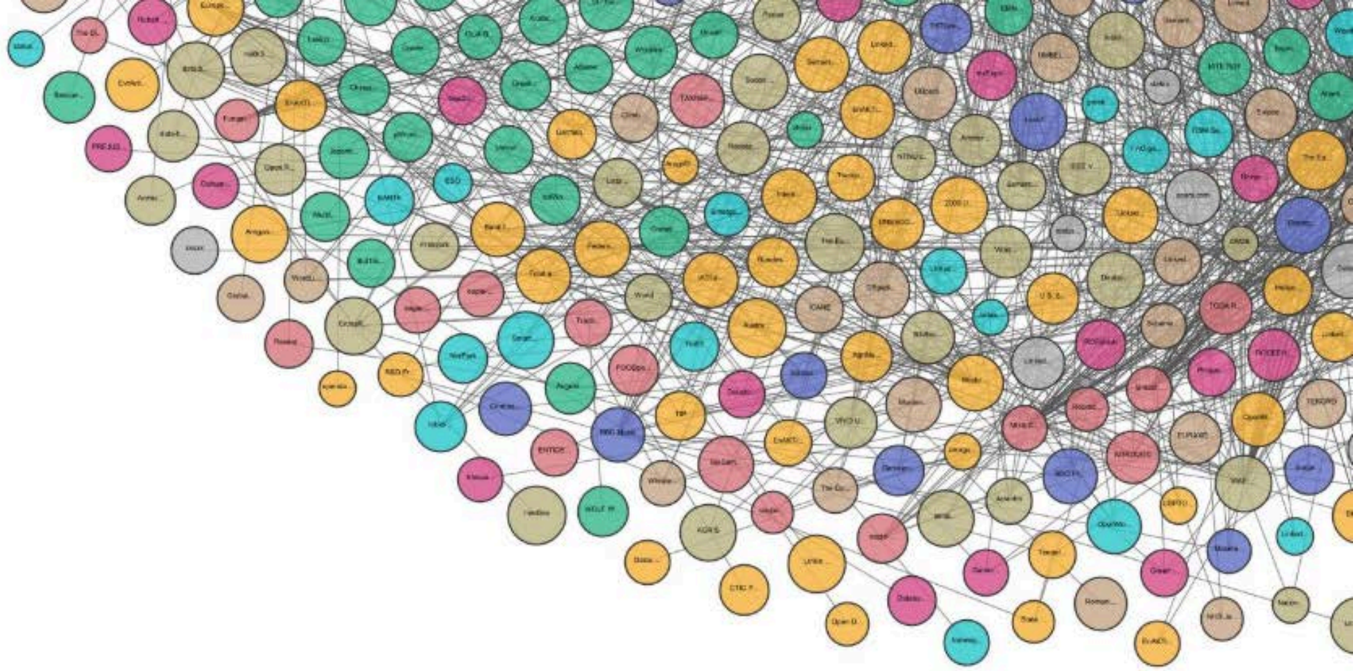


Links in Large Integrated Knowledge Graphs: Analysis, Refinement, and Domain Applications

