SAND2012-1799C

The Inherent Community Structure in Real-World Graphs

Ali Pinar, C. Seshadri, and Tamara G. Kolda

Sandia National Labs

U.S. Department of Energy
Office of Advanced Scientific Computing Research

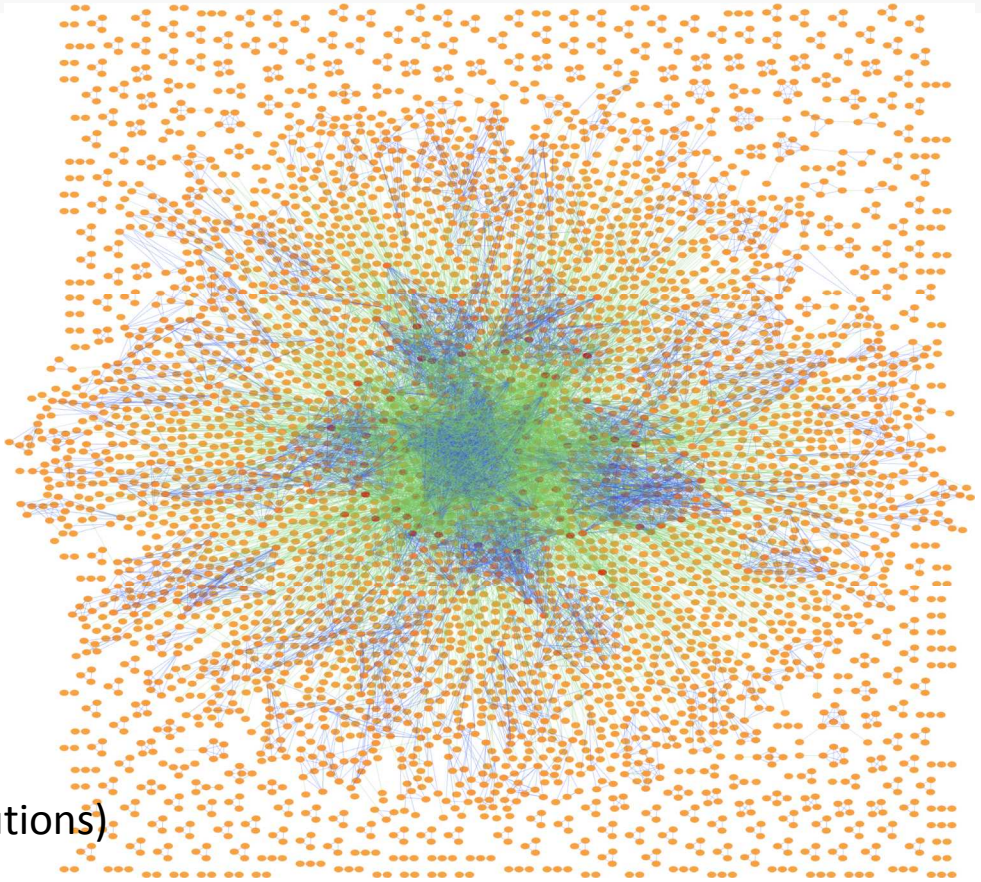
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Chicken or the egg?

- Finding communities in real–world graphs has been and still is a very active research area.
- Conventional wisdom:
  - Complex networks have communities.
- Alternative view:
  - Communities is one of the key ingredients turn an arbitrary network into a complex network.
- The goal of the talk is to show the importance of communities in characterizing/modeling a network.

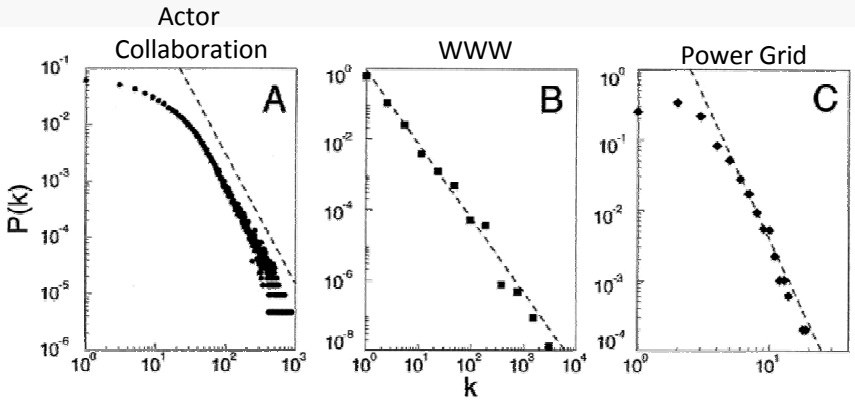
Why Model Massive Graphs?

