Analyzing, Modeling, and Generating BIG Networks

Tamara G. Kolda, Ali Pinar, C. “Sesh” Seshadhri,
Todd Plantenga, Nurcan Durak, Christine Task

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

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

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.

Networks are Everywhere!

- Ubiquitous, diverse
  - Computer traffic
  - Social networks
  - Biological signaling
  - Communications
  - Physical interaction
  - Citation
  - Purchasing
  - Etc.

Sample Co-Authorship for Gene Golub Papers

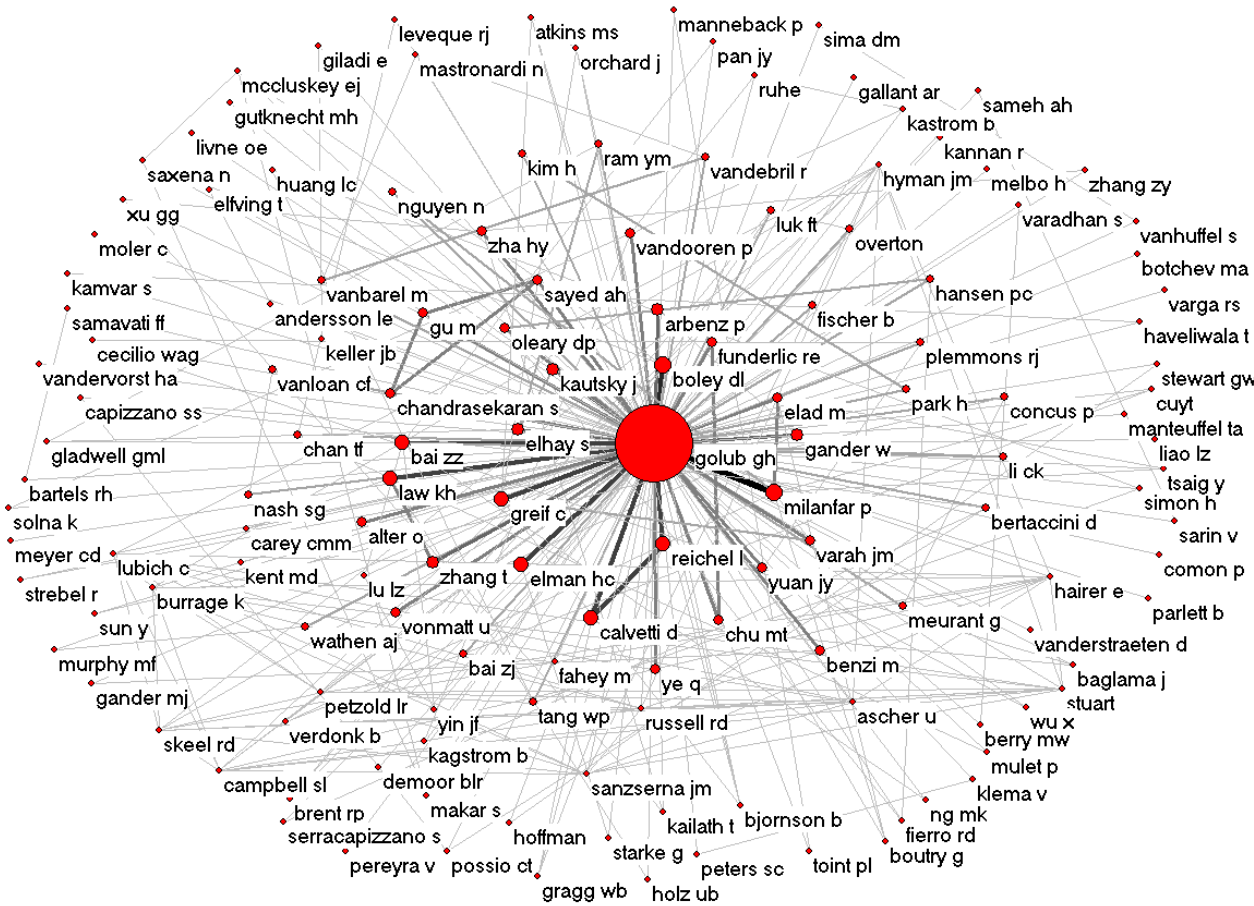


Image Credit: KU Leuven
http://www.esat.kuleuven.be/scd/golub/network.php

Networks are Growing

- Regular networks
  - O(100) nodes, O(10k) edges
  - O(KB)
  - Analysis trivial
- big networks
  - O(10k) nodes, O(1M) edges
  - O(1MB), fits on a laptop
  - Analysis easy
- Big networks
  - O(1M) nodes, O(100M) edges
  - O(1GB), fits on a workstation
  - Analysis expensive
- BIG networks
  - O(100M) nodes, O(1B) edges
  - O(10-100GB), requires cluster or supercomputer
  - Analysis really expensive

Many past studies have drawn conclusions from relatively small networks.

Example "big" network: Co-authorship Network from ArXiv High Energy Physics 9k nodes, 25k edges

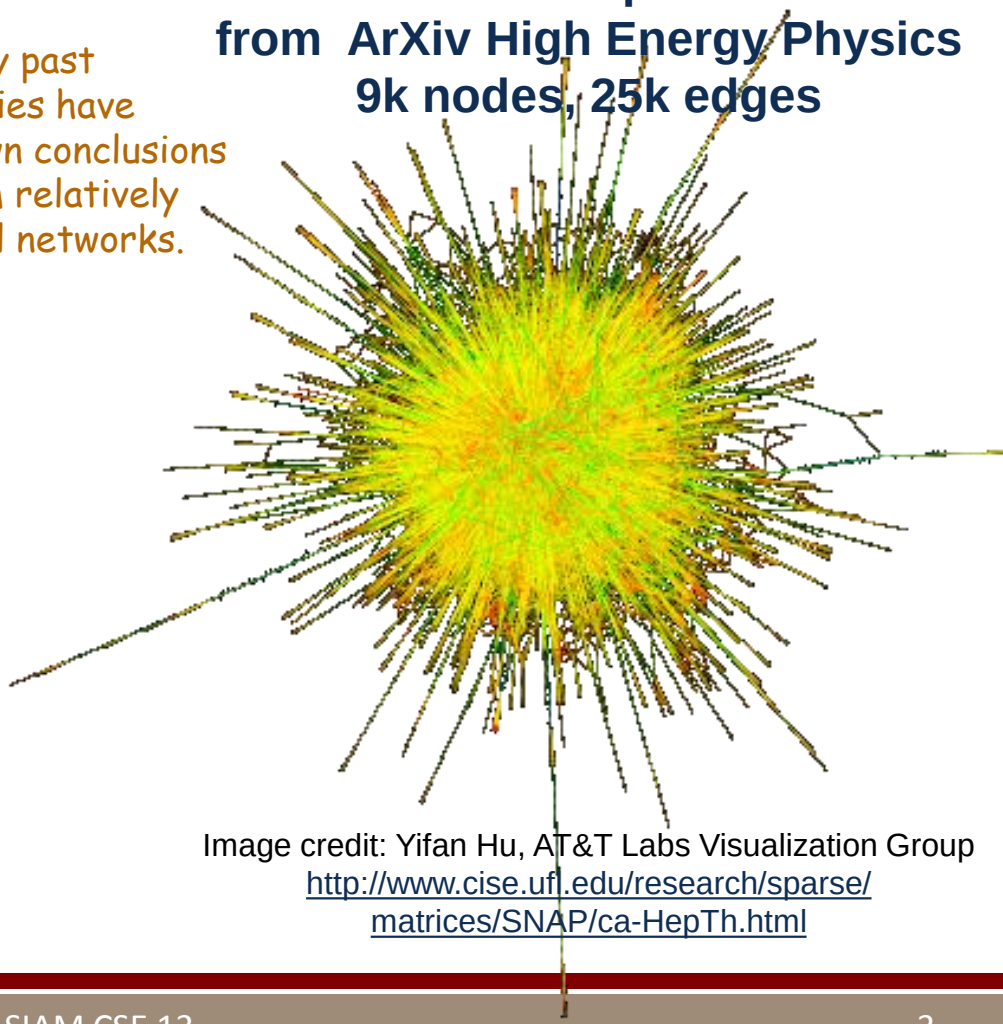


Image credit: Yifan Hu, AT&T Labs Visualization Group
http://www.cise.ufl.edu/research/sparse/matrices/SNAP/ca-HepTh.html

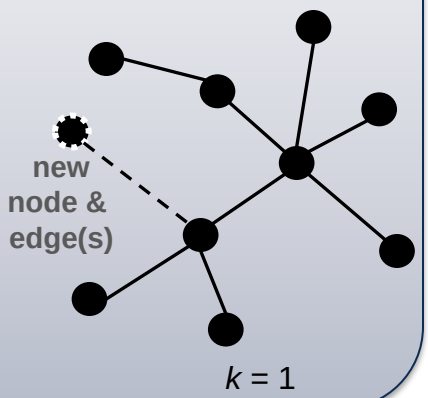
A Good Network Model…

- 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)

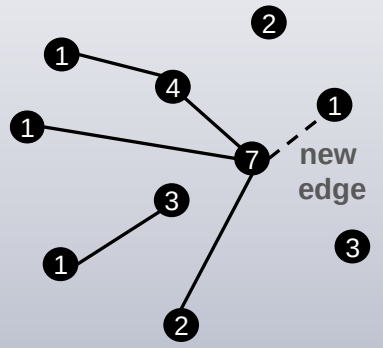
- New nodes joins graph one at a time, in sequence
- Each new node chooses *k* new neighbors, according to degree
- Node degrees updated after each addition – *Rich get richer!*



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



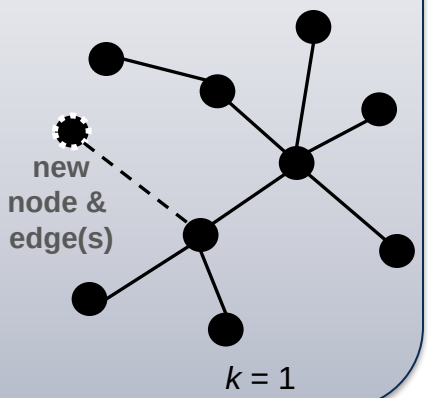
A Good Network Model…

- 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)

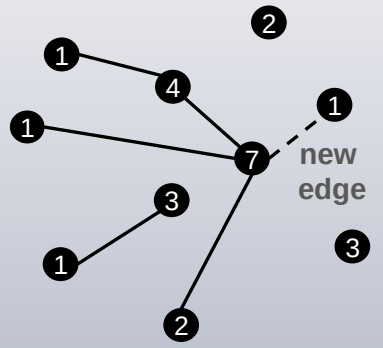
- New nodes joins graph one at a time, in sequence
- Each new node chooses *k* new neighbors, according to degree
- Node degrees updated after each addition – *Rich get richer!*



Structure-driven models

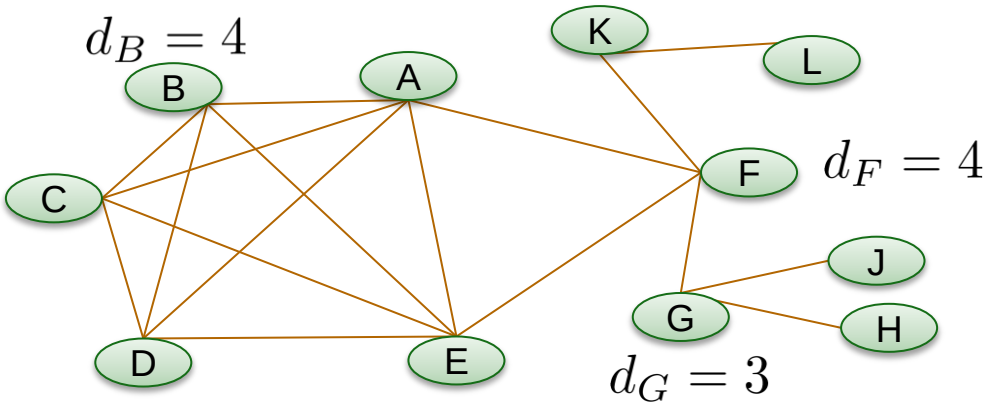
Example: CL (aka Configuration) (Chung & Lu, PNAS, 2002)

- 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.



