Dataset Announcement: MS-BioGraphs, Trillion-Scale Public Real-World Sequence Similarity Graphs

Mohsen Koohi Esfahani[1,3], Paolo Boldi[2], Hans Vandierendonck[1], Peter Kilpatrick[1], and Sebastiano Vigna[2]

[1]*Queen's University Belfast, United Kingdom*
[2]*Università degli Studi di Milano, Italy*
[3]*University of Sistan, Iran*

https://blogs.qub.ac.uk/DIPSA/MS-BioGraphs

Abstract

Progress in High-Performance Computing in general, and High-Performance Graph Processing in particular, is highly dependent on the availability of publicly-accessible, relevant, and realistic data sets.

*In this paper, we announce publication of **MS-BioGraphs**, a new family of publicly-available real-world edge-weighted graph datasets with up to 2.5 trillion edges, that is, 6.6 times greater than the largest graph published recently.*

We briefly review the two main challenges we faced in generating large graph datasets and our solutions, that are, (i) optimizing data structures and algorithms for this multi-step process and (ii) WebGraph parallel compression technique. We also study some characteristics of MS-BioGraphs.

The datasets and the complete version of this paper are available on https://blogs.qub.ac.uk/DIPSA/MS-BioGraphs.

***Index Terms*—Graph Datasets, High-Performance Graph Processing, High-Performance Computing, Biological Networks, Sequence Similarity Graph**

1. HPC Challenges and Our Solutions

Processing Model. In contrast to the MPI model, we search for a distributed processing model that *(i) dynamically adjusts the degree of parallelism (i.e., the number of machines/processors involved in the processing) and (ii) does not restrict the size of processed data to the total memory of the cluster* while machines have access to a shared storage that hosts the datasets and the intermediary data.

We deploy a distributed model in which algorithms are designed as a number of sequential steps with parallel workloads per step. In each step, machines contribute to the total processing independently from each other and the input and output data for each processing slot is loaded from and stored to the shared storage. So, machines only communicate (a) to the shared storage to retrieve/store data and (b) to the scheduler to receive a partition of a task or to inform completion of a partition.

In this way, each machine requires a memory size that is enough to complete processing a partition. This facilitates processing the datasets whose sizes are greater than the available memory. Moreover, as the machines do not communicate with each other, each step can be started as soon as at least one machine becomes available and new machines can join/leave a running step. This (i) relaxes the assumption of permanent availability of a fixed number of resources during the whole execution time, (ii) minimizes the waiting time, and (iii) optimizes cluster utilization.

Parallelizing Graph Compression. As MS-BioGraphs have binary sizes of up to 20 TeraBytes, it is necessary to compress them to make their storage, transfer over the network, and processing more efficient.

To that end, we used the WebGraph framework[1] [2], an open-source graph compression framework that has been continuously maintained and updated during the last 20 years. This framework provides a highly-efficient graph compression and includes a rich set of graph operations and analytics. Moreover, the users of languages and frameworks with WebGraph support, such as Hadoop, C++, Python, and Matlab, benefit from direct access to MS-BioGraphs.

We extended the WebGraph framework in two directions: in the first phase, we extended labelled graphs to support parallel compression of the underlying graph. In the second phase, we partially violated the decoupled design of labelled graphs in WebGraph, adding to the compression phase of the main storage format class of WebGraph, `BVGraph` (that compresses and stores the underlying graph), an option to store the labels at the same time.

2. Generating MS-BioGraphs

Inspired by HipMCL [1], we use the Metaclust dataset[2] [7] that contains 1.7 billion protein sequences.

We collected all similarities produced by the LAST sequence alignment algorithm[3] [3]. We selected LAST as aligner as it shows better single-machine performance and has been widely used and maintained since its publication in 2011. Sequence matching by LAST is performed in two steps: (i) creating a database from sequences using a program called `lastdb` and (ii) aligning the sequences of a file against the created database using `lastal` that outputs the matched sequences and their scores.