- Enable sharing of surrogate data
  - Computer network traffic
  - Social networks
  - Financial transactions
- Insight into…
  - Generative process
  - Community structure
  - Comparison
  - Evolution
  - Uncertainty
- Testing graph algorithms
  - Scalability
  - Versatility (e.g., vary degree distributions)
  - Verification & validation

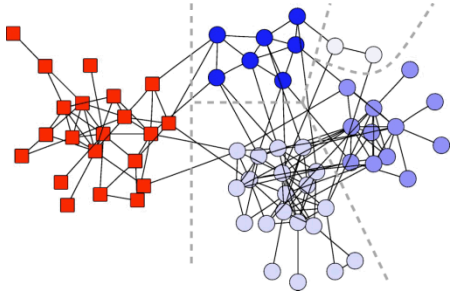


Block Two-Level Erdös-Rényi (BTER) graph; image courtesy of Nurcan Durak.

Model Desiderata



A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. Science, 286(5349):509-512, 1999.



M.E.J. Newman and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004.

http://www.graph500.org/

- Capture heavy-tailed degree distribution
  - Not necessarily exactly power law
- Capture community structure
  - Measured indirectly through clustering coefficient, *k*-cores, and other measures
- Able to "fit" real-world data
  - Reproduce degree distribution
  - Reproduce community structure

Scales to 1T nodes

- Motivated by GRAPH500 benchmark
- Typically also need for randomized fitting procedures

Clustering Coefficients

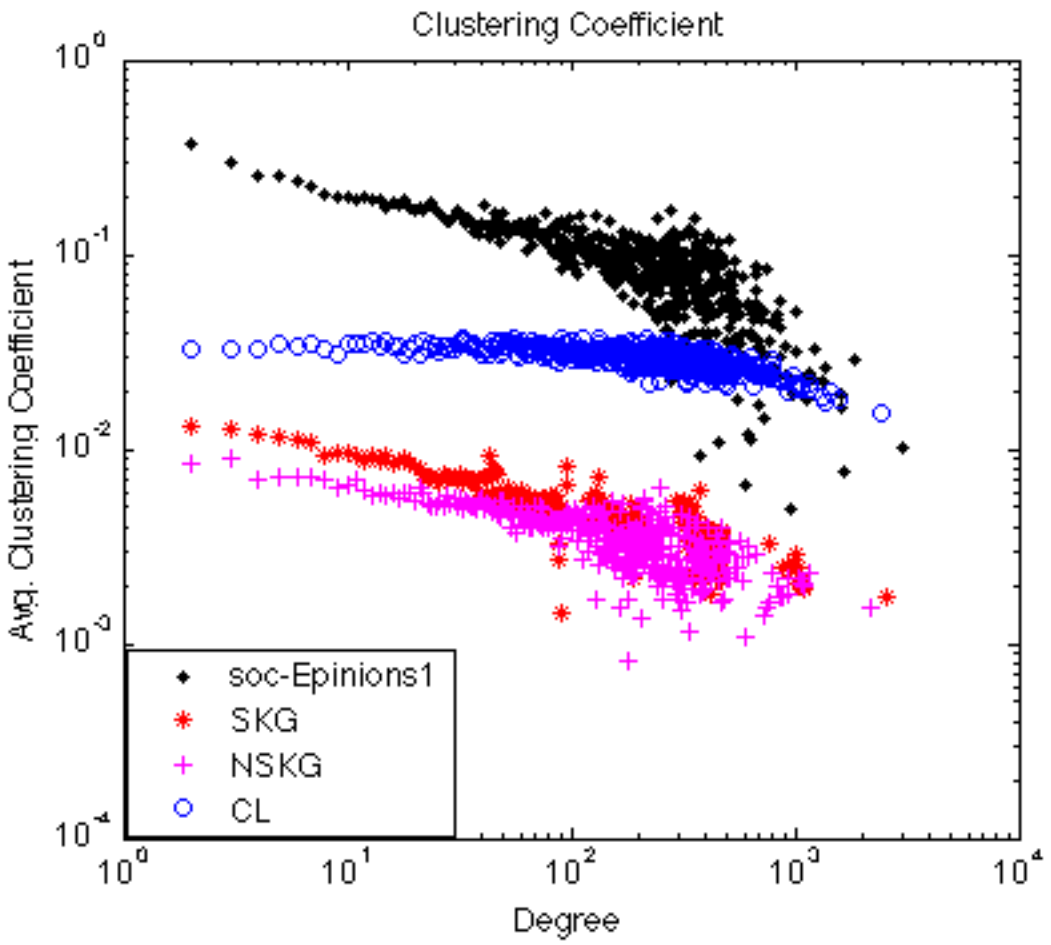
Clustering Coefficient

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

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

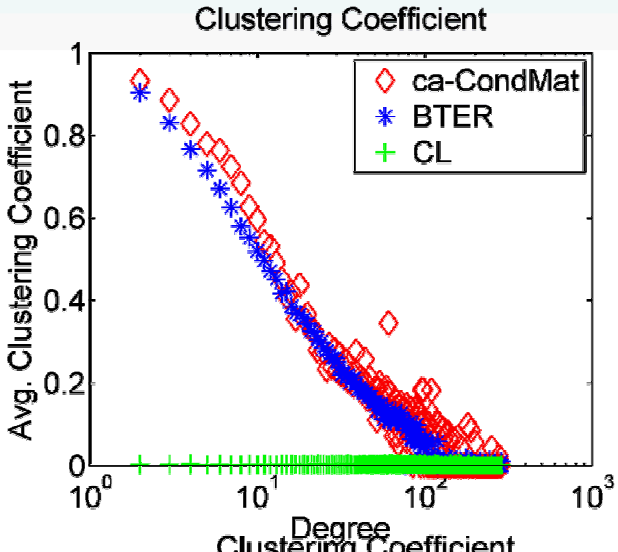
Global Clustering Coeff.

