Exploring Importance Measures for Summarizing RDF/S KBs

Alexandros Pappas, Georgia Troullinou, Giannis Roussakis
Haridimos Kondylakis, Dimitris Plexousakis

Institute of Computer Science, FORTH

Structure

Problem Definition

Explosion of the Web Data and the associated Linked Open Data creates:

1. Huge volume of data stored as graphs.
2. Extremely complex schemas.

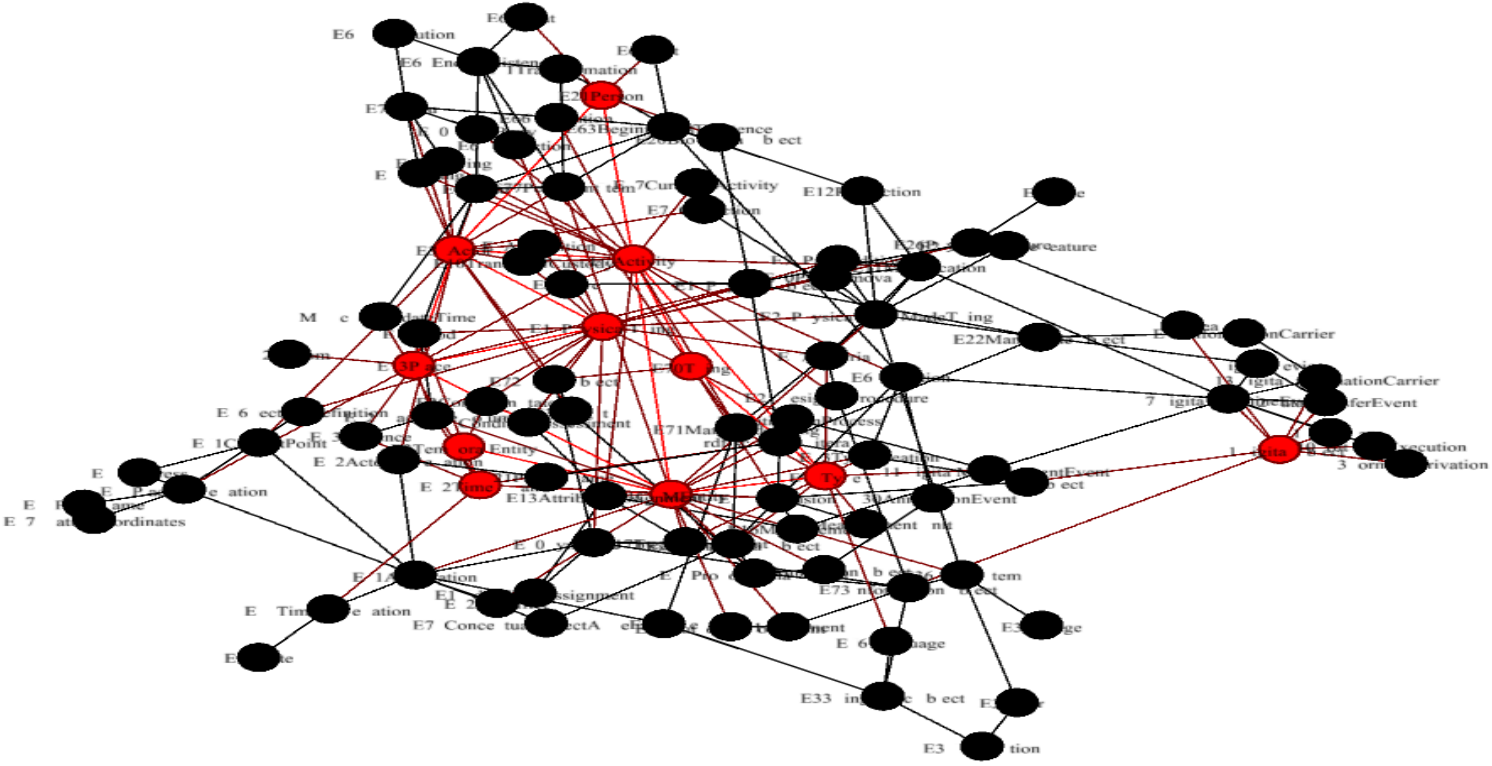
Problems:

