SAND2012-3662P

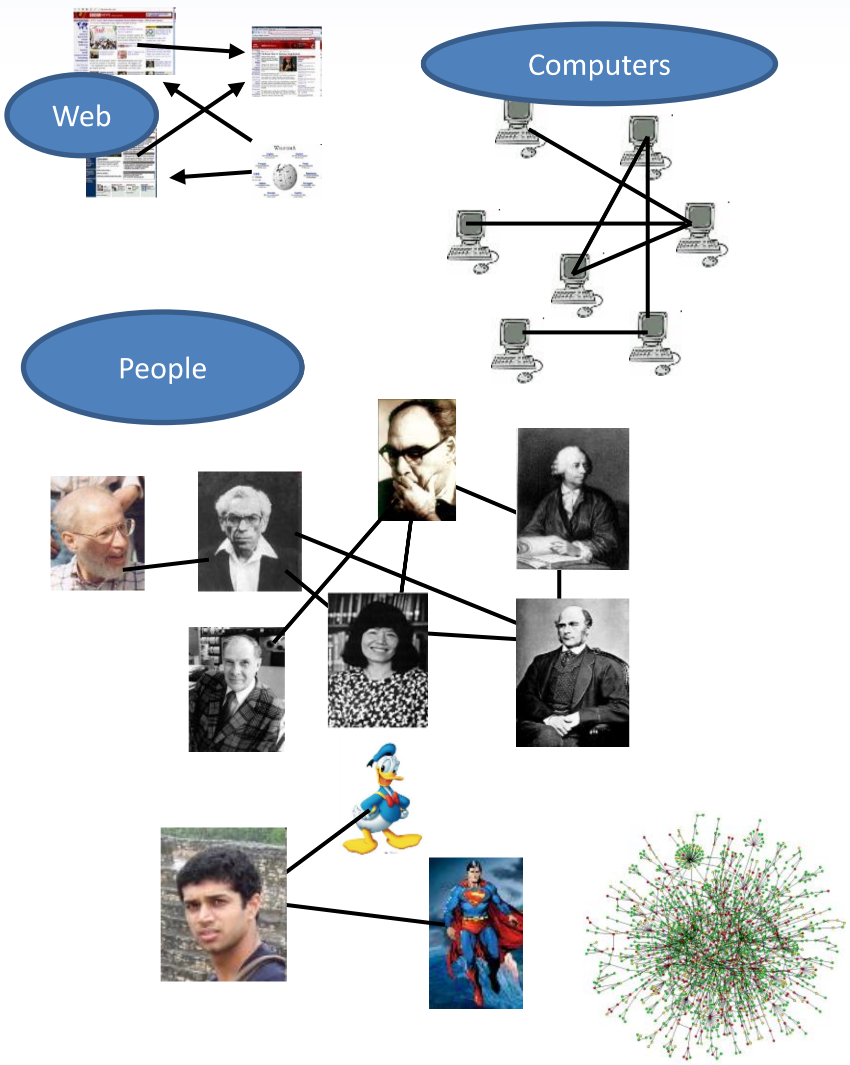
How to make sense of massive, real-world networks

Seshadhri Comandur (C. Seshadhri)
Informatics and Systems Assessment - 08966

CIS External Panel Review
May 8-10, 2012

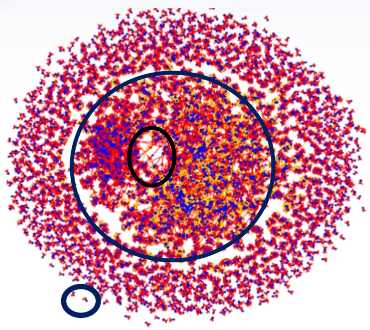
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