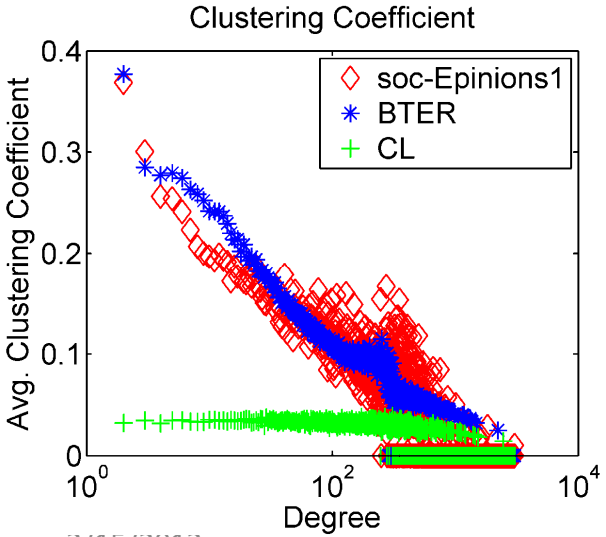
$$C = \frac{\sum_i t_i}{\sum_i \binom{d_i}{2}}$$

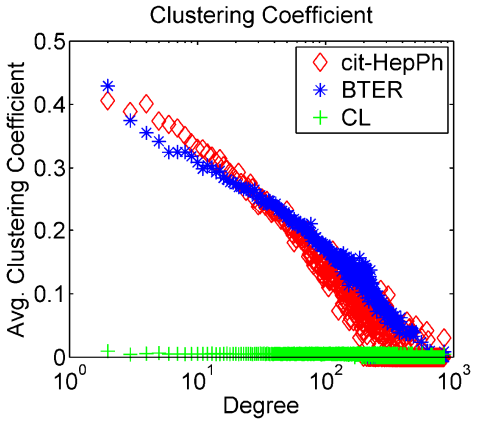


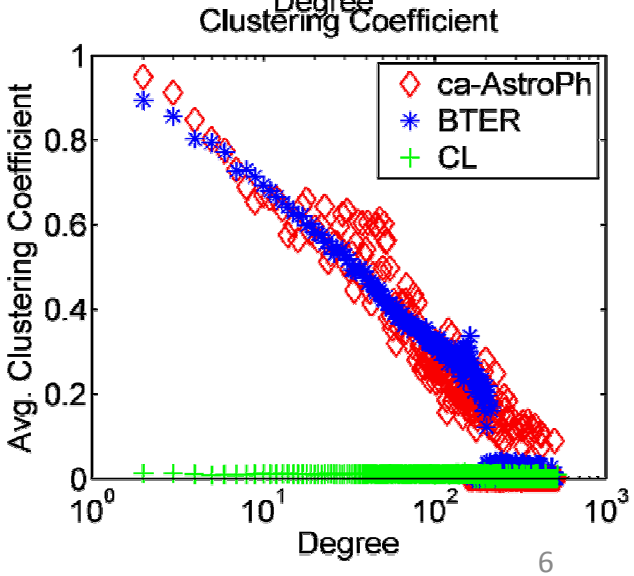
Clustering coefficients observations

Clustering coefficients are higher for lower degree vertices.