Shuai Wang

ORCID: 0000-0002-1261-9930
Date: 17 December 2025

Image retrieved from https://lod-cloud.net/. Accessed on 1 July 2025

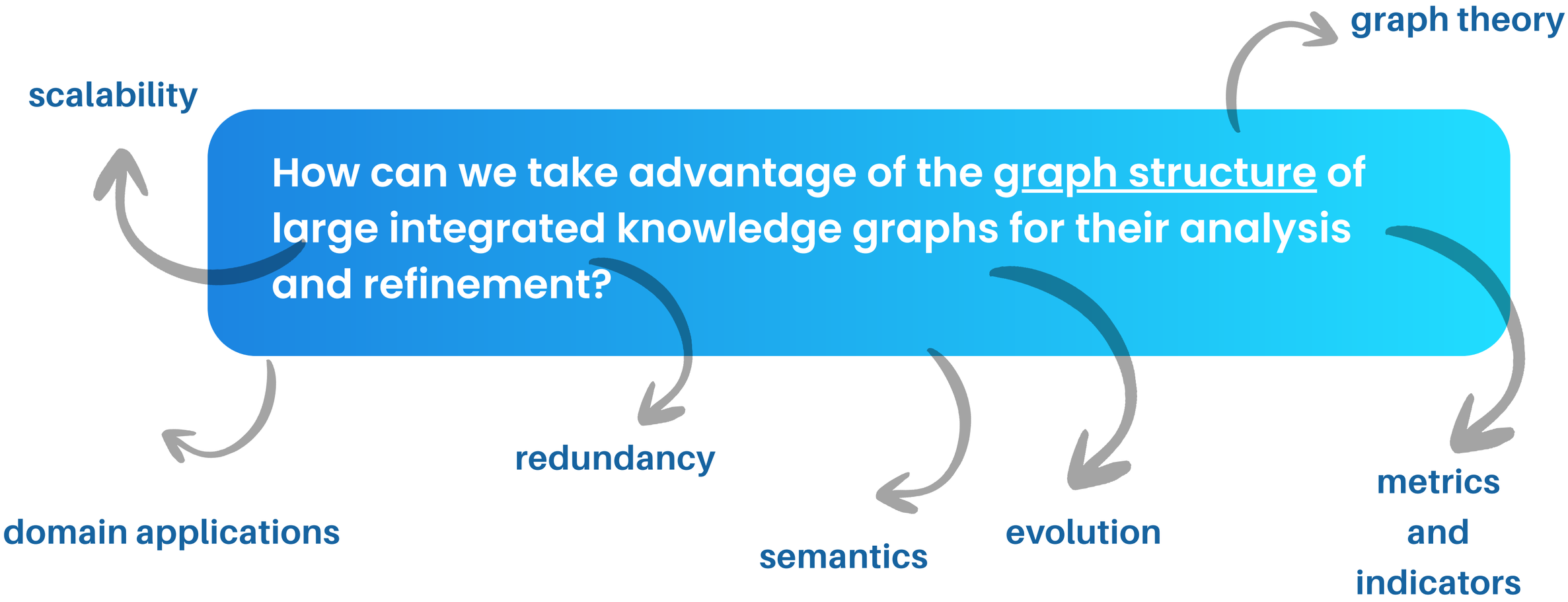
Motivation

- In my thesis, I studied knowledge graphs that are linked data. Data graphs are excluded.
- Integrating knowledge graphs of linked data published on the semantic web can result in (inherited) errors.
- Previous work showed that errors could be amplified in graphs due to the logical properties.
- The semantics in knowledge graphs imply structural properties (e.g., class subsumption), vice versa.
- Example: graphs corresponding to transitive, asymmetric, and irreflexive relations should not have cycles at all.

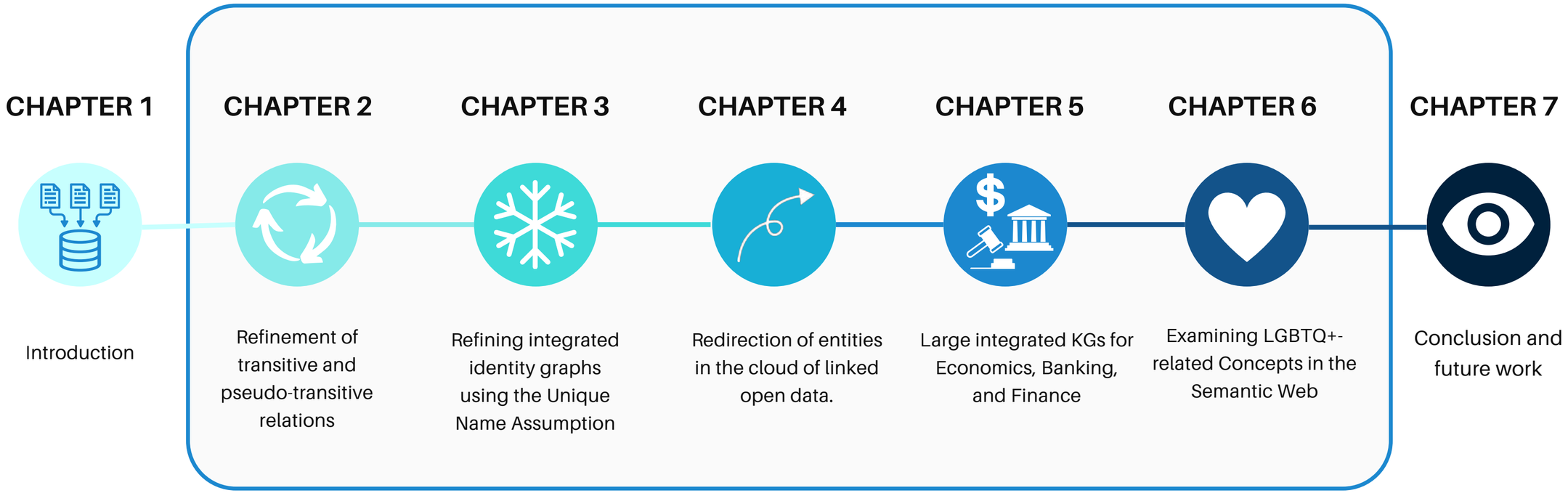
Research Question

How can we take advantage of the graph structure of large integrated knowledge graphs for their analysis and refinement?

