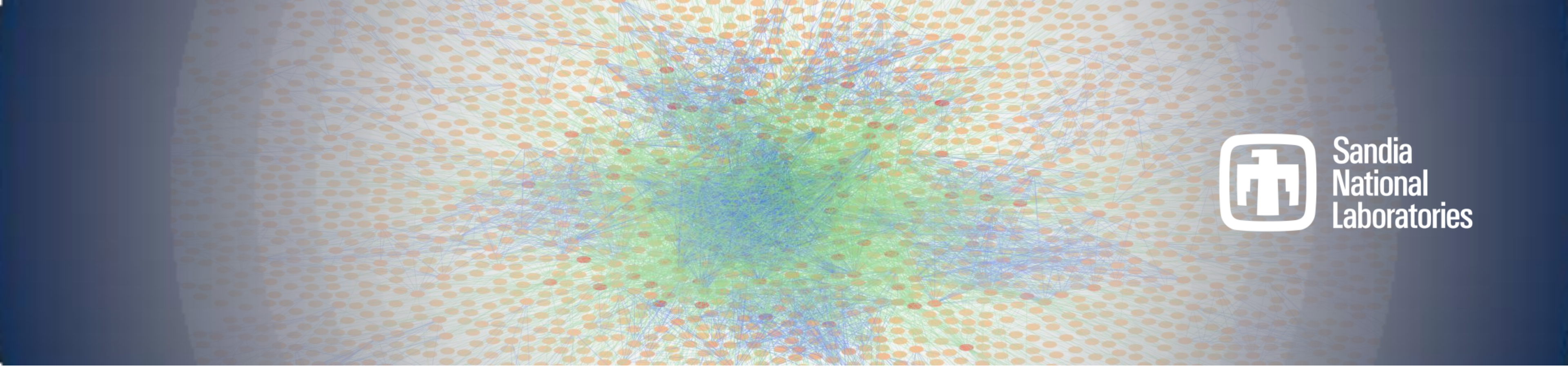


Generating Large Graphs for Benchmarking

Ali Pinar, Tamara G. Kolda,C. Seshadhri, Todd Plantenga

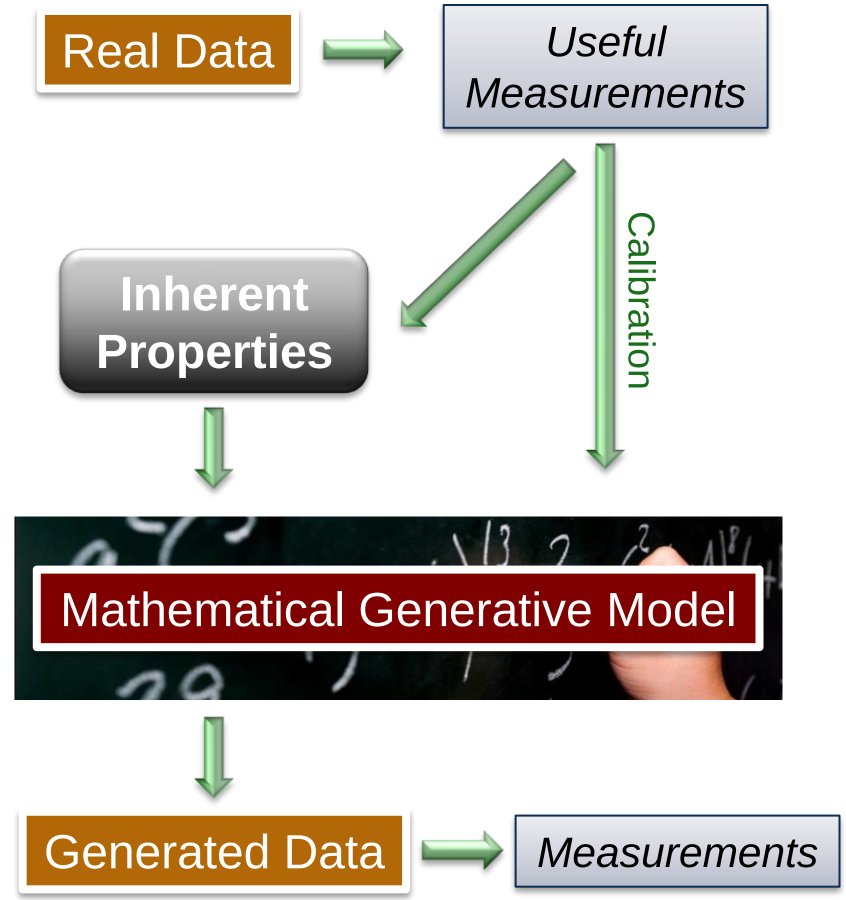
U.S. Department of Energy
Office of Advanced Scientific Computing Research

U.S. Department of Defense
Defense Advanced Research Projects Agency

Modeling graphs is a crucial challenge

Sandia National Laboratories

- Our understanding of network structure is still limited.
  - We do not have the first principles.
- Why model graphs?
  - Real data will rarely be available.
  - Understanding normal helps identifying abnormal.
  - Benchmarking requires controlled experiments.
- Challenges
  - *Data analysis:* Identifying metrics that can help in characterization (e.g., degree distribution, clustering coefficients)
  - *Theoretical analysis:* Understanding the structure inferred by these metrics
  - *Algorithms:* Designing algorithms to compute these metrics, generate graphs, etc.



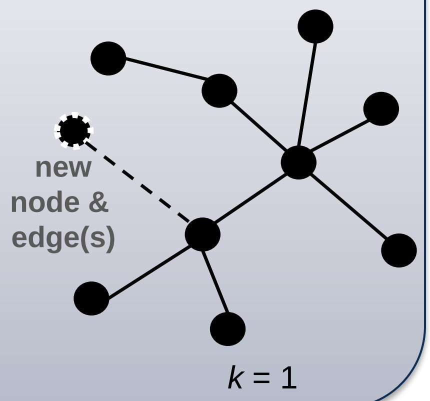
A Good Network Model…

- Encapsulates underlying driving principals
  - "Physics"
- Captures measurable characteristics of real-world data
  - Degree distribution
  - Clustering coefficients
  - Community structure
  - Connectedness, Diameter
  - Eigenvalues
- Calibrates to specific data sets
  - Quantitative vs. qualitative
  - Surrogate for real data, protecting privacy and security
  - Provides results "like" the real data
  - Easy to share, reproduce
- Yields understanding
  - Serve as null model
  - Statistical sampling guidance
  - Predictive capabilities

Story-driven models

Example: Preferential Attachment (Barabasi & Albert, Science,1999)

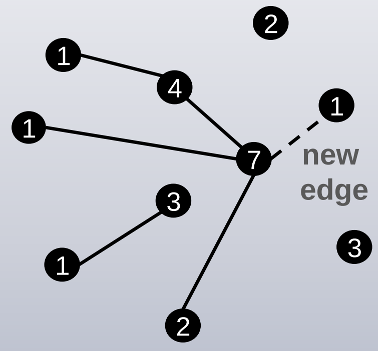
- New nodes joins graph one at a time, in sequence
- Each new node chooses *k* new neighbors, according to degree
- Node degrees updated after each addition – *Rich get richer!*



Structure-driven models

Example: CL (aka Configuration) (Chung & Lu, PNAS, 2002)

- Desired node degrees specified in advance
- New edges inserted, choosing endpoints by desired degree
- Higher-degree nodes are more likely to be selected