1. Difficult to comprehend.
2. Limiting the exploration and the exploitation potential of the information.

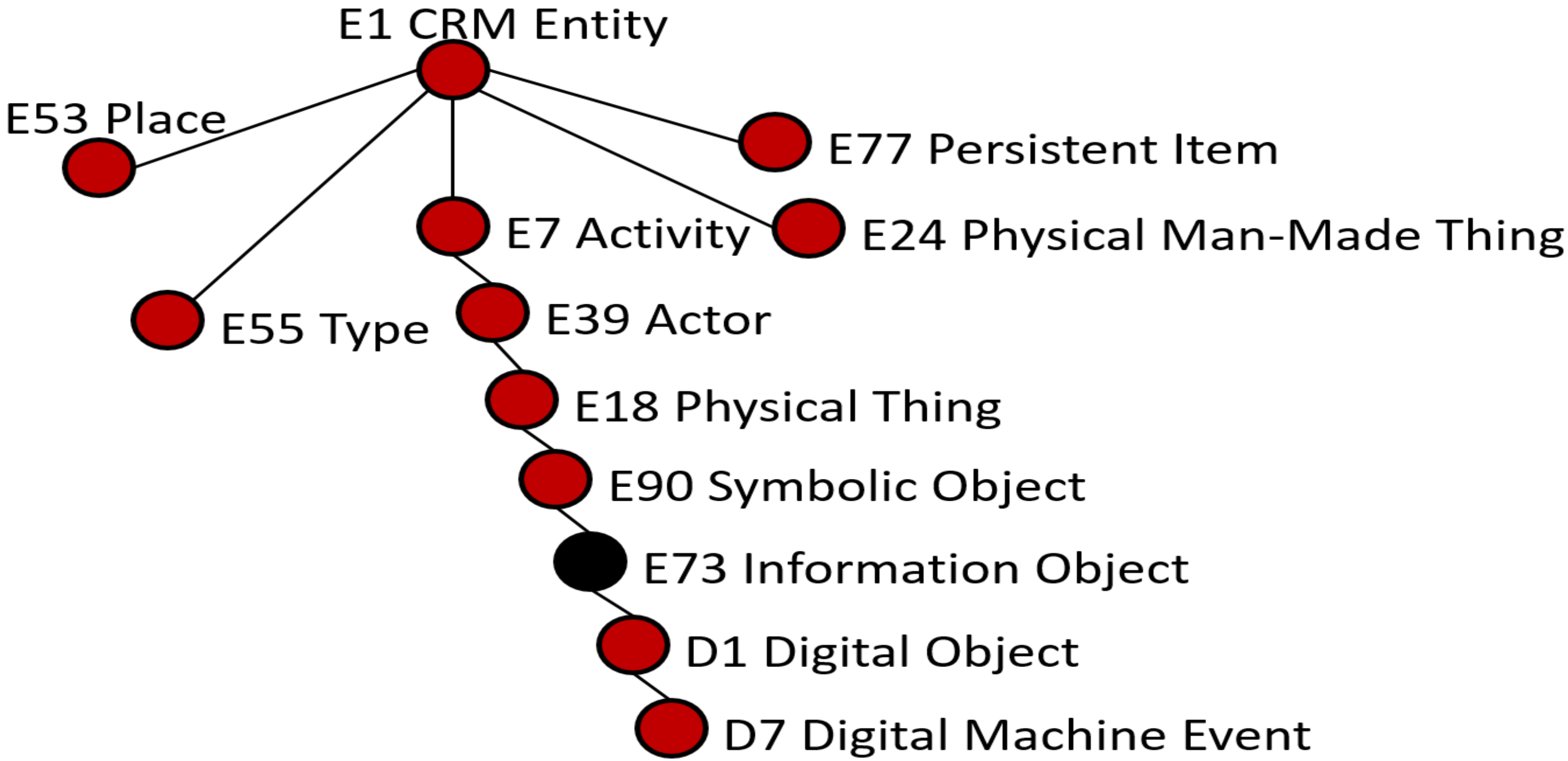
The idea

Summarization

The process of distilling knowledge from an ontology in order to produce an abridged version.



Original Graph


E1 CRM Entity
E53 Place
E77 Persistent Item
E7 Activity
E24 Physical Man-Made Thing
E55 Type
E39 Actor
E18 Physical Thing
E90 Symbolic Object
E73 Information Object
D1 Digital Object
D7 Digital Machine Event

Summary

Central questions to the process of summarization

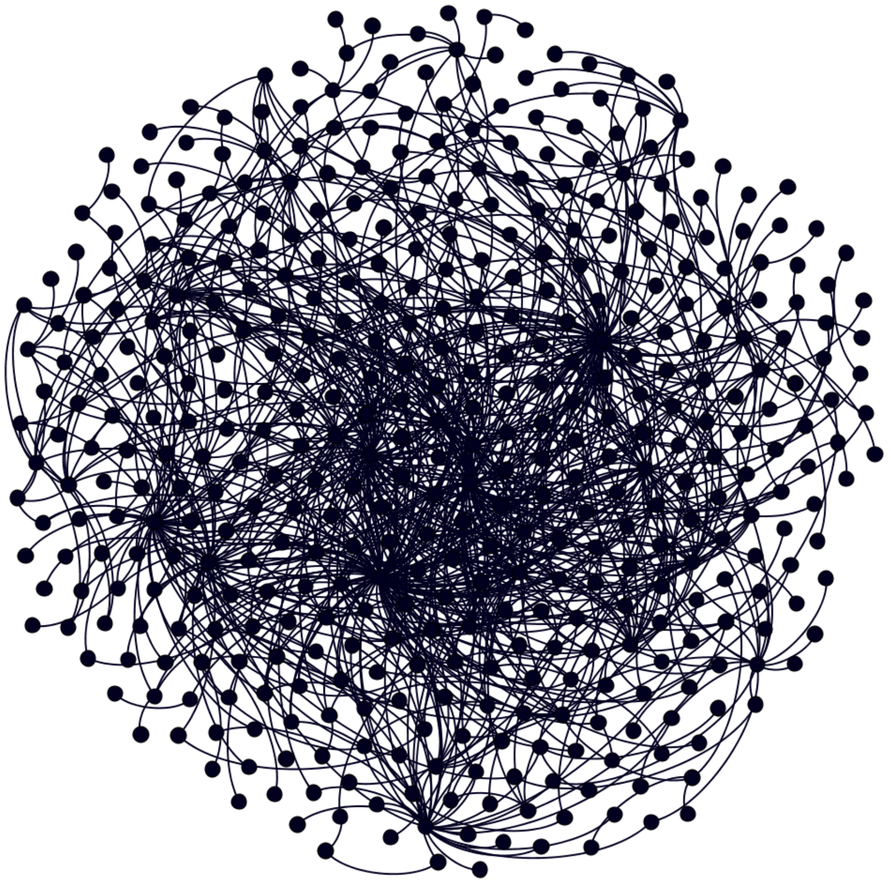
1. How to identify the most important nodes
2. How to link those nodes to produce a valid sub-schema graph.

Importance Measures

Importance measures, produce rankings which seek to identify the role and importance of any vertex in a graph.

"There is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is little agreement on the proper procedure for its measurement "