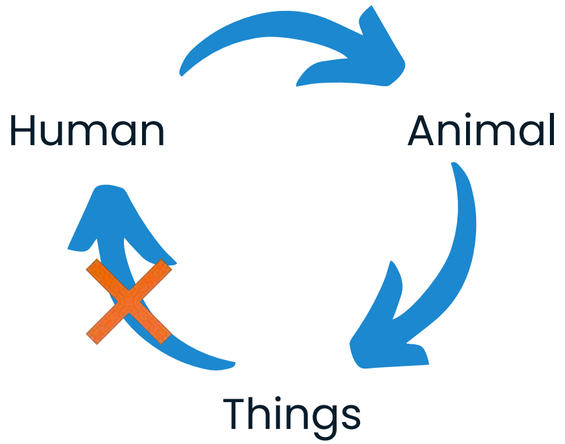
Research Question



Thesis Outline



Ch2: Transitive and (Pseudo-)transitive relations



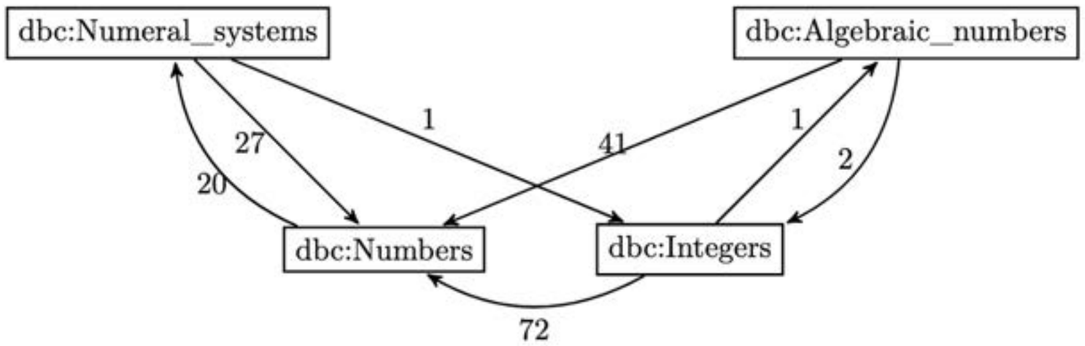
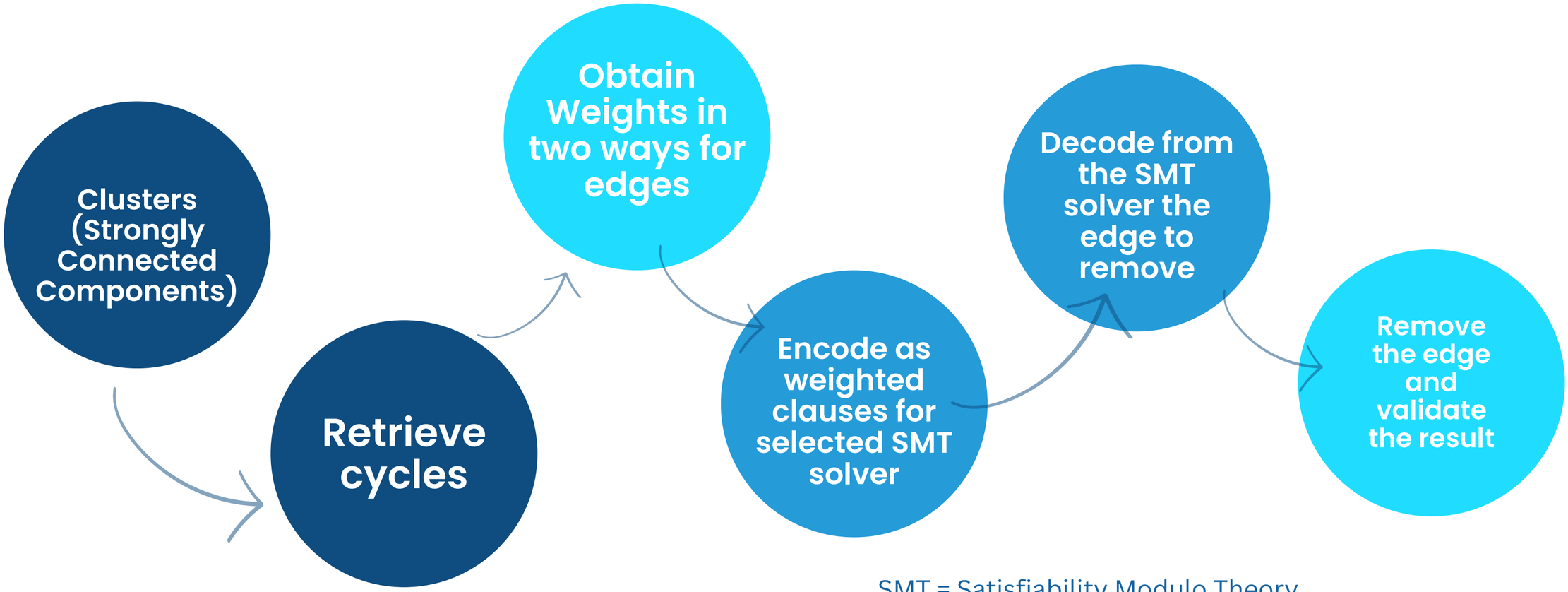


Figure 6: An example subgraph of `skos:broader` with weights.

- (Pseudo)transitive relations
- Complex nested cycles (LOD-a-lot)
- An algorithm to make knowledge graphs acyclic
- preserving as much information as possible
- Multilingual, no reference to textual info, etc.

	rdfs:subClassOf	skos:broader
Type	Transitive	Pseudo-Transitive
#Edges	4.4M	11.8M
#Nodes	3.6M	5.7M
#Nodes in the Largest SCC	837	82K
#Edges in the Largest SCC	1.4K	365.9K