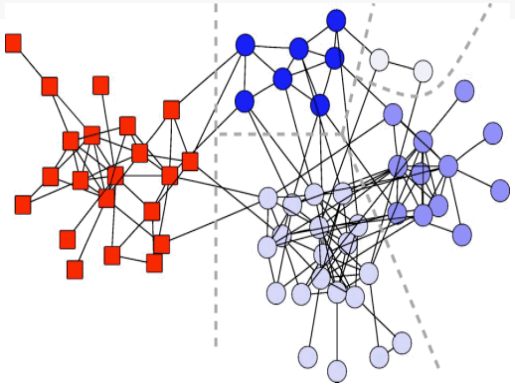




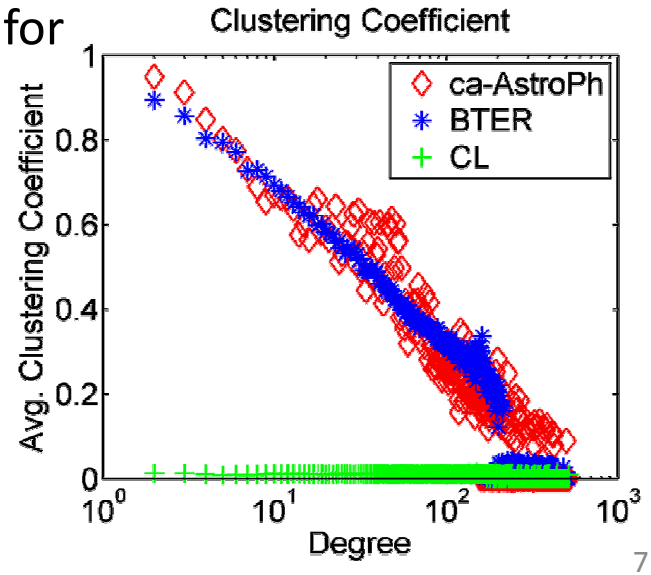


Community Structure in Graphs

- Numerous community finding algorithms exist
  - Difficult to validate
  - Trouble in finding full range of sizes
- Instead, use related measures like clustering coefficient
  - Triangles arise because of community structure
- What "community" structure must be present to ensure a high clustering coefficient, especially for low-degree nodes?
  - How many communities?
  - What do they look like?



M.E.J. Newman and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004.



Building the basis for a model

Seshadhri, Kolda, & Pinar, arxiv:1112.3644

Empirical Observations

- Degree distributions are heavy tailed.
- Clustering coefficients are highest for small degree vertices.

Theoretical Analysis

Theorem: *If a community has s edges then there must be Ω(√s) vertices with degree Ω(√s).*

We are not only trying to build a formal model, we are trying to formalize the model building process itself.

Hypothesis: Real-world interaction networks consist of a scale –free collection of dense Erdős-Rényi graphs.

Verifying the model

Hypothesis: Real-world interaction networks consist of a scale –free collection of dense Erdős-Rényi graphs.

Prediction: the number of communities grows with the number of nodes.

Verification: The sizes of the communities are small, and do not grow (or grow very slowly) for larger graphs. (e.g., Dunbar number).

Prediction: Within a community the degree variance should be small.

Verification: Vertex degree is correlated to the average degree of neighbors.

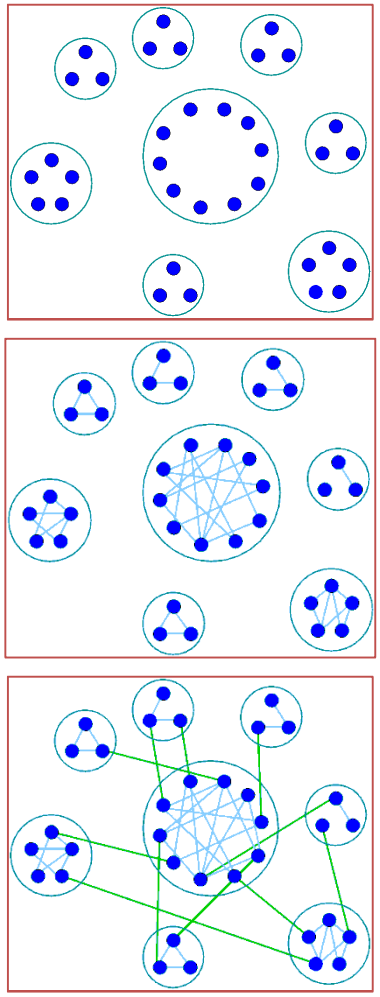
BTER: A New Model with Explicit Community Structure

