European Open Web Index: Large Complex Graph Visualization

Pavlína Smolková
pavlina.smolkova@vsb.cz
IT4Innovations, VSB - Technical University of Ostrava
Ostrava, Czech Republic

Kateřina Slaninová
katerina.slaninova@vsb.cz
IT4Innovations, VSB - Technical University of Ostrava
Ostrava, Czech Republic

Abstract

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. Considering the possibilities offered by recent supercomputing power, we have examined current algorithms suitable for the visualization of large-scale networks and were able to visualize networks in sizes ranging from hundreds of thousands to millions of nodes. The experiments were performed on Karolina supercomputer. We have visualized the European Open Web Index produced by the OpenWebSearch.eu project. The complexity of the problem is discussed in the context of performance and computation power needed for the visualization of such (extreme) large-scale graphs.

CCS Concepts

• **Computing methodologies** → *Machine learning*; *Modeling and simulation.*

Keywords

Complex Networks, Graph Visualization, Open Web Index

ACM Reference Format:
Pavlína Smolková and Kateřina Slaninová. 2025. European Open Web Index: Large Complex Graph Visualization. In *Proceedings of The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '25).* ACM, New York, NY, USA, 3 pages.

1 Introduction

Visualization of complex networks allows for intuitive recognition of structures such as communities, central nodes, or important connections between parts of the network.

Recent years, approaches describing networks through vector representations (embeddings) have become popular (e.g. DeepWalk [11])

To represent domain or topic representation of Open Web Index, we have used the approach to represent nodes by thematic information as an additional layer of context above the basic topological relationships (TENE [15], ASNE [5], and MUSAE [12] methods were tested).

There are well known visualization methods based on force-directed layouts like Fruchterman-Reingold algorithm (FRL) [1] or ForceAtlas2 [3]. Other approach could be two-phase methods according to the community structure [6], or methods based on a computed multidimensional embedding.

Dimensionality reduction methods are then applied to map the vector of the graph into 2D or 3D space (e.g. t-SNE [13], PCA [7] or UMAP [8] and specialized in dimensionality reduction for graph structures t-SGNE [4] or LocalMap [14]).

Current tools for visualization of large graphs have their limitations. BigGraphViz is able to visualize over 3 million nodes, 34 million edges [9]. Gephi allows to process typically up to 300,000 nodes, 1,000,000 edges [10].

The OpenWebSearch.eu[1] project aims to develop a European infrastructure for independent web search [2]. The project is creating a European Open Web Index (OWI)[2], which serves as a framework for indexing web content to ensure free, open, and impartial access to information.

2 Methodology

A comparative analysis of 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.

Multiple current visualization tools and algorithms were explored to capture how communities change over time in large-scale, highly fragmented networks derived from OWI data.

For OWI dataset (300,000 nodes), TENE was selected as the representative embedding-based approach and compared with the Cosmograph force-directed layout algorithm.

The quality of the layout was assessed by measuring how well nodes can be classified based solely on their positions (Liblinear library). A 9:1 train-test split was applied and performance was evaluated using accuracy and weighted F1 score on the top 20 Louvain communities and 16 predefined topics. Under a uniform random guessing strategy across 20 classes, the baseline accuracy is estimated at 5%.

To evaluate scalability, we tested several 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. Benchmarks were concluded on synthetic sparse graphs with properties similar to those observed in the OWI dataset.

SC '25, St. Louis, MO, USA

[1] OpenWebSearch.eu: https://openwebsearch.eu
[2] OWI: https://openwebindex.eu

3 Results

3.1 Comparative Analysis: Community Detection Algorithms

Regarding the comparative analysis of multiple node embedding algorithms using the t-SNE algorithm, ASNE produced the most balanced performance, achieving 68.19% accuracy in topic classification and 36.68% Louvain accuracy. However, it required substantial computation time (61 seconds) and its effectiveness significantly declined after dimensionality reduction. Other algorithms such as DeepWalk and MUSAE yielded relatively good accuracy on Louvain-based classification but did not perform well in topic classification. TENE performed best in topic classification (93.08% accuracy). Dimensionality reduction lowered this accuracy to 45.04%. TENE was also one of the fastest tested methods, completing in just 5 seconds.

For OWI dataset, TENE was selected as the representative embedding-based approach and compared with the Cosmograph force-directed layout algorithm. TENE, when reduced to three dimensions, achieved the highest topic classification accuracy (42.00%). For Louvain-based evaluation, the best performance was observed using the base version of Cosmograph algorithm without specific clustering force (referred here as Cosmos-classic), achieving with 33.18%. The most balanced results were then achieved, when a clustering force based on Topic classification was added (Cosmos-Topic), having 32.09% topic accuracy and 30.15% Louvain accuracy.

3.2 Scalability

For parallelized CPU-based t-SNE, the openTSNE library provided the best results, achieving linear complexity, improving upon O(N log N) complexity of standard scikit-learn implementation, and proved efficient for graphs up to 2 million nodes.

GPU-accelerated t-SNE using the cuML library significantly improved the performance. On the Karolina supercomputer, the same algorithm processed 2 million nodes in just 111 seconds and remained under 10 minutes even for 5 million nodes, though embedding size have risen to 8 GB at that point.