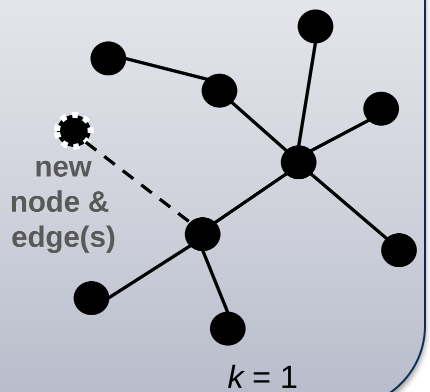
A Good Network Model…

- Encapsulates underlying driving principals
  - "Physics"
- Captures measurable characteristics of real-world data
  - Degree distribution
  - Clustering coefficients
  - Community structure
  - Connectedness, Diameter
  - Eigenvalues
- Calibrates to specific data sets
  - Quantitative vs. qualitative
  - Surrogate for real data, protecting privacy and security
  - Provides results "like" the real data
  - Easy to share, reproduce
- Yields understanding
  - Serve as null model
  - Statistical sampling guidance
  - Predictive capabilities

Story-driven models

Example: Preferential Attachment (Barabasi & Albert, Science,1999)

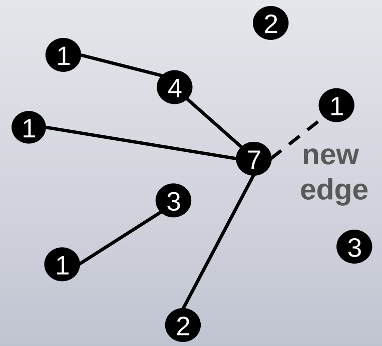
- New nodes joins graph one at a time, in sequence
- Each new node chooses k new neighbors, according to degree
- Node degrees updated after each addition – *Rich get richer!*



Structure-driven models

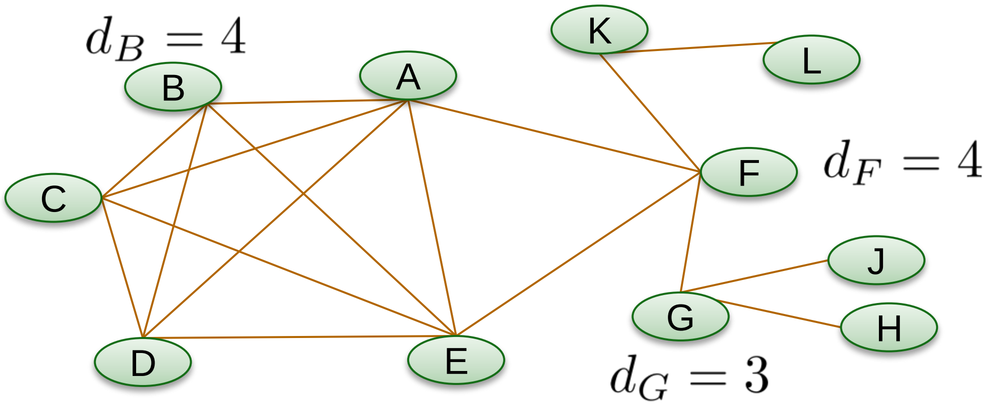
Example: CL (aka Configuration) (Chung & Lu, PNAS, 2002)

- Desired node degrees specified in advance
- New edges inserted, choosing endpoints by desired degree
- Higher-degree nodes are more likely to be selected



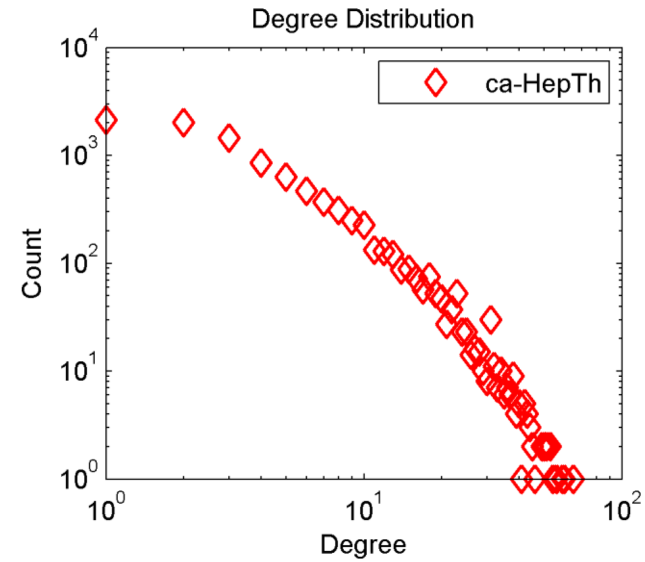
Degree Dist. Measures Connectivity

The ***degree distribution*** is one way to characterize a graph.



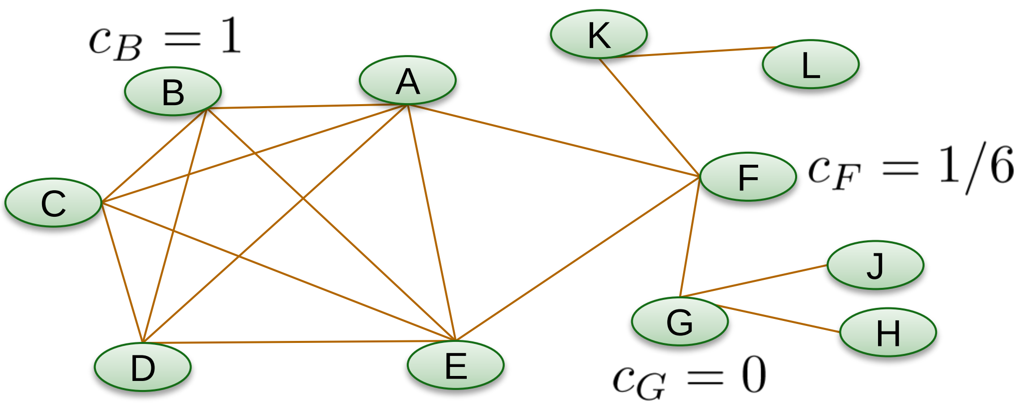
$$G = (V, E)$$
$$n = |V| = \text{number of nodes}$$
$$m = |E| = \text{number of edges}$$
$$V_d = \{i \mid d_i = d\} = \text{set of nodes of degree } d$$
$$n_d = |V_d| = \text{number of nodes of degree } d$$

Barabasi & Albert, Science, 1999: *"A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution"*



Clustering Coeff. Measures Cohesion

The ***clustering coefficient** measures the rate of wedge closure*.