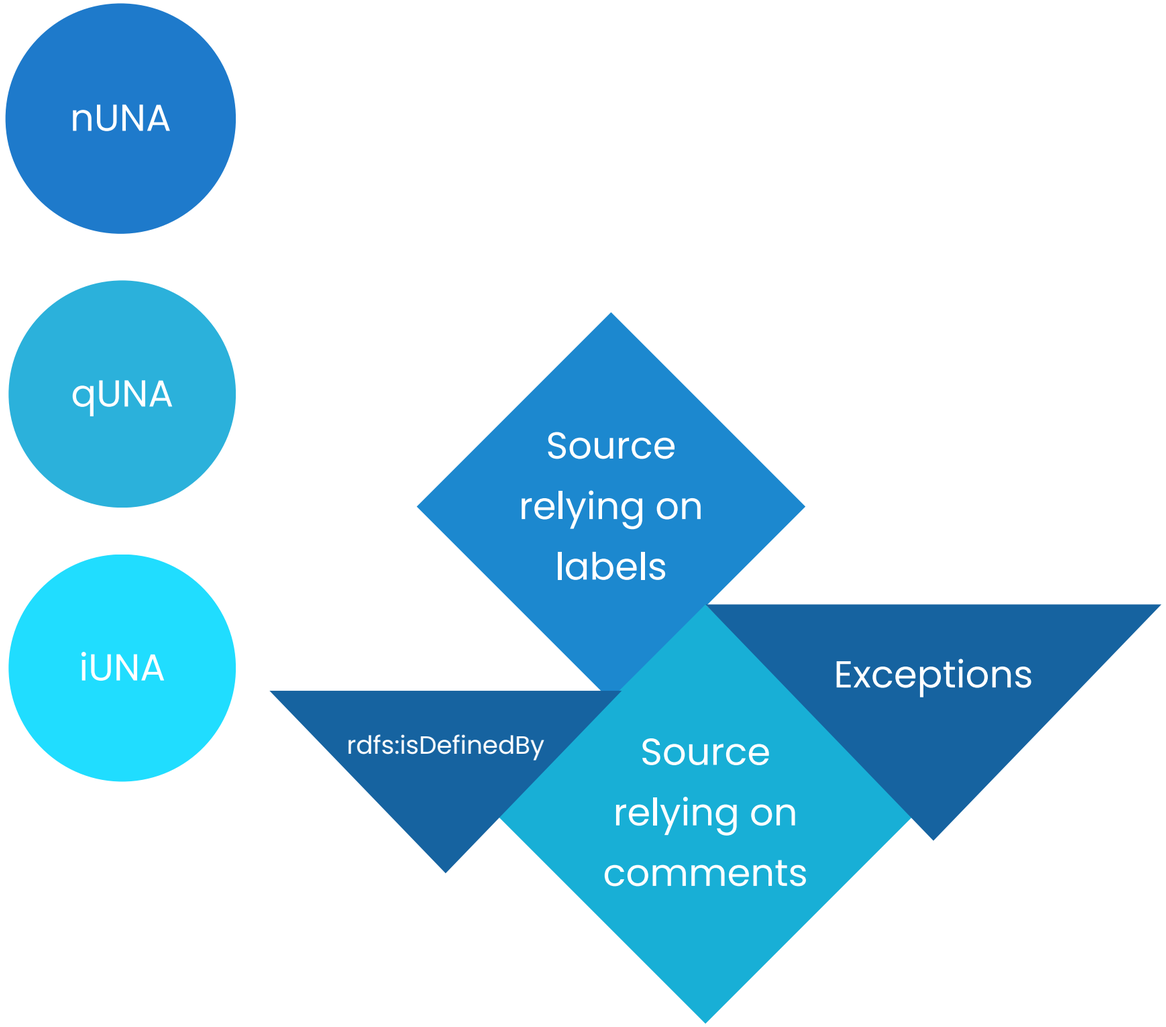
Ch2: Transitive and (Pseudo-)transitive relations