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

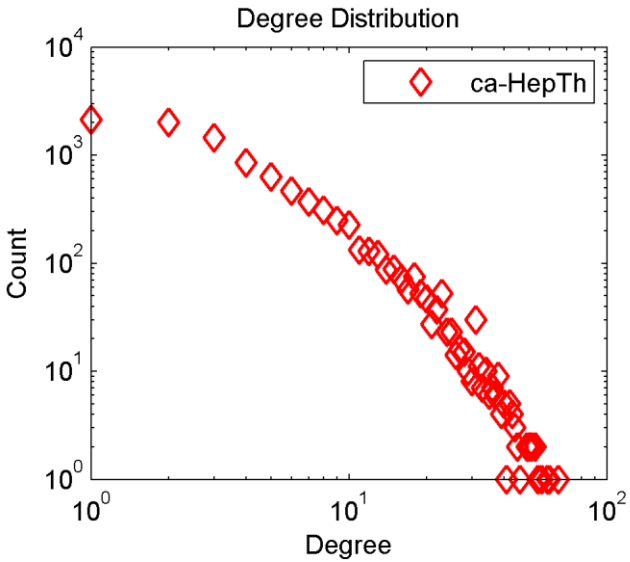
$$G = (V, E)$$

$$n = |V| = \text{number of nodes}$$

$$m = |E| = \text{number of edges}$$

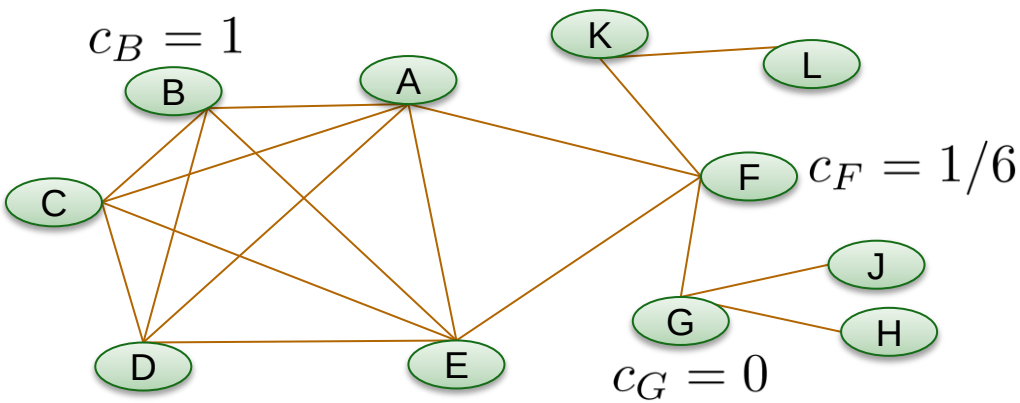
$$V_d = \{i \mid d_i = d\} = \text{set of nodes of degree } d$$

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



Clustering Coeff. Measures Cohesion

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

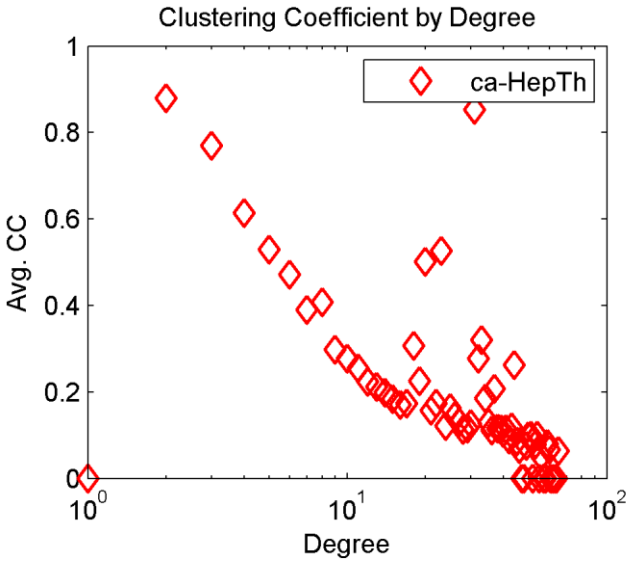


In social networks, the clustering coefficients decrease smoothly as the degree increases. High degree nodes generally have little social cohesion.

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

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

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

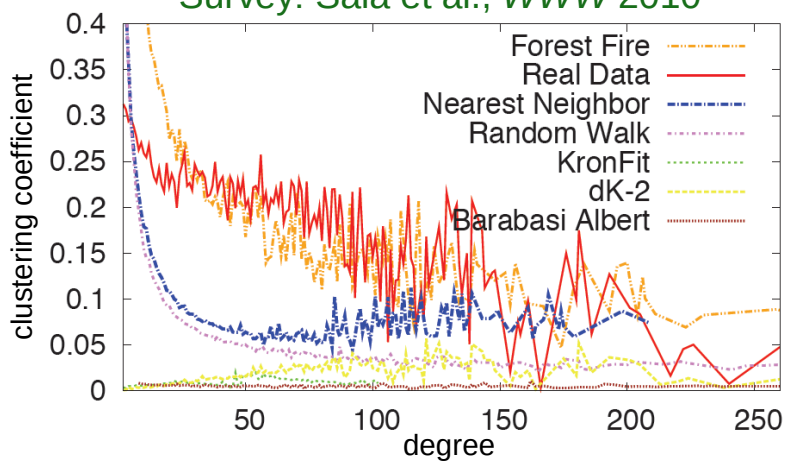


Current State-of-the-Art Falls Short

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

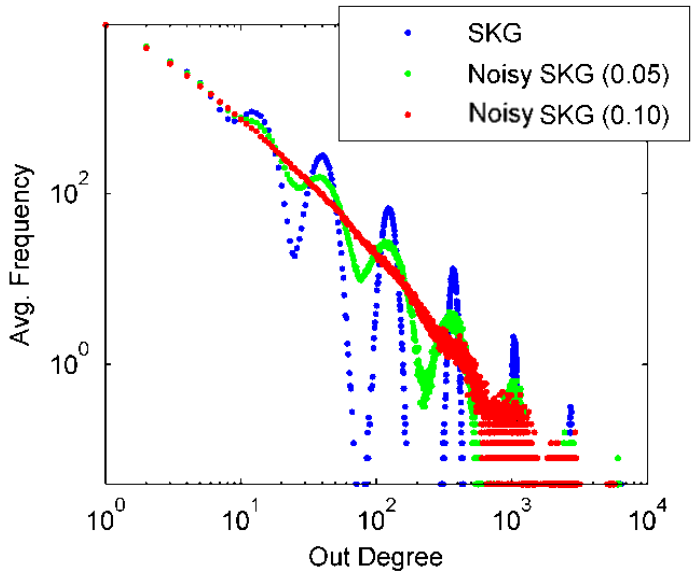


Structure-Driven Models

- 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, *JACM* 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

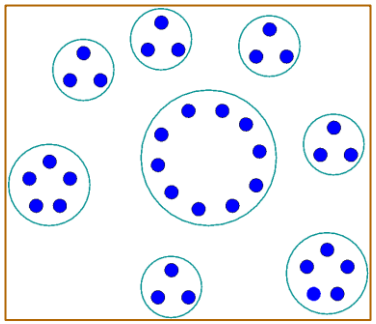
$$\bar{V} \subset V, \bar{E} \subset E$$
$$\text{Prob}\left((i,j) \in \bar{E} \mid i, j \in \bar{V}\right) \propto \text{constant}$$

New Story

Node joins a graph, but doesn't follow preferential attachment.
(1) First, joins an "affinity block" and makes within-block connections
(2) Second, makes outside-block connections (preferential attachment)

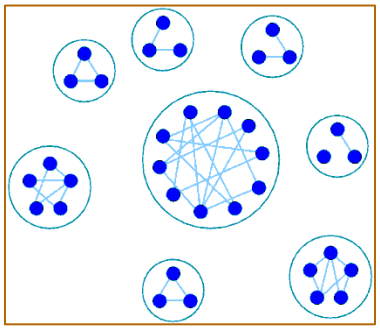
Seshadhri, Kolda, Pinar, *Phys. Rev. E* 2012

BTER: Block Two-Level Erdös-Rényi



Preprocessing

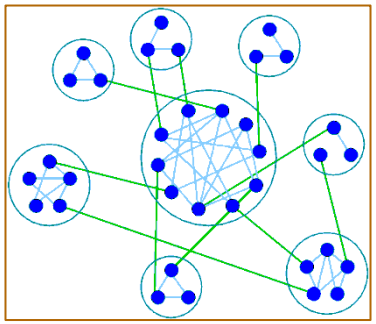
- Create affinity blocks of nodes with (nearly) same degree, determined by **degree distribution**
- Connectivity per block based on **clustering coefficient**
- For each node, compute desired
  - within-block degree
  - excess degree