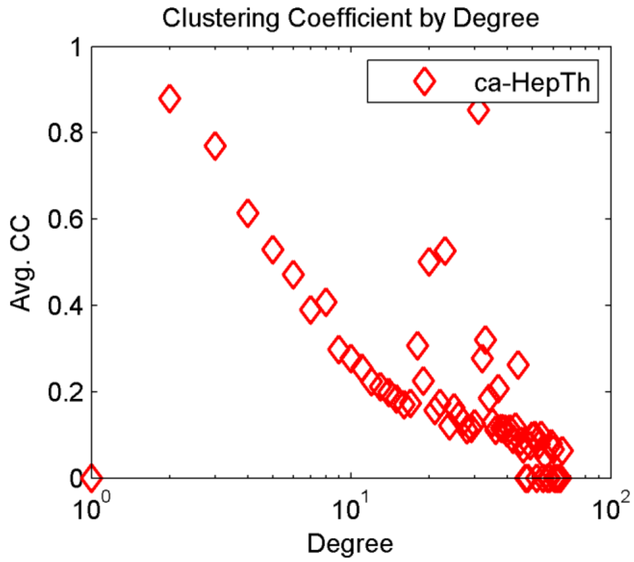


$$c_i = \frac{\#\text{ closed wedges centered at node } i}{\#\text{ wedges centered at node } i}$$

$$c_d = \frac{1}{n_d} \sum_{i \in V_d} c_i = \text{average for nodes of degree } d$$

$$c = \frac{3 \times \#\text{ triangles in graph}}{\#\text{ wedges in graph}}$$

In social networks, the clustering coefficients decrease smoothly as the degree increases. High degree nodes generally have little social cohesion.

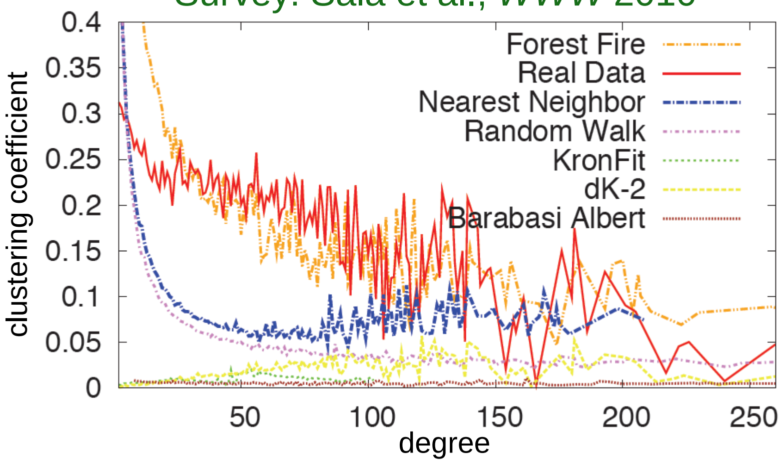


Current State-of-the-Art Falls Short

Story-Driven Models

- Examples
  - Preferential Attachment
    - Barabasi & Albert, *Science* 1999
  - Forest Fire
    - Leskovec, Kleinberg, Faloutsos, *KDD* 2005
  - Random Walk
    - Vazquez, *Phys. Rev. E* 2003
- Pros & Cons
  - Poor fits to real data
  - Expensive to calibrate to real data
  - **Do not scale** – inherently sequential

Survey: Sala et al., *WWW* 2010

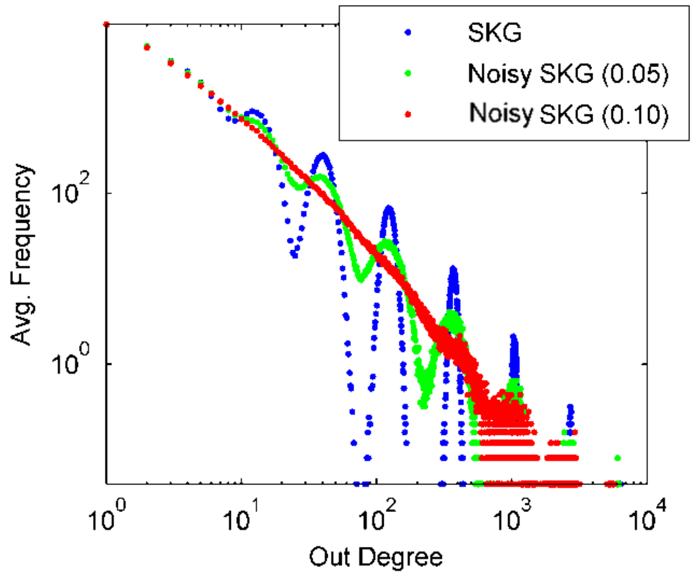


Structure-Driven Models

- Examples
  - CL: Chung-Lu; aka Configuration Model, Weighted Erdös-Rényi
    - *PNAS* 2002
  - SKG: Stochastic Kronecker Graphs; R-MAT is a special case
    - Leskovec et al., *JMLR* 2010; Chakrabarti, Zhan, Faloutsos, *SDM* 2004
    - *Graph 500 Generator!*
- Pros & Cons
  - **Do not capture clustering coefficients**
  - SKG expensive to calibrate
  - Scales – generation cost O(m log n)
  - CL & SKG very similar in behavior
    - Pinar, Seshadhri, Kolda, *SDM* 2012

Stochastic Kronecker Graph (SKG) as Graph 500 Generator

- Pros
  - Only 5 parameters
    - 2x2 generator matrix (sums to 1)
    - $n = 2^L$ = # nodes
    - $m = 16n$ = # edges
  - O(m log n) generation cost
  - Edge generation fully parallelizable
    - Except de-duplication
- Cons
  - Oscillations in degree distribution (fixed by adding special noise)
  - Limited degree distribution (noisy version is lognormal)
  - Half the nodes are isolated!
  - Tiny clustering coefficients!



L	Isolated	d_{avg}
26	51%	32
29	57%	37
32	62%	41
36	67%	49
39	71%	55
42	74%	62

Seshadhri, Pinar, Kolda, *Journal of the ACM,* April 2012

The Physics of Graphs

Random graph:

(1) Formed according to CL Model

(2) "High" clustering coefficient

$\Downarrow$

Thm: Must contain a "substantive" subgraph that is a **dense Erdös-Rényi graph**.

$\Downarrow$

A heavy-tailed network with a high clustering coefficient contains many Erdös-Rényi **affinity blocks**. (The distribution of the block sizes is also heavy tailed.)

CL Model

$$G = (V, E) \quad \{d_i\}_{i \in V} \text{ (prescribed)}$$

$$\text{Prob}\left((i,j) \in E \mid i, j, \in V\right) \propto d_i \cdot d_j$$

Global Clustering Coefficient

$$c = \frac{3 \times \#\text{ triangles in graph}}{\#\text{ wedges in graph}}$$

Dense Erdös-Rényi Subgraph

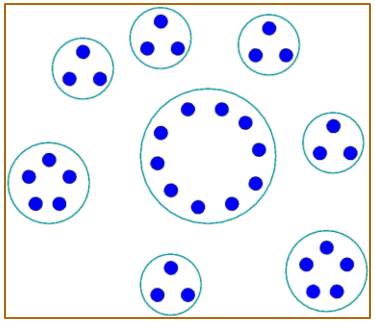
$$\bar{V} \subset V, \bar{E} \subset E$$

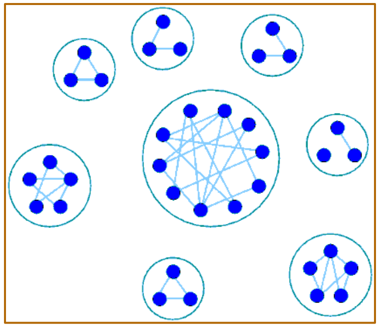
$$\text{Prob}\left((i,j) \in \bar{E} \mid i, j \in \bar{V}\right) \propto \text{constant}$$

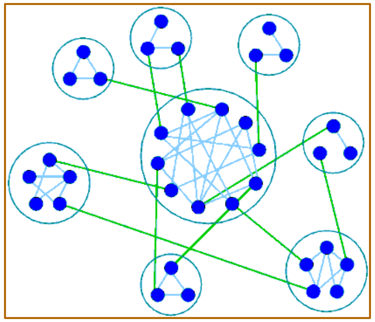
Basic measurements lead to inferences about larger structures (communities) that are consistent with literature.

