Bioinformatics Summer School Report

Name: Hassan Ali Alsayhah
Internship Organization: Data Science Institute, UofT
Supervisor's Name: Charlie Boone, Duncan Forster
Project title: GANLA: GPU Accelerated Network Layout Algorithm

1. Introduction

Network layout algorithms are used to visualize networks in a human friendly manner for further data exploration. Currently, the Kamada-Kawai algorithm is commonly used. However, it does not scale to large networks, such as those derived from human functional genomics data.

The objective is to develop a fast and scalable version of Kamada-Kawai taking advantage of graphics processing unit (GPU) hardware acceleration, called GPU Accelerated Network layout Algorithm (GANLA).

2. Materials and Methodology

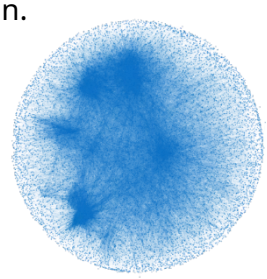
1. Implementing a version of Kamada-Kawai algorithm that computes the layout positions for all nodes in parallel.
2. Modifying the algorithm to use mini-batching technique by separating the nodes to random batches.
3. Applying GANLA on different networks from real datasets.
4. Analyzing the time and memory performance of GANLA using varied structured graphs.

Materials:

Python is mainly used for programming. The algorithm was implemented using common libraries like pytorch, networkx, parser, matlab, etc.

Cytoscape application was used for visualization and comparison.

3. Results and Discussion



After applying it to a real dataset. The next figure represents

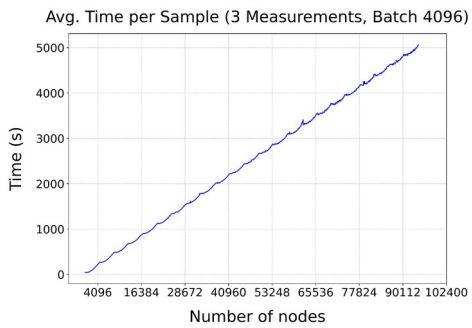
Huttlin-2021 dataset[1], which was laid out using GANLA.

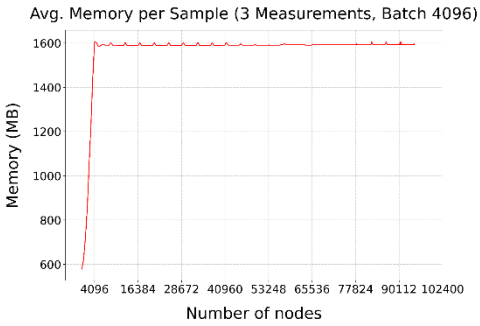
It consists of ~10,000 human proteins and ~70,000 interactions between them, with proteins represented as nodes and interactions as edges. The network provides insights into how human proteins relate to and cluster under various conditions.

After analyzing the behavior of GANLA:

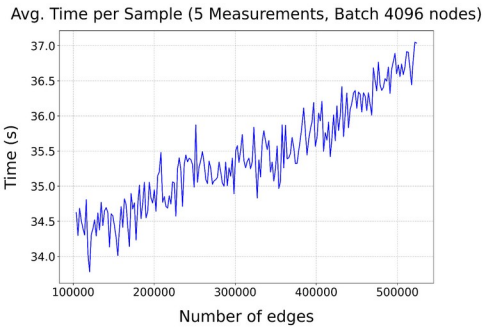
Note: The memory shown is each sample's average maximum memory.

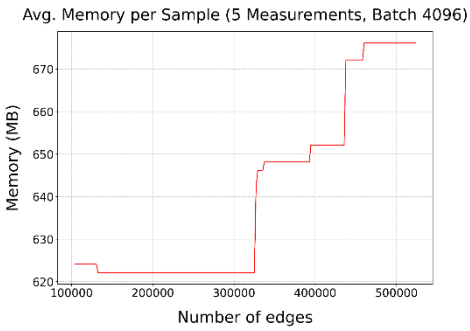
200 distinct samples each were split up into 4,096 nodes and has 130,816 edges.



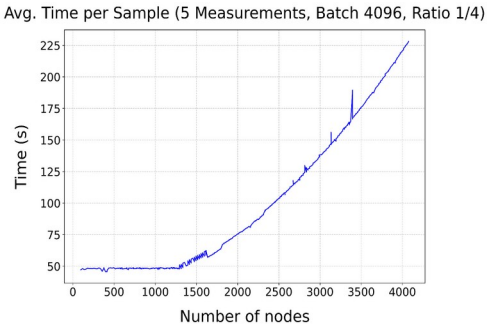


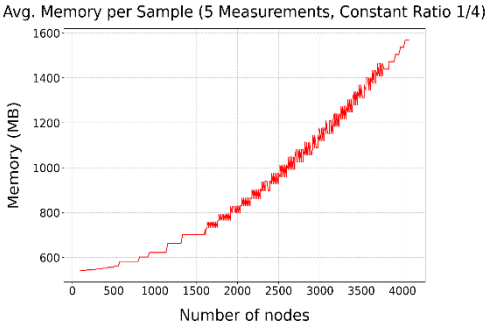
206 distinct samples each have 1,024 nodes with a batch size of 4,096 nodes.





200 samples, each with a 1:4 node-to-edge ratio and a batch size of 4,096 nodes. Plots with edge counts has shown similar patterns.





When representing the above data based on number of edges, they show the same behavior, especially that they have a constant ratio.

GANLA is promising, but it should be compared to the original algorithm.

4. Skills and Knowledge Gained

In conclusion, Working in UofT in the research and development area has exposed me more about the academic filed and the challenges faced during researching and filling a research gap. Moreover, the experience of my training to be in Canada taught me the significant of academia. Furthermore, I learned about the field of lay outing networks and the importance of them with a deeper understanding of Kamada-Kawai and GANLA. Also, I had an eye-opening on various python libraries, such as networkx, PyG, Parser, and more. At the end, I gained the skills of creating a scientific abstract and poster. In summary, I improved GANLA by adding mini-batch technique, created all needed data, finished the performance analysis, and handled real dataset to test GANLA.

5. Challenges and Solutions

Computational challenges: Analyzing the behavior of GANLA was too slow, so I have divided the structured data on many jobs to finish faster.

Memory Limitation: A graph was too large to fit in RAM, so we reduced many edges because they were too many for the dataset.

6. Conclusion

In conclusion GANLA can scale to larger networks than the CPU version of Kamada-Kawai algorithm and is promising to be faster, enabling the exploitation of larger and more complex biological networks relevant to human functional genomics. In the future, GANLA will be Benchmarked with the original version and applied on a large network of ~1M nodes.

7. Acknowledgments

- The GANLA: GPU Accelerated Network Layout Algorithm is supported by the Data Sciences Institute, University of Toronto.
- KASUT Academy is financially supporting Hassan Alsayhah.
- Compute Canada provided the compute power.

8. References (if applicable)

1. Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Guha Thakurta, S., Zhang, T., Rad, R., Pan, J., Nusinow, D. P., Paulo, J. A., Schweppe, D. K., Pontano Vaites, L., Harper, J. W., & Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell, 184(11), 3103-3121.