Phase 1

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

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



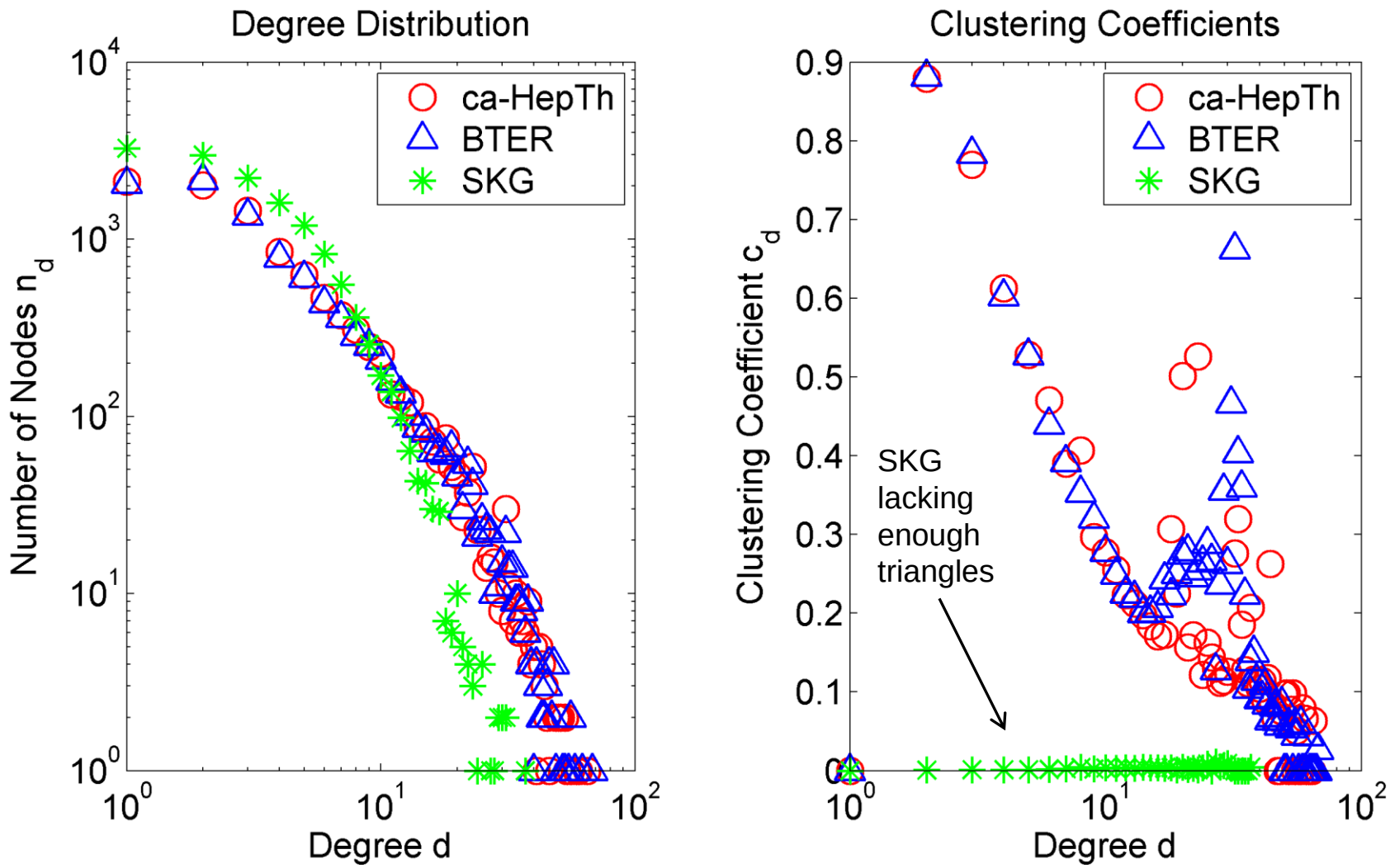
Phase 2

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

Occurring independently

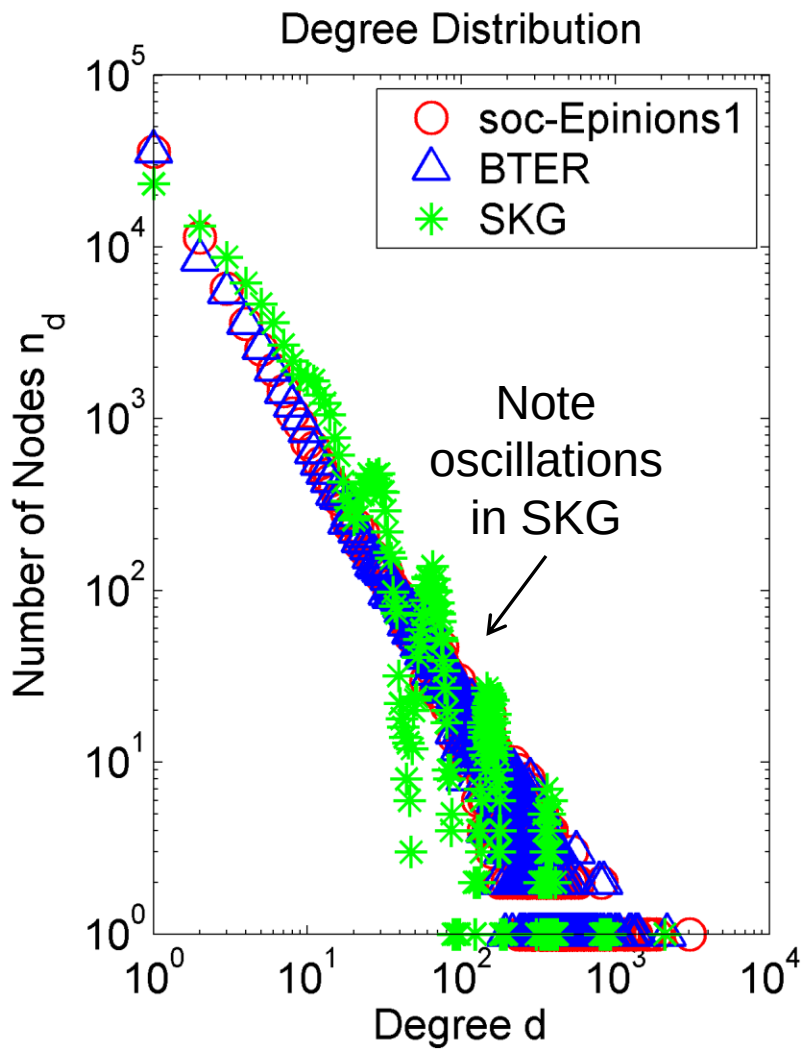
Seshadhri, Kolda, Pinar, *Phys. Rev. E* 2012
Kolda, Plantenga, Pinar, Seshadhri, arXiv 2013