Seshadri, Kolda, Pinar, *Phys. Rev. E,* 2012

BTER: Block Two-Level Erdös-Rényi







Preprocessing

- Create affinity blocks of nodes with (nearly) same degree, determined by **degree distribution**
- Connectivity per block based on **clustering coefficient**
- For each node, compute desired
  - within-block degree
  - excess degree

Phase 1

- Erdös-Rényi graphs in each block
- Need to insert extra links to insure enough *unique* links per block

$$w_b = \binom{n_b}{2} \ln \left(\frac{1}{1-\rho_b} \right)$$

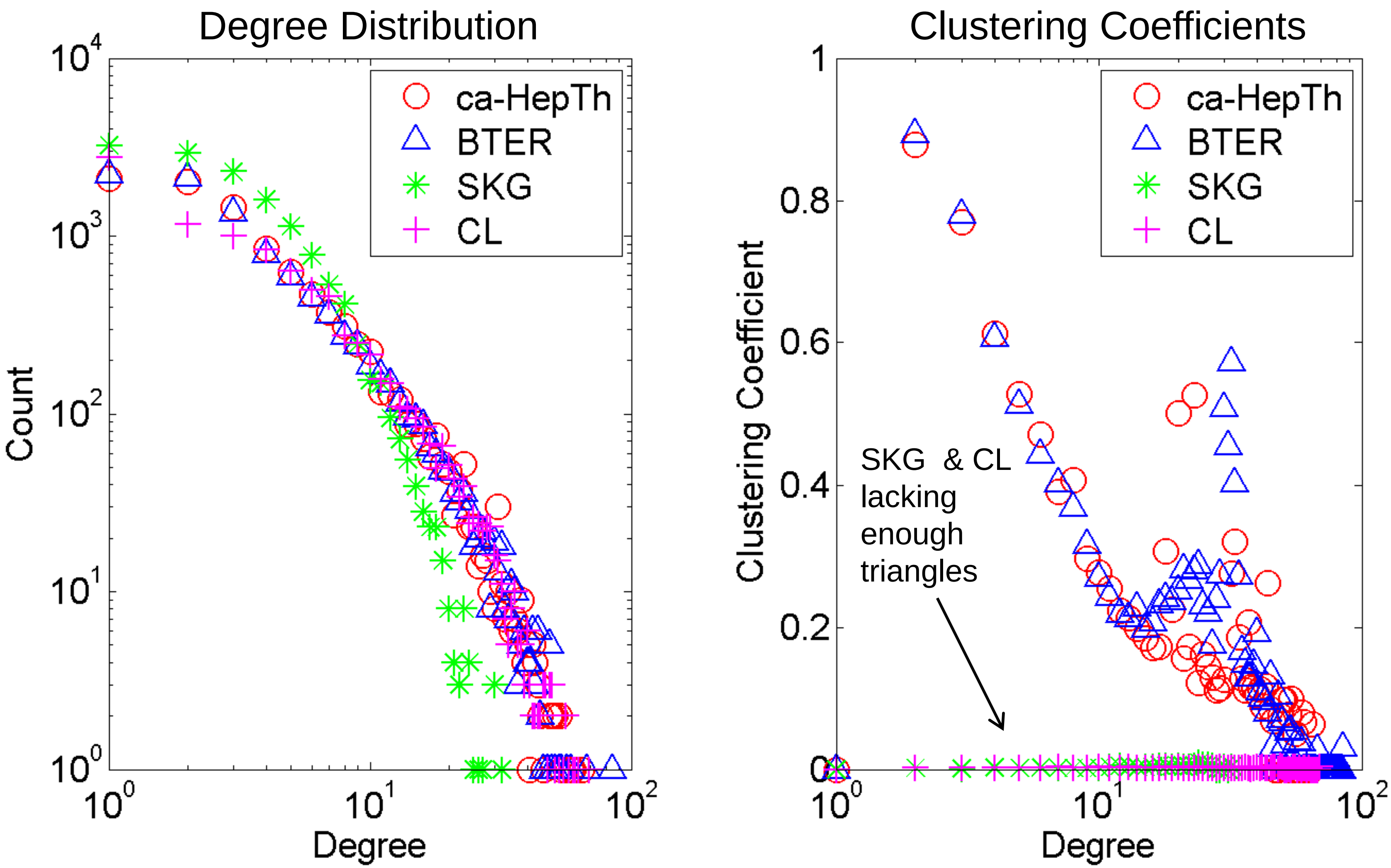
Phase 2

- CL model on excess degree (a sort of weighted Erdös-Rényi)
- Creates connections across blocks

Occurring independently

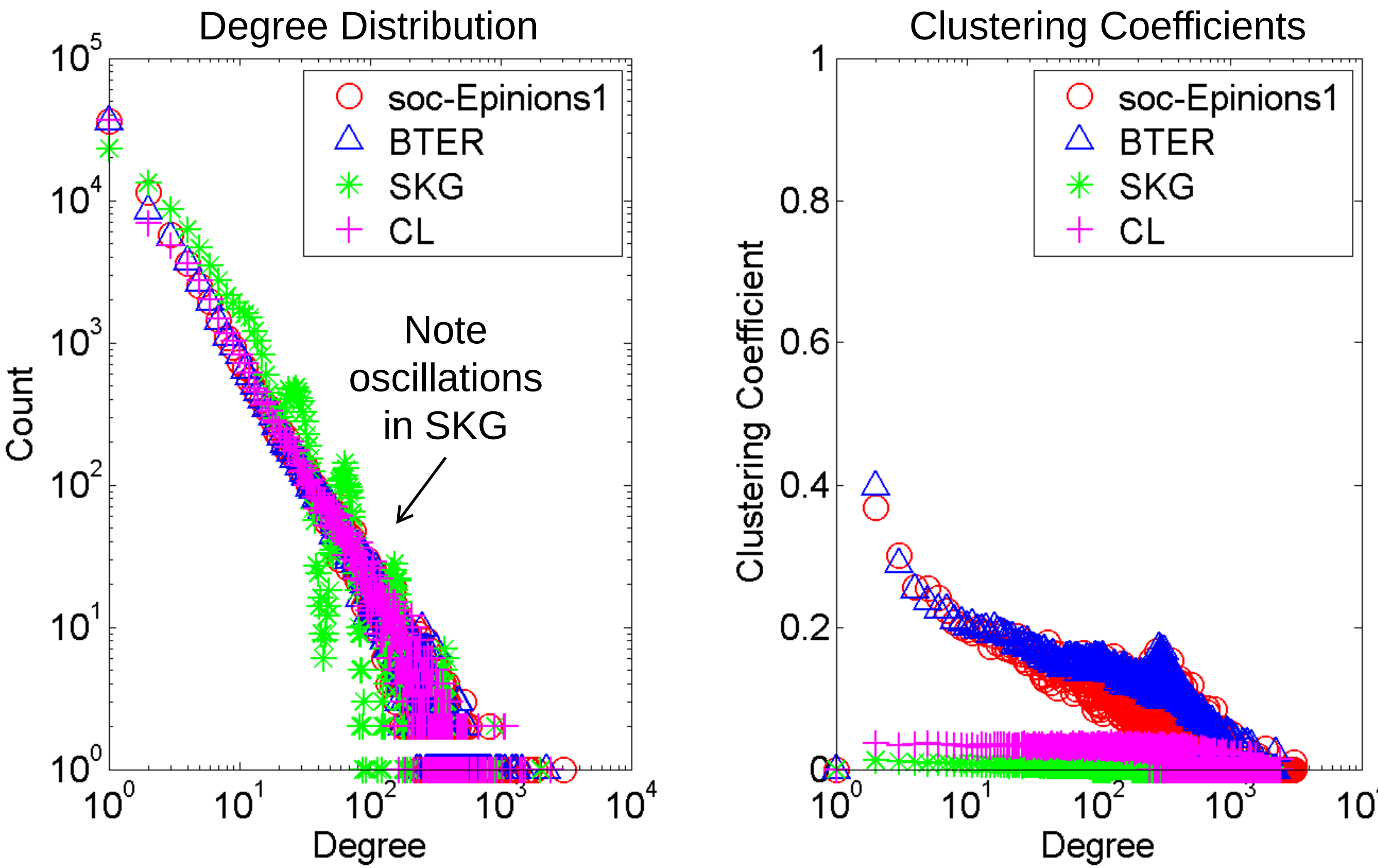
Seshadhri, Kolda, Pinar, *Phys. Rev. E,* 2012
Kolda, Pinar, Plantenga,, Seshadhri, arXiv:1302.6636, Feb. 2013