Seshadhri, Kolda, & Pinar, arxiv:1112.3644

Preprocessing: Create explicit communities

Phase 1: Erdös-Rényi graphs in each community

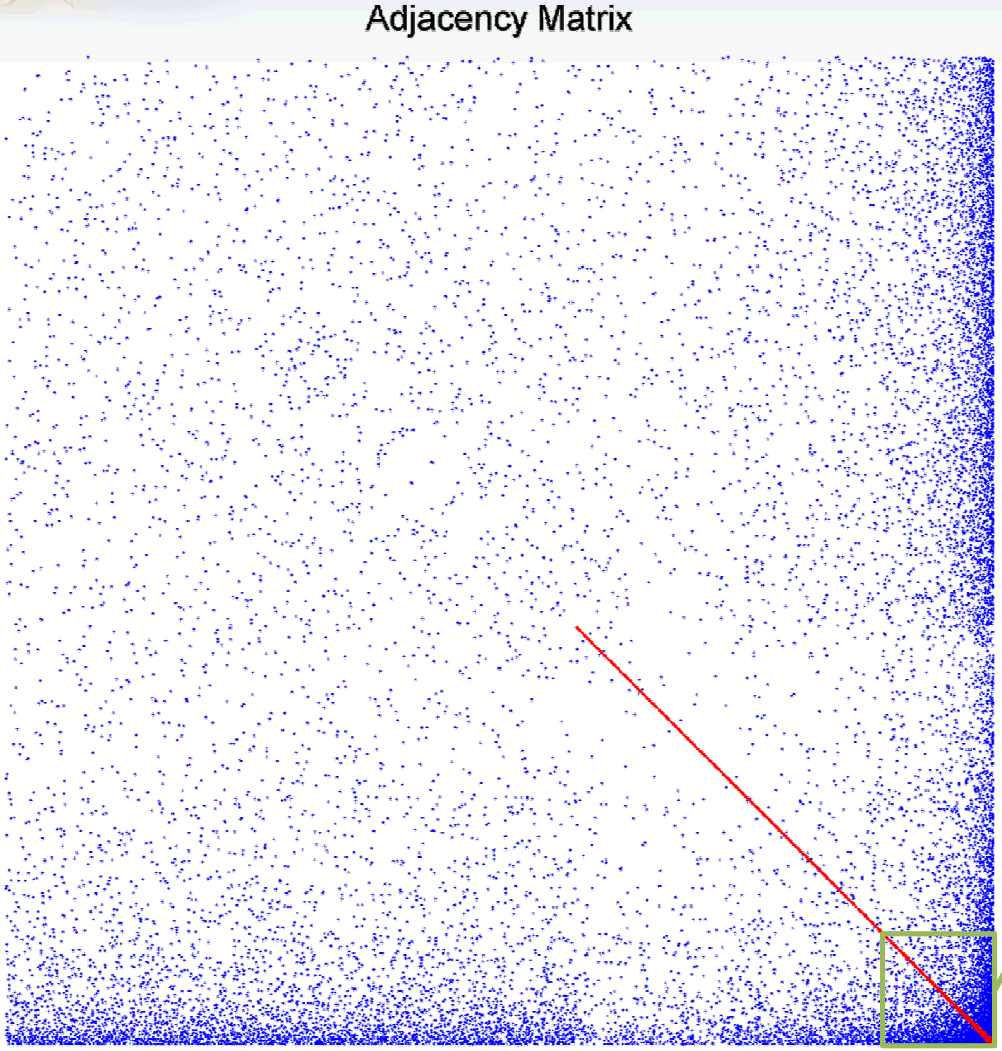
Phase 2: CL model on "excess" degree



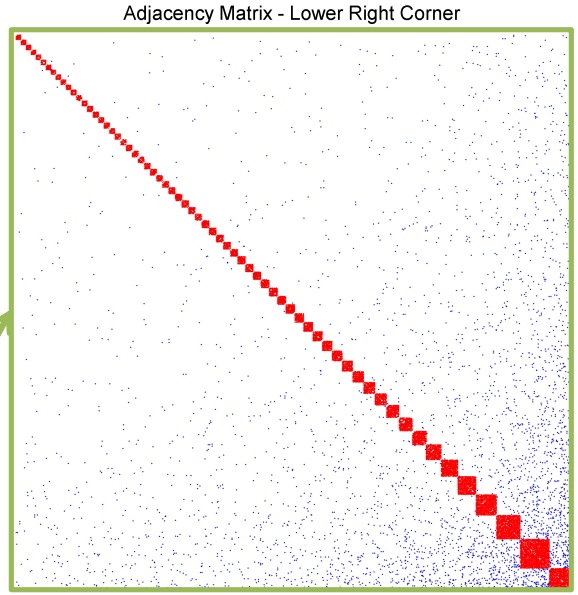
Hypothesis: Real-world interaction networks consist of a scale –free collection of dense Erdős-Rényi graphs.

- **Preprocessing:** Generate communities
  - Determined by **desired degree distribution**
  - All nodes have (close to) the same degree
  - Size of cluster = min degree + 1
- **Phase 1:** Generate ER graph on each community
  - User must **specify connectivity coefficient** for each community, $\frac{1}{2}_k$
  - We use a function of the min degree in the community, d_k
- **Phase 2:** Generate CL graph on "excess" degree
  - $e(i) = d(i) - \frac{1}{2}_k d_k$ where vertex i is in community k