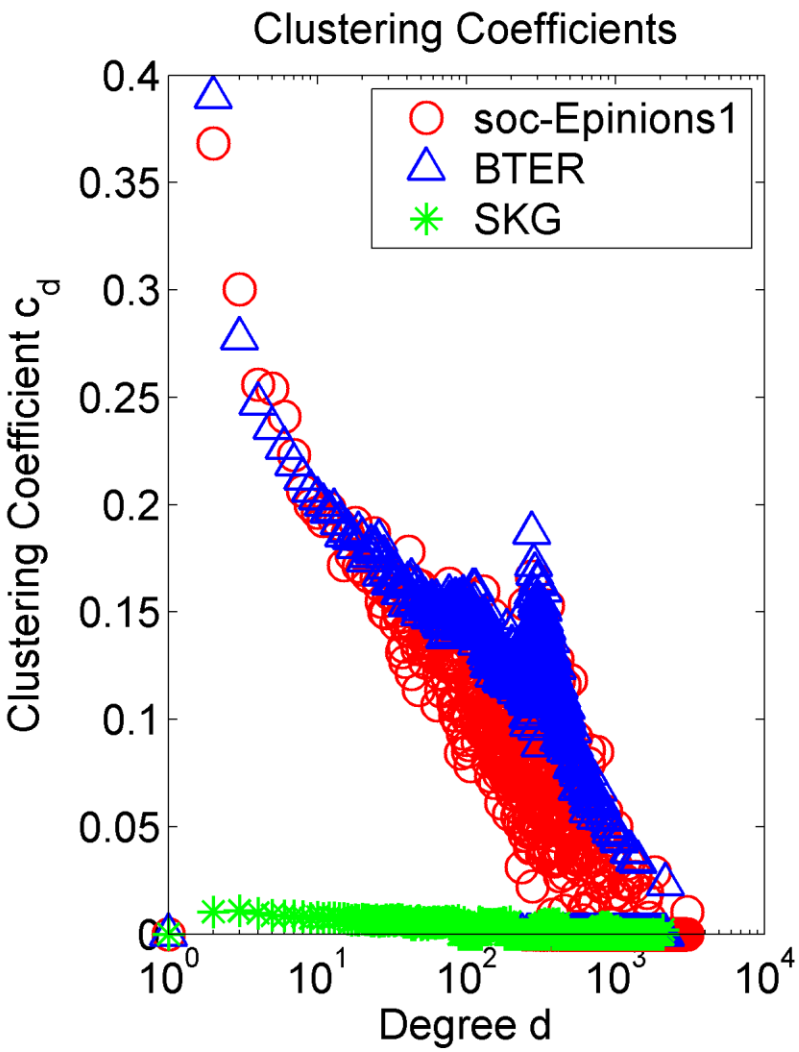
BTER vs. SKG: Co-authorship



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

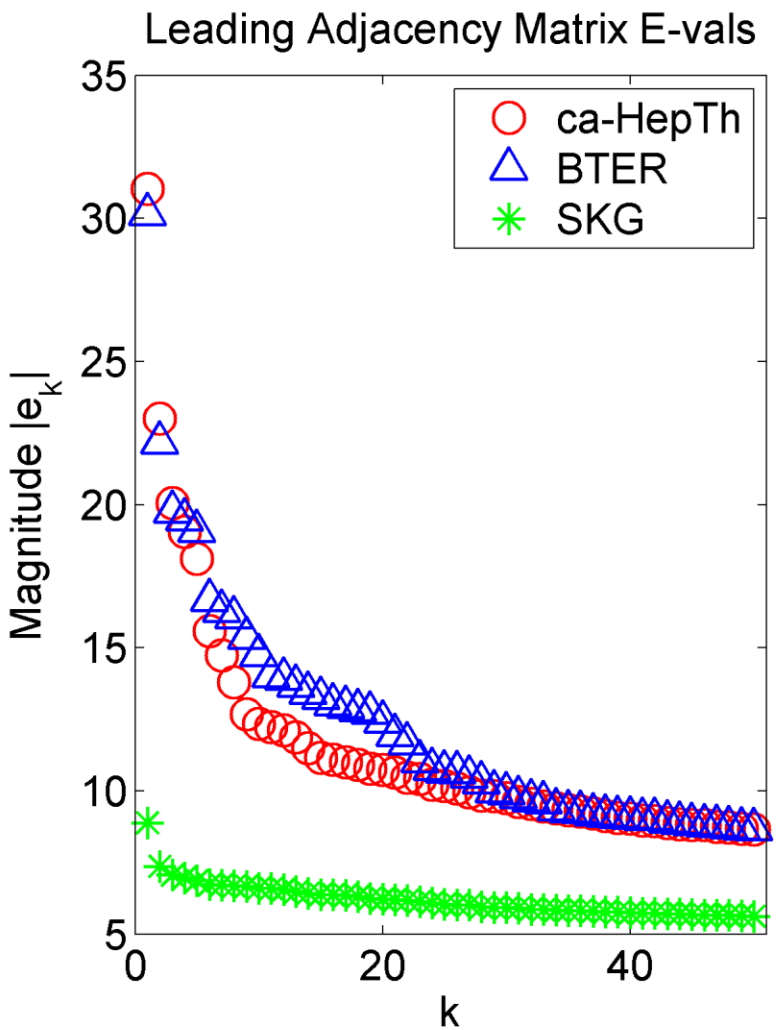
BTER vs. SKG: Social Website

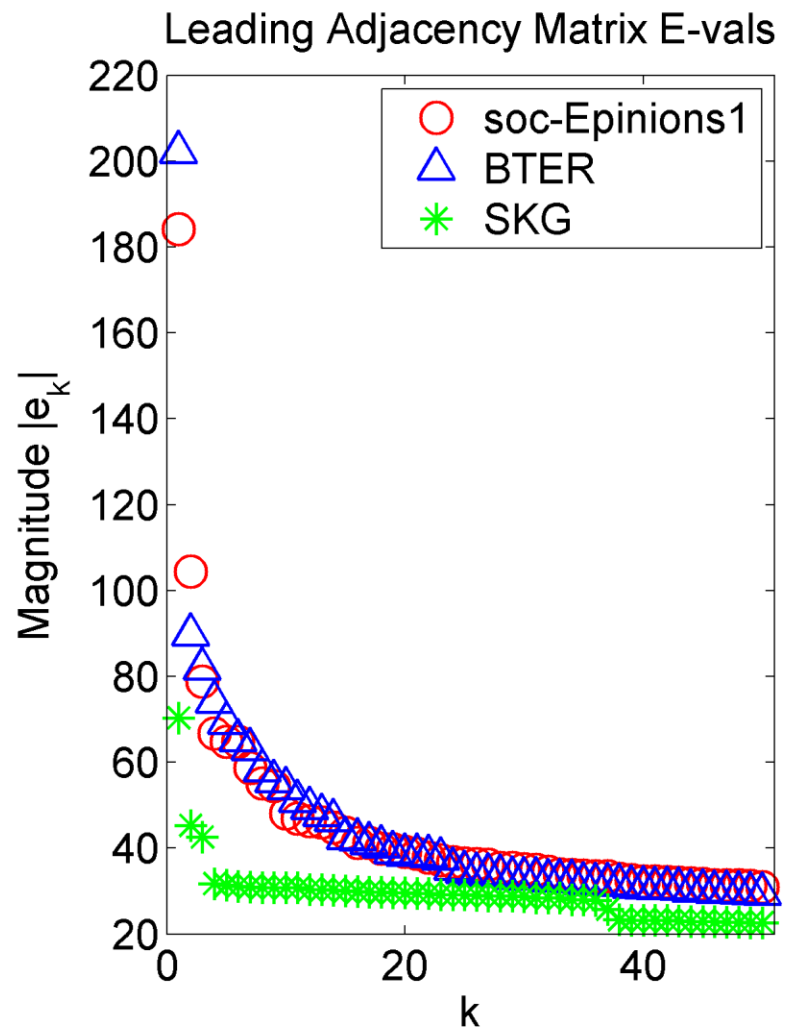




SKG parameters from Leskovec et al., *JMLR* 2010

Community Structure of BTER Improves Eigenvalue Fit





BTER for BIG Networks

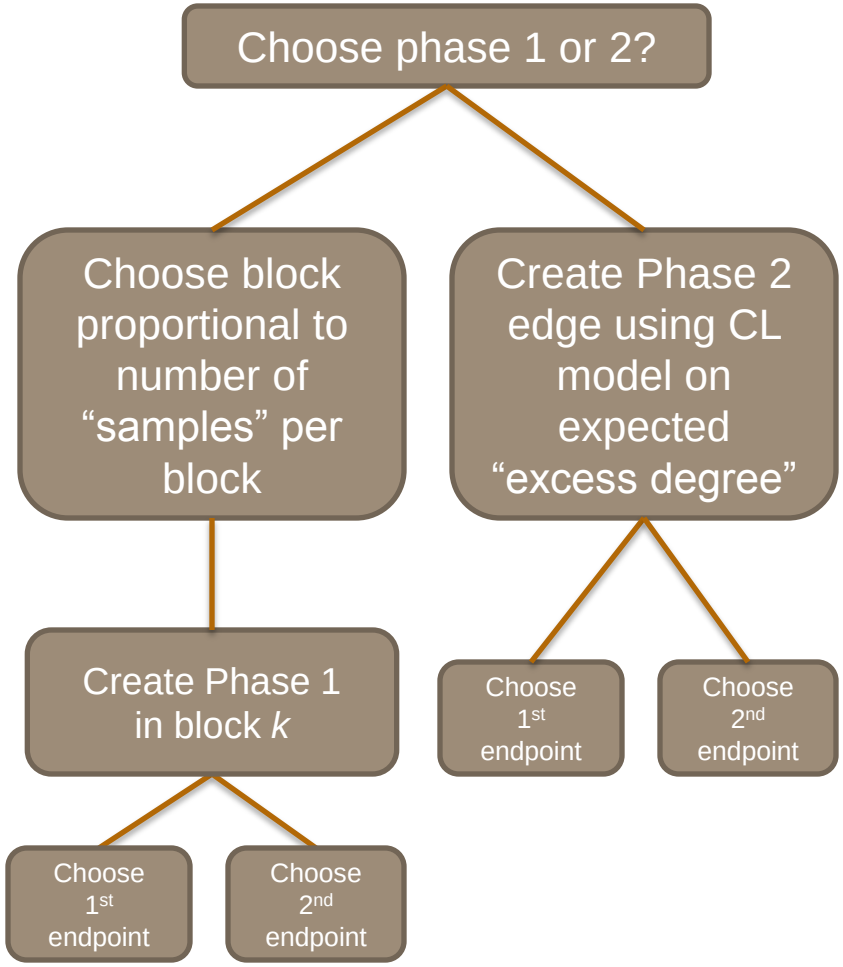
- Need **degree distribution**
  - Calculate explicitly for real data (d_{max} parameters)
  - Can provide a formula, e.g., power law (1-2 parameters)

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

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

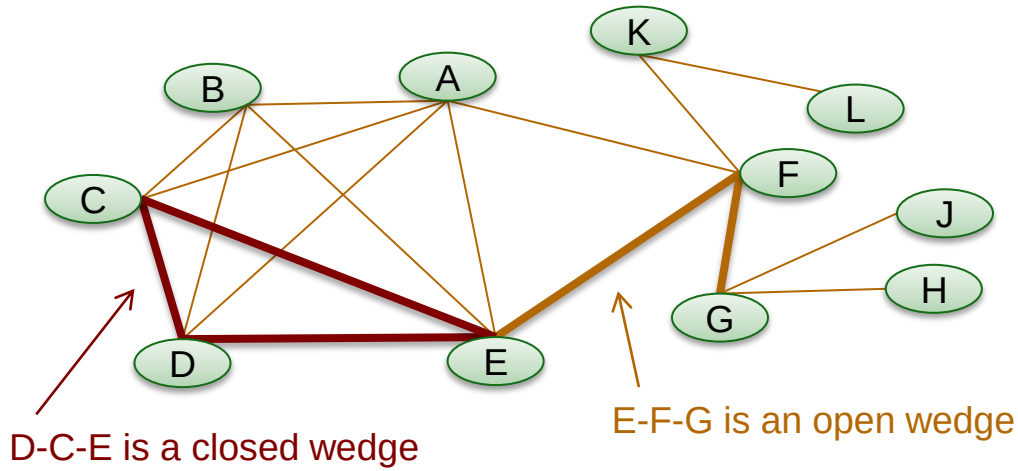
$$c_d = \frac{\#\text{ closed wedges centered at nodes of degree } d}{\#\text{ wedges centered at node of degree } d}$$

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



Clustering Coefficients for BIG Graphs

c = fraction of wedges that are closed



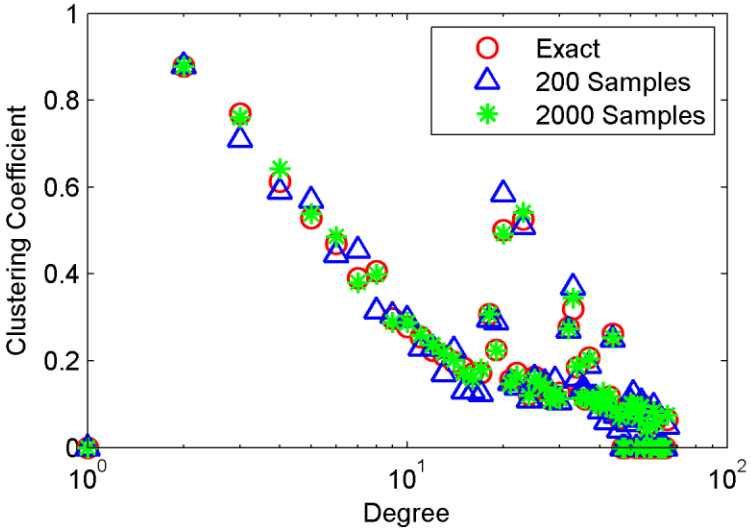
<u>Enumeration</u>: Find *every* wedge. Check if each is closed.
c = # closed wedges / # wedges

<u>Sampling</u>: Sample a few wedges (uniformly). Check if each is closed.
c ≈ # closed sampled wedges / # sampled wedges

Seshadhri, Pinar, Kolda, *SIAM Intl. Conf. Data Mining* 2013

Benefits of Wedge Sampling

- Bounded error for specified sample size and desired confidence
- Work is O(# edges) vs O(# wedges)
- **1000X average speedup** versus enumeration, k = 32,000 (ϵ = 0.011)
- Faster than edge sampling (Doulion) and less variance
- Can also compute clustering coefficient per degree



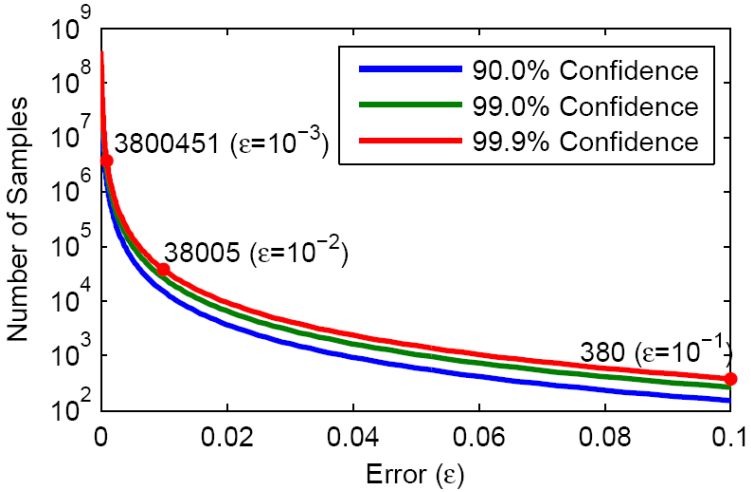
Bounded Error: Hoeffding's Inequality

Theorem: (Hoeffding 1963) Let $X_1, X_2, \ldots, X_k \in [0,1]$ be independent random variables. Define the sample mean: $\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$

Let μ be the true mean. Then for $\epsilon \in (0,1-\mu)$,

$$\text{Prob}\left\{|\bar{X} - \mu| \geq \epsilon\right\} \leq \delta \equiv 2\exp(-2k\epsilon^2)$$