BTER vs. SKG: Co-authorship



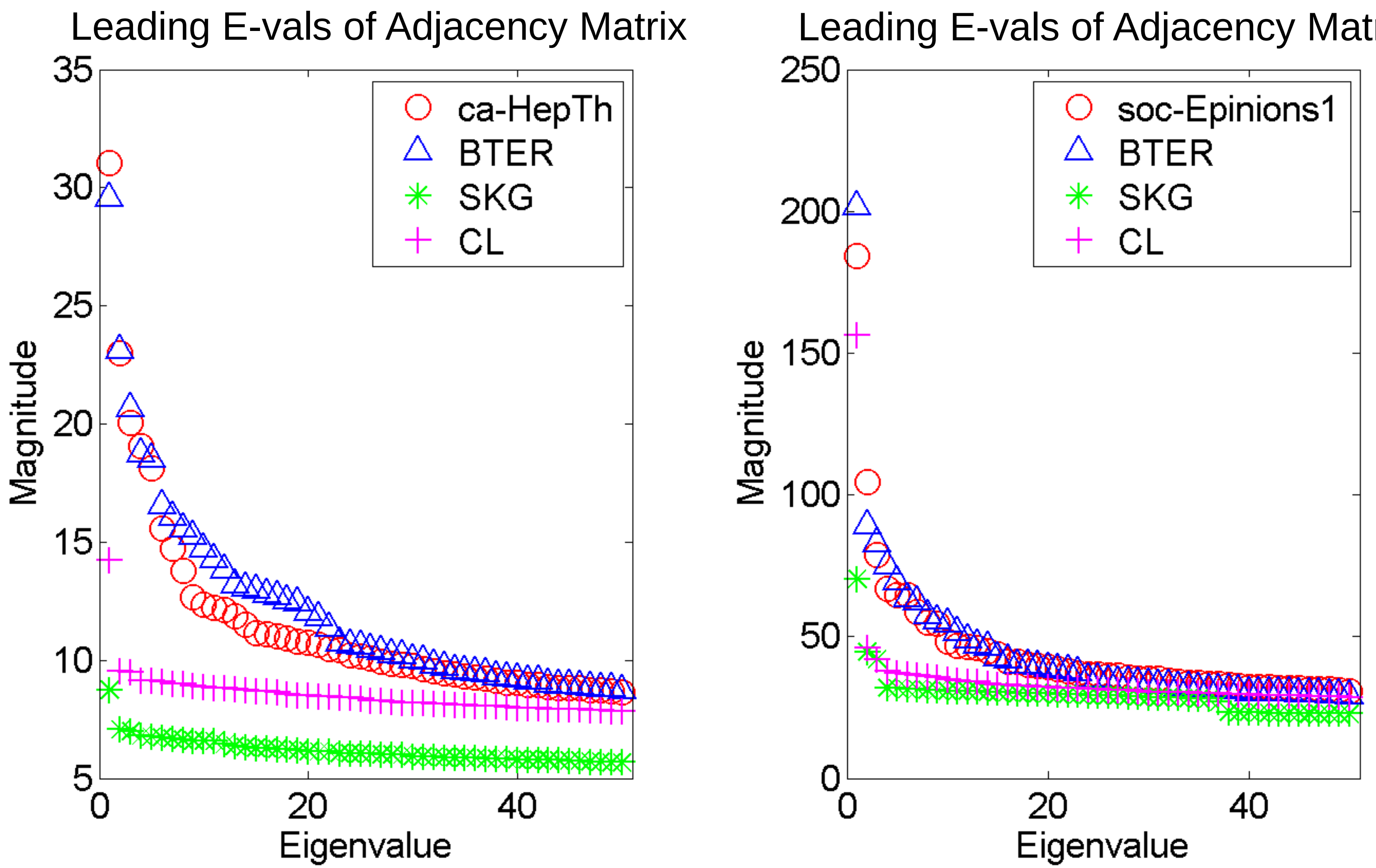
SKG parameters from Leskovec et al., *JMLR*, 2010