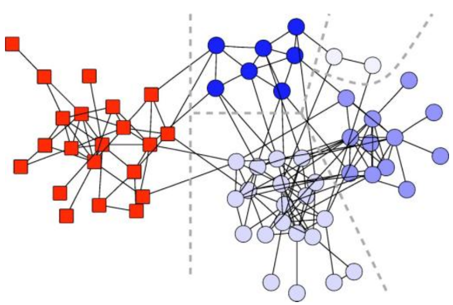
Massive networks are everywhere

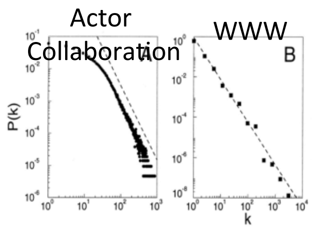


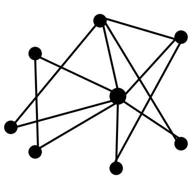
- Many real world interactions/phenomena expressed as “graphs”
  - Vertices represent “entities”, edges are “connections”
  - Massive (n = 10,000 to millions) and sparse (no. of edges < 10 n)
- A very simple, intuitive way of representing data
- Sandia’s interests in graphs: communication, computer networks (cyber), supply chains, counter-terrorism…
  - Generally understanding complex phenomena
  - A very powerful “modeling tool” for entities and interaction (e.g. agent based models)
  - DARPA BAA: “Graph–theoretic Research in Algorithms and the PHenomenology of Social networks (GRAPHS)”
- We need capability in analyzing, processing, generally “understanding” these graphs
  - Want to analyze the graph to understand underlying process

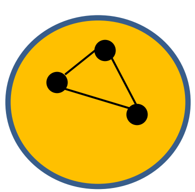
Challenges in analyzing real graphs









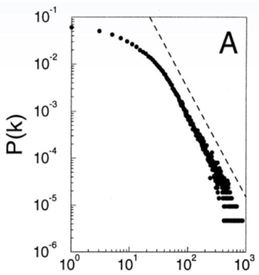


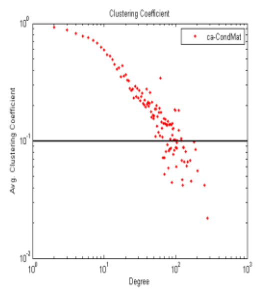
- Kinds of questions and their applications
  - Are there common patterns we can identify? (Communication patterns in email)
  - Is there some notion of “normal” and “abnormal” structure? (Anomaly detection in supply chain networks)
  - How does this evolve over time? Can we track this evolution? (Situational awareness in cyber data)

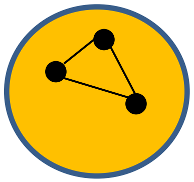
- Graphs are extremely large, and we lack scalable algorithms
- Real graphs have peculiar properties
  - [Barabasi-Albert 98, Watts-Strogatz 97, Newman-Girvan 04]
- Graph modeling is a concrete approach for understanding graphs
- How does one create a synthetic graph that looks “real”?
  - (We’ve been asked this quite often.)
  - For testing algorithms
  - For validating hypotheses
  - For understanding properties of interest
- Graph500 supercomputer benchmark is a relevant example

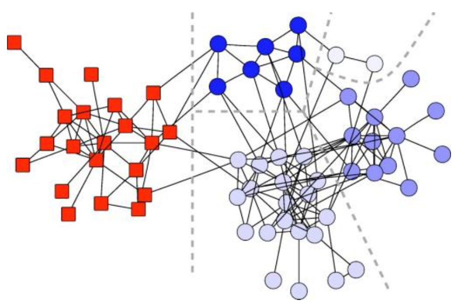
Broader research perspective of graph modeling

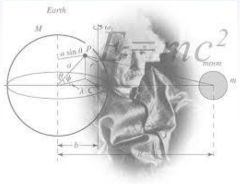
Sandia National Laboratories











- Physicists/social scientists ask:
  - What kind of physical/social processes produce these graphs?
- Computer scientists/engineers ask:
  - How can we find special structures (e.g. communities)? How to generate “benchmark” graphs for testing algorithms?
- Mathematicians ask:
  - Can we formally prove theorems about these graphs? “Because graph is heavy tailed, eigenvalues must be like…”
- Graph modeling intimately related to all these questions
- [Watts-Strogatz 97, Barabasi-Albert 98, Kumar et al 00, Chakrabarti-Faloutsos-Zhan, 04 Leskovec et al 05, Bickel-Chen 06]…
- “All models are wrong, but some are useful” – George Box
- Good models help in design of faster algorithms

