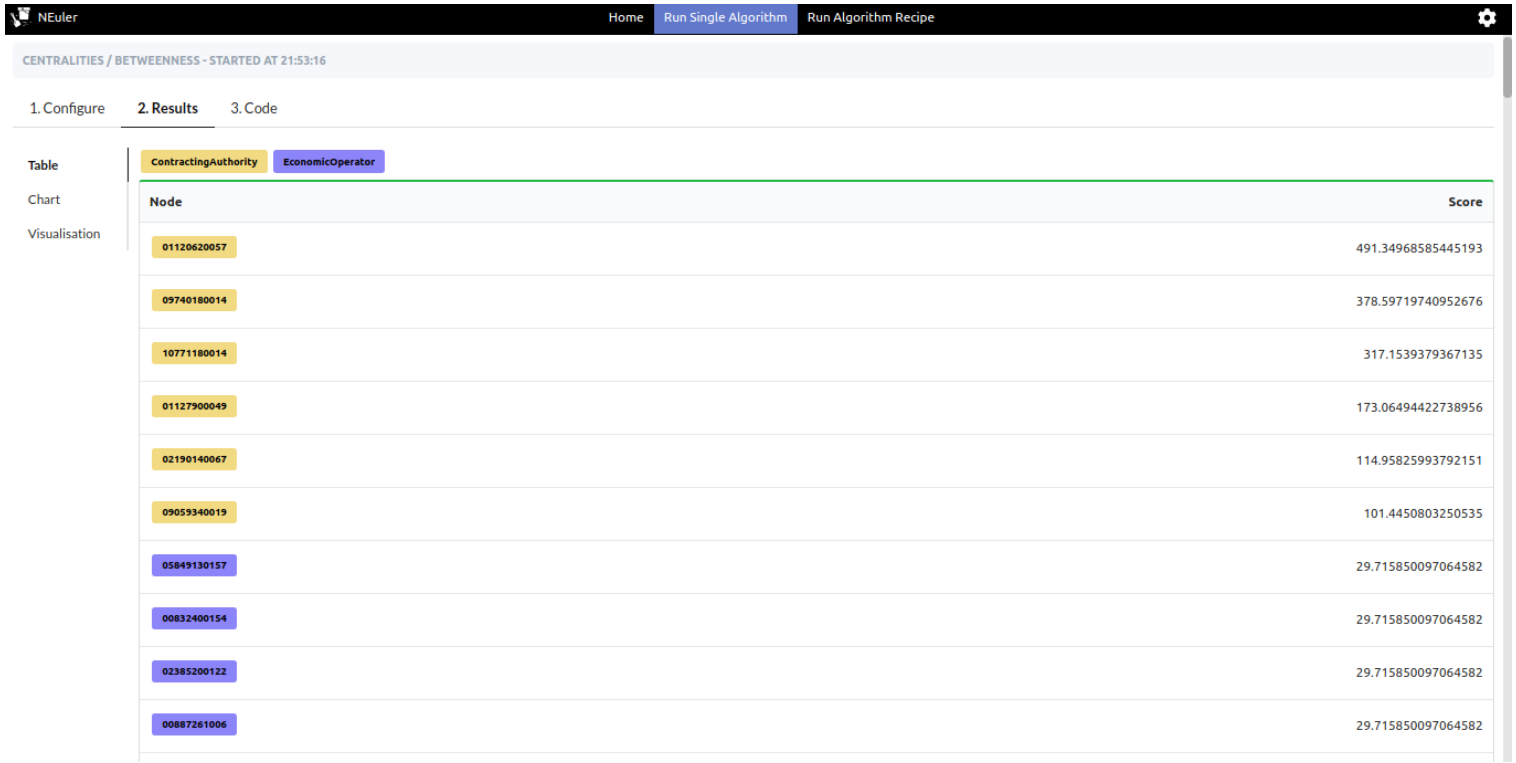
Community	Communities	Size	Nodes
242	238,242	38	10771180014, 01127900049, 01120620057, 02190140067, 09238800156, 08082461008, 09012850153, 02774840595, 00832400154, 05903120631, 11206730159, 00488410010, 05912670964, 02385200122, 04411460639, 11264670156, 10172190018, 02006400960, 07123400157, 11654150157, 06032681006, 00747170157, 01781570591, 11187430159, 00735390155, 10051170156, 02061610792, 00421210485, 06522300968, 02466440167, 00503151201, 00807970157, 05849130157, 00887261006, 02707070963, 00696360155, 02689300123, 09284460962
307	307,307	12	09740180014, 01542210222, 00674840152, 00212840235, 03524050238, 01423300183, 00907371009, 02344710484, 00803890151, 03432221202, 01513360345, 06209390969
242	242,242	5	01522670031, 09059340019, 00076670595, 09699320017, 02883250017
261	261,261	5	00514490010, 07820120017, 03747970014, 03717670016, 06484280018
343	343,343	4	02078000037, 00967720285, 01481120697, 00468910070

First five communities detected by the Louvain algorithm: green nodes are contracting authorities; orange nodes are economic operators (the communities are in descending order according to their "Size" - number of nodes -).

Analysis on the Graph

Graphical representation of the community detection: light blue nodes are the contracting authorities; economic operators are the blue nodes.

Analysis on the Graph



First ten nodes with higher *centrality* score: yellow nodes are contracting authorities; orange nodes are economic operators. The "Score" on the right side indicates how central the node is; nodes are sorted in descending order according to this value.

Analysis on the Graph

Graphical representation of the betweenness centrality: light blue nodes are contracting authorities; blue nodes are economic operators.

Conclusions and future work

- In this paper, we explored the possibility of the integration by IR of two information sources (ANAC and IAJ) about procurement using common data in both datasets.
- By fitting models on observed data applying the principle of MLE, we estimated the probability that a company awards a tender and the number of participation.
  - These are the ingredients for the identification of the companies that cause the highest number of litigation whose elimination could drastically improve the justice overload.
- We applied also graph analytic to identify the communities formed by the public contractors and economic operators with recurrent procurement.

Conclusions and future work

- As future work, we plan to:
  - study the use of Legal BERT to search within the judgments for named entities such as the names of the economic operators that were excluded from the tender selection (thus not tracked in the ANAC dataset) in order to create a graph database of the economic operators that may appear in the appeals despite an unsuccessful bid.
  - After labeling procurements with appeals, evaluate supervised machine learning models.

Thank you!