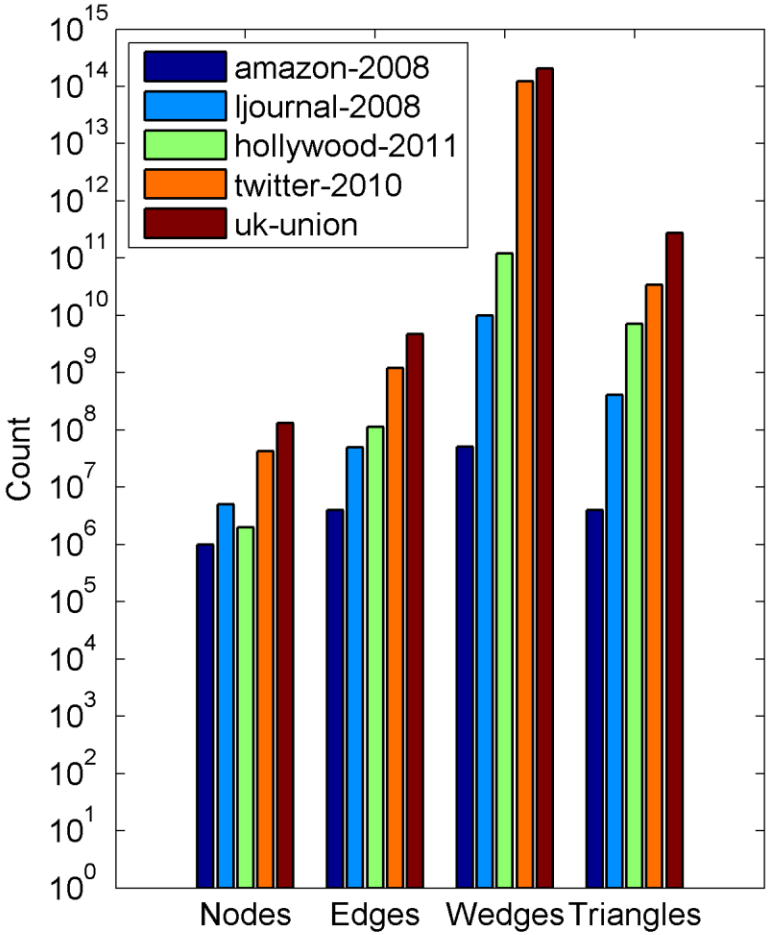
Hence, for a given error ϵ and confidence 1-δ, we just need to set $k = \lceil 0.5\epsilon^{-2}\log(2/\delta)\rceil$



Seshadhri, Pinar, Kolda, *SIAM Intl. Conf. Data Mining*, 2013

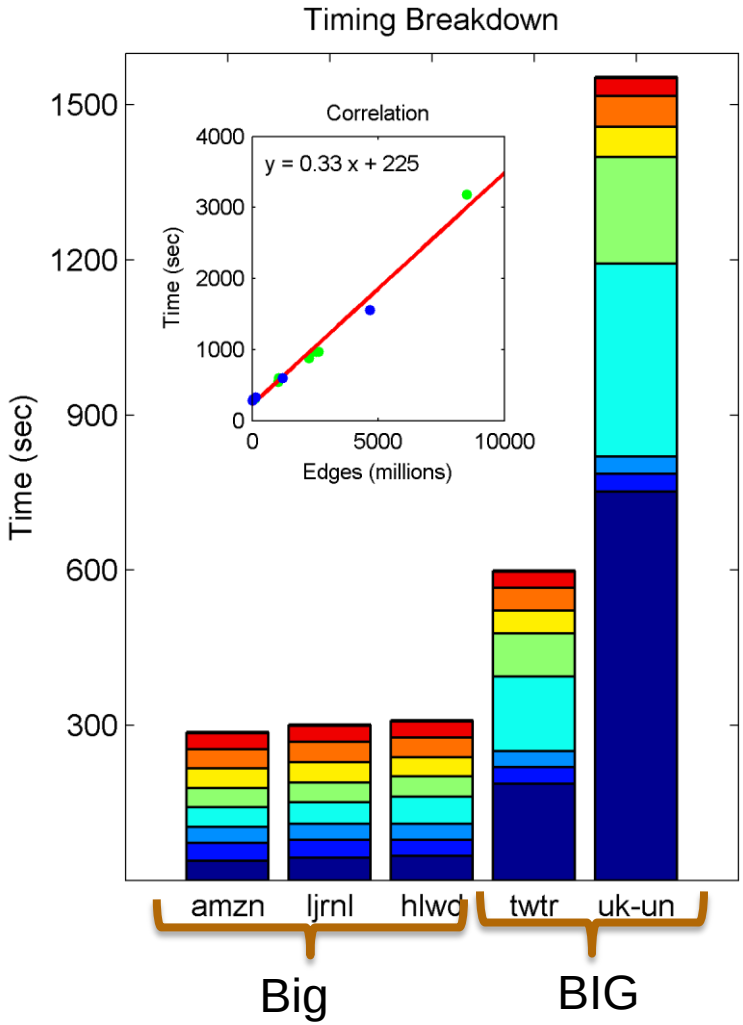
Data Sets and Machines

- 5 real-world networks
  - Source: Laboratory for Web Algorithms (Italy)
  - Largest: 132M nodes, 4.6B edges
  - Observe: # wedges $\gg$ # edges!
- Compute Servers
  - Big Memory Server: SGI Altix UV 10
    - 32 cores (4 Xeon 8-core 2.0GHz processors)
    - 512 GB DDR3 memory
  - Distributed Server: 32-Node Hadoop Cluster
    - 32 x Intel 4-Core i7 930 2.8GHz CPU = 128 cores
    - 32 x 12GB = 384GB memory
    - 32 x 4 2TB SATA disks = 256TB disk storage



Wedge Sampling for BIG Graphs

- 32-node Hadoop cluster results using wedge sampling
  - Logarithmic bins
  - 2000 samples per bin
- "BIG" graphs benefit from Hadoop
  - Merely "Big" graphs don't require Hadoop – just shown as examples
- Compare twitter times
  - Sampling: 10 mins on 32-node Hadoop cluster
  - Enumeration: 483 mins on 1636-node Hadoop cluster
    - Suri & Vassilvitskii, 2011
  - Enumeration: 180 mins on 32-core SGI, using 128GB RAM
    - by Jon Berry, 2013
- No comparisons for uk-union due to its size



Kolda, Pinar, Plantenga, Seshadhri, Task, arXiv:1301.5886, 2013

BTER for BIG Networks

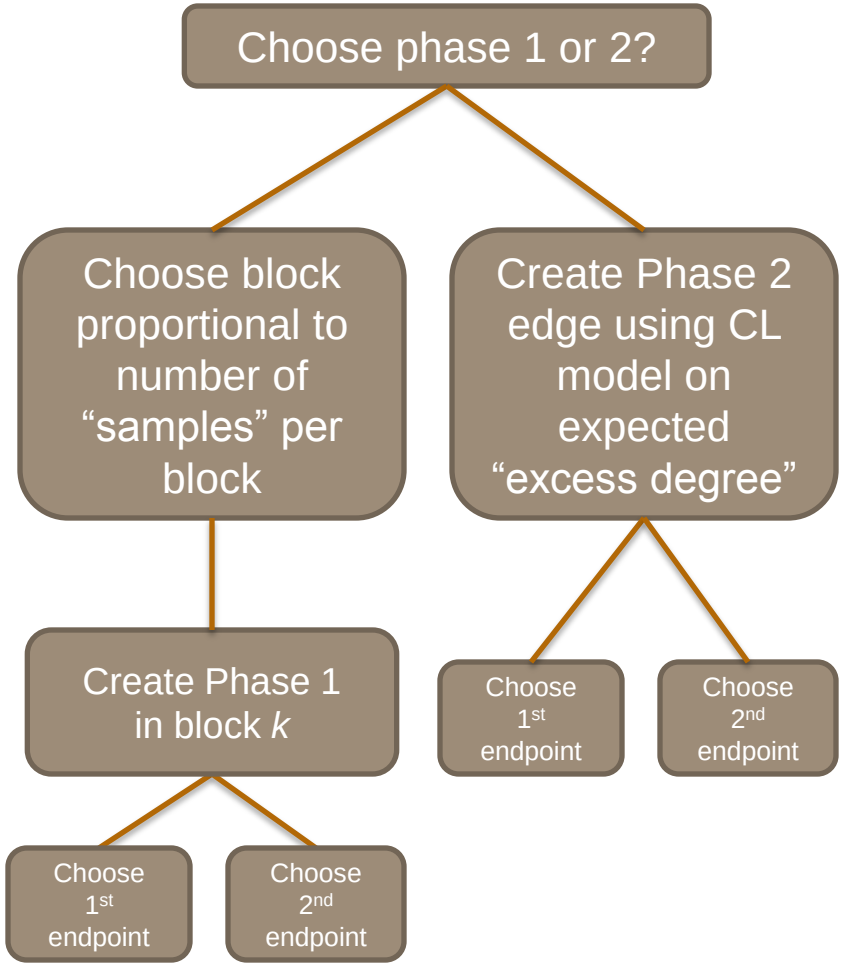
- Need **degree distribution**
  - Calculate explicitly for real data (d_{max} parameters)
  - Can provide a formula, e.g., power law (1-2 parameters)

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

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

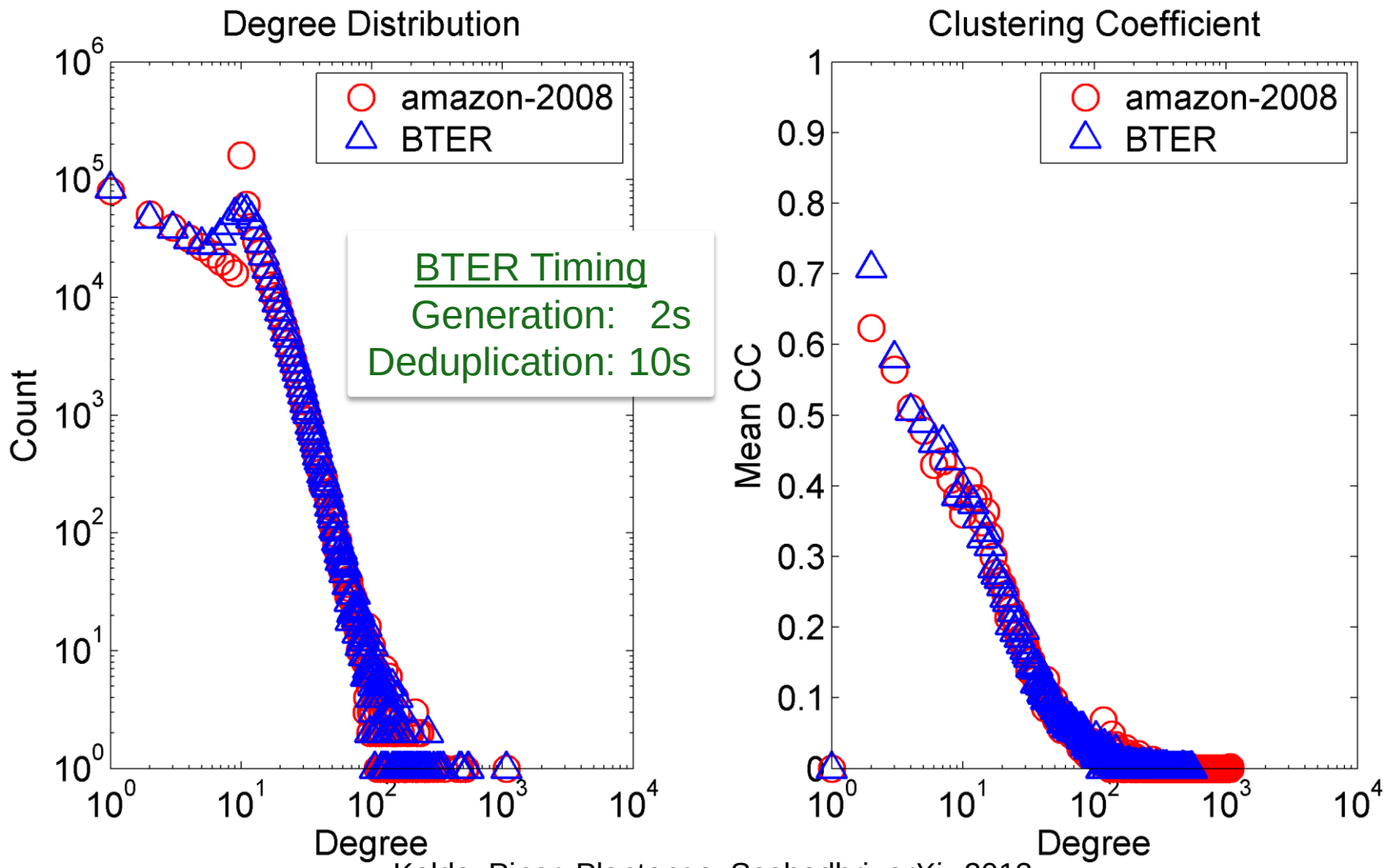
$$c_d = \frac{\#\text{ closed wedges centered at nodes of degree } d}{\#\text{ wedges centered at node of degree } d}$$

