What Would a Graph Look Like in This Layout? A Machine Learning Approach to Large Graph Visualization

Ziqi Zhang, Manyao Peng

November 7, 2018

https://github.com/allhailjustice/vis_class_project

1 Background and Motivation

Our motivations for choosing this project are following:

1. Graph is often used to abstract data. In many research fields, the problem of data analysis can be transformed into graph analysis that focuses on the topological structure of graph. One of the important parts of this project is to measure the topological similarity between graphs, which is exact what we are interested in.

2. This project combines visualization and machine learning. It works on predicting visualiztion performance on specific layouts for graph, using machine learning strategy. By doing this project, we may get further inspriation on how to apply machine learning approaches to visualization.

3. So far, the ways we have learnt in class to evaluate visualization layouts are more like intuitive senses rather than numerical metrics. However, the latter one can lead to automatic evaluation process withou human inspection, which is very convenient for the cases have large amount of graphs. This project uses several aesthetic criteria and metrics to evaluate visualization performance which can be a supplement for the course.

2 Objectives

As a result of the project, given an input graph and a layout method, we expect to achieve

1. Predict what its layout look like without calculating the actual layout.

2. Estimate the aesthetic metrics without calculating actual layout.

3 Data

We will use the data in The SuiteSparse Matrix Collection (formerly the University of Florida Sparse Matrix Collection) [2], which is available at https:

//sparse.tamu.edu/. We will download the symmetric matrices with $10^2 \leq n.rows, n.cols \leq 10^3$ (considering our computational resource, matrices that are too large are abandoned) using the provided MATLAB Interface (ssget). Beside real-world data, we may also generate some synthetic data in this project in case that the amount of real-world data is not sufficient for training machine learning models.

4 Data Processing

The non-binary matrices will be binarized with threshold 0; the asymmetric matrices will be augmented as $\begin{bmatrix} A & 0 \\ 0 & A^T \end{bmatrix}$.

5 Must-Have Features

- Computational efficiency: Considering that we will use data driven machine learning approach, and our data are large scale sparse matrices, we must do the computation efficiently since we have limited time and resource for this project.
- Accuracy: For the aesthetic metrics, we want to accurately estimate the layout's aesthetic metrics without computing the layout so that difference between measured values (ground truth) and estimated values is small enough. Therefore, we must select reasonable features and machine learning approaches.

6 Project Schedule

Week of Nov. 5 :

1. Collect real-world data and pre-process it.
2. Generate synthetic data.

Week of Nov.12:

1. Implement program based on Random Walking algorithm to sample graphlet frequencies and then scale the frequency vector

Week of Nov. 19:

1. Implement program based on machine learning method (KNN) to estimate what a graph would look like in specific layouts.

Week of Nov. 26:

1. Implement program based on machine learning method (SVR etc.) to estimate Aesthetic Metrics for given graph under specific layouts.

Week of Dec. 3:

1. Results collection and analyze.
2. Evaluation

7 Techniques

- Machine learning approach: In the chosen paper, the authors use Support Vector Regression as framework, similarity matrix derived from pair-wise inner product between the topological feature of instances as input feature. Beside reproduce experiments in the paper, we may also try other frameworks like neural networks, and other input features like directly use topological feature of each instance as the input.
- Data encoding: We will encode data by its topological features which are frequencies of graphlets (selected small, induced, and non-isomorphic subgraph patterns in graph [5]). To count the frequencies, we may use Random Walk Sampling[1].
- Aesthetic metrics: We will follow the experiment setting of the chosen paper, using Crosslessness[6], Minimum angle metric[6], Edge length variation[4], and Shape based metric[3] as aesthetic metrics, which are prediction objective for our machine learning models.
- Layout methods: We will select several layout methods from Force-direct method, Dimension reduction based method, Spectral method, Clustering based method, and Multi-level method.

References

[1] Xiaowei Chen, Yongkun Li, Pinghui Wang, and John Lui. A general framework for estimating graphlet statistics via random walk. *Proceedings of the VLDB Endowment*, 10(3):253–264, 2016.

[2] Timothy A Davis and Yifan Hu. The university of florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS)*, 38(1):1, 2011.

[3] Peter Eades, Seok-Hee Hong, Karsten Klein, and An Nguyen. Shape-based quality metrics for large graph visualization. In *International Symposium on Graph Drawing and Network Visualization*, pages 502–514. Springer, 2015.

[4] Stefan Hachul and Michael Jünger. Large-graph layout algorithms at work: An experimental study. *J. Graph Algorithms Appl.*, 11(2):345–369, 2007.

[5] Oh-Hyun Kwon, Tarik Crnovrsanin, and Kwan-Liu Ma. What would a graph look like in this layout? a machine learning approach to large graph visualization. *IEEE transactions on visualization and computer graphics*, 24(1):478–488, 2018.

[6] Helen C Purchase. Metrics for graph drawing aesthetics. *Journal of Visual Languages & Computing*, 13(5):501–516, 2002.