-Freeman 1978



Graph Instance

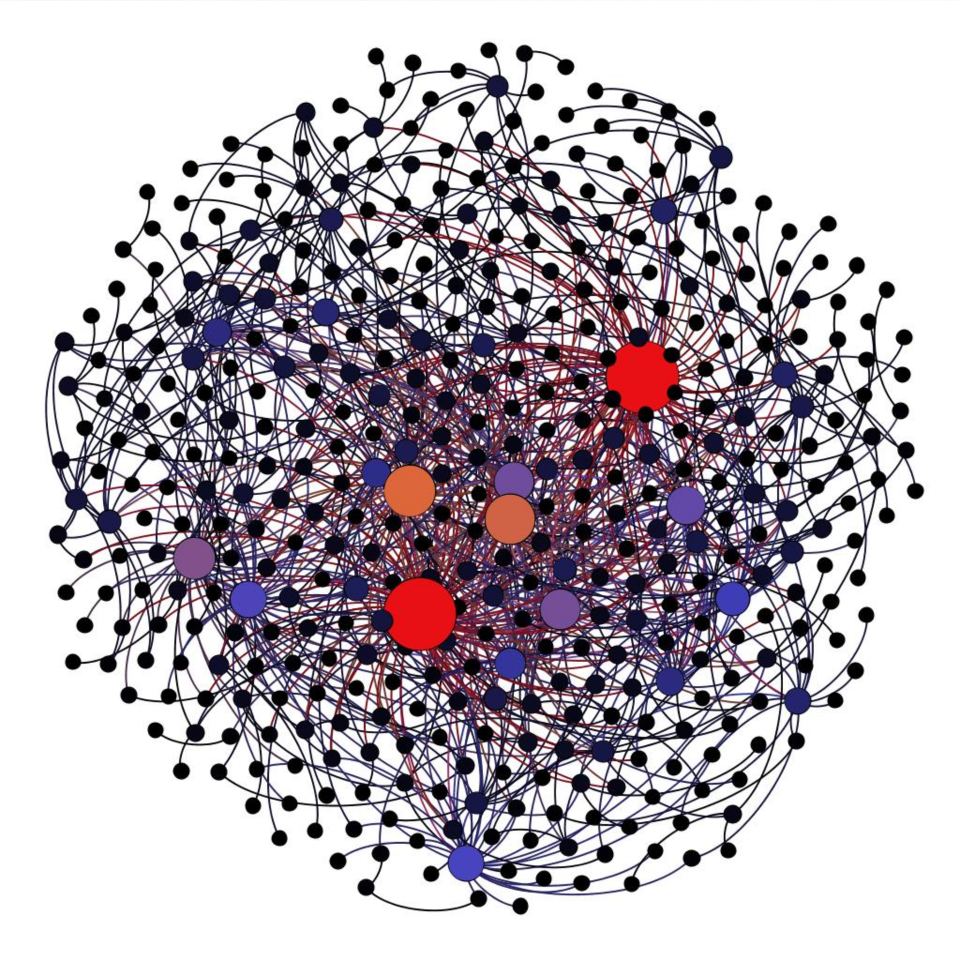
Nodes Ranking

Color:

Min size: 4 Max size: 20

Tool:

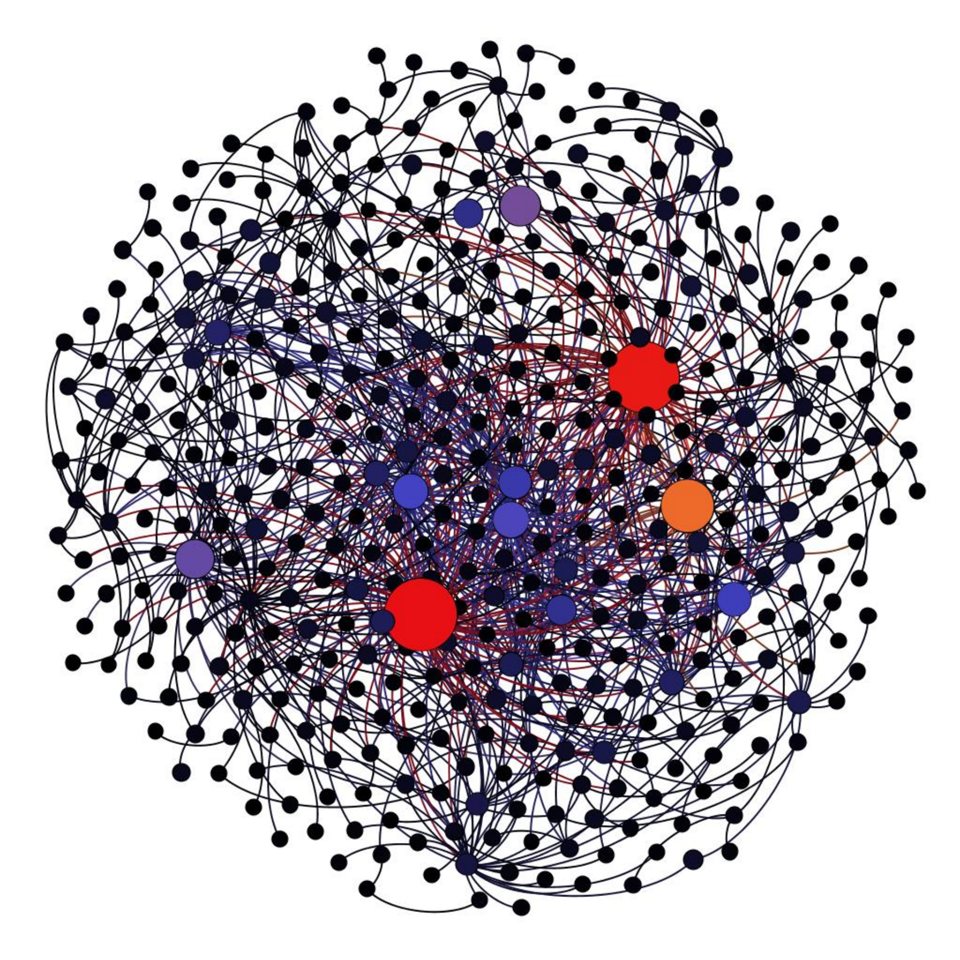
Gephi
makes graphs handy



CENTRALITIES (1/6)

Degree

Counts the number of edges incident to a node.