BTER vs. SKG: Social Website

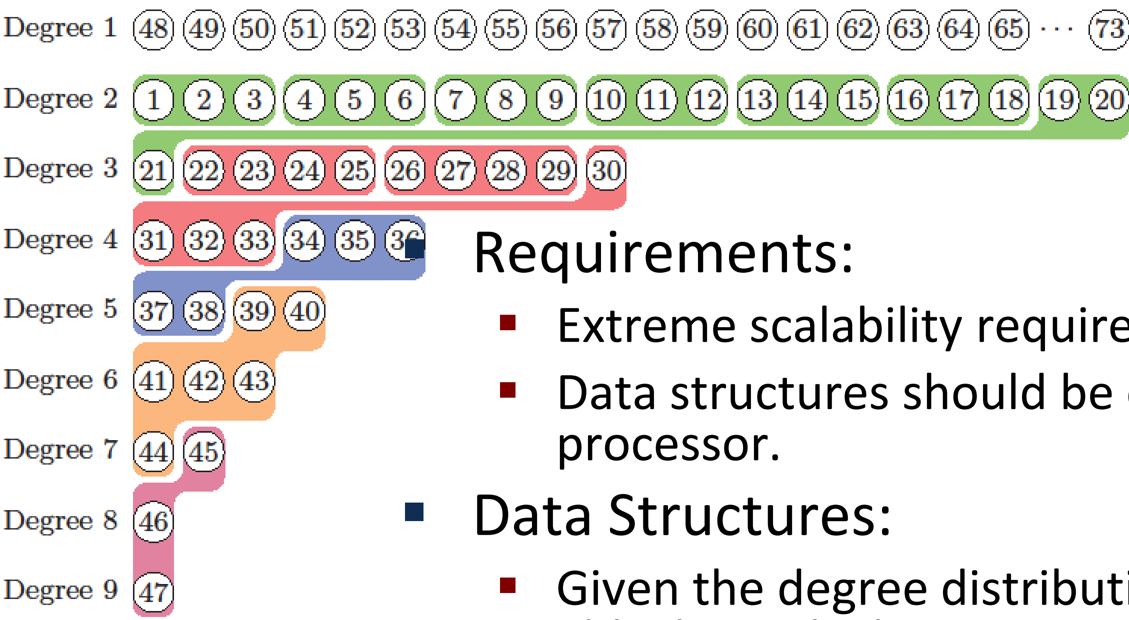


SKG parameters from Leskovec et al., *JMLR,* 2010

Community Structure of BTER Improves Eigenvalue Fit



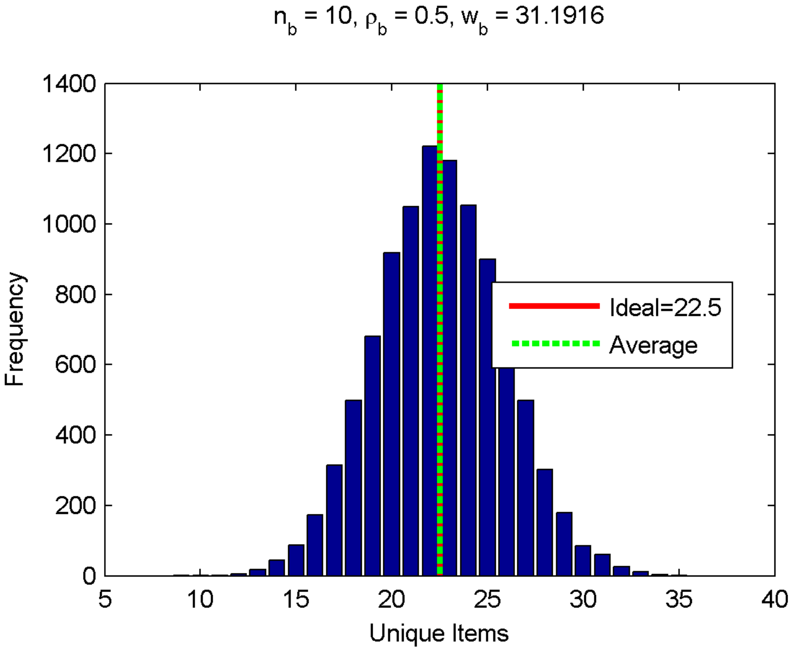
Making BTER Scalable



- Requirements:
  - Extreme scalability requires independent edge insertion.
  - Data structures should be o(|V|) to be duplicated at each processor.
- Data Structures:
  - Given the degree distribution, compute <block size, #blocks>, which requires O(dmax) memory.
  - Given the clustering coefficients, compute the number of edges per block, hence the phase 1 degrees.
  - Given Phase 1 degrees, we can compute residual (Phase 2) degrees.
- Challenge: Adjust for repetitions

Adjusting for repeated edges

- Parallel edge insertion leads to multiple edges.
- This is negligible if edge probabilities are small.
  - This is the case for SKG, CL
  - But not for BTER.
- BTER has dense blocks, hence many repeats.



- We had extra edges to guarantee the number of unique items is as expected.
  - Coupon collector problem.

$$w_b = \binom{n_b}{2} \ln(1/(1-\rho_b)).$$

BTER for BIG Networks

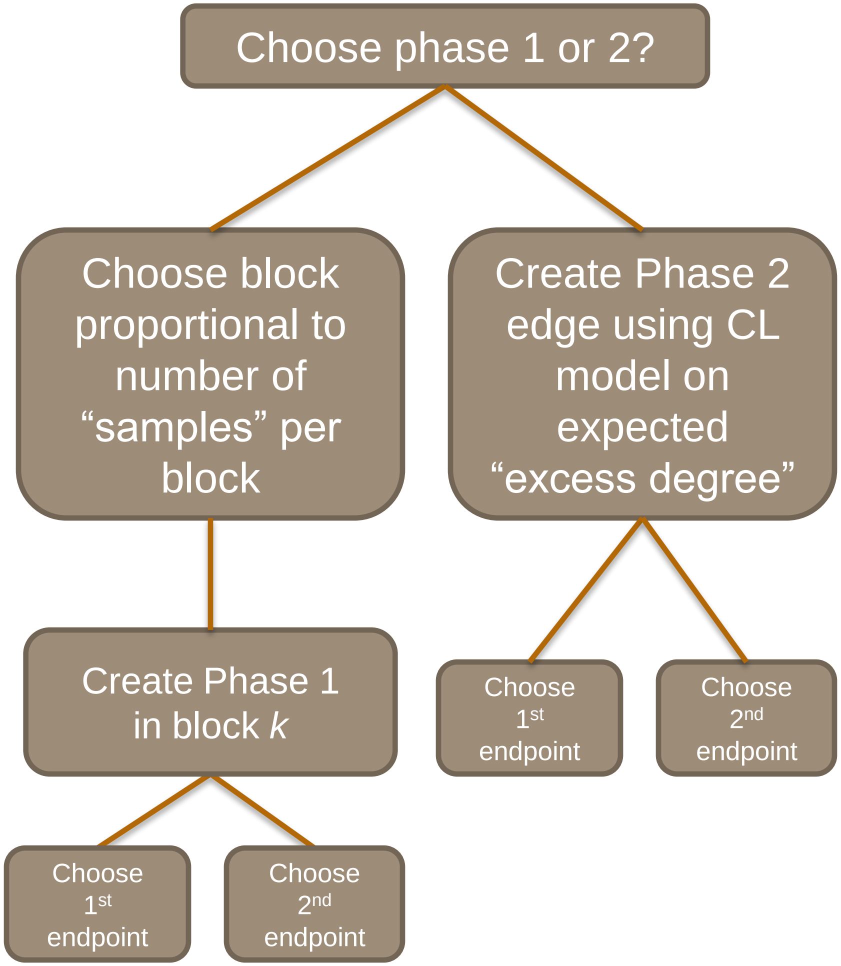
- Need **degree distribution**
  - Calculate explicitly for real data (d_{max} parameters)
  - Can provide a formula, e.g., power law (1-2 parameters)

$$n_d = |V_d| = \text{number of nodes of degree } d$$

- Need to specify **clustering coefficients** per degree
  - Calculate explicitly for real data (d_{max} parameters)
  - Can provide an arbitrary formula (1-2 parameters)