To create MS-BioGraphs, we compute all-against-all matching of the sequences. Since sequence similarity is

1. https://webgraph.di.unimi.it/
2. https://metaclust.mmseqs.com/2018_06/metaclust_all.gz
3. https://gitlab.com/mcfrith/last

2835-2238/23/©2023 IEEE
https://doi.org/10.1109/IISWC59245.2023.00029

TABLE 1: MS-BioGraphs Statistics

Name	Directed	|V| (M)	|E| (B)	Filtering Intention	Max. Deg.		Weight		Zero Deg.		Avg. Deg.		Weak. Con. Comp.		Size (GB)	
					In(K)	Out(K)	Min.	Max.	In(M)	Out(M)	In	Out	Count(M)	Max. Size(%)	Base	Labels
MS	No	1,757.3	2,488.0	-	814.9		98	634,925	6.4		1,415.8		148.9	99.95	6,843.6	4,696.0
MS200	No	1,414.4	502.9	0.200|E|, W	745.7		460	634,925	0.0		355.6		338.3	96.61	1,362.7	1,119.6
MS50	No	585.6	124.7	0.050|E|, W	507.8		900	634,925	0.0		213.1		155.3	81.95	327.1	303.1
MS1	No	43.1	2.6	0.001|E|, W	14.2		3,680	634,925	0.0		61.7		15.7	4.66	6.1	7.7
MSA500	Yes	1,757.3	1,244.9	$ID_{neigh} \leq ID_v$	229.4	814.4	98	634,925	6.4	16.8	711.0	715.3	148.9	99.94	3,502.2	2,351.8
MSA200	Yes	1,757.3	500.4	0.200|E|, VRW	658.8	709.1	98	634,925	6.4	7.4	285.8	286.0	221.5	99.29	1,455.2	1,033.7
MSA50	Yes	1,757.3	125.3	0.050|E|, VRW	543.1	297.9	98	634,925	6.4	8.5	71.6	71.7	363.1	94.15	385.2	268.3
MSA10	Yes	1,757.3	25.2	0.010|E|, VRW	207.2	62.0	98	634,925	6.4	9.9	14.4	14.4	628.5	61.72	84.0	57.3

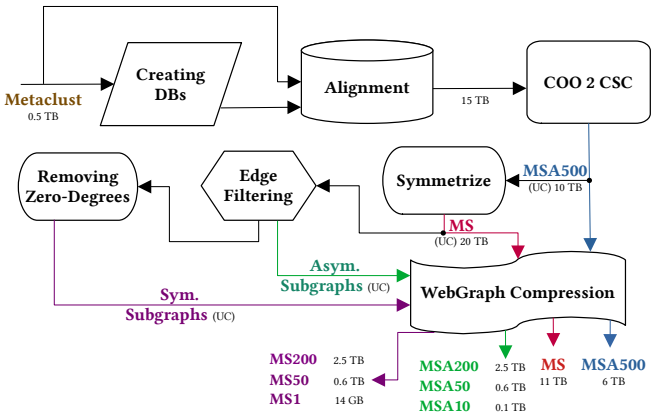


Figure 1: Creation Steps (UC: uncompressed)

a symmetric relation, instead of matching each pair of sequences twice, we match each sequence only to sequences with lower IDs. This produces a directed weighted graph whose symmetric version represents all the matches and their scores.

We have the following steps as depicted also in Figure 1. First, we need to create LAST databases using `lastdb` and then call `lastal` to create the similarities, i.e., the asymmetric graph in the coordinate format (COO).

The next step is converting the COO graph to the Compressed Sparse Columns (CSC) [6] format which is followed by symmetrizing and compression. We also create some subgraphs to support research studies with different graph size and direction requirements. Therefore an "Edge Filtering" step is required to create subgraphs and we need to remove zero-degree vertices.

3. Characteristics of MS-BioGraphs

Naming. The name of each graph is started by two characters *M* and *S* as initials of Metaclust (as the source dataset) and Sequence similarity (as the real-world domain of the graph), respectively. The name of the directed subgraphs has a third character *A* that indicates the graph is asymmetric. The name of subgraphs is followed by up to *3 digits* that show the relative-size of the subgraph in comparison to the MS graph, multiplied by a thousand.

Column 5 of Table 1 summarizes the naming scheme. For the undirected subgraphs MS200, MS50, and MS1 the weight of edges (shown as W in the table) has been considered as the filtering metric. For the directed subgraphs MSA200, MSA50, and MSA10 the vertex-relative weight (shown as VRW in the table) has been used as sampling metric.