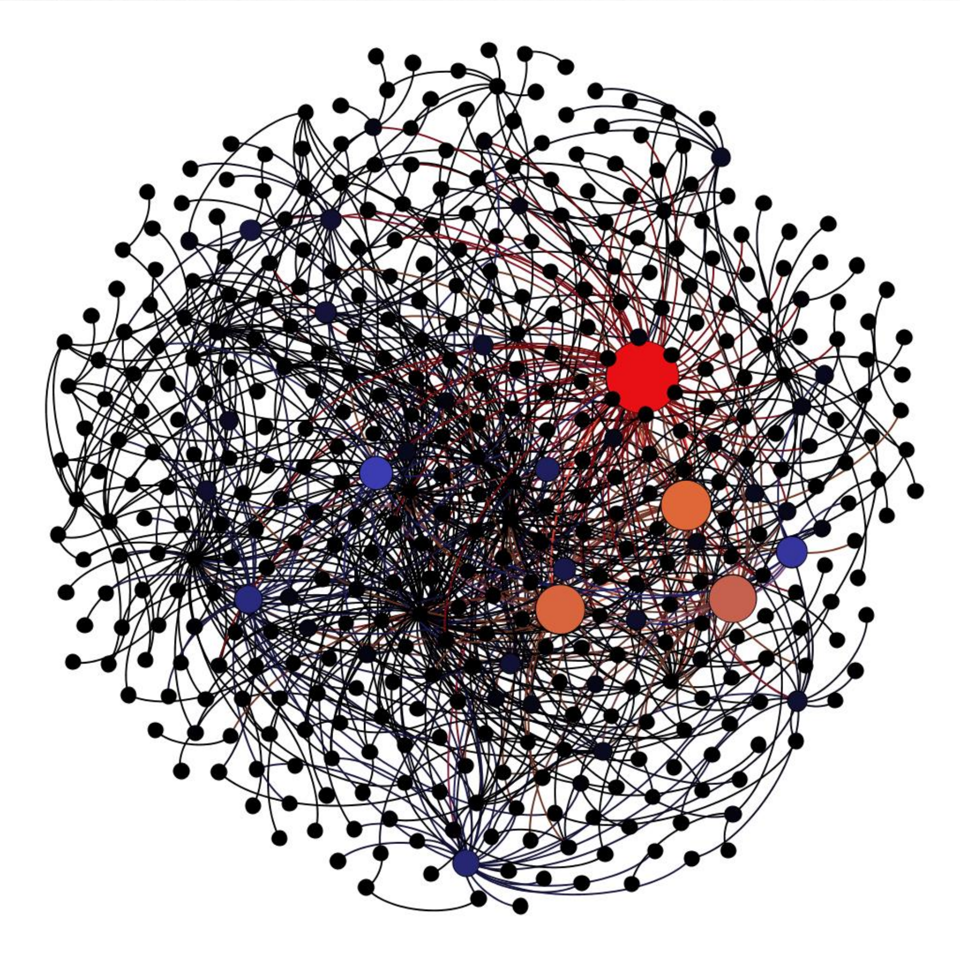


Ego

Defines how important a node is to his neighborhood.

The importance of a node v to a neighbor node u is defined as:

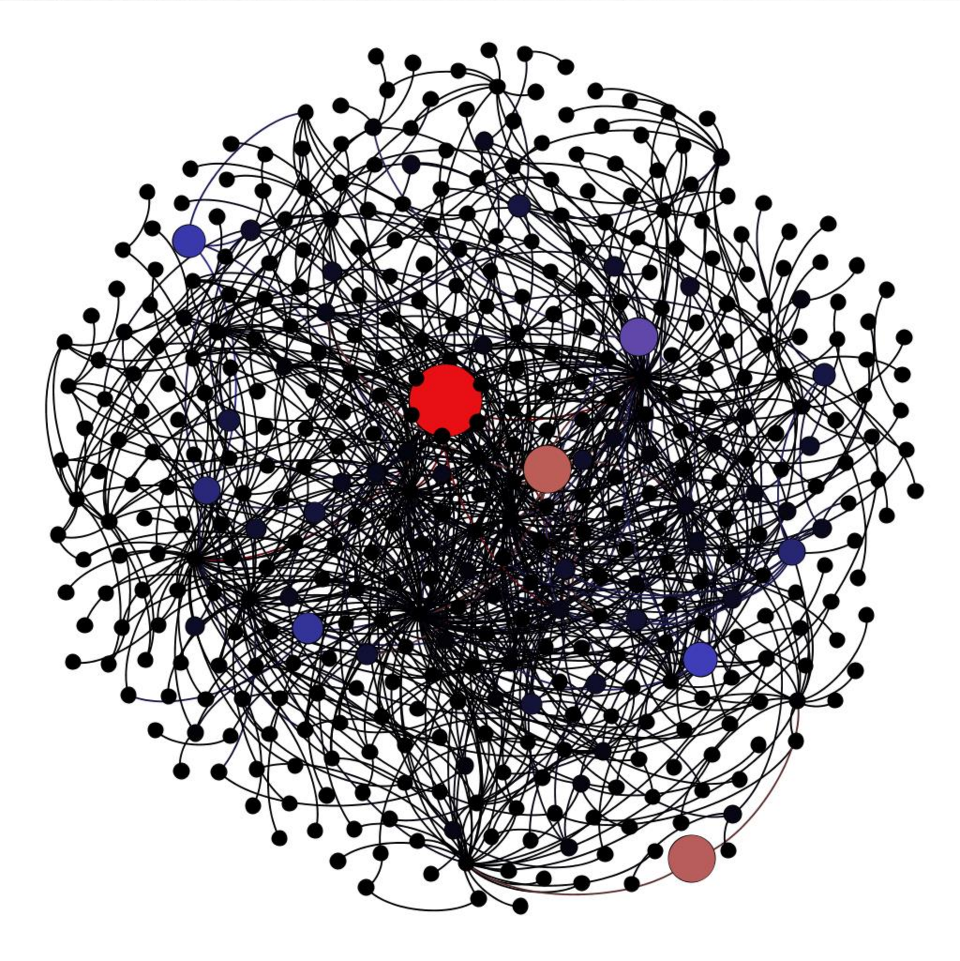
$$Im(v,u) = \frac{1}{Degree(u)}$$



CENTRALITIES (3/6)

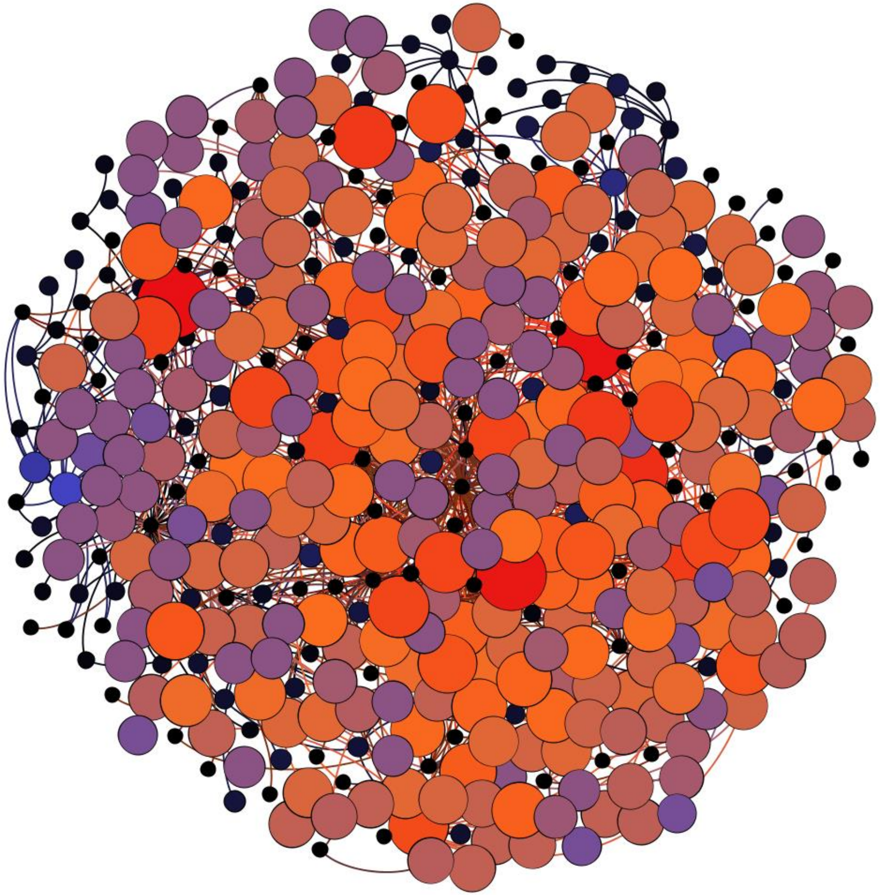
Betweenness

Quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.