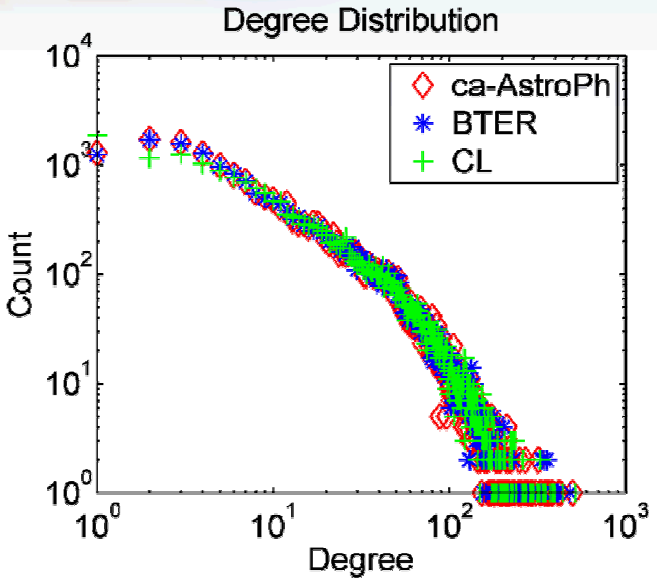
Visualization of BTER Adjacency Matrix

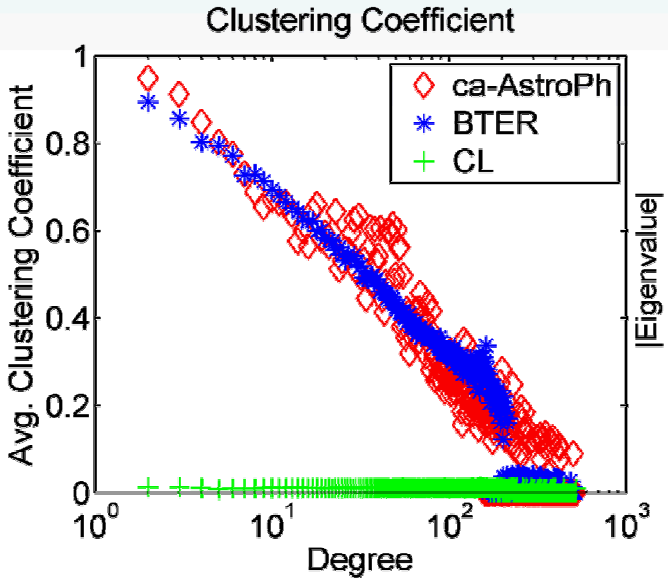


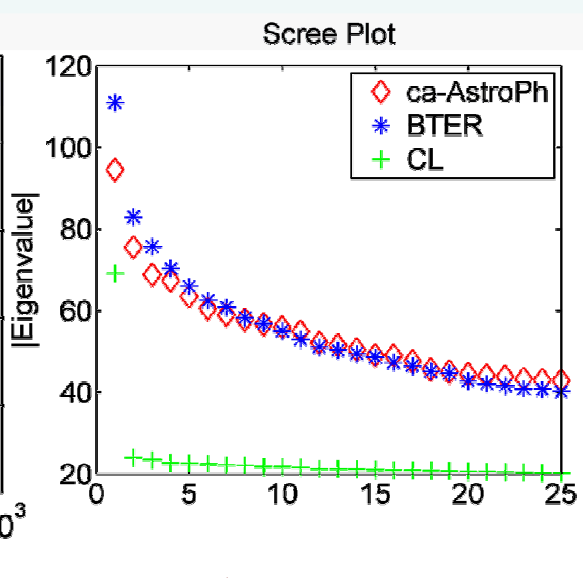
Red = Phase 1
Blue = Phase 2



Co-authorship (ca-AstroPh)