Statistics. Table 1 shows the general statistics of the MS-BioGraphs.

Degree Distribution. Figure 2 shows the degree distribution of the MS graph. The Frequency degree distribution plot shows that the *MS graph has a skewed degree distribution.* The Fibonacci Binned plot [8] shows that the degree distribution does not follow a particular known degree distribution, especially given that two changes of concavity are observed.

We identify that the MS graph has a steep slope on the Cumulative Edges plot that indicates more than 98% of edges are incident to the vertices with degrees 100 to 50K. As such, the low-degree vertices (degrees $\leq$ 100) and very high-degree vertices (degrees $\geq$ 50K) hardly contribute to the total edges in MS.

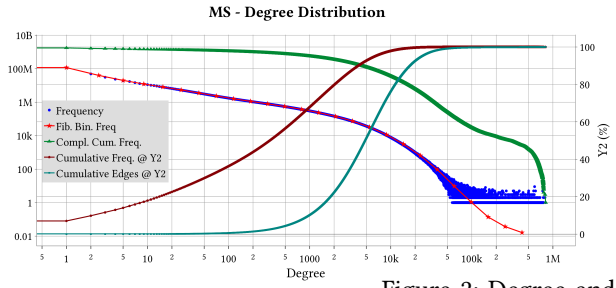
To identify the connection between vertices, we use the *degree decomposition* plot [5] in Figure 2. This shows the low-degree vertices (vertices with degree 1–100) of the MS graph do not contribute to the higher vertex classes. This is in contrast to social networks and web graphs whose low-degree vertices are the main constituents of all vertex classes [5]. Moreover, in MS graph high-degree vertices are tightly connected to each other. The similar trend has been observed in social networks [5].

This *tight connection between high-degree vertices and its coincidence with their high cumulative frequency introduces a new structure of real-world skewed graphs* with obvious differences to the previously studied ones such as web graphs and social networks [5].

Weight Distribution. Figure 2 shows the weight distribution of the MS graph and their Cumulative Frequency plots. The plots indicate that *weights do not have a random distribution and follow a skewed distribution with a tail close to power-law distribution.*

Weakly-Connected Components. Figures 3 and 4 show the component size distribution for symmetric and asymmetric MS-BioGraphs, respectively. The plots indicate *a power-law size distribution and a very-high degree of connectivity in MS and also large subgraphs.*

Push vs. Pull Locality. The Push vs. Pull Locality metric [5] considers the cumulative effectiveness of the in-hubs in comparison to the out-hubs in an asymmetric graph. Figure 6 illustrates it for the MSA200 and shows that the push locality curve is very close to the pull locality