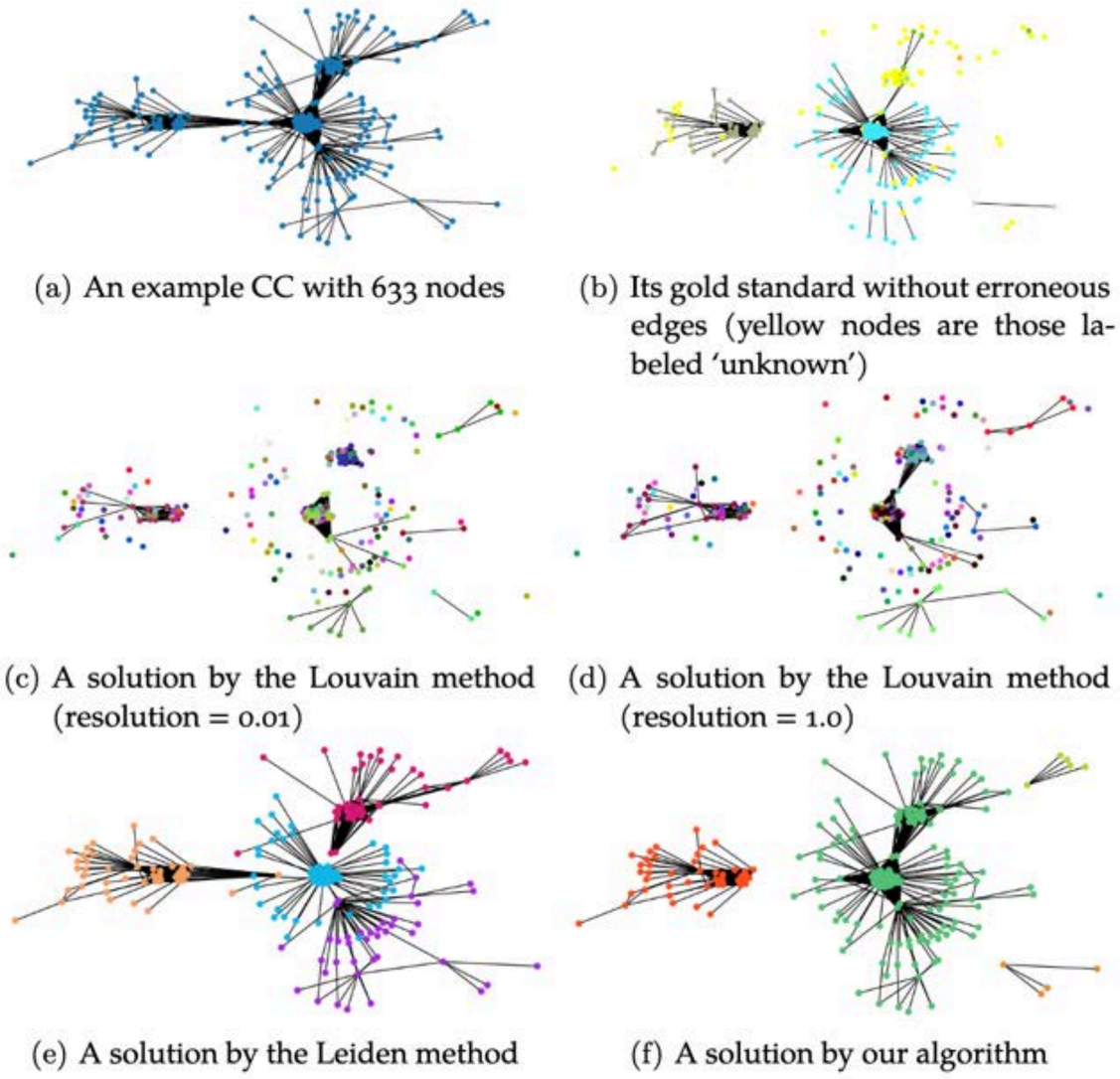


SMT = Satisfiability Modulo Theory

Algorithm Design

Ch3: Identity Graphs & Unique Name Assumption





(a) An example CC with 633 nodes

(b) Its gold standard without erroneous edges (yellow nodes are those labeled 'unknown')

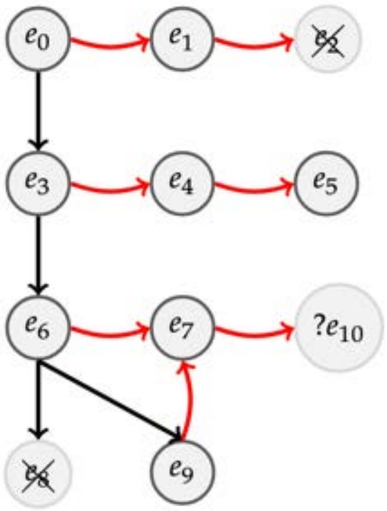
(c) A solution by the Louvain method (resolution = 0.01)

(d) A solution by the Louvain method (resolution = 1.0)

(e) A solution by the Leiden method

(f) A solution by our algorithm

Ch4: Examining the Evolution of Entities through Redirection



- Not found (HTTP 400+)
- Timeout
- Redirected until not found
- Error
- ...

- Outdated resources are typically redirected to new locations (19.4% of entities no longer exist on the web after 2 years).
- Semantics in redirect?
- 38-53% of entities are redirected. Redirection chains can involve as many as 9 entities.
- Estimate that between 45.1% and 83.2% of such links can preserve identity.

