VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

European Open Web Index: Large Complex Graphs Visualization

Pavlína Smolková, Kateřina Slaninová (IT4Innovations, VSB - Technical University of Ostrava)

Motivation

Visualization and processing of (extreme) large-scale networks is a challenging task due to unique characteristics such as load imbalance, lack of locality, and access irregularity.

- Supercomputing power with current algorithms for the visualization of large-scale networks.
- Visualization of networks in sizes ranging from hundreds of thousands to millions of nodes

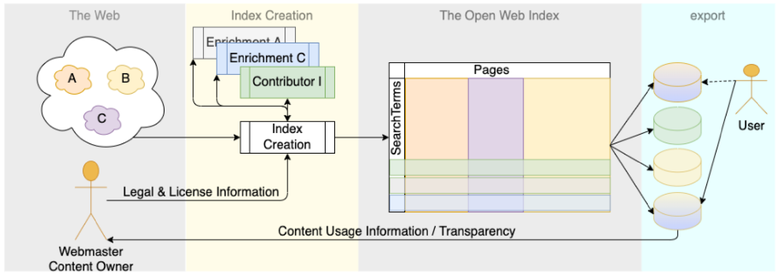
OpenWebSearch.eu project* aims to develop a European infrastructure for independent web search. The project is creating a European Open Web Index (OWI) which contributes to Europe's digital sovereignty and strives to ensure free, open and impartial access to information.

Open Web Index

A web index and **federated infrastructure across existing Europe HPC and data centers** allows query-based access and filtering of web documents at scale for a large variety of web-data driven services.

- First step for moving from the search engine oligopoly to a search engine market.
- Sovereign and open access to the internet and contribution to a fair, open, diverse and free Web.
- Index stores well-structured open web data, making it available for search applications and LLMs.

Figure: Open Web Index Creation



Community Detection Algorithms

A comparative analysis of multiple node embedding algorithms was concluded using a small synthetic graph of 5,000 nodes, both with and without subsequent dimensionality reduction using the t-SNE algorithm.

Table: Comparison of Embedding Methods

Embedding Method	Topic [%]		Louvain Community [%]		Time [s]
	Accuracy	F1	Accuracy	F1	
Deepwalk / + t-SNE	4.72 / 4.84	4.64 / 2.79	28.68 / 16.36	27.50 / 11.29	22 / 23
TENE / + t-SNE	**93.08** / 45.04	**91.43** / 33.84	17.88 / 13.16	7.12 / 12.14	5 / 15
MUSAE / + t-SNE	5.36 / 4.76	5.33 / 2.59	28.52 / 13.16	25.42 / 10.74	100 / 19
ASNE / + t-SNE	68.20 / 5.04	68.15 / 3.18	**36.68** / 16.48	**33.82** / 14.59	61 / 20
FRL	4.88	1.22	18.64	4.49	0.82

* Granitzer, M. et al.: Impact and development of an Open Web Index for open web search (2024) Journal of the Association for Information Science and Technology, 75 (5), pp. 512 - 520. DOI: 10.1002/asi.24818.

Community Detection Algorithms: Layout Comparison

Multiple current visualization tools and algorithms were explored to capture how communities change over time in large-scale, highly fragmented networks derived from OWI dataset (300,000 nodes). We have compared embedding-based approach (TENE) with Cosmograph force-directed layout algorithm.

- TENE (reduced to 3 dimensions) achieved the highest topic classification accuracy 42.00%.
- The most balanced results achieved, when a clustering force based on topic classification added in Cosmos having 32.09% topic accuracy and 30.15% Louvain accuracy.

Table: Community Detection Algorithms: Layout Comparison

Layout	Topic [%]		Louvain Community [%]		Time [s]
	Accuracy	F1	Accuracy	F1	
Random	7.48	3.92	10.46	1.51	–
TENE/t-SNE 2D	29.16	10.76	21.89	8.03	982
TENE/t-SNE 3D	**42.00**	18.75	**23.91**	9.29	1,003
FRL	9.61	4.95	30.20	11.66	28
Cosmos-classic	24.17	7.50	33.18	17.35	11
Cosmos-Louvain	12.49	4.74	34.79	18.27	23
Cosmos-topic	**32.09**	22.11	**30.15**	14.82	24

Figure: Community Detection Algorithms: Layout Comparison of Topic Visualization - Cosmos, weak clustering (left), Cosmos, strong clusterring (middle), TENE + t-SNE + Three.js (right)

Comparison of Layout Algorithms Performance

Open-source implementations of force-directed layout algorithms and t-SNE that use distributed or GPU accelerated approach and focus on visualization of large datasets were tested. Benchmarks were concluded on synthetic sparse graphs with properties similar to those observed in the OWI dataset with size of graph containing 500k, 1M, 2M, 5M, and 50M nodes.

Table: Computation Time of Layout Algorithms Depending on Dataset Size (in seconds), columns represent number of nodes in the graph

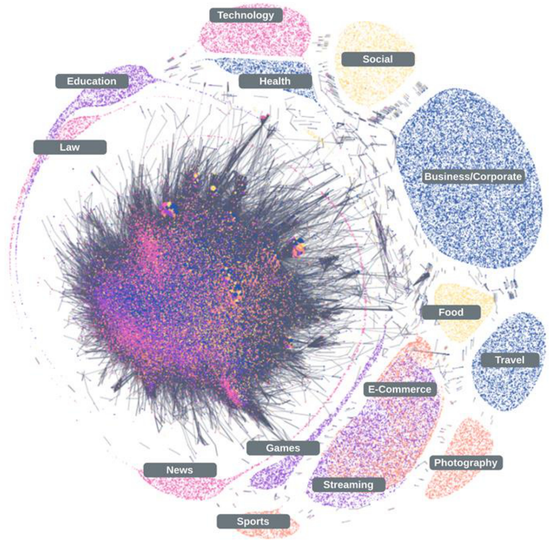
Algorithm	500k	1M	2M	5M	50M
openTSNE (CPU)	597	968	2,073	–	–
cuML/t-SNE (Colab)	40	145	509	2,813	–
cuML/t-SNE (Cluster)	16	39	111	581	–
Cosmograph	34	65	106	284	–
cuGraph (Colab)	15	28	79	260	–
cuGraph (Cluster)	13	25	43	84	533

Visualization of OWI

Open-source implementations of force-directed layout algorithms and t-SNE that use distributed or GPU accelerated approach and focus on visualization of large datasets were tested. Benchmarks were concluded on synthetic sparse graphs with properties similar to those observed in the OWI dataset with size of graph containing 500k, 1M, 2M, 5M, and 50M nodes.

- OWI from 29 April 2025 containing 41M nodes and 230M edges,
- 192,303 components and 197,180 Louvain communities classified based on website topics,
- CuGraph forceAtlas algorithm, which ran by 23 minutes on Karolina GPU node, used as base layout for aggregated visualization with Cosmos gl library.

Figure: Visualization of topics crawled by Open Web Index from 29 April 2025. Graph contains 41M nodes and 230M edges



Conclusion

- We plan further extension of the algorithms by using algorithms for the prediction of the community evolution over time or to leverage other appropriate network properties such as PageRank.
- We also plan to investigate different approaches for the aggregation of vertices by domain or similarity, which would allow for more efficient visualization of even larger graphs.

- Code used for the visualization of Open Web Index is published on GitHub.
- Video with the dynamic visualization of Open Web Index crowling is published on the same repository.

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014).

openwebsearch.eu
ows@openwebsearch.eu
www.linkedin.com/company/openwebsearch-eu
suma-ev.social/@openwebsearcheu

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140).