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



Kolda, Pinar, Plantenga, Seshadhri, *arXiv* 2013

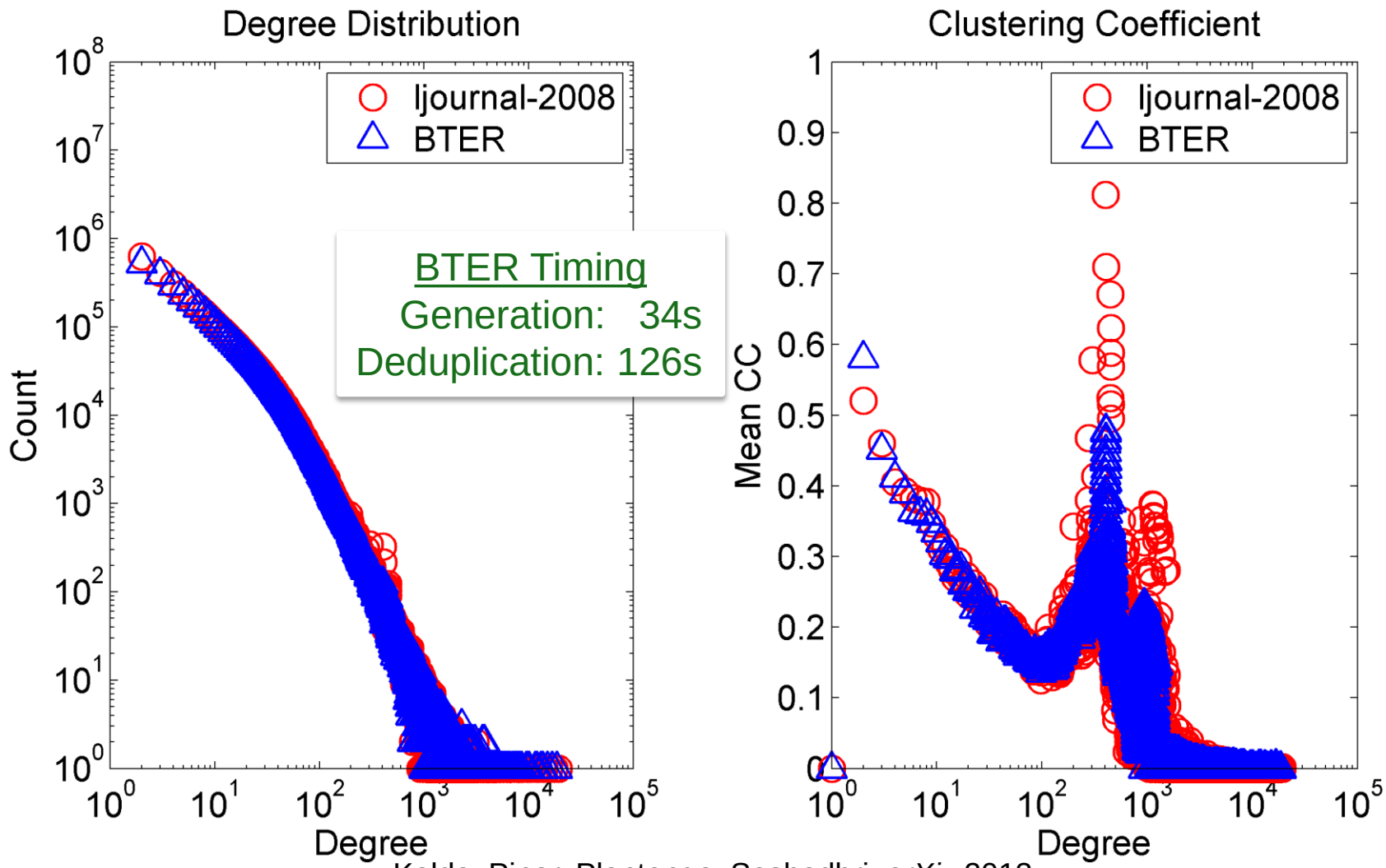
BTER Results: amazon-2008 (4M edges)



BTER Timing
Generation: 2s
Deduplication: 10s

Kolda, Pinar, Plantenga, Seshadhri, *arXiv* 2013

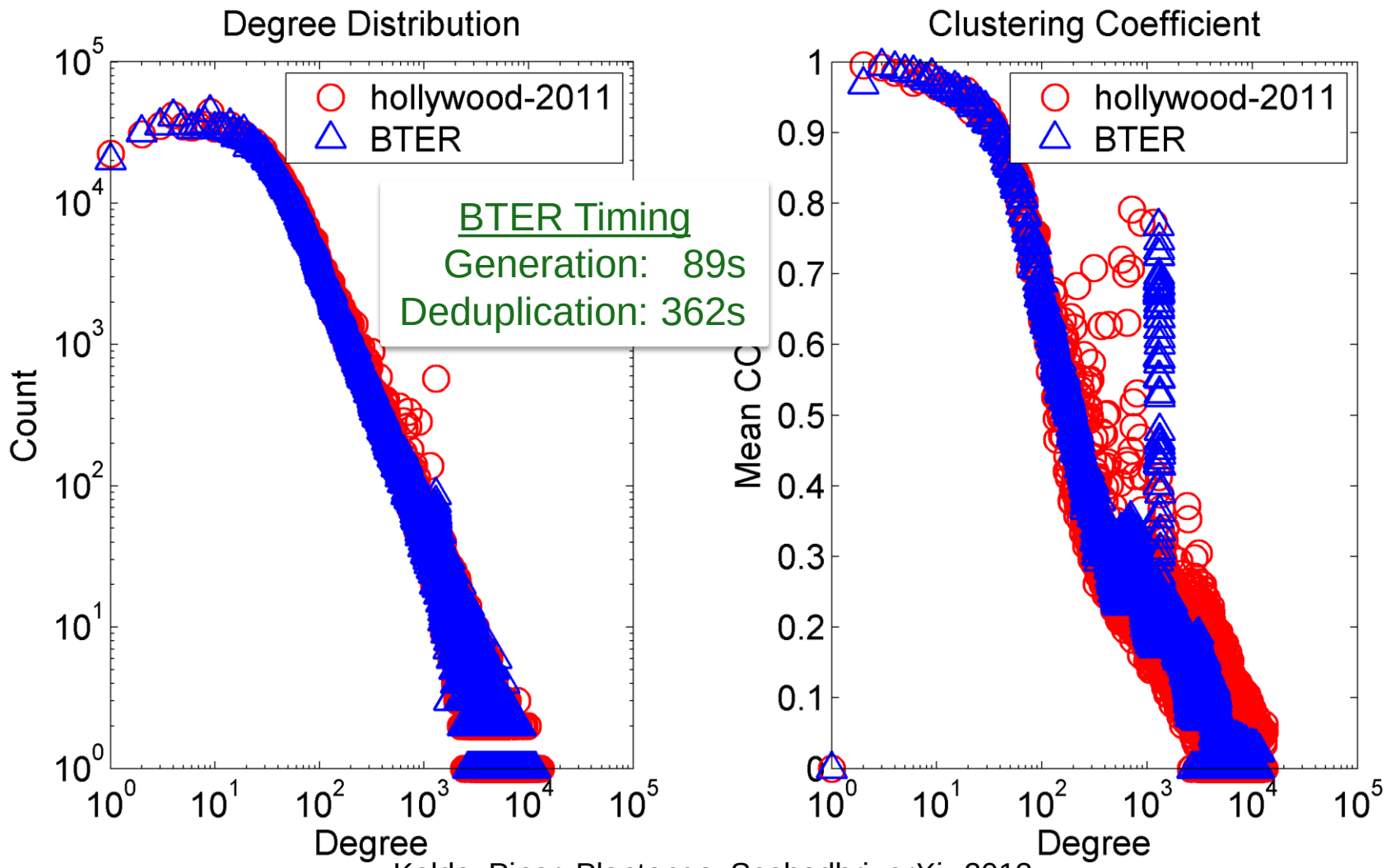
BTER Results: ljournal-2008 (49M edges)



BTER Timing
Generation: 34s
Deduplication: 126s

Kolda, Pinar, Plantenga, Seshadhri, *arXiv* 2013

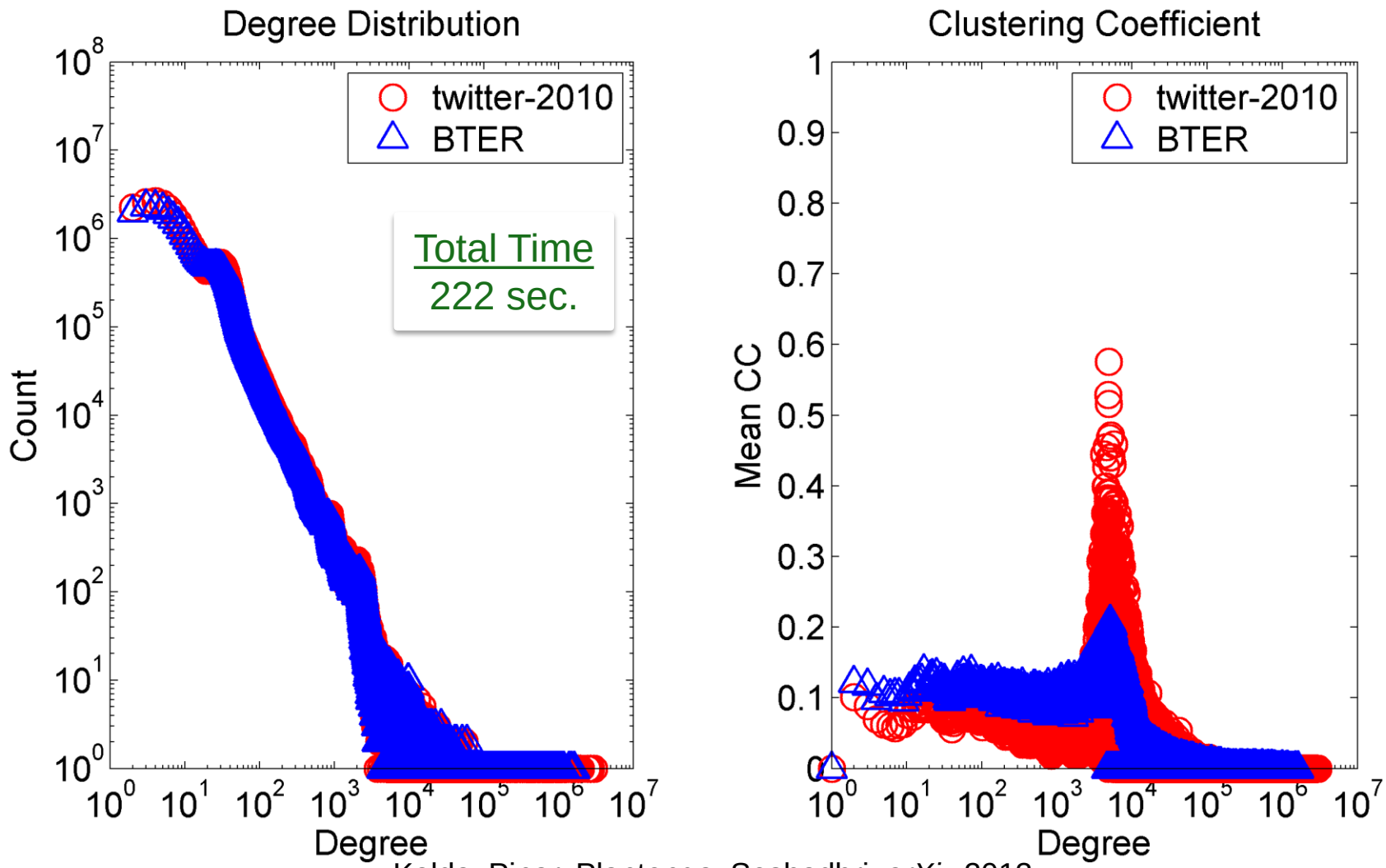
BTER Results: hollywood-2011 (114M edges = 3.4GB sparse matrix)