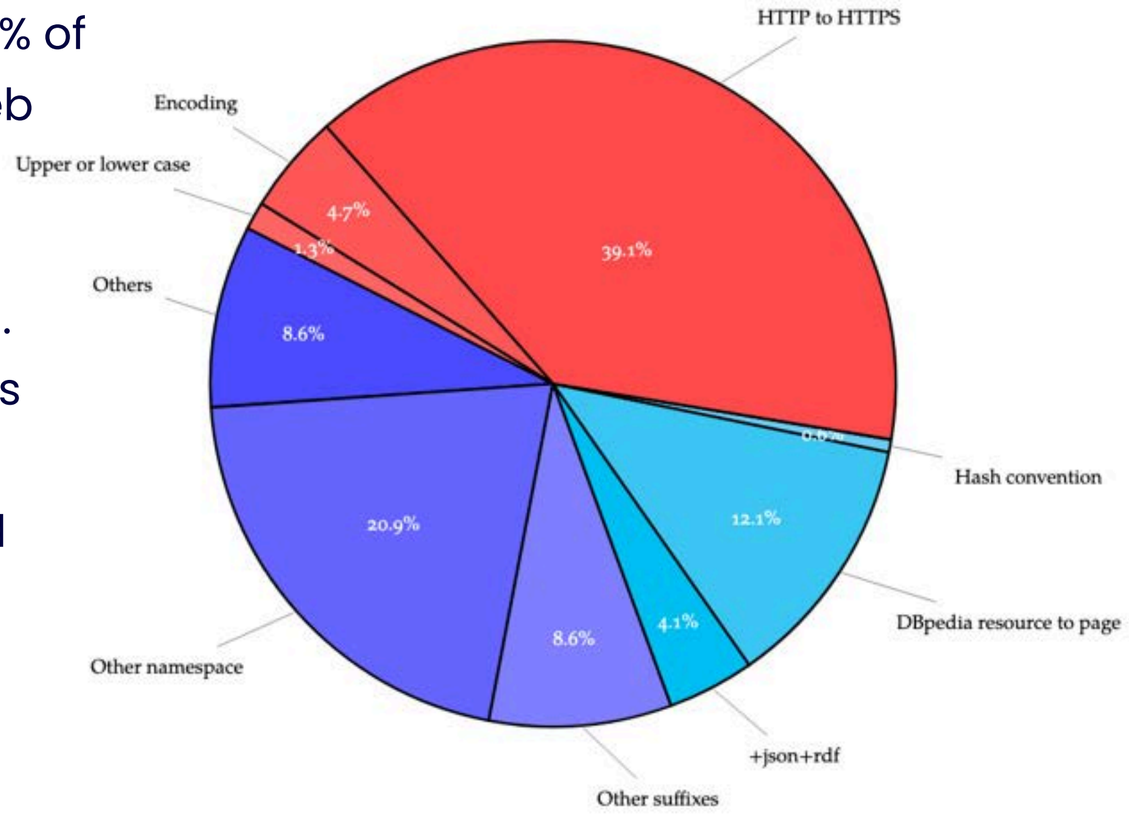


Figure 15: Proportion of redirection behavior among sampled entities

TWO DOMAIN APPLICATIONS

- If not the biggest few KGs (e.g., DBpedia), would the problem be different?
- What happens if we limit the integrated KGs to some restricted domains?
- What are the links in the resulting integrated graphs like?
- What is the scale of the problem of refinement?
- Do they need some new algorithms for domain-specific problems?
- Integrating some knowledge graphs of selected domains to form a larger KG.

Ch5: Links in Integrated KGs in Economics, Finance, and Banking

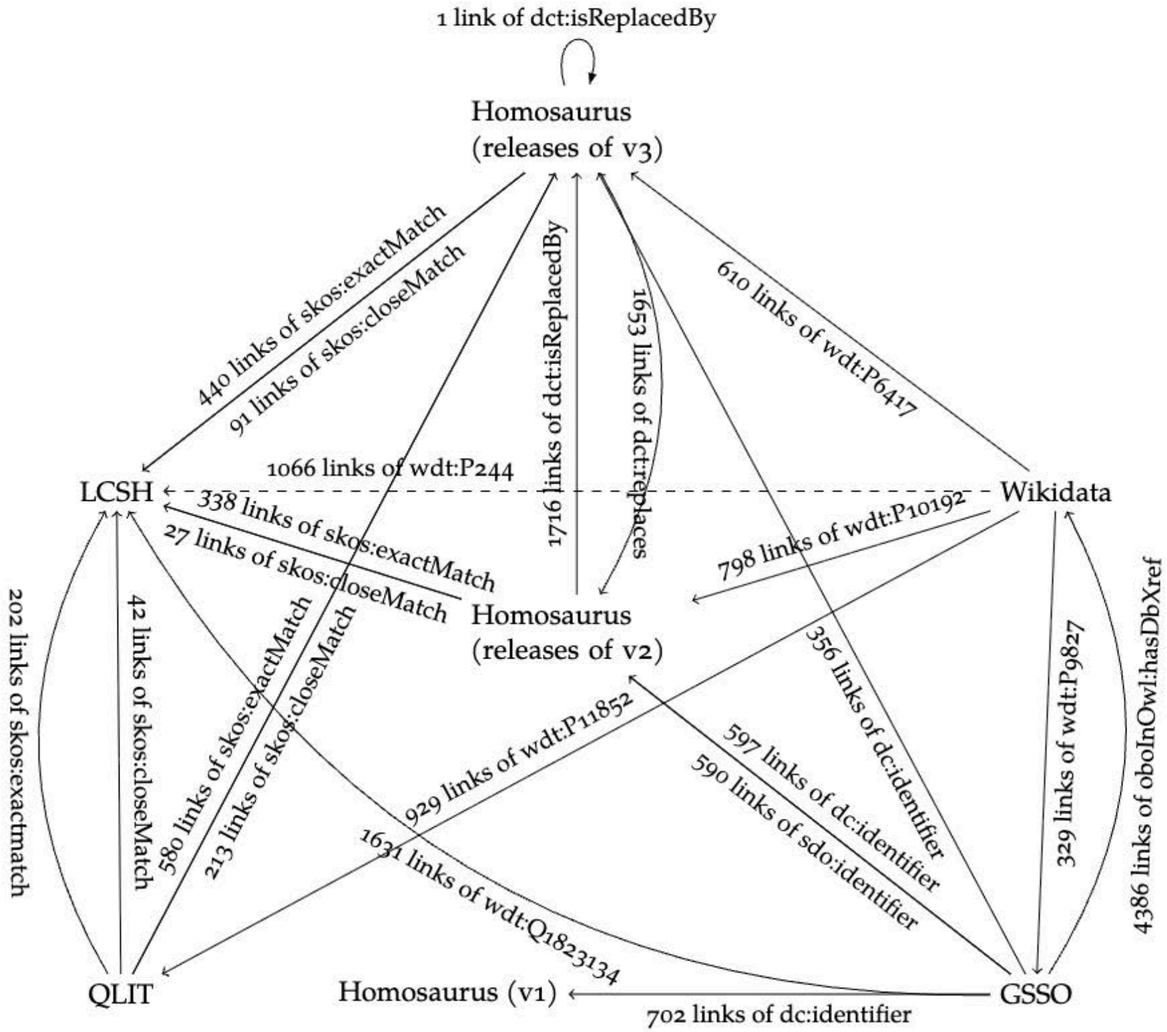
- More **skos:relatedMatch** and **skos:closeMatch** than **owl:sameAs.**
- A “large” connected component with **15** entities about the telephone system, telecommunications engineering, telecommunications, etc.
- Similar to the experiments before, larger clusters are more likely to have errors.
- Smaller and simpler. Possible for manual refinement because the entities represents concepts that are well defined.
- No cycle of rdfs:subClassOf found.
- No cycle with more than two entities for lkif-core:component, fro:divides20, and its inverse fro:divided_by.