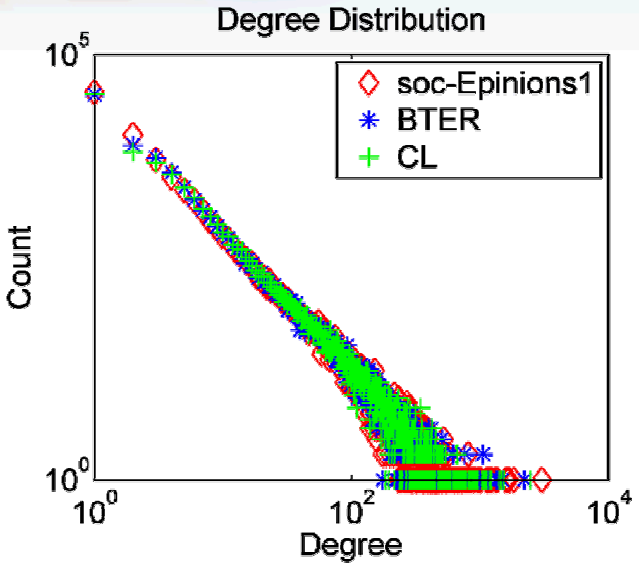


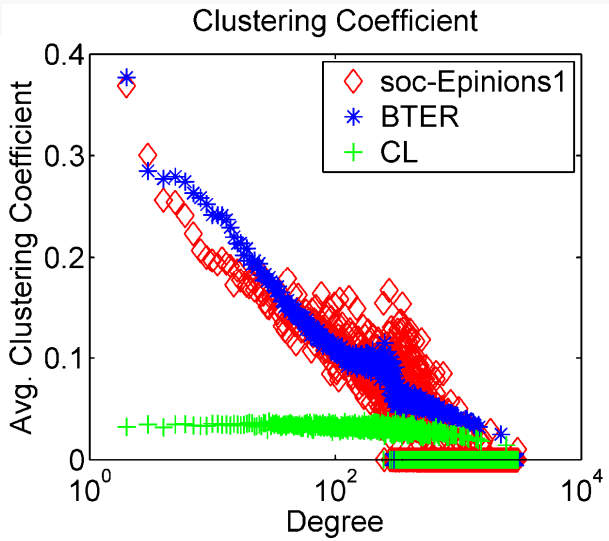


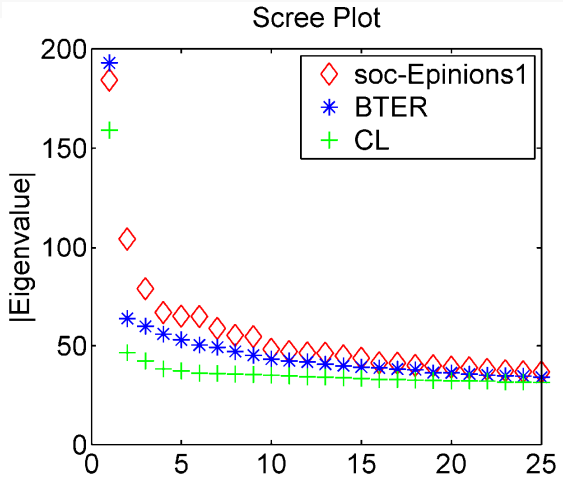
- 18,771 nodes 396,100 edges; based on arxiv repository
- Global clustering coefficient
  - Original: 0.32, BTER: 0.31, CL: 0.01
- Normalized size of the largest connected component
  - Original: 0.95, BTER: 0.86, CL: 1.0

Eigenvalues are not determined by degree distribution

Trust Network





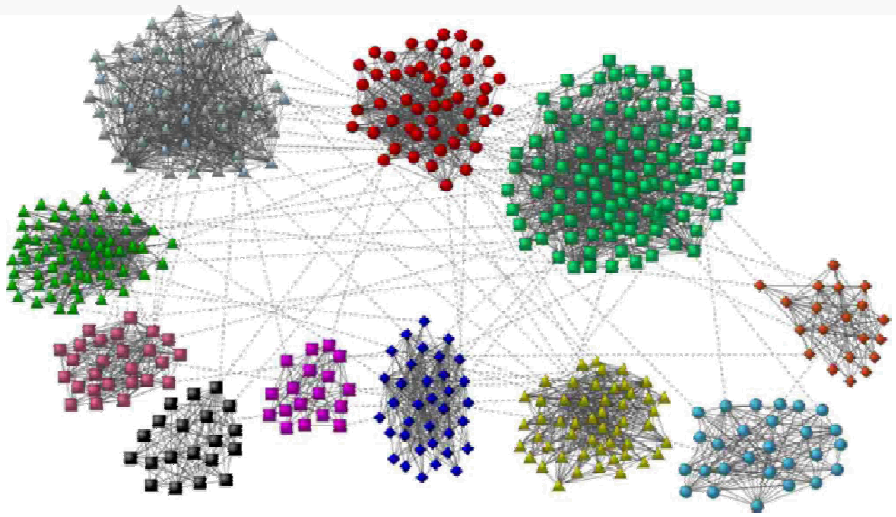


- 75,879 vertices, 811,480 edges; based on Epinion web site; edges represent trust between two users
- Global clustering coefficient
  - Original: 0.07, BTER: 0.07, CL: 0.03
- Normalized size of the largest connected component
  - Original: 1.0, BTER: 0.96, CL: 0.98

BTER provides good matches to real data, even when the clustering coefficients are not too high

Relationship of BTER and LFR Models

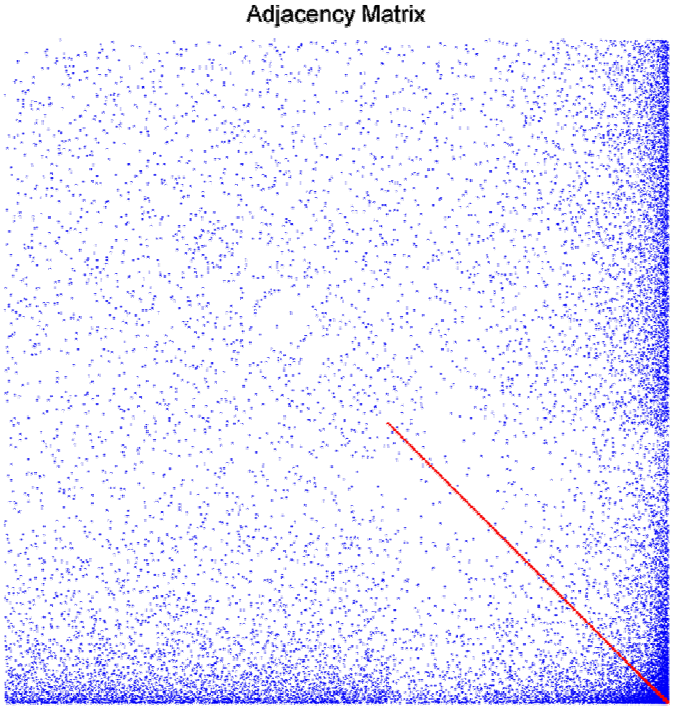
- Both explicitly insert communities
- Community structure
  - BTER: generated automatically according to degree distribution
  - LFR: power law distributed
- Assignment of nodes to communities
  - BTER: Determined during community building phase
  - LFR: Random assignments; any node can go into a community where the size is higher than its "internal degree"
- Internal vs. external links
  - BTER: Varies by node
  - LFR: constant proportions for all nodes
- Community sizes
  - BTER: All sizes down to 3 nodes
  - LFR: Minimum community size specified by user
- Scalabilty
  - FLR community assignment procedure is not obviously parallelizable



