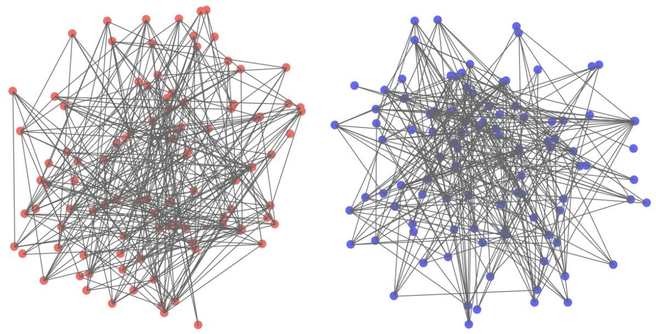
Group 21: Graph Matching and Similarity

Team Members: Louis Hu, Fayfay Ning, Kieran Lele
Project Leads: Jiaming Xu, Sophie Yu

https://github.com/louishu17/fasciaGraphSimilarity

Abstract



Problem:

- Want to know how similar two networks are, however, when given two large networks, it is hard to tell vertex correspondence. Two isomorphic networks may look drastically different.
- Enumerating every possible vertex correspondence takes n! combinations, which is too exhaustive as n increases to 1000 or 10000.

Goal:

- Quantify the similarity in structure between two large networks

Solution:

- Compute similarity based on occurrence of trees
- Since directly counting of tree size k takes O(n^k) runtime complexity, we estimate the similarity statistic via color coding algorithm

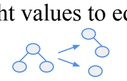
Applications:

- Online social networks, protein-protein interaction network alignment, and shape matching in computer vision

Methods

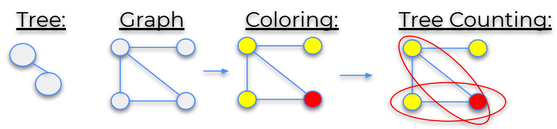
Tree Counting:

- Randomly assign color to each node in given graph
- Center graph by assigning weight values to edges
- Partition trees into smaller trees



- Use dynamic programming to count colorful trees (bottom-up counting via tree partitions)

Example: counting colorful trees with 2 nodes (colorful edges)



Similarity Score:

- Find all non-isomorphic trees of a given size
- Determine the counts for each of those trees in both graphs
- Take dot product of the vectors containing tree counts and normalize

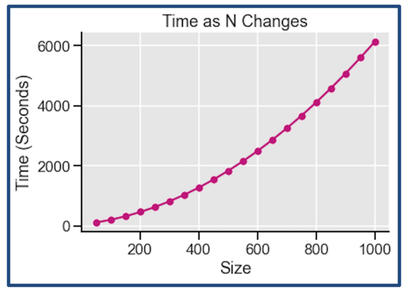
Sampling:

- Sample a specified number of nodes from each graph, and from that node sampling, sample a certain number of edges
- Calculate similarity scores between pairs of these same-sized samples

Tools:

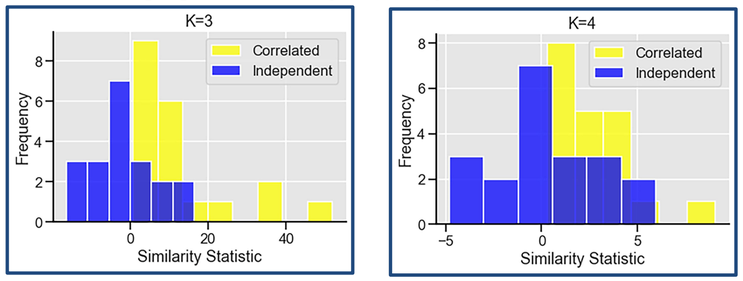
- C++: Fast Approximate Subgraph Counting and Enumeration (FASCIA) package from Penn State University [1]
- Duke Compute Cluster: OpenMP and job arrays

Results



Average runtimes of color coding algorithm over 1000 different random colorings of increasing Erdos-Renyi graph sizes with edge probability 0.001

- Color coding algorithm scales quadratically as network size increases



Similarity score distributions for 20 pairs of correlated and 20 pairs of independent Erdos-Renyi graphs as K increases from 3 to 6

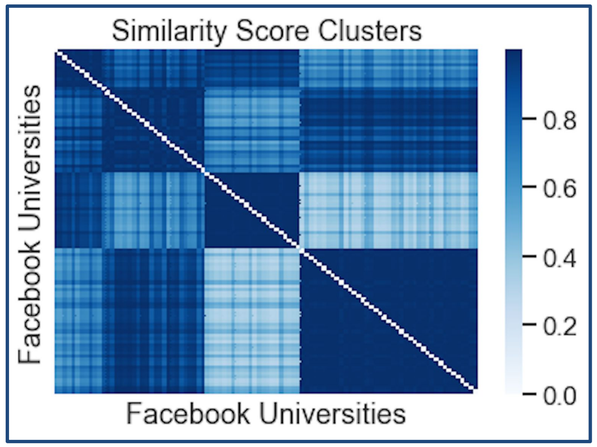
- Almost complete separation between the 20 correlated and 20 independent Erdos-Renyi graphs over 1000 different random colorings with size of 1000 nodes and edge probability of 0.001 at K=6, where K is the number of edges of the tree, which provided the best results to runtime tradeoff

	MIT 1	MIT 2	UChicago 1	UChicago 2
MIT 1	--	0.99999879	0.83558504	0.830918848
MIT 2	--	--	0.835511608	0.830846652
UChicago 1	--	--	--	0.999963312
UChicago 2	--	--	--	--

Normalized similarity scores for subsampled Facebook networks

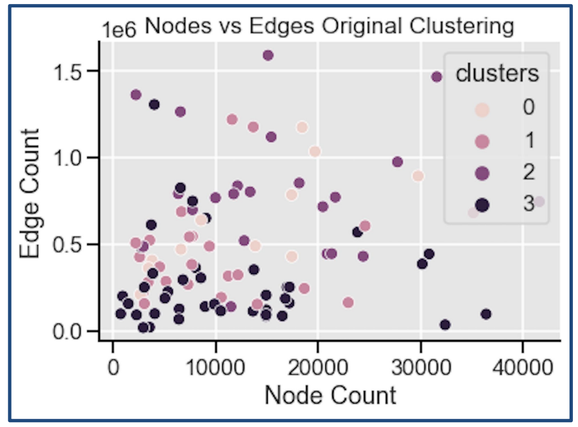
- Samplings from the same universities have higher similarity than samplings from different universities

Pairwise Similarity among Universities



K Means clustering results of 98 different university's Facebook social networks

- The algorithm is able to make large differentiations when determining the similarity of structure for 98 university's Facebook networks



Nodes vs Edges with shading of original clustering

- While edges and nodes seem to be the two characteristics of universities that most closely determine similarity in structure, there is no one single factor (e.g region. enrollment, state) that explains similarity between networks.
- This reaffirms the need for the color coding and tree counting as similarity is difficult to find otherwise

References

1. G. M. Slota and K. Madduri, "Fast Approximate Subgraph Counting and Enumeration," 2013 42nd International Conference on Parallel Processing, 2013, pp. 210-219, doi: 10.1109/ICPP.2013.30.
2. Alon, Noga et al. "Biomolecular network motif counting and discovery by color coding." *Bioinformatics (Oxford, England)* vol. 24,13 (2008): i241-9. doi:10.1093/bioinformatics/btn163
3. C Mao, Y Wu, J Xu, SH Yu, "Counting trees and testing correlation of unlabeled random graphs," preprint, 2021.