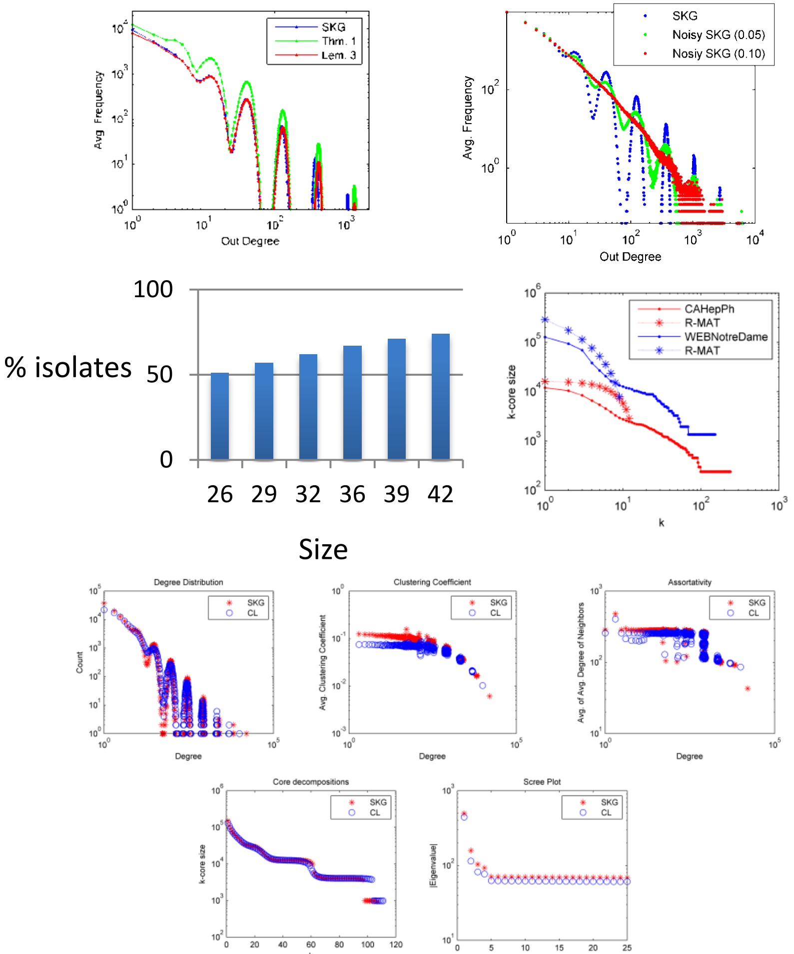
Our work

- Understanding current models
  - Mathematical analysis of RMAT/SKG graph model (Graph500)
  - Connections of SKG to conceptually cleaner CL model
- New models and generation methods
  - BTER, a new scalable model with provably good properties
  - Theorems about convergence of MCMC methods to general graphs
- Faster algorithms to process graphs
  - Sampling methods to count triangles
  - Speeding up agglomerative clustering schemes

Analysis of SKG

Seshadhri, Pinar & Kolda; short version in ICDM11
"An in-depth analysis of Stochastic Kronecker Graphs"

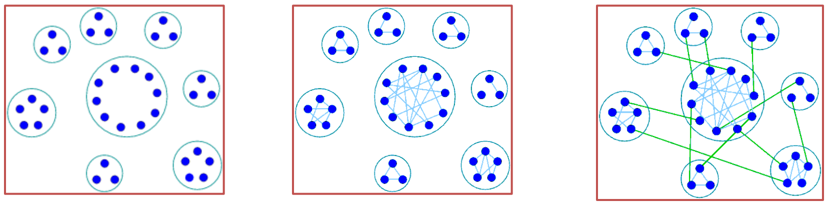
Pinar, Seshadhri & Kolda; short version in SDM11
"The similarity between SKG and CL models"

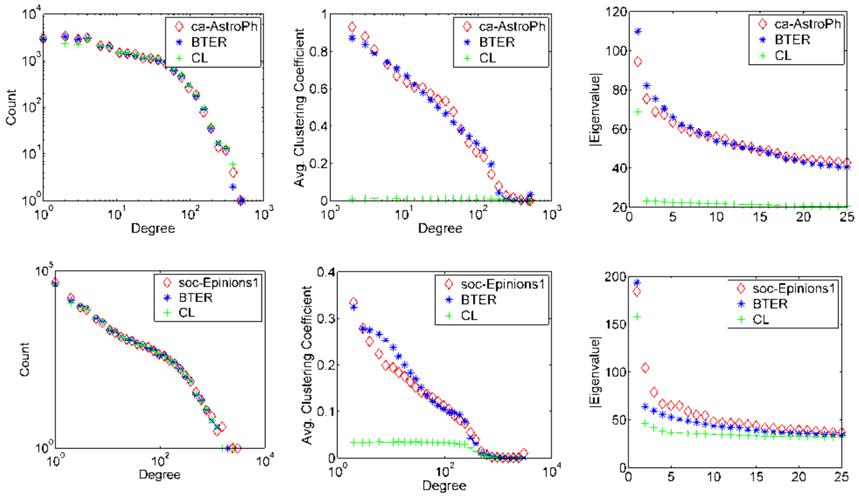


- SKG/R-MAT is Graph500 benchmark
  - Very important for HPC applications
  - Very poorly understood model
- Complete analysis of degree distribution
  - Standard degree distribution has large oscillations
  - Refute unsubstantiated claims made about dd
  - Define a "fixed" version of SKG. Prove that is gives a lognormal distribution
- Many isolated vertices in Graph500
- Graph500 benchmark changed because of this work
- Actually, SKG modeled quite well just Chung-Lu – a much simpler model
  - Ironically, SKG thought to be "realistic" but not CL
- Detailed studies of probability matrices underlying SKG and CL – much deeper connection