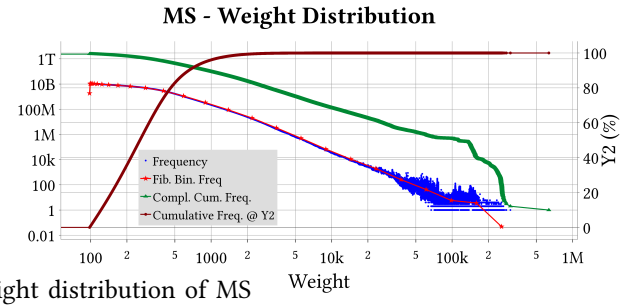


Figure 2: Degree and weight distribution of MS

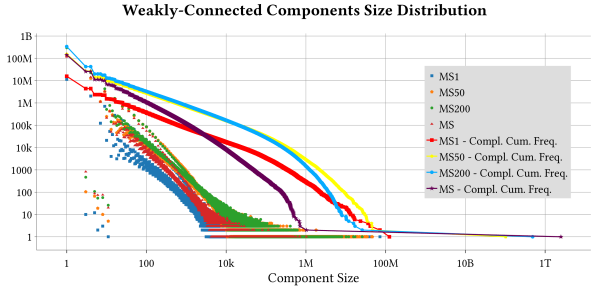


Figure 3: WCC of symmetric graphs

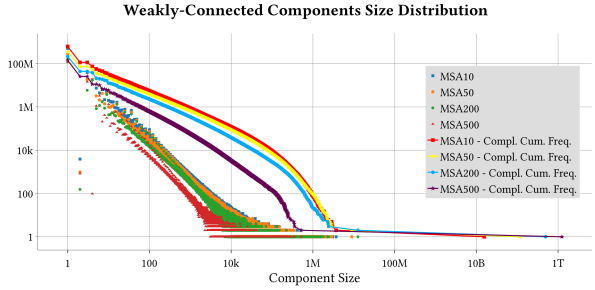


Figure 4: WCC of asymmetric graphs

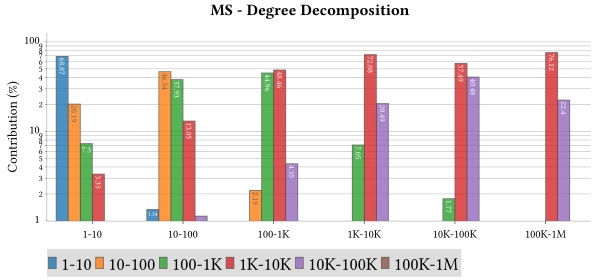


Figure 5: MS degree decomposition

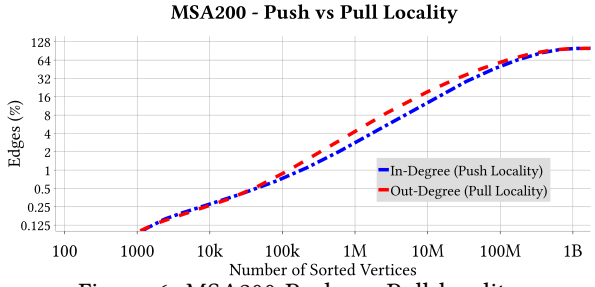


Figure 6: MSA200 Push vs. Pull locality

curve. *MS-BioGraphs, in contrast to social networks and web graphs, demonstrate the same Push and Pull Locality.*

4. Conclusion

To provide a more effective HPGP research environment by accessing realistic and updated datasets with a better coverage of various application-domains, we announce the **MS-BioGraphs**, a family of sequence similarity graphs with up to 2.5 trillion edges which is 6.6 times greater than the previous largest real-world graph.

Our study of MS-BioGraphs' characteristics shows a skewed degree distribution and a particular graph structure that makes their structure very different from web graphs and social networks.

The full version of this paper [4] presents (i) a comprehensive discussion on the necessity and importance of the updated real-world public datasets, (ii) a detailed explanation of the process-wide engineering and design of the required data structures and algorithms for generation steps of MS-BioGraphs, and (iii) a comparative study of MS-BioGraphs characteristics to other real-world graphs.

Acknowledgements

This work was partially supported by (i) the High Performance Computing center of Queen's University Belfast and the Kelvin-2 supercomputer (UKRI EPSRC grant EP/T022175/1) and (ii) the SERICS project (PE00000014) under the NRRP MUR program funded by the EU - NGEU. First author was also supported by a scholarship from the Department for the Economy, Northern Ireland and Queen's University Belfast.

References

[1] A. Azad, G. Pavlopoulos, C. Ouzounis, N. Kyrpides, and A. Buluc, "HipMCL: a high-performance parallel impl. of the markov clustering algorithm for large-scale networks," *Nucleic Acids Research*, 2018.

[2] P. Boldi and S. Vigna, "The webgraph framework I: Compression techniques," in *WWW'04*. ACM, 2004.

[3] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith, "Adaptive seeds tame genomic sequence comparison," *Genome research*, vol. 21, no. 3, 2011.

[4] M. Koohi Esfahani, P. Boldi, H. Vandierendonck, P. Kilpatrick, and S. Vigna, "MS-BioGraphs: Sequence similarity graph datasets," *CoRR*, vol. abs/2308.16744, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2308.16744

[5] M. Koohi Esfahani, P. Kilpatrick, and H. Vandierendonck, "Locality analysis of graph reordering algorithms," in *IISWC'21*. IEEE, 2021.

[6] Y. Saad, "Sparskit: a basic tool kit for sparse matrix computations," 1994.

[7] M. Steinegger and J. Söding, "Clustering huge protein sequence sets in linear time," *Nature Communications*, vol. 9, 06 2018.

[8] S. Vigna, "Fibonacci binning," *CoRR*, vol. abs/1312.3749, 2013.