BTER Timing
Generation: 89s
Deduplication: 362s

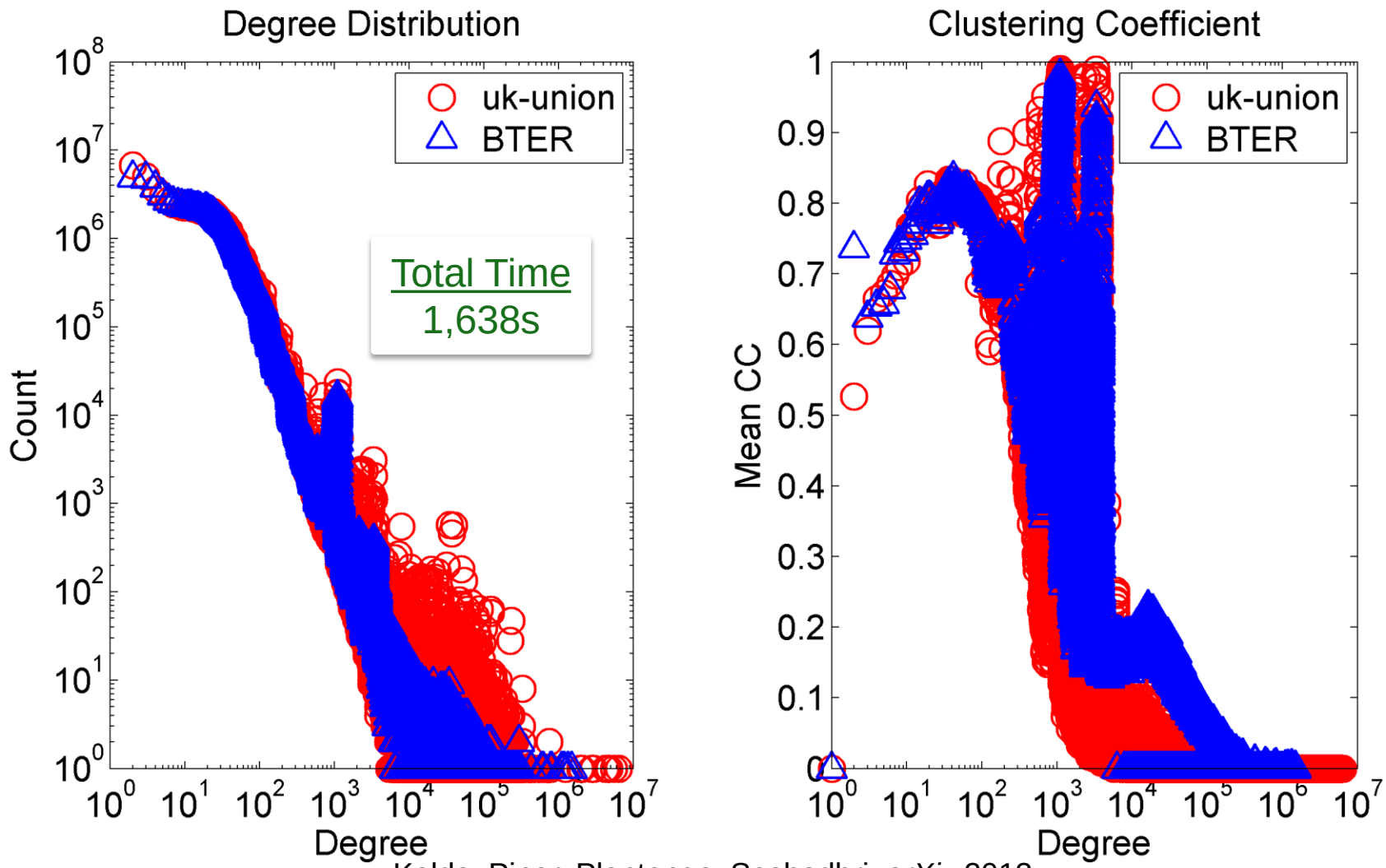
Kolda, Pinar, Plantenga, Seshadhri, *arXiv* 2013

BTER Hadoop Results: twitter-2010 (1.2B edges)



Kolda, Pinar, Plantenga, Seshadhri, *arXiv* 2013

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



Kolda, Pinar, Plantenga, Seshadhri, *arXiv* 2013

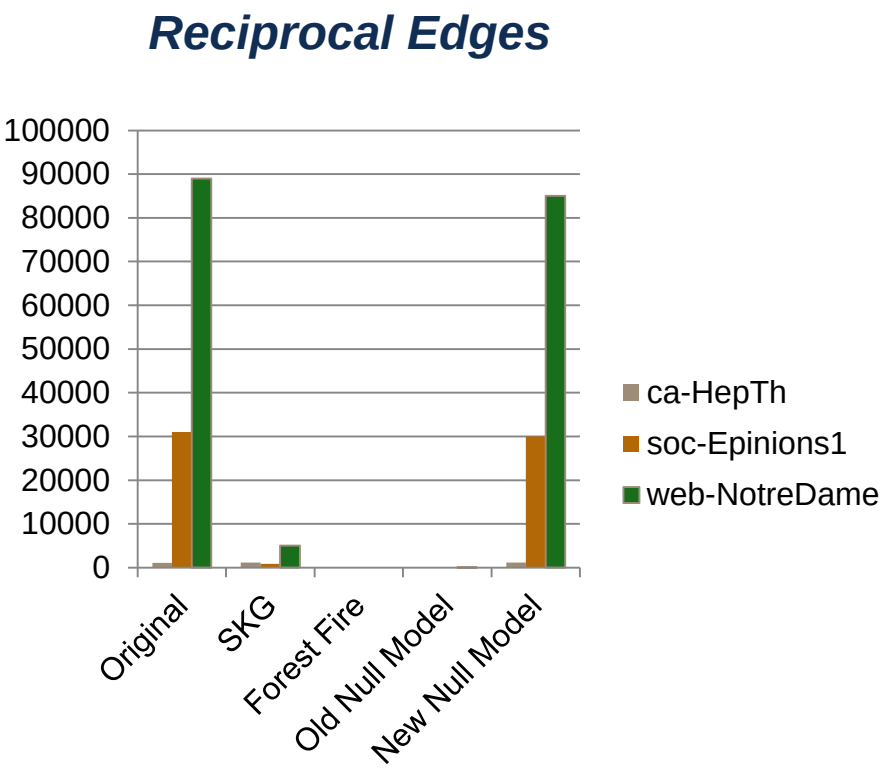
Measurable Characteristics of Directed Graphs: Reciprocation

Directed Edges

One-Way Edge

Reciprocal Edge

- Degree distribution
  - In-degree
  - Out-degree
  - Reciprocal degree(!)
- Many directed networks have an *undirected* network at their heart
- Most "directed" models ignore reciprocation
  - SKG has very little reciprocation
  - Forest Fire (FF) has <u>no</u> reciprocation

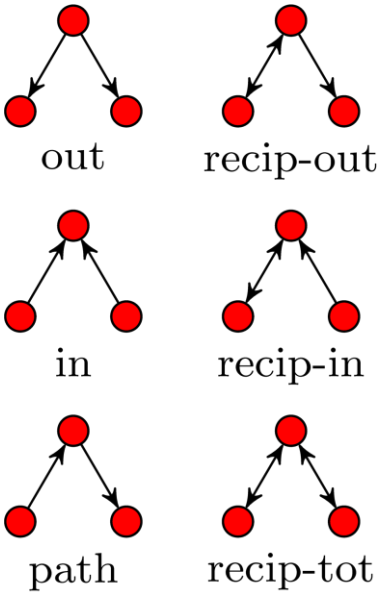


Data from Stanford Network Analysis Project (SNAP)

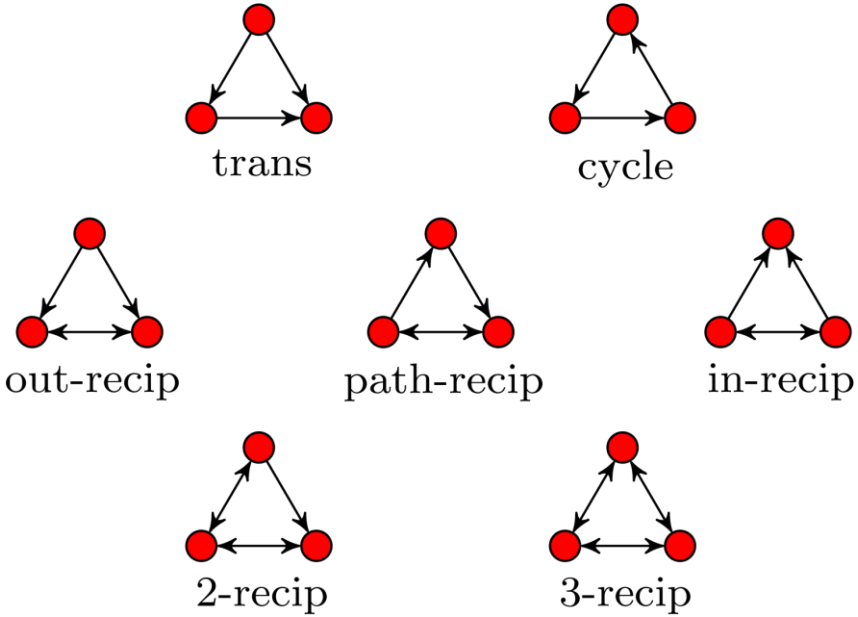
Durak, Kolda, Pinar, Seshadhri, *IEEE Network Science Workshop 2013*