The BTER model

Seshadhri, Pinar & Kolda; Physical Review E, 2012
“Community structure and scale-free collections of Erdos-Renyi graphs”

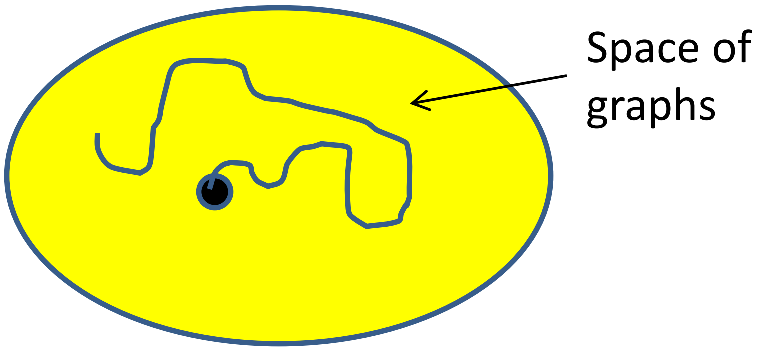


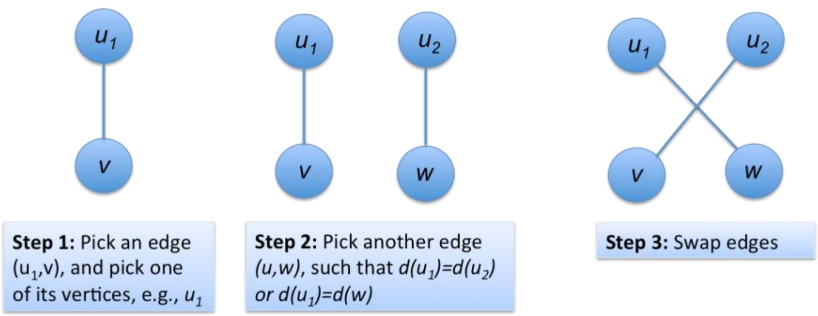


- Can we construct a scalable model with heavy tailed degree distribution and many triangles (large clustering coeff)?
- No current model satisfies these
- We define formal notion of community structure
- Use extremal combinatorics to show this implies presence of dense Erdos-Renyi graph
  - Well known that ER graphs are not realistic
  - But we show that a properly chosen “collection” of these
- Construct the Block-Two-level Erdos Renyi (BTER) model
- Provable properties; transparent model
- We are currently building scalable implementation of this model

Sampling graphs using MCMC

Ray, Pinar, and Seshadhri; short version in WAW12, “Are we there yet? When to stop a Markov chain while generating random graphs”

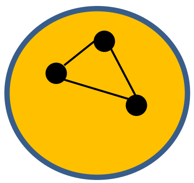


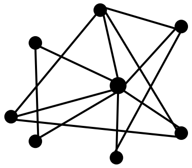


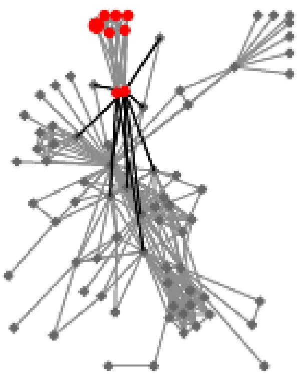
- A common method to generate random graphs with a given degree distribution is to use MCMC methods
  - Usually to generate “similar” graphs from a given graph
  - Start with G, perform a series of random rewirings
- But how many rewirings to perform?
  - Basically mixing time of a large Markov Chain
- Theoretical bounds infeasible: “run 10^{10} steps to generate graph with 10,000 edges”
- Practical methods: “run for some number of steps and hope for the best”
- We bridge this gap: a theoretical analysis giving practical bound
- We prove that 10|E| steps enough for most practical purposes