Ch6: Examining LGBTQ+-related Concepts in the Semantic Web: Identity-related Link Discovery, Concept Change, Ambiguity, and Multilingual Information Enrichment

KG construction

1. Relation extraction
2. Refinement (wd:Q1823134->wdt:P244)
3. Redirection

- skos:exactMatch/closeMatch
- dc:identifier
- dct:isReplacedBy/replaces
- oboInOwl:hasDbXref
- wdt:P6417/P10192/etc.
- redirection (61+2): meta:redirectedTo
- ~~owl:sameAs~~



19.2K entities, 17.8K links.

6.4K WCCs. The largest WCCs are with 45, 36, 36, 35 entities, respectively.

Ch6: Examining LGBTQ+-related Concepts in the Semantic Web: Identity-related Link Discovery, Concept Change, Ambiguity, and Multilingual Information Enrichment

Scenario 1: Link Discovery

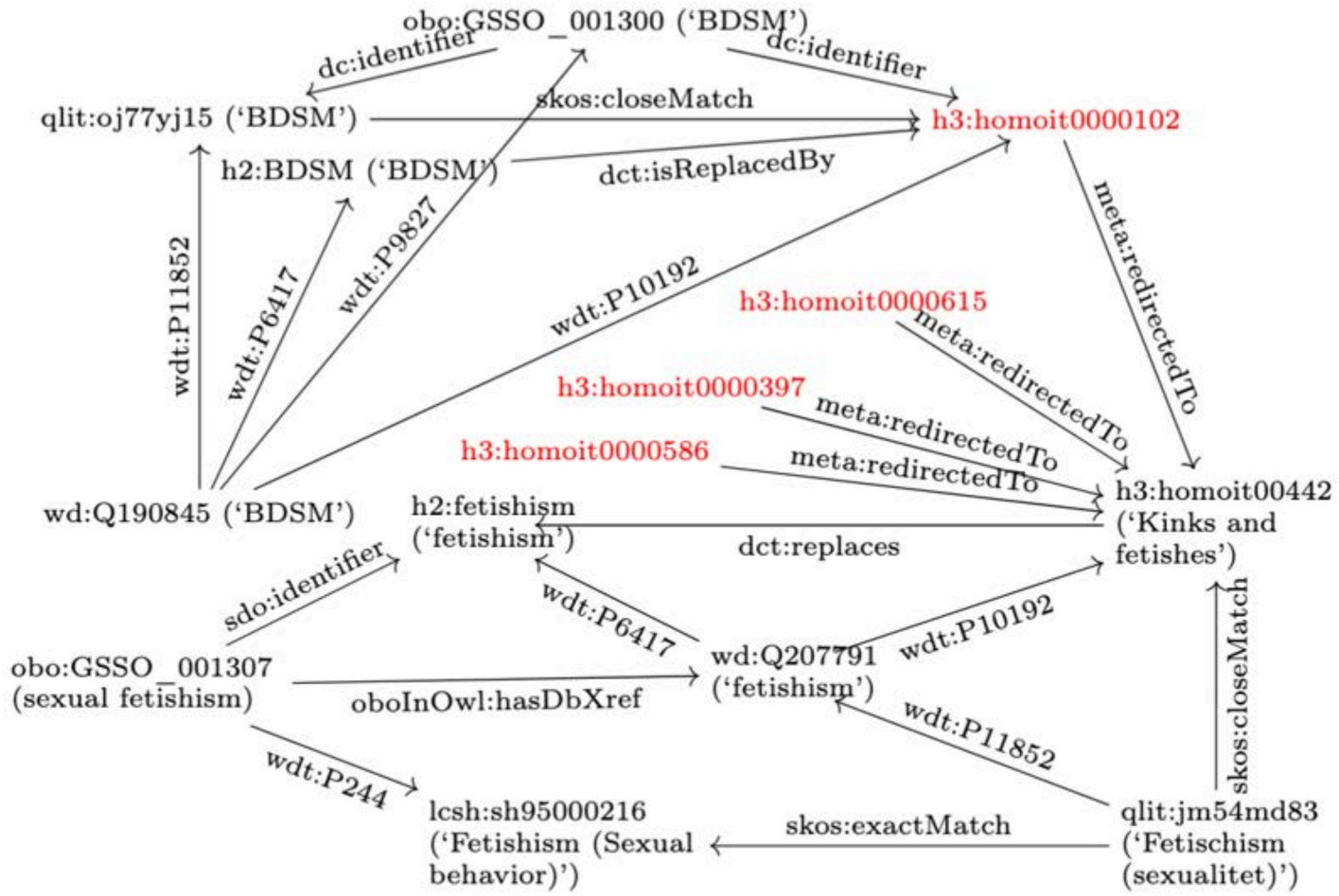
Scenario 3: Multilingual Information Reuse

- Not to start from scratch when constructing a linked vocabulary of a given language
- Reuse multilingual labels from other resources (1-to-1 mapping)
- To not miss their alternative labels.

