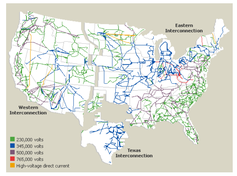
Models for Generating Large Realistic Graphs

Ali Pinar, C. Seshadhri, Tamara Kolda, Jaideep Ray, Cindy Phillips, Janine Bennett
Sandia National Laboratories

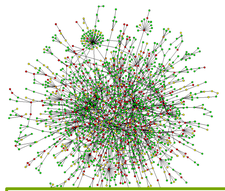
SAND2011-7627C

Networks are everywhere

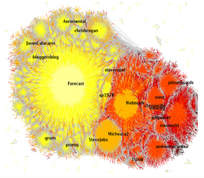
Networks are recognized as the standard tool to model complex interconnected systems.



U.S. Power Grid [GENI]



Yeast protein interactions [Bordalier institute]



Twitter social network [Akshay Java, 2007]

Physical Networks: Power, water, fracture, communication networks

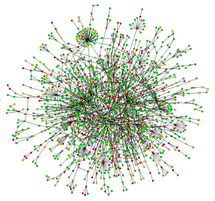
Functional Networks supply chains, chemical reaction networks, regulatory networks

Interaction Networks: cybersecurity, social networks, epidemics

Need for network models underlie many challenges

Our goal is to design models that can describe a graph with a small number of parameters. Such models will be instrumental for:

- ❖ Insights into
  - ❖ generative process
  - ❖ graph properties (e.g., eigenvalue distribution)
  - ❖ evolution
- ❖ Design and analysis of algorithms and architectures
  - ❖ Alternative to worst-case analysis
  - ❖ Rigorous studies of heuristics
  - ❖ Runtime analysis of algorithms
  - ❖ Benchmarking computers
- ❖ Comparing graphs
- ❖ Sharing of realistic but non-sensitive data
- ❖ Statistically significant graph mining
- ❖ Model validation
- ❖ Network inference

$f(x) =$ 

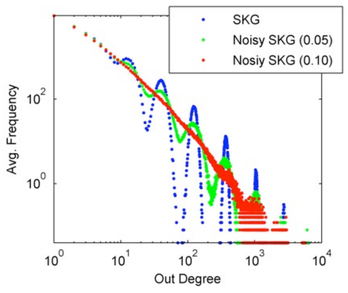
In-depth analysis of stochastic Kronecker graphs

Stochastic Kronecker Graph (SKG) has been chosen to generate graphs for the GRAPH500 supercomputer benchmark. It is favored for small number of parameters, ease of implementation, full parallelism, an the assumption that graphs generated by these models resemble real world graphs.

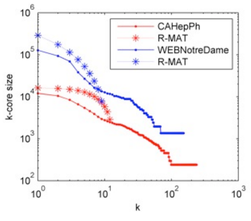
Our analysis of this model provided several important results:

✳ We proved that *SKG cannot generate power-law or lognormal degree distributions*, because its degree distribution oscillates between a lognormal and an exponential.

✳ We improved the degree distributions by a properly adding noise. *We theoretically and empirically showed that this method yields lognormal degree distributions.*



✳ SKG creates notably fewer vertices than intended. For Graph500 parameters 50-75% of the vertices are isolated. We developed techniques to estimate this number for given model parameters, and predict which vertices will be isolated.



✳ Core sizes of SKG graphs are significantly lower than those of the graphs they model, which is a sign of teh poor community structure.

GRAPH 500

Graph500 benchmark is modified to use the noisy SKG model that we have proposed. Our tools are also being used to design future benchmarks.

Requirements for a good model

- *Flexibility in degree distribution* There is no single distribution that works for all graphs, thus a good model should be able to generate graphs with a variety of degree distributions.
- *Communities and high clustering coefficients* Graphs are known to have many small communities and high clustering coefficients for vertices of all degrees. (i.e., If (u,v) and (v,w) are edges, probability of (u,w) should be high.) This is a major shortcoming of all scalable graph models.
- *Small-world diameter* Many real graphs have amazingly short distances between most pairs of vertices.
- *Scalability and parallelizability* *We need to generate extremely large instances in an efficient way.*

Clustering Coefficient

$$cc_i = \frac{t_i}{\binom{d_i}{2}}$$

t_i = # triangles at vertex i
d_i = degree of vertex i

The Blocked Two-Level (BTER) Graph Model

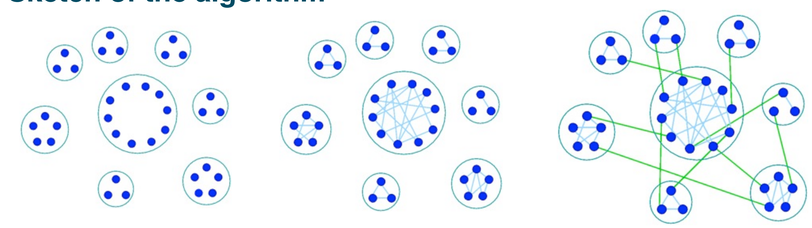
Core idea of BTER

High clustering coefficients for high degree vertices imply fairly dense clusters. High clustering coefficients for low degree vertices imply many small communities. We explicitly account for this factor. Skewed degree distributions leave enough edges after building these dense blocks to satisfy small world property.

BTER Parameters

Degree distribution or a description of it (e.g., the power-law coefficient) guides the generation.
Density parameter controls the density of the smallest blocks.
Density decay parameter controls how fast the densities of the blocks decrease with increasing block sizes.