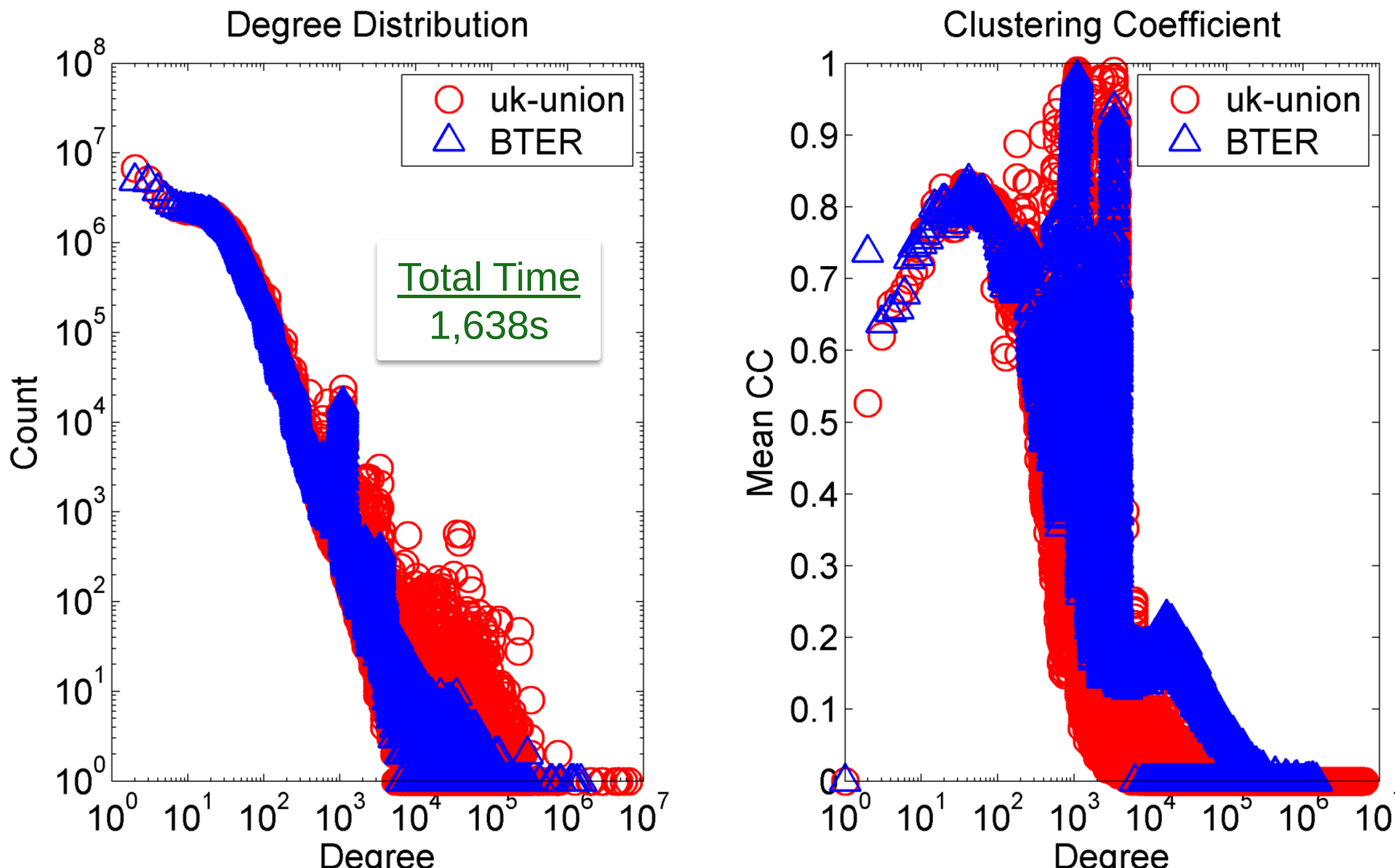
$$c_d = \frac{\#\text{ closed wedges centered at nodes of degree } d}{\#\text{ wedges centered at node of degree } d}$$

- Cost per edge is $O(\log d_{max})$
- Edge generation is parallelizable
- Requires de-duplication (like SKG)



Kolda, Pinar, Plantenga, Seshadhri, arXiv:1302.6636, 2013

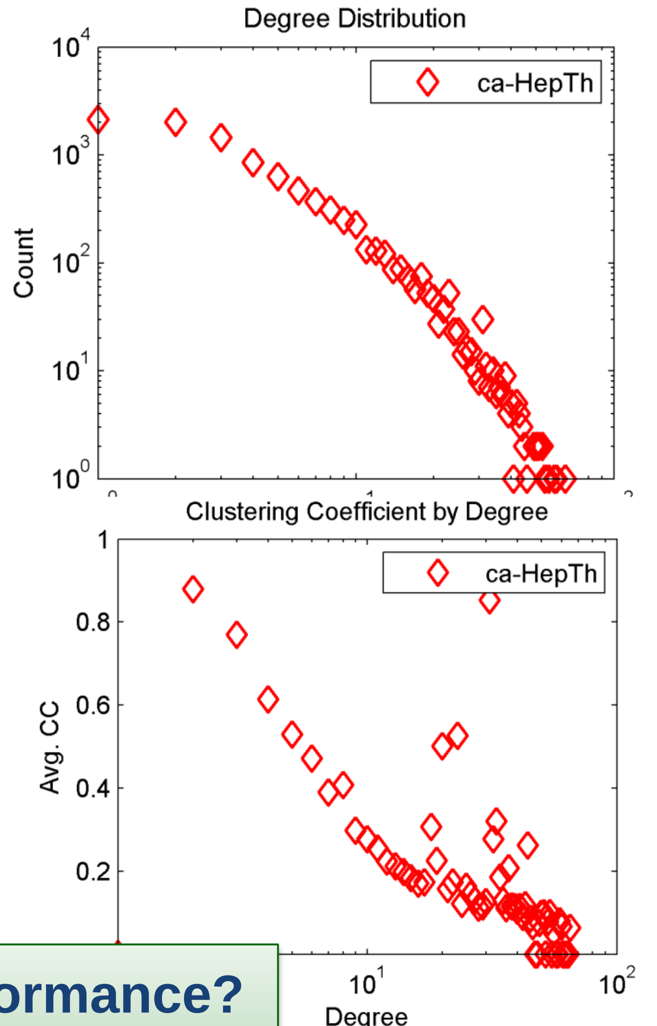
BTER Hadoop Results: uk-union (4.6B edges)



Kolda, Pinar, Plantenga, Seshadhri, arXiv:1302.6636, 2013

Choosing BTER parameters for benchmarking

- BTER can regenerate graphs with specifed parameters.
  - Parameters are provided by an existing graph.
  - Benchmarking requires non-existent graphs.
- Parameters for benchmarking
  - Should be realistic
  - Should be tunable for performance analysis.
- We want to control
  - #vertices, #edges, maximum-degree, cohesiveness.
- Challenges:
  - What is a good degree distribution?
  - What is a good clustering coefficient curve?

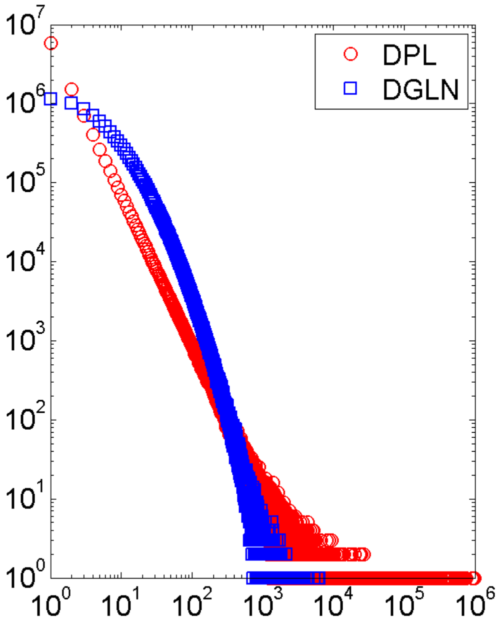


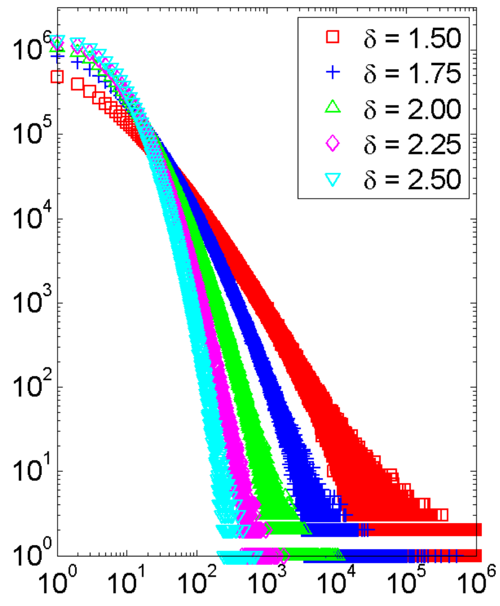
Discussion topic: What else does affect performance? What else would you like to control?