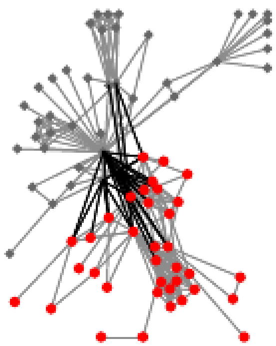
Faster algorithms for massive graphs

Seshadhri, Pinar & Kolda; submitted
"Fast Triangle Counting through Wedge Sampling"





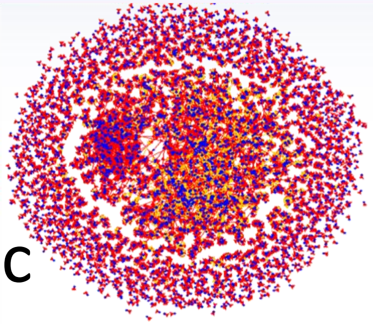




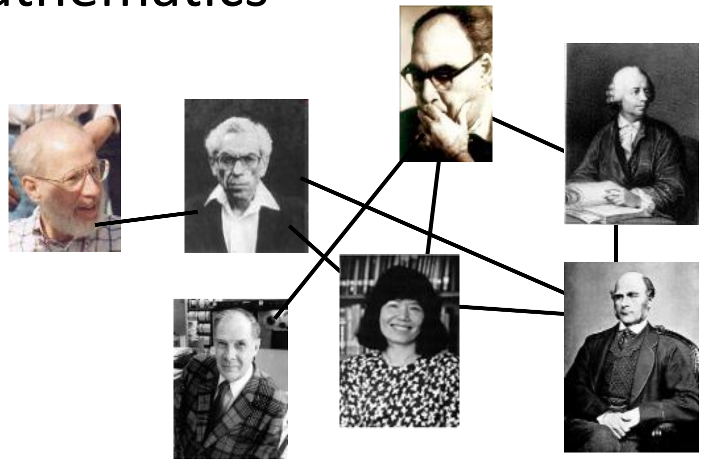
Gleich & Seshadhri; submitted
"Neighborhoods are good communities"

- Counting triangles is a very important task, but graphs are becoming larger and larger
- We give simple, sampling based method to approximate number of triangles
  - Formal accuracy/time tradeoff
  - Provable guarantees of behavior
  - Works well in practice
- We prove: Abundance of triangles means it's easy to find "communities"
  - Based on theorem relating conductance to clustering coefficients
- We use theorem to speed up agglomerative community finding algorithms
  - How to find right starting seed?

In conclusion…



- Study of massive graphs a deeply scientific endeavor with deeply relevant applications
- Lot of exciting research in Sandia on this topic
  - Funded by LDRDs, ASCR Applied Mathematics Program, will get DARPA funding



- Impact
  - Graph 500 and benchmarking
  - Many applications within Sandia
  - Theoretical results with practical impact (7+ research papers in last 2 years)

Supplementary Information

People involved

- Jonathan Berry, 1464
- Janine Bennett, 8953
- Richard Chen, 8954
- Nurcan Durak, 8965
- David Gleich, Purdue U.
- Tamara Kolda, 8966
- Richard Lehoucq, 1444
- Vitus Leung, 1464
- Ali Pinar, 8965
- Cynthia Phillips, 1465
- Todd Plantenga, 8958
- Jaideep Ray, 8954
- David Robinson, 1464
- Matthew Rocklin, U. Chicago
- Isabelle Stanton, UC Berkeley
- Yevgeniy Vorobeychik, 8953