Bridging Centrality

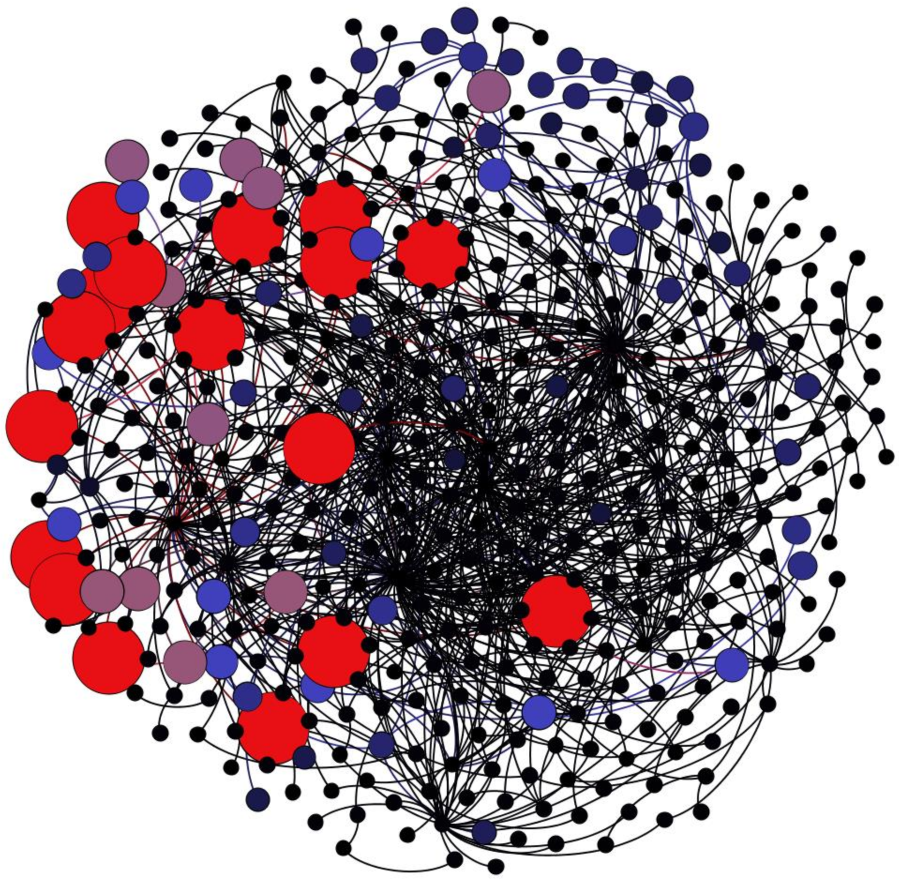
Identifies the information flow and the topological locality of a node in a Graph.



CENTRALITIES (5/6)

Harmonic

Identifies the centrality of a node in a Graph, in terms of distance.



Radiality

Average tendency to node proximity or isolation.

Summarized Importance Value

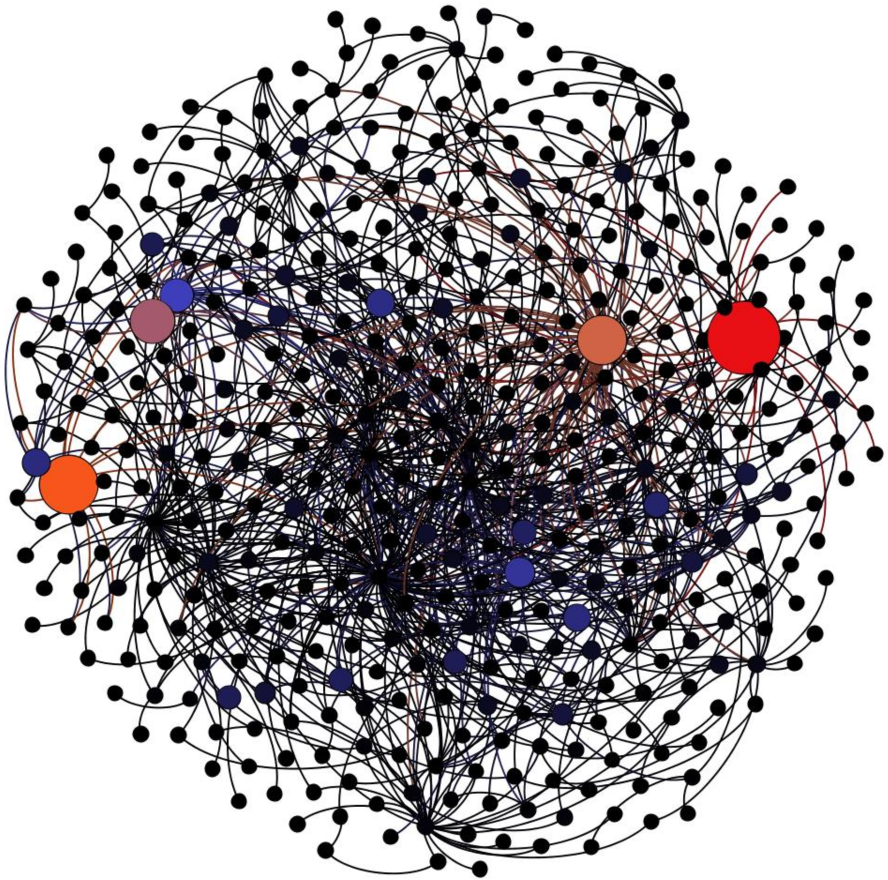
- Normalization of importance Value is defined as:

$$normal\big(IM_i(v)\big) = \frac{IM_i(v) - min\big(IM_i(g)\big)}{max\big(IM_i(g)\big) - min\big(IM_i(g)\big)}$$

- Similarly, we normalize the number of instances ($Inst(v)$) that belong to a schema node.