Triangles in Directed Networks

Directed Wedges



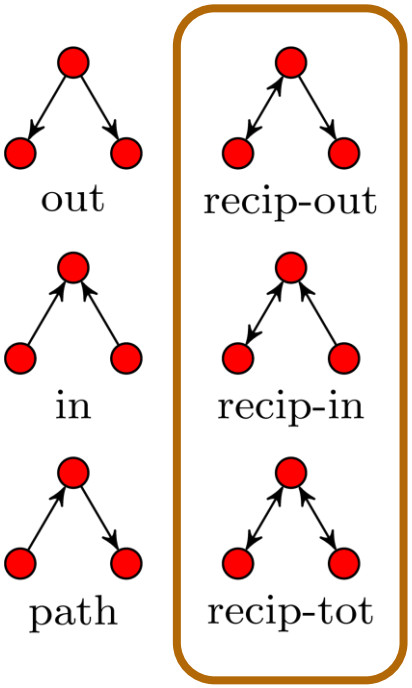
Directed Triangles



Seshadhri, Pinar, Durak, Kolda, *arXiv 2013*

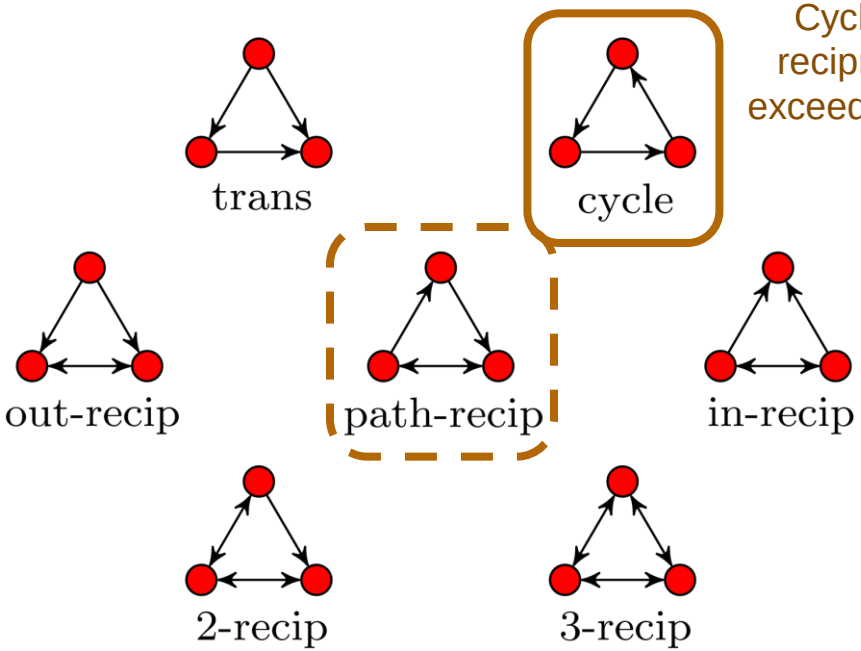
Hints on the "Physics"…

Directed Wedges



Wedges with reciprocal edges are much more likely to close in social and web networks.

Directed Triangles

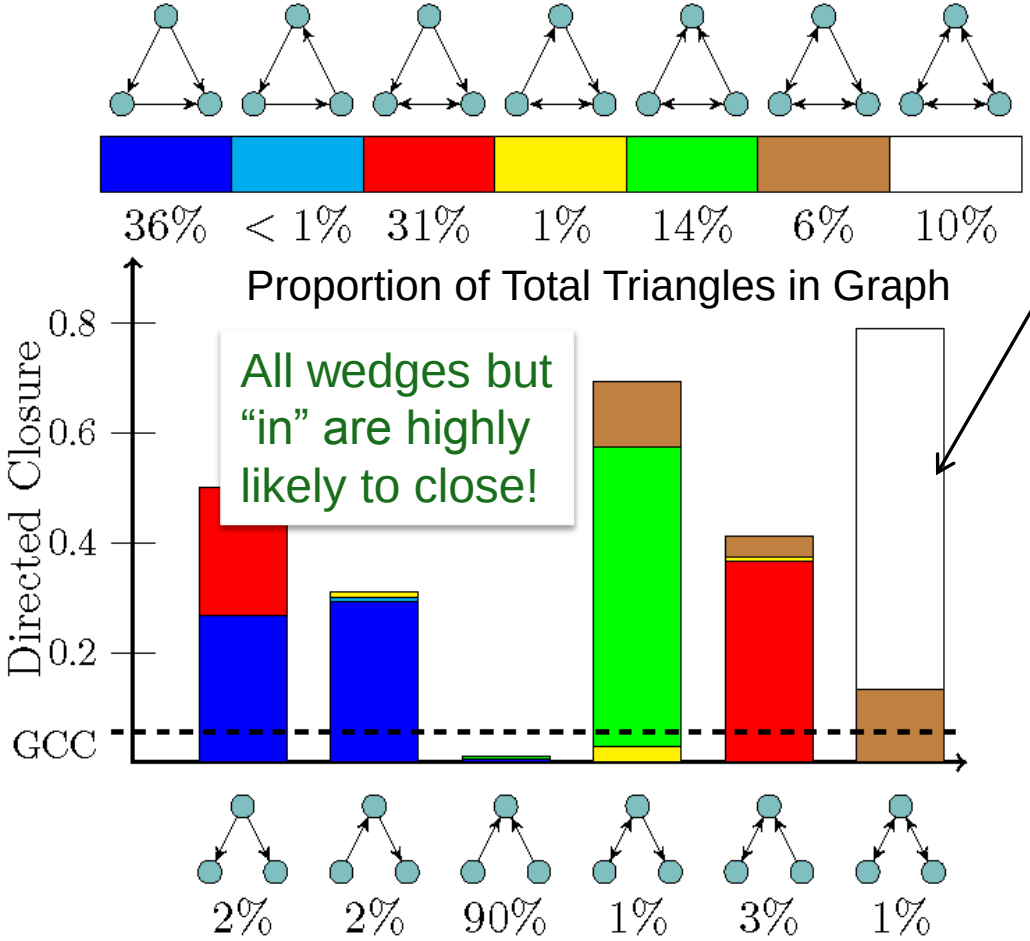


Cycles w/o reciprocation exceedingly rare

Seshadhri, Pinar, Durak, Kolda, *arXiv 2013*

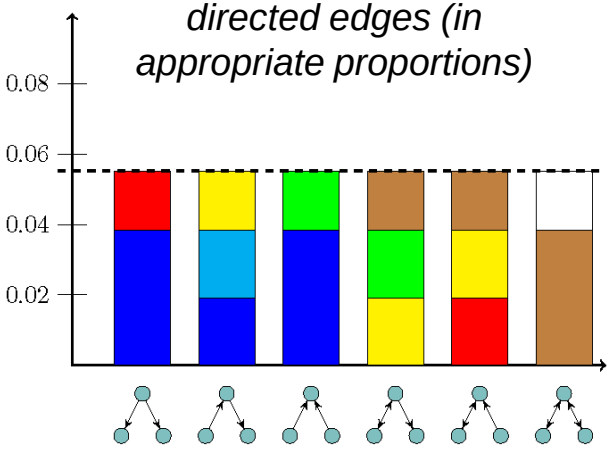
Wedges and Triangles in Web Network: web-Google

web-Google: 876K nodes, 5.1M edges, reciprocation = 31%, GCC=0.055



Proportion of these wedges that closed into that color triangle.

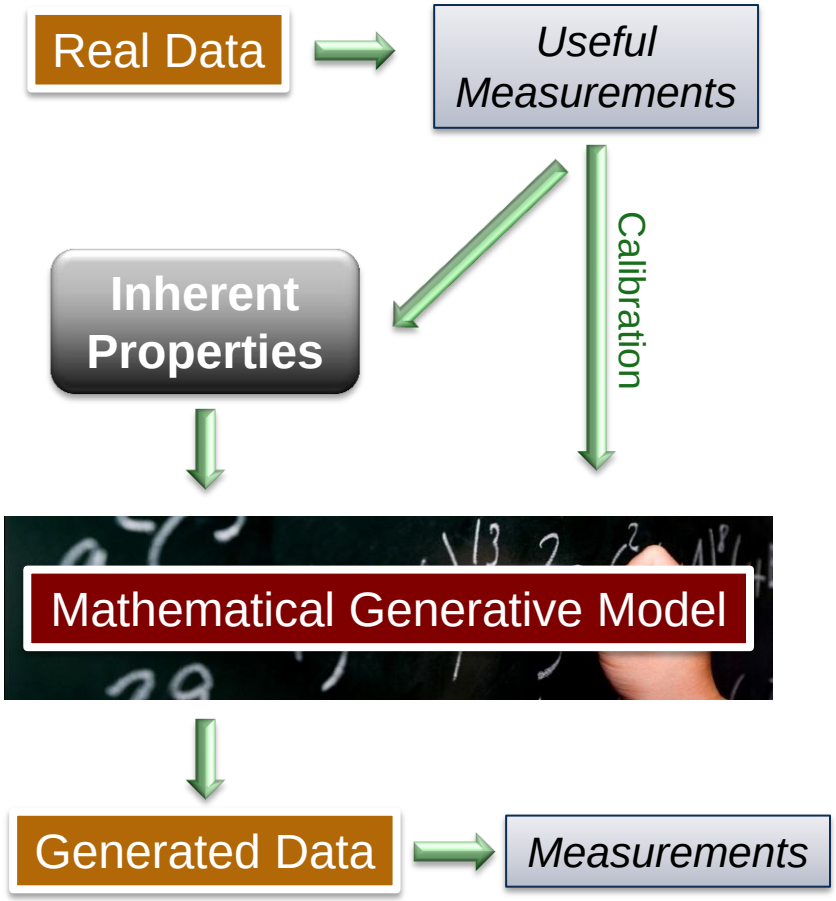
Contrast to proportions for a graph with randomly directed edges (in appropriate proportions)



Data from SNAP

Analysis, Models, and Generators for BIG Networks

- Generating networks requires more than stories – it requires math!
  - Capture "physics"
  - Can be calibrated
  - Scale
  - Yield understanding
- Computing cluster coefficients
  - Statistically sound wedge-sampling approach for BIG networks
- BTER model
  - Theory says: Graphs with high clustering coefficients <u>must</u> have "affinity blocks"
  - Can be used as a generator for benchmarks
- Directed graphs
  - Reciprocation is high in many networks
  - Triangle behavior is not random – reciprocation is extremely important for wedge closure



References

- SKG Analysis: C. Seshadhri, A. Pinar and T. G. Kolda. ***An In-Depth Analysis of Stochastic Kronecker Graphs***, JACM, in press (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***, *coming soon to arXiv, 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***, *coming soon to arXiv, 2013*

- *For copies: Tammy Kolda, tgkolda@sandia.gov*

Please consider submitting papers to the SISC Special Section on Planet Earth and Big Data!

Friday, March 1
MS250
Analysis and Modeling of Static and Dynamic Networks
1:00 PM - 3:00 PM
Room: Commonwealth Ballroom B - Concourse Level