List of Publications

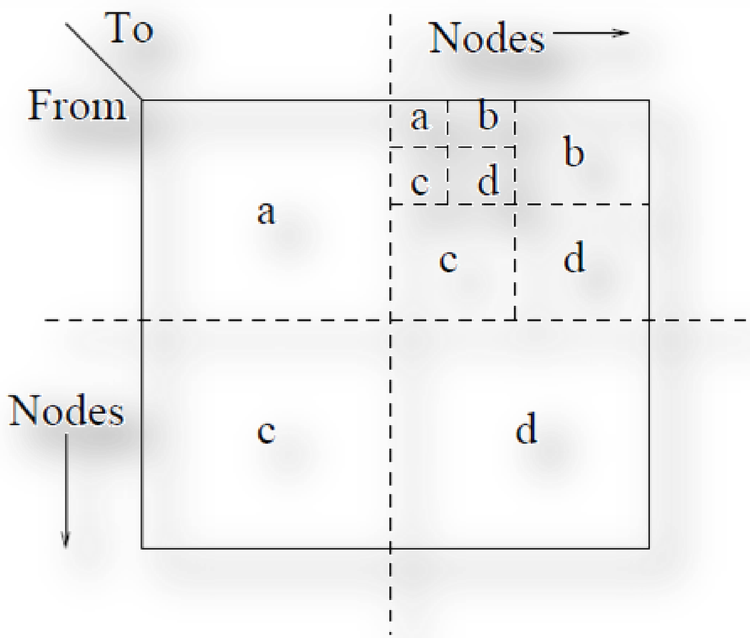
- C. Seshadhri, A. Pinar, T. G. Kolda, “An In-Depth Analysis of Stochastic Kronecker Graphs”, International Conference of Data Mining (ICDM), 2011
- A. Pinar, C. Seshadhri, T. G. Kolda, “The Similarity between Stochastic Kronecker and Chung-Lu Graph Models”, SIAM Conference on Data Mining (SDM), 2011
- J. Berry, B. Hendrickson, R. LaViolette, and C. Phillips, “Tolerating the community detection resolution limit with edge weighting," Physical Review E, 2011.
- C. Seshadhri, T. G. Kolda, A. Pinar, “Community Structure and Scale-free Collections of Erdos-Renyi Graphs”, Physical Review E, 2012
- J. Ray, A. Pinar, C. Seshadhri, “Are we there yet? When to stop a Markov chain while generating random graphs”, Workshop on Algorithms and Models for the Web Graph (WAW), 2012
- C. Seshadhri, A. Pinar, T. G. Kolda, “Fast Triangle Counting through Wedge Sampling”, Arxiv Tech Report
- D. F. Gleich, C. Seshadhri, “Neighborhoods are Good Communities”, Arxiv Tech Report
- J. Berry, D. Nordman, L. Fostvedt, C. Phillips, C. Seshadhri, A. Wilson, “Listing triangles in expected linear time on power law graphs with exponent at least 7/3”, in preparation

Graph 500 Model: R-MAT/Stochastic Kronecker (SKG)

Chakrabarti , Zhan, & Faloutsos, SDM04; Leskovec et al., JMLR, 2010

- R-MAT/SKG Inputs
  - L = # of levels
  - T = 2 x 2 generator matrix (entries sum to 1)
  - M = # edges

- SKG Edge Insert Procedure
  - Choose a quadrant of the adjacency matrix proportional to entries of T
  - Repeat for a total of L times to land at a single entry of the matrix



Graph 500 Parameters
- T = [0.57, 0.19, 0.19, 0.05]
- L $\in$ {26, 29, 32, 26, 39, 42}
- M = $16 \cdot 2^L$

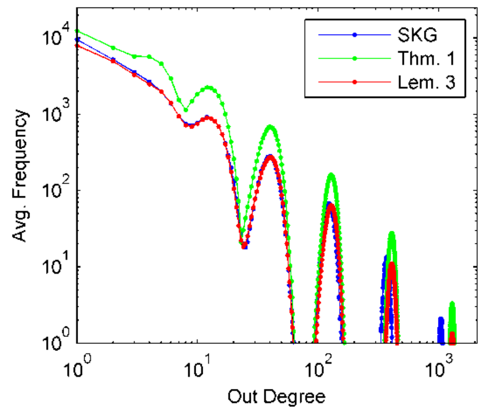
Degree Distribution of SKG

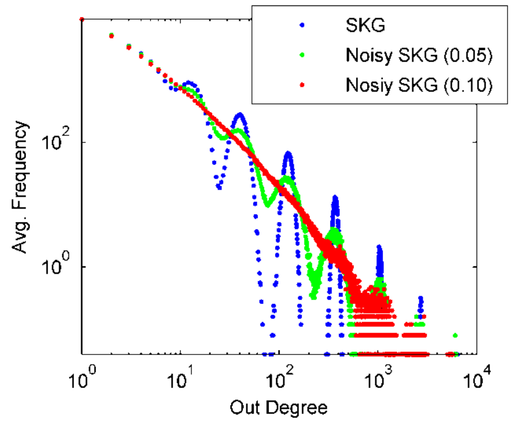
Seshadhri, Pinar & Kolda, arXiv:1102.5046, Sept 2011; short version to appear in ICDM11

- Standard degree distribution has large oscillations
  - Pretty much unexplained
  - Lot's of well…bogus claims made about dd
- We give an accurate and easily computable formula for degree distribution
  - Theorem: oscillates between lognormal and exponential
- We can actually fix the problem
  - Define a noisy version of SKG
  - Prove that is gives a lognormal distribution

$$T_i = \begin{bmatrix} a - \frac{2\mu_i a}{a+d} & b + \mu_i \\ b + \mu_i & d - \frac{2\mu_i d}{a+d} \end{bmatrix}$$

SKG for Graph 500 for L=16





Isolates in SKG for Graph 500

Seshadhri, Pinar & Kolda, arXiv:1102.5046, Sept 2011; short version to appear in ICDM11

- An incredibly huge number!