- The summarized importance value of each node is defined as:

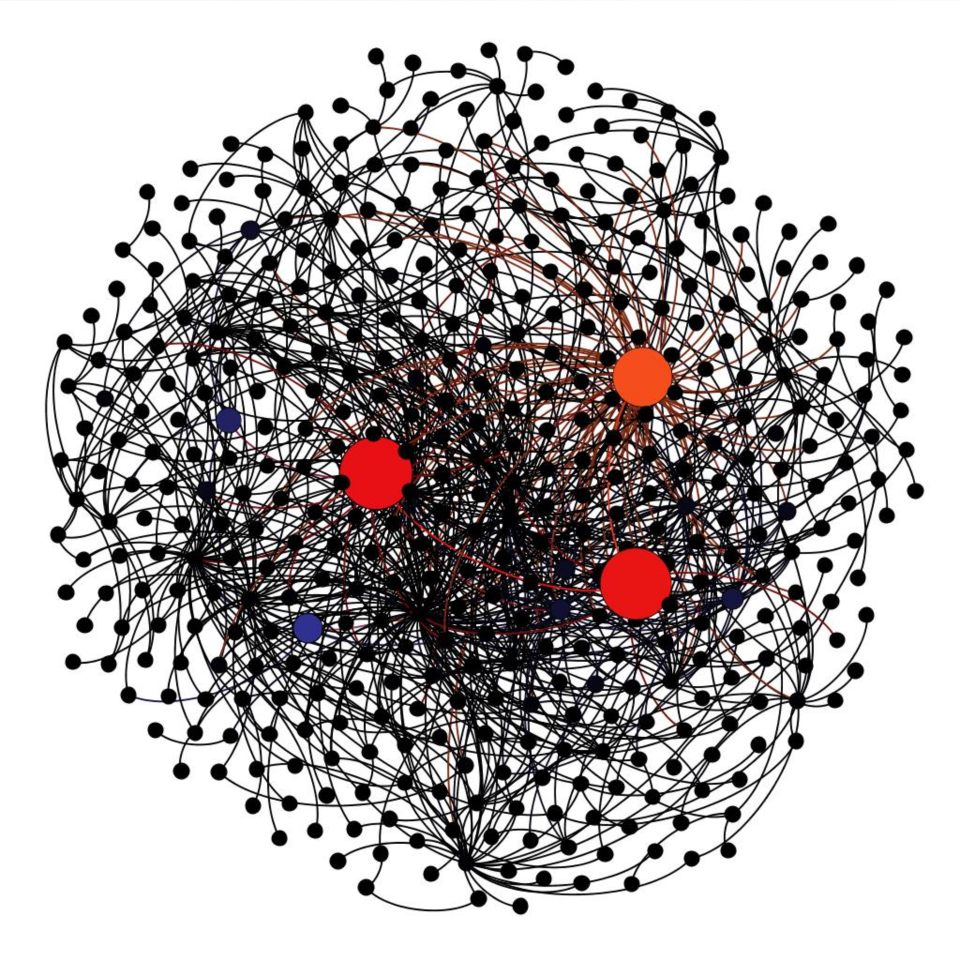
$$SIM_i(v) = normal\big(IM_i(v)\big) + normal\big(Inst(v)\big)$$



Related Work - Relevance

Determined by:

- Connectivity in the schema
- Cardinality of the instances



Hybrid CENTRALITIES (2/2)

Related Work – KCE importance

Determined by:

- Connectivity in the schema
- Cardinality of the instances
- Psycho-linguistic criteria

Construction of the RDF/S Summary Schema Graph

Focus on the paths that link the most important nodes:

The Graph Steiner Tree Problem

The idea

The Steiner Tree Problem

Given an undirected graph $G = (V, E)$, with edge weights $w: E \rightarrow R^{+}$ and a node set of terminals $S \subseteq V$, find a minimum-weight tree $T \in G$ such that $S \subseteq V_t$ and $E_t \subseteq E$.

Algorithms, Approximation & Heuristics

1. SDISTG
2. CHINS
3. HEUM

Worst case bound of 2 $Deviation = \frac{(Z_t - Z_{opt})}{Z_{opt} \times 100}$

Where Z_t and Z_{opt} denotes the objective function values of a feasible solution and an optimal solution respectively.

Complexity of Algorithms

Algorithm	Un-weighted graph
MST	$O(\|V + E\|)$
SDISTG	$O(Q \times \|V + E\|)$
CHINS	$O(Q \times \|V + E\|)$
HEUM	$O(V \times \|V + E\|)$

The competitor - The Maximum-Cost Spanning Tree (MST)

Given an undirected graph $G = (V, E)$, with edge weights $w: E \to R^+$ find a spanning tree $T \in G$ of maximum total edge cost, where $E_t \subseteq E$.

Evaluation - Data Sets

- Dbpedia 3.8: 359 classes, 1323 properties and more that 2.3M instances
- Dbpedia 3.9: 552 classes, 1805 properties and more than 3.3M instances

Ontology	Density	Diameter	Avg path length
DBpedia 3.8	0.00472	9	3.80
DBpedia 3.9	0.00298	13	4.36

Evaluation - Gold Standard

Frequency: To identify the most important nodes of those ontologies we rely on the corresponding SPARQL endpoint query logs, created by users queries

- DBpedia 3.8 more than 50K queries
- DBpedia 3.9 more than 110K queries

Evaluation - Measures

1. Evaluation of Measures for Assessing vertices' Importance
    - Spearman's rank correlation coefficient
    - The Similarity Measure

2. Evaluating Summaries
    - Graph edit distance
    - Additional vertices Introduced

Spearman’s rank correlation coefficient

Statistical dependence between the ranking of:

- Important measures
- Gold Standard

The Similarity Measure

Ontologies can be compared at two different levels:

- The number of classes that exists in gold standard
- For those that does not exists, asses their distance in the taxonomic structure

Graph edit distance

- Measure of similarity between two graphs.