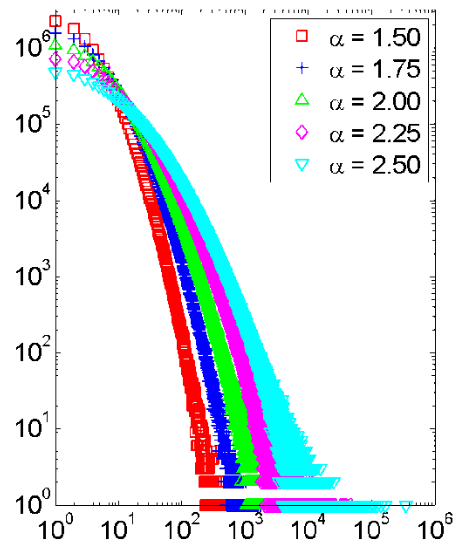
What is a good degree distribution model?

- **Myth:** Real graphs have power-law degree distribution.
- **Common-wisdom:** Not really, but they are okay.
- **Reality:** Power-law graphs are not good for benchmarking.
- **Proposed:** generalized log normal

$$n_d \propto \exp\left[-\left(\frac{\log d}{\alpha}\right)^{\delta}\right]$$



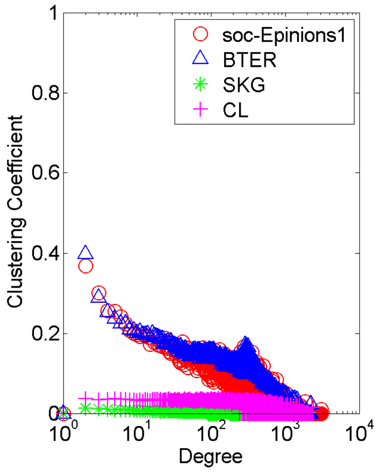


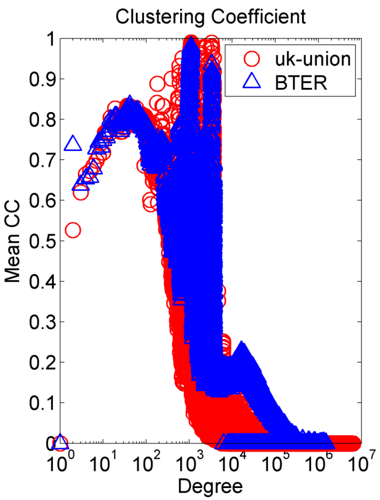


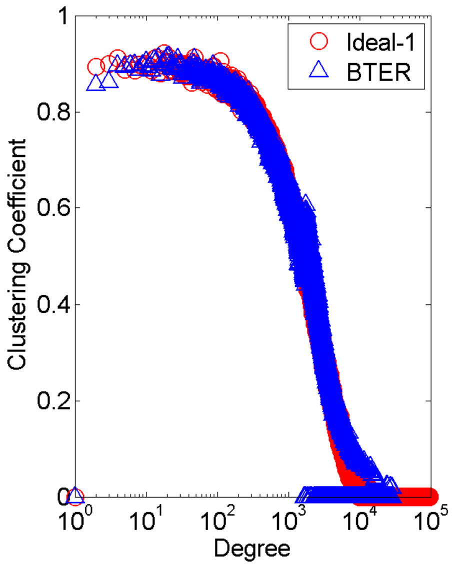
What is a good clustering coefficient curve?

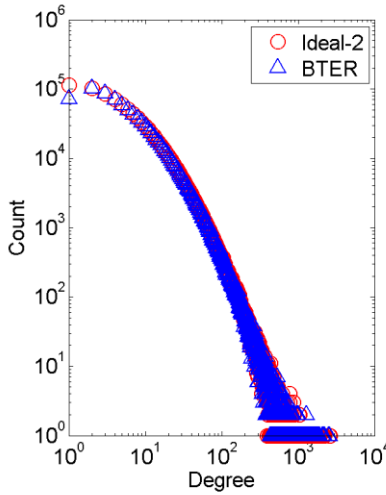
- Clustering coefficient curves come in all sorts and shapes.
- Difficult to see a pattern
- Proposed method:
  - Can control the maximum and the global clustering coefficient.

$$\bar{c}_d = c_{\max} \exp(-(d-1)\cdot\xi)$$









Conclusions and Future Work

- Generators are crucial for benchmarking (scalability, sensitivity).
  - Current generators are and future generators will be imperfect.
  - One has to understand the underlying graphs before drawing conclusions.
- Block Two-level Erdos Renyi model improves the state of the art.
  - is based on theoretical analysis.
  - matches degree distribution and clustering coefficients.
  - allows scalable graph generation.
- For benchmarking,
  - Generalized lognormal distributions provide realistic and realizable degree distributions.
  - We proposed reasonable clustering coefficient distributions.
- Codes are available: http://www.sandia.gov/~tgkolda/feastpack

References

- SKG Analysis: C. Seshadhri, A. Pinar and T. G. Kolda. ***An In-Depth Analysis of Stochastic Kronecker Graphs***, Journal of the ACM, Apr 2013 (preprint: arXiv:1102.5046)
- Wedge Sampling: C. Seshadhri, A. Pinar and T. G. Kolda, ***Triadic Measures on Graphs: The Power of Wedge Sampling***, Proc. SIAM Intl. Conf. on Data Mining (SDM'13), Apr 2013 (preprint: arXiv:1202.5230)
- Wedge Sampling MapReduce: T. G. Kolda, T. Plantenga, C. Task, A. Pinar, and C. Seshadhri, ***Counting Triangles in Massive Graphs with MapReduce***, arXiv:1301.5887, Jan 2013
- BTER Model: C. Seshadhri, T. G. Kolda and A. Pinar. ***Community structure and scale-free collections of Erdös-Rényi graphs***, Physical Review E 85(5):056109, May 2012, doi:10.1103/PhysRevE.85.056109
- Scalable BTER Model: T. G. Kolda, A. Pinar, T. D. Plantenga, and C. Seshadhri, ***A Scalable Generative Graph Model with Community Structure,*** arXiv:1302.6636, Feb 2013
- Directed Graph Models: N. Durak, T. G. Kolda, A. Pinar, and C. Seshadhri, ***A scalable directed graph model with reciprocal edges***, IEEE Network Science Workshop, May 2013 (preprint: arXiv:1210.5288)
- Directed Triangles: C. Seshadhri, A. Pinar, N. Durak, T. G. Kolda, ***The Importance of Directed Triangles with Reciprocity: Algorithms and Patterns***, arXiv:1302.6220, Feb 2013

- *For copies or information about job openings: Ali Pinar apinar@sandia.gov*

THE END