MISSISSIPPI STATE UNIVERSITY™
DISTRIBUTED ANALYTICS AND SECURITY INSTITUTE

Analysis of Synthetic Graph Generation Methods for Directed Network Graphs

Spencer Callicott, Mentors: Dr. Stefano Iannucci, Stefano Cordio

Introduction

Historically, scientific experiments have been conducted to generate scale-free network graphs based on structure. Metrics used to measure veracity ensure the integrity of a scale-free algorithm given a seed.

However, studies do not explore the performance benefits/drawbacks of specific algorithms running on Apache Spark and GraphX.

Recognizing the lack of performance benchmarks demands ensuring accuracy through experimenting.

Objectives

- Create a synthetic data-set generator using a real-data seed. Different data can be used as seed. (e.g., network traces, system logs, performance counters)
- Data-set represented as property-graph (directed multi-graph with structured labels).
- Workload implemented on top of Apache Spark (GraphX for data persistence)
- Performance measurement of graph generation algorithms to determine the most effective approach at generating large, synthetic network graphs

Glossary

Network Graph – A graph of connected nodes and edges with both node properties and edge weights.

Degree Distribution – Probability that a Node has a certain degree over the whole network.

Veracity – Measurement of how similar a certain graph matches another graph.

Euclidian Distance – Difference between two graphs

PageRank – algorithm developed by Google to rank nodes that are more likely to be visited due to degree.

Hypothesis

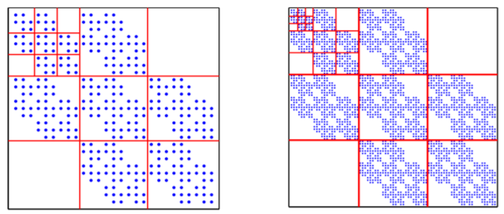
Synthetic Graphs will maintain the properties of their parent graphs by mirroring the structure and distribution of the seed graphs. Also, the algorithms will scale across a cluster, allowing for faster performance given more computational power.

Methods

Method One – Barabási–Albert Model

At each iteration *t*, a new node is added to the graph and connected to an existing node *j* with probability:

$p_j = \frac{d_j(t)}{\sum_i d_i(t)}$, where $d_i(t)$ is the degree of node *i* at iteration *t*.



(a) K_3 adjacency matrix (27×27) (b) K_4 adjacency matrix (81×81)

Method Two – Kronecker Expansion

At each iteration *k* the *k-th* Kronecker adjacency matrix *K* is computed as:

$$K_k = \mathrm{K}_1 \otimes K_1 \otimes \ \ldots \otimes K_1$$

where $\otimes$ is the Kronecker product operator.

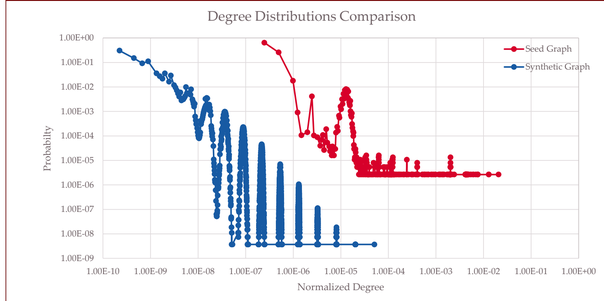
Kronecker product of a 2x2 matrix $\mathrm{K}_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$:

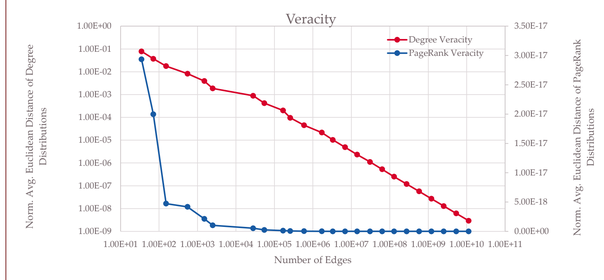
$$\mathrm{K}_2 = \mathrm{K}_1 \otimes \mathrm{K}_1 = \begin{pmatrix} a\mathrm{K}_1 & b\mathrm{K}_1 \\ c\mathrm{K}_1 & d\mathrm{K}_1 \end{pmatrix}$$

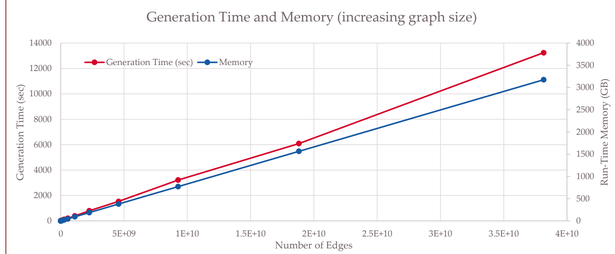
Base (naïve) algorithm with complexity $O(N^2)$, with *N* number of vertices.

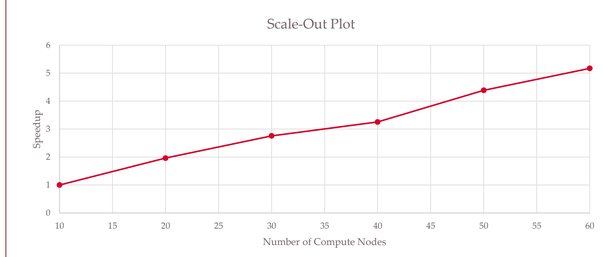
Advanced Stochastic algorithm with complexity $O(E)$, with *E* number of edges.

Results - Kronecker









Conclusion

After the initial results we found Kronecker to be a much faster algorithm than Barabási–Albert method. This is due to BA being a linear algorithm and unable to scale across a cluster.

Kronecker correctly models the internal structure of the seed graph accurately (Graph 1,2). The distance between the two graphs decreases as the generation size increases. This shows that the Kronecker algorithm is extremely accurate at generating larger graphs than the seed

Also, the advanced Kronecker algorithm scales well across clusters, showing a linear speedup when given more cluster resources.

Further Application

Evaluate performance bottlenecks in specific system architectures through testing the scalability when generating graphs.

Compare performance to additional scale-free network models such as Mediation-driven Attachment or Non-linear BA Preferential Attachment.

Use Kronecker expansion to model future networks given a current-day seed and then simulate cyber attacks on the network given a distribution of attacks.

References

- Swedish DoD dataset. ftp://download.iwlab.foi.se/dataset/smia2011/Network_traffic/
- Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. “Kronecker graphs: An approach to modeling networks.” The Journal of Machine Learning Research, 11:985–1042, 2010.
- Reinert, Gesine. "Barabási-Albert Random Graphs, Scale-free Distributions and Bounds for Approximation through Stein's Method." *ORA - Oxford University Research Archive*. N.p., 01 Jan. 1970. Web. 28 Feb. 2017.

Download the source for CSB at:
https://github.com/msstate-dasi/csb/