- Count the minimum number of operations required to transform a graph G into (a graph isomorphic to) G'

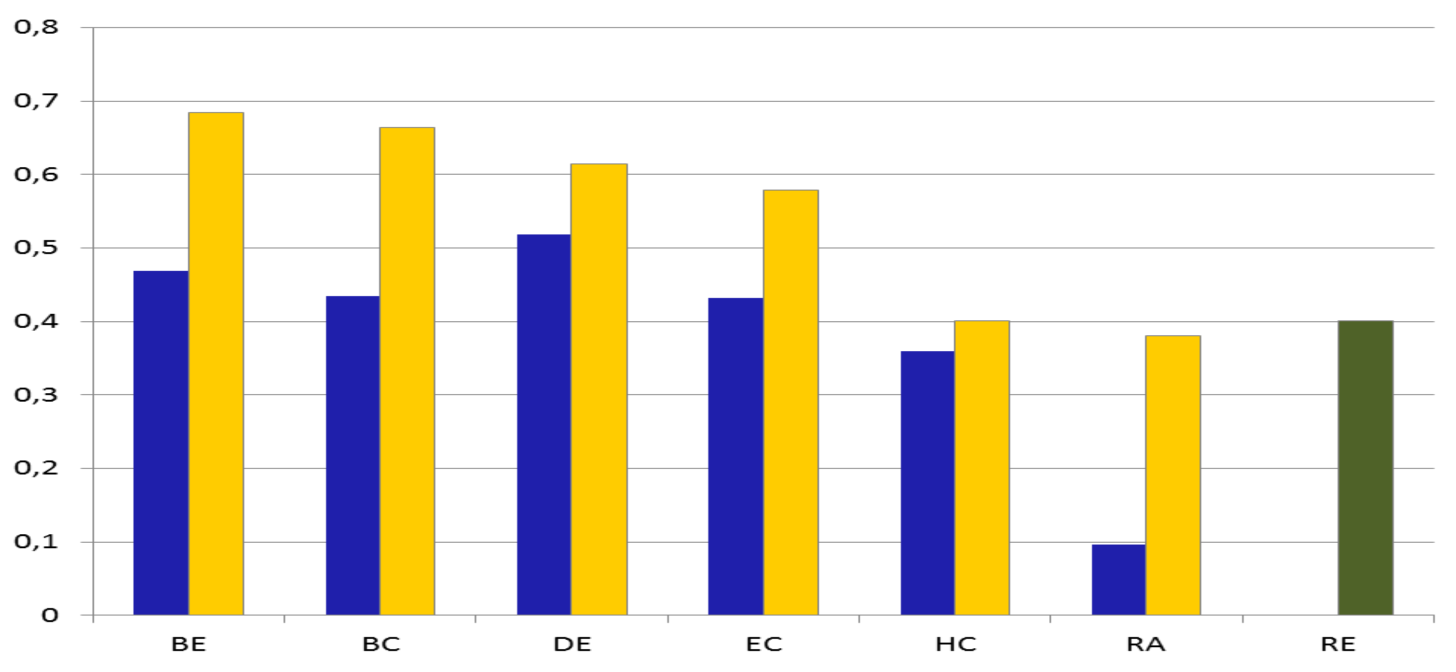
Additional vertices Introduced

Overhead imposed by the algorithms for linking the
most important vertices in terms of the additional vertices that are introduced.

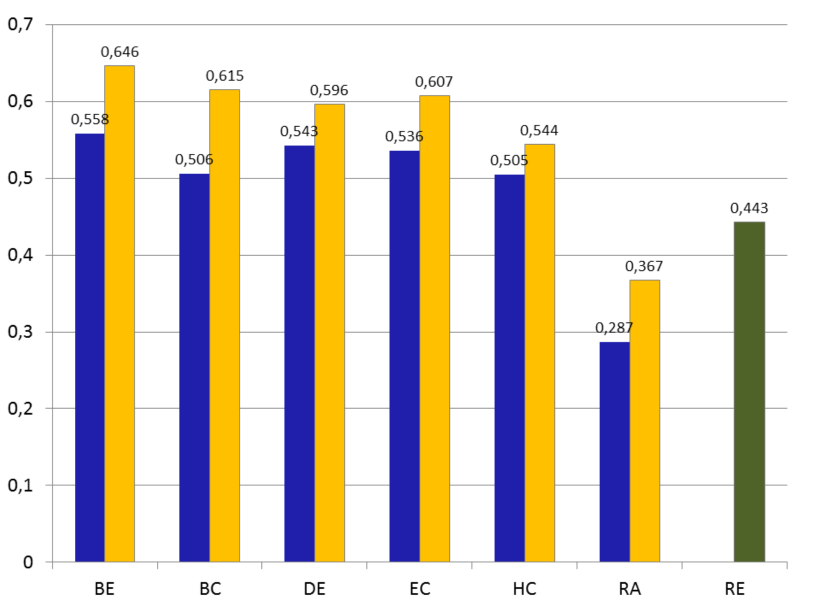
Execution Time

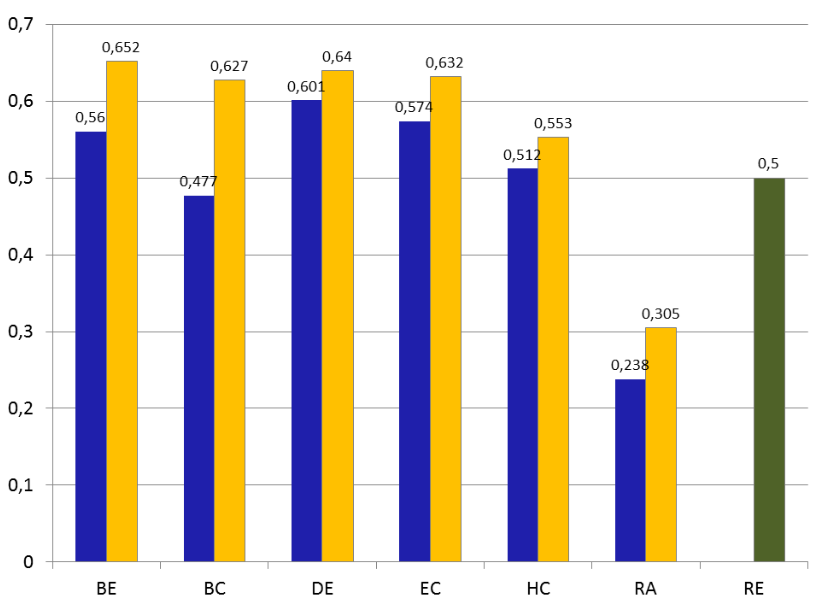
- Performance and scalability with graph sizes
- Average time of 50 executions