- Number of isolates is

$$I = \sum_{r=-L/2}^{L/2} \binom{L}{L/2+r} \exp(-2\lambda\tau^r),$$

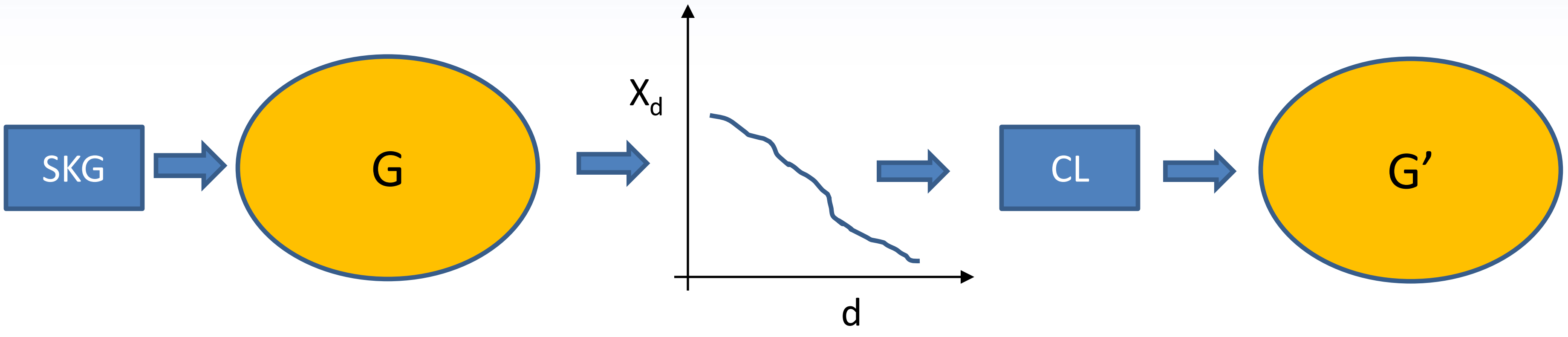
$$\tau = (a+b)/(1-(a+b))$$

$$\lambda = \frac{M}{N}[4(a+b)(1-(a+b))]^{L/2}$$

- Impacts benchmark because number of nodes is less than anticipated and average degree is much higher!

L	Isolated Nodes	Avg. Degree
26	51%	32
29	57%	37
32	62%	41
36	71%	55
39	71%	55
42	74%	62

CL from SKG



- G' is CL graph from degree distribution of G
- How similar is G' to G?

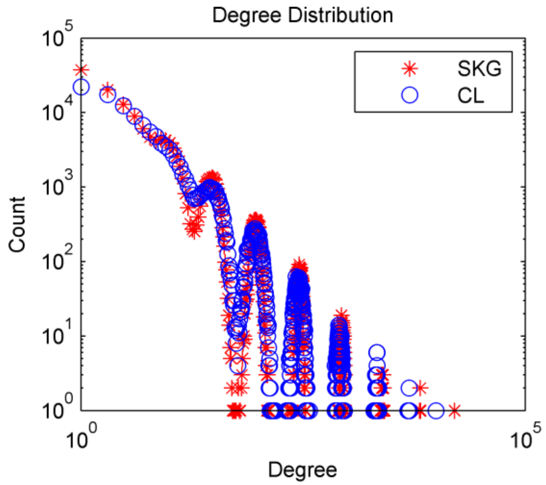
SKG vs CL

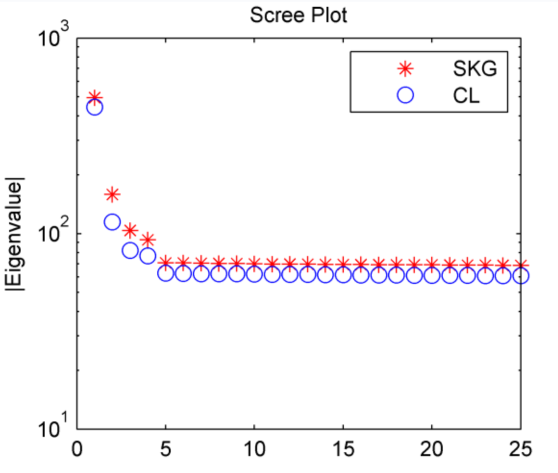
Pinar, Seshadhri, & Kolda, preprint, Oct 2011