Lancichinetti, Fortunato, & Radicchi, Benchmark graphs for testing community detection algorithms, *Phys. Rev. E,* **2008**

Observations on BTER

- The current model is only a first order approximation.
- Requires desired degree distribution
  - Approximation can be used to save space
- Phase 1: Communities
  - All nodes have the same (expected) degree; easy generation of dense subgraphs
  - But there are ways we could allow the communities to be heterogeneous
  - Community edge density is a parameter which may be tuned to fit real data
- Phase 2: Uses expected excess degree
  - Enables “streaming edge” generation
- BTER edge generation is fully parallelizable
  - community membership for each node
  - edge density for each community
  - excess degree (in expectation) for each node

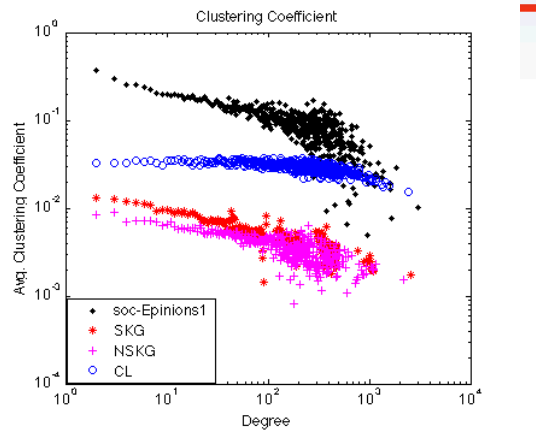


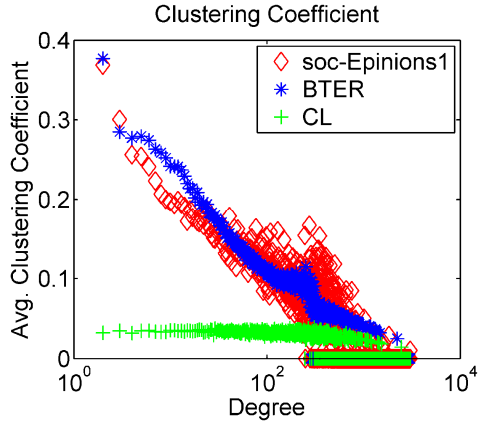
Concluding Remarks

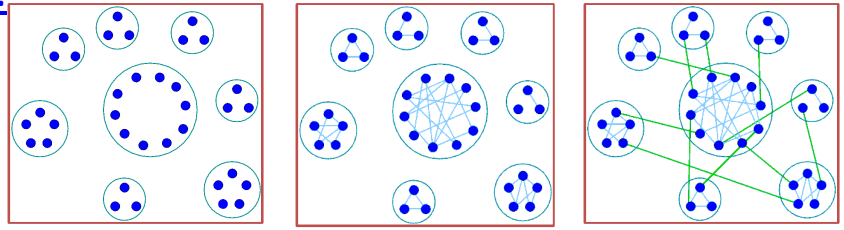
- Modeling of graphs underlie many challenges for principled graph analysis.
- The challenge is not in building a formal model, but formalizing the modeling process itself.
- We propose the Block Two-Level Erdös-Rényi (BTER)
  - New theory says there must be many dense subgraphs for high clustering coefficient
  - New BTER model explicitly creates dense communities using ER
  - Exceptional similarities to real data in terms of clustering coefficients and eigenvalues
- The code is available at

http://www.sandia.gov/~tgkolda/bter_supplement/.

- For more information,
  - Ali Pinar apinar@sandia.gov







Relevant Publications

- Modeling of graphs
  - C. Seshadhri, T. Kolda, and A. Pinar, “The Blocked Two-Level Erdos Renyi Graph Model,” submitted for journal publication
  - C. Seshadhri, A. Pinar, and T. Kolda, “An In Depth analysis of Stochastic Kronecker Graphs," submitted for journal publication.
  - A. Pinar, C. Seshadhri, and T. Kolda, “The Similarity of Stochastic Kronecker Graphs to Edge-Configuration Models,” submitted for conference publication.
  - C. Seshadhri, A. Pinar, and T. Kolda, “An In Depth study of Stochastic Kronecker Graphs,” to appear in Proc. Int. Conf. on Data Mining (ICDM).
- Sampling Graphs
  - I. Stanton and A. Pinar, “Constructing and uniform sampling graphs with prescribed joint degree distribution using Markov Chains,” submitted for journal publication.
  - I. Stanton and A. Pinar, “Sampling graphs with prescribed joint degree distribution using Markov Chains,” Proc. ALENEX 11.
- Community structure
  - M. Rocklin, and A. Pinar, “On Clustering on Graphs with Multiple Edge Types,” submitted for journal publication.
  - 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).
  - 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).

Supplementary Material

Model building approach: find features that restrict the space

- Empirical observations
  - Skewed degree distributions
  - High clustering coefficients
  - Ex