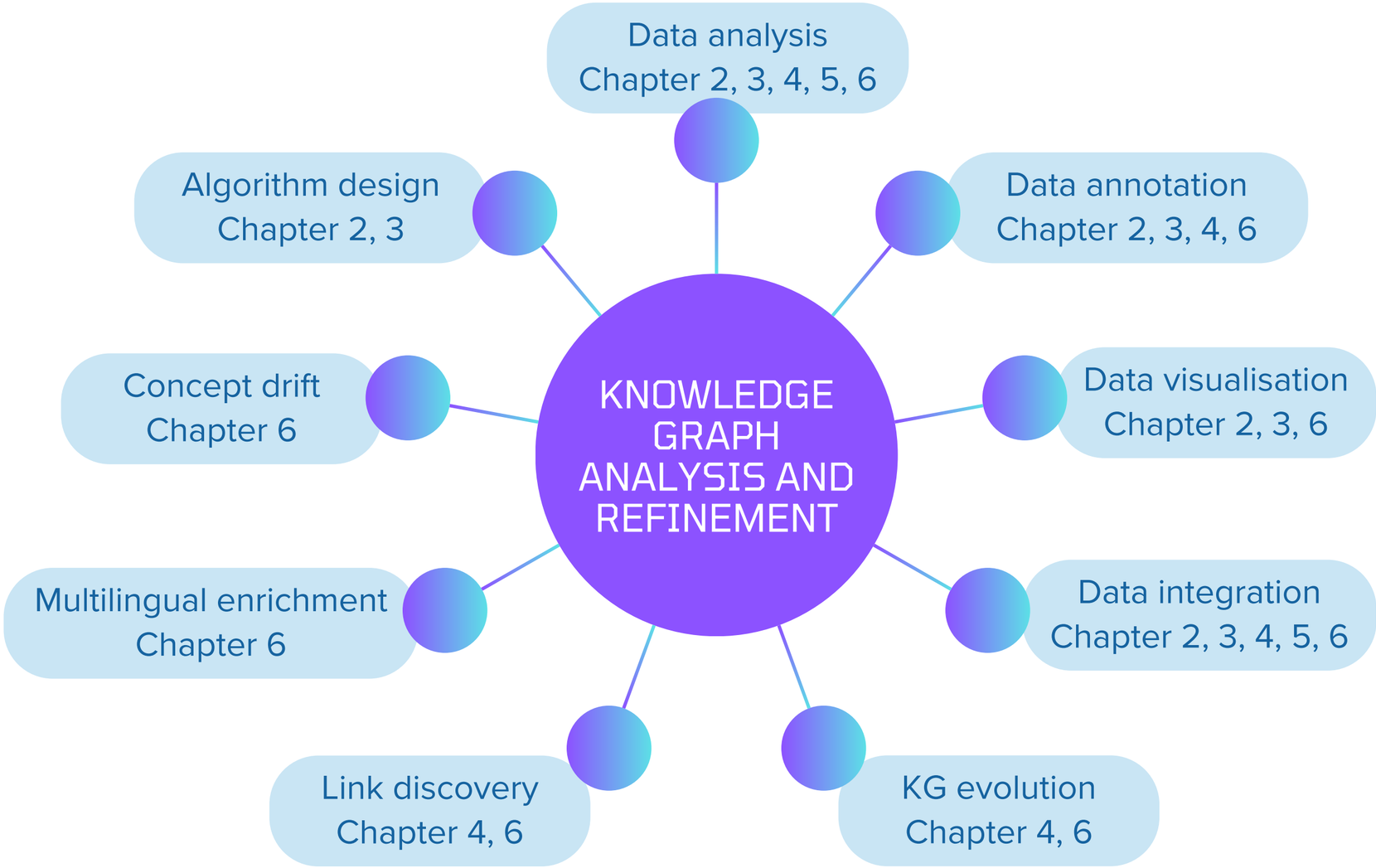
	English	Spanish	Chinese	Portuguese	German
Number of labels	1,696	951	893	881	824
Number of one-to-one mappings	429	333	287	299	298
Avg. labels per entity	3.94	2.86	3.11	2.95	2.77

Scenario 2: Concept Drift and Change

"elephant in the room"



Conclusion


KNOWLEDGE GRAPH ANALYSIS AND REFINEMENT
Data analysis
Chapter 2, 3, 4, 5, 6
Algorithm design
Chapter 2, 3
Data annotation
Chapter 2, 3, 4, 6
Concept drift
Chapter 6
Data visualisation
Chapter 2, 3, 6
Multilingual enrichment
Chapter 6
Data integration
Chapter 2, 3, 4, 5, 6
Link discovery
Chapter 4, 6
KG evolution
Chapter 4, 6

Thank you all!

Prof. Frank van Harmelen
Vrije Universiteit Amsterdam

Dr. Peter Bloem
Vrije Universiteit Amsterdam

Dr. Joe Raad
Université Paris-Saclay

https://shuai.ai

and many more...