The layout capabilities of the Cosmograph library were tested using its Python API for easier integration. It successfully handled graphs with up to 5 million nodes in approximately 5 minutes, including real-time rendering. However at this scale, the frame rate dropped significantly and RAM usage spiked to 12 GB. Cosmograph relies on WebGL and performs layout computations directly in shaders, which limits performance to the capabilities of the local client GPU.

In contrast, cuGraph uses CUDA-based GPU acceleration and can fully leverage high-performance GPUs in cluster environments. It outperformed all other tested solutions in both speed and scalability, processing graphs with up to 50 million nodes in under 9 minutes while maintaining memory usage below 10 GB.

4 Conclusion

Based on our findings, we have been able to visualize the OWI from 29 April 2025 containing 41M nodes and 230M edges. We have identified 192,303 components and 197,180 Louvain communities based on website topics. The graph layout was counted by CuGraph forceAtlas algorithm by 23 minutes on Karolina GPU node (IT4Innovations) used as base layout for aggregated visualization with Cosmos gl library.

Table 1: 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

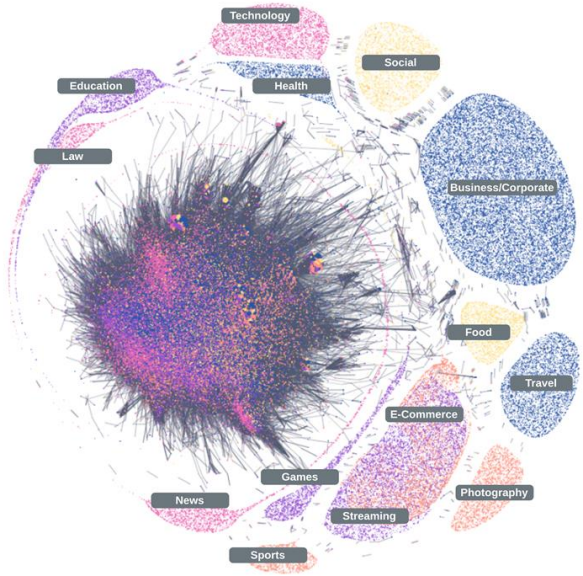

Figure 1: Visualization of topics crawled by Open Web Index from 29 April 2025. Graph contains 41M nodes and 230M edges.

Acknowledgments

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).

References

[1] Thomas M. J. Fruchterman and Edward M. Reingold. 1991. Graph drawing by force-directed placement. *Softw. Pract. Exper.* 21, 11 (Nov. 1991), 1129–1164. doi:10.1002/spe.4380211102

[2] Michael et al. Granitzer. 2024. Impact and development of an Open Web Index for open web search. *Journal of the Association for Information Science and Technology* 75, 5 (2024), 512 – 520. doi:10.1002/asi.24818

[3] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network

Visualization Designed for the Gephi Software. *PloS one* 9 (06 2014), e98679. doi:10.1371/journal.pone.0098679

[4] X. Li, Y. Yiao, and Y. Zhou. 2023. Efficiently Visualizing Large Graphs. Online. https://arxiv.org/pdf/2310.11186 arXiv:2310.11186v1 [cs.LG].

[5] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2018. Attributed Social Network Embedding. *IEEE Transactions on Knowledge and Data Engineering* 30, 12 (Dec. 2018), 2257–2270. doi:10.1109/tkde.2018.2819980

[6] Kwan-Liu Ma and Chris W. Muelder. 2013. Large-Scale Graph Visualization and Analytics. *Computer* 46, 7 (2013), 39–46. doi:10.1109/MC.2013.242

[7] Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (PCA). *Computers & Geosciences* 19, 3 (1993), 303–342. doi:10.1016/0098-3004(93)90090-R

[8] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML] https://arxiv.org/abs/1802.03426

[9] Ehsan Moradi and Debajyoti Mondal. 2021. BigGraphVis: Leveraging Streaming Algorithms and GPU Acceleration for Visualizing Big Graphs. arXiv:2108.00529 [cs.DC] https://arxiv.org/abs/2108.00529

[10] G.A. Pavlopoulos, D. Paez-Espino, N.C. Kyrpides, and I. Iliopoulos. 2017. Empirical Comparison of Visualization Tools for Larger-Scale Network Analysis. *Adv Bioinformatics* 54, 2, Article 1278932 (2017), 50 pages. doi:10.1155/2017/1278932

[11] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*. ACM, 701–710. doi:10.1145/2623330.2623732

[12] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. 2021. Multi-scale Attributed Node Embedding. arXiv:1909.13021 [cs.LG] https://arxiv.org/abs/1909.13021

[13] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. Online, 2579–2605 pages. http://jmlr.org/papers/v9/vandermaaten08a.html

[14] Yingfan Wang, Yiyang Sun, Haiyang Huang, and Cynthia Rudin. 2024. Dimension Reduction with Locally Adjusted Graphs. arXiv:2412.15426 [cs.LG] https://arxiv.org/abs/2412.15426

[15] Shuang Yang and Bo Yang. 2018. Enhanced network embedding with text information. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 326–331.