Models for Generating Large Realistic Graphs

Scalable Methods for Representing, Characterizing, and Generating Large Graphs

Ali Pinar (speaker), C. Seshadhri, Tamara G. Kolda
Sandia National Laboratories P.O. Box 969, MS 9159, Livermore, CA 94550

Despite their growing importance as the standard model for interconnected systems, our understanding of the graphs is still limited. Most notably we do not have models that can characterize these graphs. Our core thesis is that a graph may have a phenotype and a genotype. The phenotype of a graph corresponds to all the measurable properties, such as the degree distribution, clustering coefficient, etc. The genotype, on the other hand, refers to its organizational principles and has to be inferred from its observable features. Most of the work to date has concentrated on the phenotype, where the challenge has been generating graphs that satisfy a set of specified metrics. In contrast, our goal is to understand the graphs genotype (its organizational principles), so that we may generate accurate models of real-world complex networks. Such models are crucial, since they can provide insights into generative processes, properties, and evolution of these graphs. Moreover, due to limitations in sharing real graphs, generative models are critical for developing better algorithms at various scales and properties. Such models will also be a critical enabler for anomaly detection on graphs and will guide statistical sampling. The goal of our project is to provide characterizations of these graphs, our efforts so far have focused on both analysis and improvements of the current models and developing new models.

Analysis of Stochastic Kronecker Graphs and Impact on GRAPH500 Benchmark. The stochastic Kronecker graph model (SKG) is widely used in the HPC community, most notably as the graph model for the Graph500 benchmark. Our rigorous analysis showed that SKG *cannot* generate a power-law degree distribution or even a lognormal distribution, but instead has an oscillatory degree distribution that is not characteristic of real-world data. Therefore, we formalized an enhanced version of the SKG model that uses random noise for smoothing, and rigorously proved that this enhancement leads to a lognormal degree distribution. Additionally, we provide a precise analysis of isolated vertices, showing that the graphs that are produced by SKG might be quite different than intended. For example, between 50% and 75% of the vertices in the Graph500 benchmarks will be isolated. Finally, we show that this model tends to produce extremely small core numbers (compared to most social networks and other real graphs) for common parameter choices. These results have been communicated to the Graph500 steering committee, and they are currently using our theorems to set the benchmark parameters. Our noisy version of SKG is likely to become the Graph500 standard for next year's benchmark.

A new model: Block Two-Level Erdös Rényi (BTER) Our proposed model is motivated by two basic observations about real-world graphs: they tend to have skewed degree distributions (few nodes with very high degrees and many nodes with small degrees) and high clustering coefficients (two nodes are more likely to be connected if they have common neighbors). We have observed that these two properties imply existence of specific structures in the graph, in the form of tightly connected communities of vertices with similar degrees. Exploiting this property, our algorithm takes a specified degree distribution as input, and starts with forming such communities. Then we add edges randomly to the graph to preserve the specified degree distribution. Our initial results show that the generated graph is strikingly similar to the original. We are currently analyzing and improving our model for even better results.