Sketch of the algorithm



Partitioning into blocks
Group vertices into bins, taking into account their degrees. The grouping is assortative to ensure high clustering coefficients.

Community Level
Add edges between vertices in the same group. Edges are added uniformly random within each group (Erdos Renyi). The density of the smallest blocks are set by the density parameters. Densities of the blocks decrease with increasing community size, which is controlled by the decay parameter.

This phase ensures high clustering coefficients.

Interconnection Level
The remaining edges are added to satisfy the specified degree distribution using the Chung and Lu model. In this model probability of an edge is directly proportional to the product of the degrees of its end vertices.

This phase ensures small-world property and satisfies the degree distribution requirement.

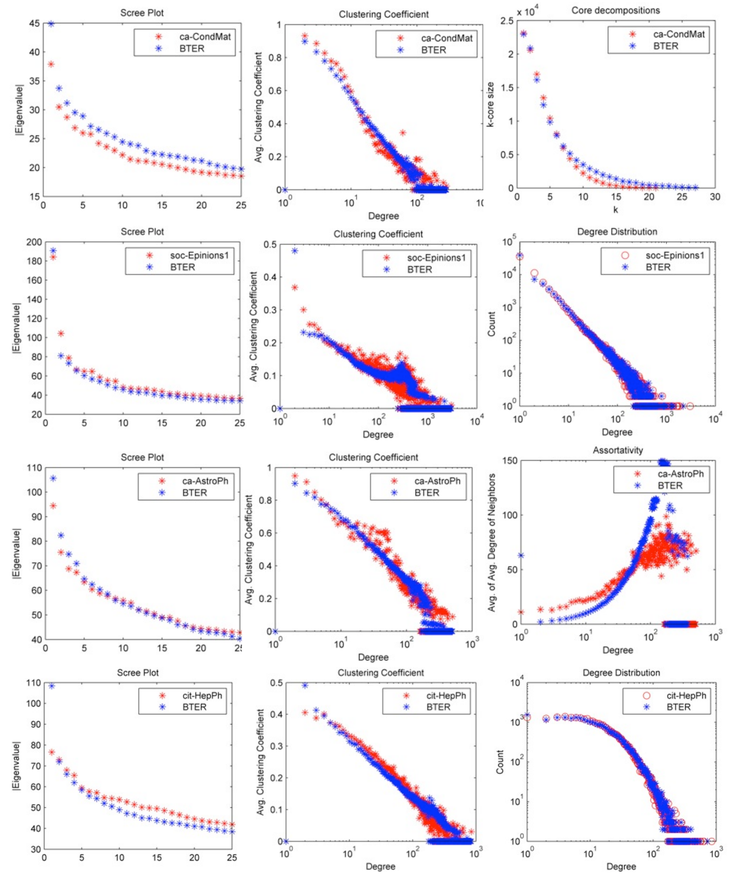
Advantages of BTER

✳ Is not restricted to any particular degree distribution
✳ Can match the large clustering coefficients observed in real graphs
✳ Accommodates a community structure with many small communities
✳ Provides a compact representation of a graph
✳ Can generate arbitrarily large graphs and is parallelizable

Future Work

✳ Model improvement Current model is our first-order approximation,and many details needs to be studied. In particular, we are working on incorporating better community structure models.
✳ Model validation Our results show that our graphs exhibit many favorable properties. We claim that any good model should be similar to BTER, and we believe we can show that.
✳ Evolution We want to extend our model to include how graphs evolve over time.
✳ Parallel implementation Our model is parallelizable, and we plan to provide such an implementation for broader adoption of our method.
✳ Applications Our model has already drawn a lot of interest from the HPC community. We plan to broaden our impact through other applications (e.g., statistical analysis, anomaly detection).

BTER accurately regenerates real-world graphs



Code Availability

The source code for the BTER graph generator is available at http://csmr.ca.sandia.gov/

Relevant Publications

1. C. Seshadhri, T. Kolda, and A. Pinar, The Blocked Two-Level Erdos Renyi Graph Model,submitted for journal publication
2. C. Seshadhri, A. Pinar, and T. Kolda, “An In Depth study of Stochastic Kronecker Graphs," submitted for journal publication.
3. I. Stanton and A. Pinar, " Constructing and uniform sampling graphs with prescribed joint degree distribution using Markov Chains," submitted for journal publication.
4. C. Seshadhri, A. Pinar, and T. Kolda, Comparison of Scalable Graph Generation Models, submitted for conference publication.
5. M. Rocklin, and A. Pinar, ``On Clustering on Graphs with Multiple Edge Types," submitted for joirnal publication.
6. E. Kayaaslan, A. Pinar, U. Catalyurek, and C. Aykanat, "Hypergraph Partitioning through Vertex Separators on Graphs", submitted for journal publication.
7. C. Seshadhri, A. Pinar, and T. Kolda, “An In Depth study of Stochastic Kronecker Graphs," to appear in Proc. Int. Conf. on Data Mining.
8. M. Rocklin and A. Pinar, ``Latent Clustering on Graphs with Multiple Edge Types," Proc. 8th Workshop on Algorithms and Models for the Web Graph (WAW11).
9. I. Stanton and A. Pinar, ``Sampling graphs with prescribed joint degree distribution using Markov Chains, " Proc. ALENEX 11.
10. M. Rocklin and A. Pinar, ``Computing an Aggregate Edge-weight function for Clustering Graphs with Multiple Edge Types", in Proc. 7th Workshop on Algorithms and Models for the Web Graph (WAW10).

Contact Information

For further details about the project, please contact Ali Pinar at apinar@sandia.gov.

Funding Statement

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.

Disclaimer of Liability