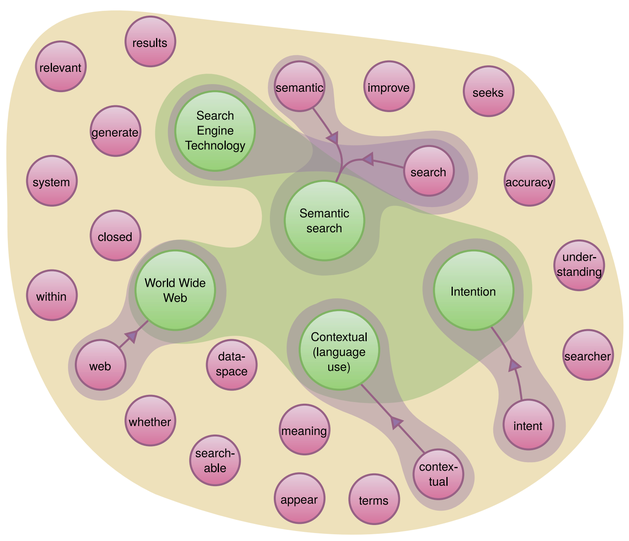
The characterization of a **hypergraph** representing **text and knowledge** shows a sparse structure, with log-normally distributed node-based node degree, and a hyperedge-based node degree following a power law.

Characterizing the Hypergraph-of-Entity Representation Model

José Devezas & Sérgio Nunes

INESC TEC and Faculty of Engineering, University of Porto

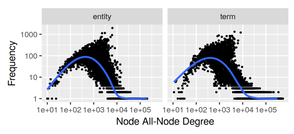
jld@fe.up.pt, ssn@fe.up.pt



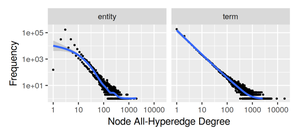
Nodes (607,213)
- *term (323,672)*
- *entity (283,541)*

Hyperedges (253,154)
- *document (7,484)*
- *related_to (7,454)*
- *contained_in (238,216)*

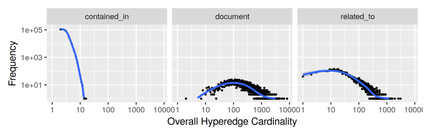
Other statistics
- Avg. degree (0.8338)
- Avg. Cl. Coef. (0.1148)
- Avg. Path Len. (8.3667)
- Diameter (17)
- Density (3.88e-06)



Node-based node degree.



Hyperedge-based node degree.



Overall hyperedge cardinality.

What is the hypergraph-of-entity? 1

- Unified model for entity-oriented search.

- Joint representation model for corpora and knowledge bases.

- Random walk score universal ranking function for:
  * Ad hoc document retrieval
  * Ad hoc entity retrieval
  * Related entity finding

Why characterize the hypergraph-of-entity? 2

- It supports decision making in the design iterations over the retrieval model.

- Statistics like the average path length will help us tune the random walk score length parameters, and the clustering coefficient will help us understand how many repeated random walks to issue.

- Understanding the evolution of the graph, as the number of documents increases, also gives us insights on how to measure the impact of the pruning that we apply to the model (e.g., removing redundancies, or retaining only document keywords).

Computing estimated statistics 3

- We approximated **shortest distances** based on random walks (Ribeiro et al., 2012) launched from multiple sampled source target nodes. We then found path intersections for pairs and merged paths, keeping only the shortest path per pair.

- We approximated **two-node clustering coefficients** (Gallagher and Goldberg, 2013) based on a set of sampled nodes and a large sample of their neighbors.

- We computed a **density** indicator for the hypergraph by analogy to its corresponding bipartite graph. If we consider $n = |V|$ vertices and $m = |E|$ hyperedges, as well as $|E_U^k|$ as the number of undirected hyperedges of cardinality k and $|E_D^{k_1,k_2}|$ as the number of directed hyperedges of tail cardinality k_1 and head cardinality k_2, density D can be calculated as follows:

$$D = \frac{2\sum_k k|E_U^k| + \sum_{k_1,k_2}(k_1+k_2)|E_D^{k_1,k_2}|}{2(n+m)(n+m-1)}$$

Discussion 4

- Few attention has been given to hypergraph characterization.

- The community is still lacking in tools to analyze hypergraphs:

  * **Visualization is a major issue**:
    # The illustrations we use here have been designed by hand mostly using Inkscape. There is no Gephi for this!
    # We used arrows made of lines that all touch at a given point, where the arrowhead is placed.

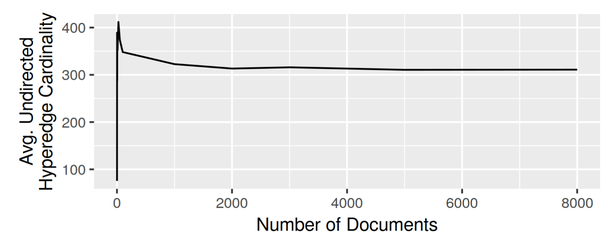
  * **There is no *de facto* library for hypergraph analysis**, similar to what igraph or NetworkX are for graphs.

  * **Few formats support hypergraphs.** GraphML does, but it only supports undirected hyperedges.

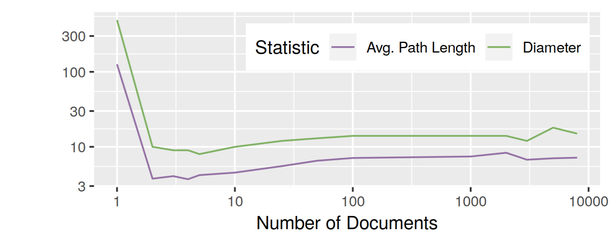
- Polyadism introduces additional complexity and calls for novel metrics that take the information within collective relations into account.

Average node-based node degree over time.

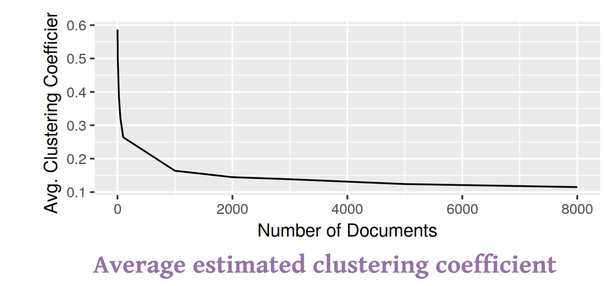
Average hyperedge-based node degree over time.



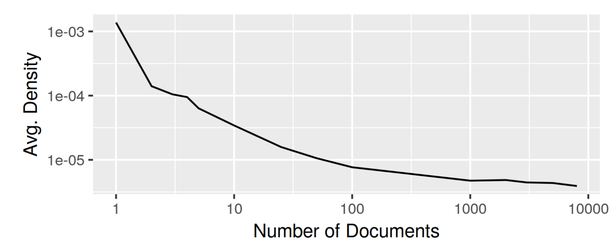
Average hyperedge cardinality over time.



Mean estimated diameter and average path length over time.



Average estimated clustering coefficient over time.



Average density over time (log-scale).

José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program Human Capital (POCH), supported by the European Social Fund and by national funds from MCTES.