Spearman’s rank correlation coefficient



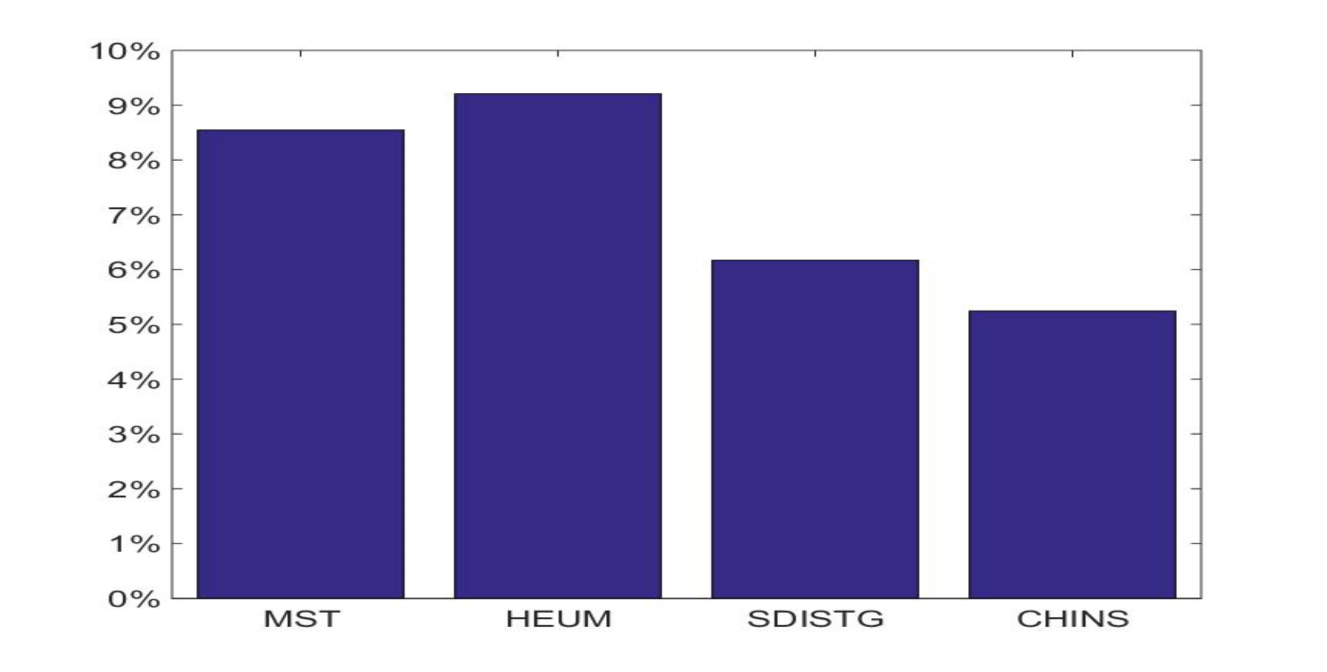
The Similarity Measure



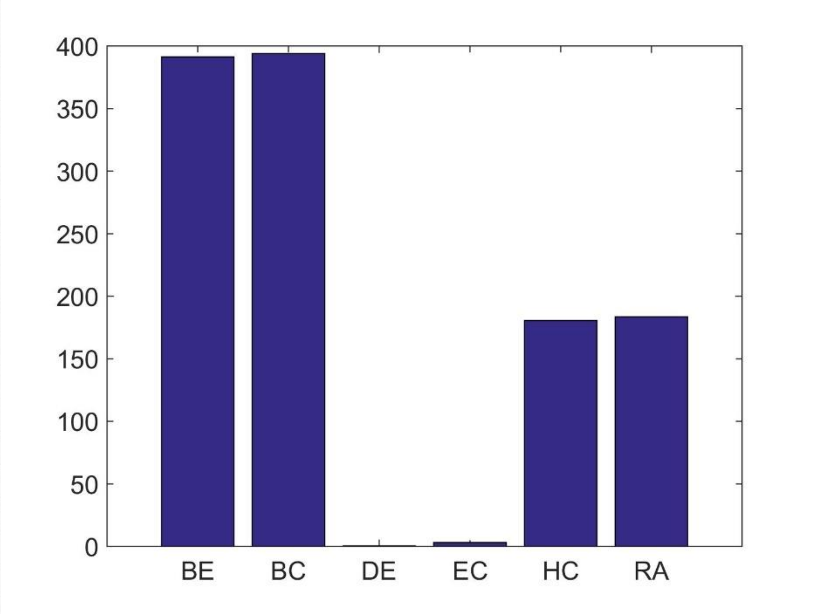


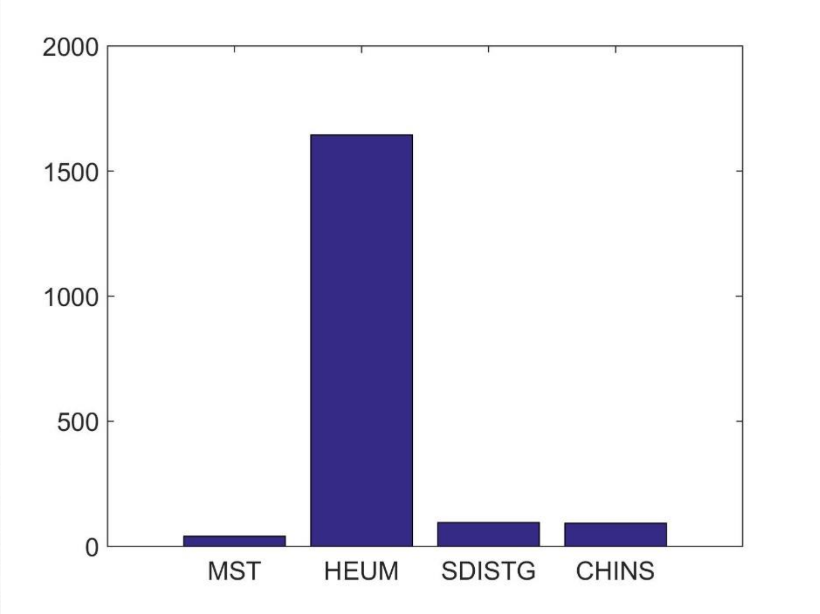
Average similarity DBpedia 3.8 and DBpedia 3.9 for a summary of 1-50%.

Additional vertices Introduced



Average Execution Time in milliseconds





Discussion & Conclusion

- Structural measures have better results in most of the cases.
  - Betweenness and the Ego Centrality

- Steiner-Tree approximation algorithms produce better summaries and introduce less additional nodes to the result schema graph
  - CHINS seems to achieve an optimal trade-off between quality and execution time.

Future Work

1. Combine the various measures in order to achieve the best results according to the specific characteristics of the input ontologies.

2. Extend our approach to handle more constructs from OWL(not only) ontologies

3. Index coverage in the sense of frequent sub-graph problem has to be further examined.

4. Smaller index size and improvement of query performance over graph databases is a growing need.

5. Exploit summaries for Query Answering

THANK YOU for your attention

"Graphs are everywhere
and their possibilities are endless."

-Mark Cryer

Questions?