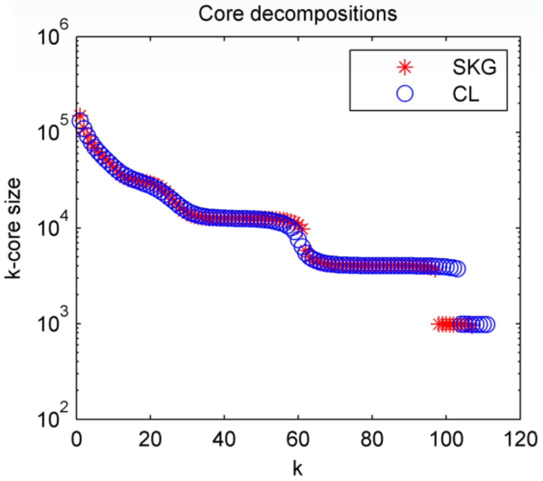
- SKG is incredibly similar to CL!
  - If we plug in an SKG degree distribution into CL, the graphs we get are incredibly similar.
- Probability matrices used by these models are almost same
  - Theorem: For certain parameter settings, models are indeed identical
- If you're going to use SKG, you might as well just use CL
  - It fits real data just as well

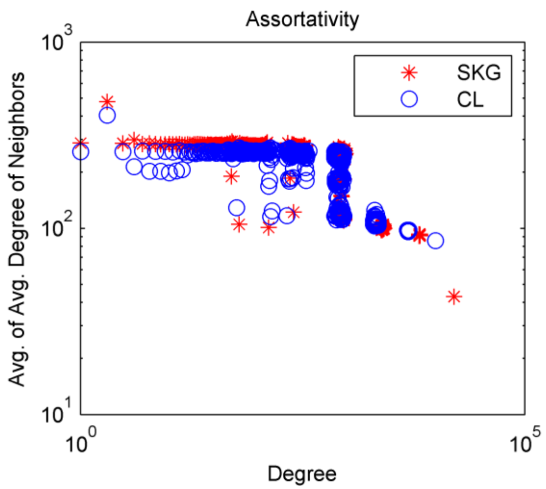
Similarity of CL to SKG for Graph 500

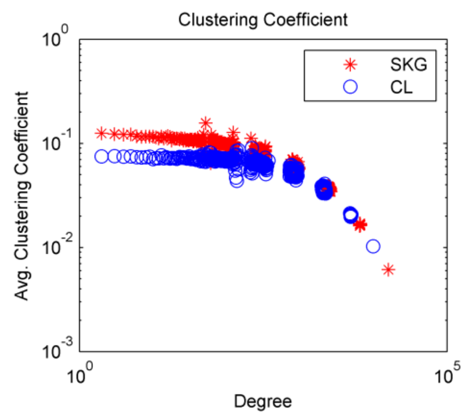
Fit CL to the degree distribution produced by SKG for Graph 500 with L = 18



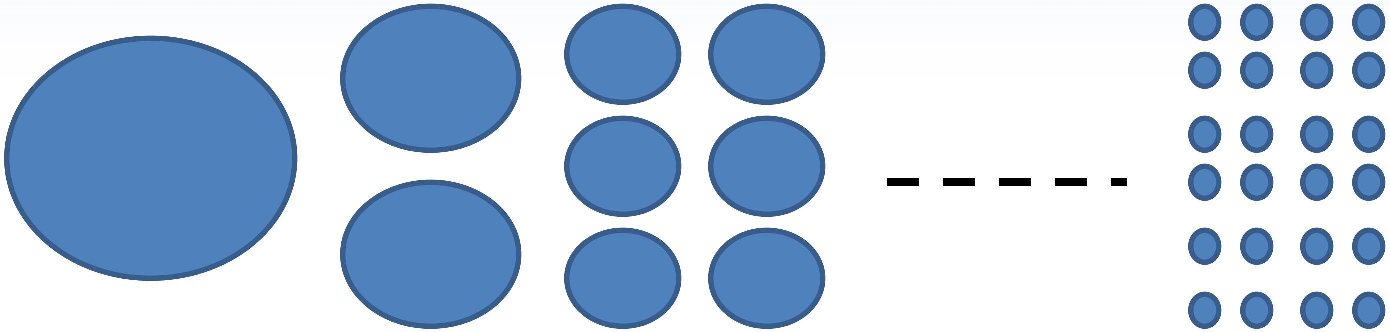








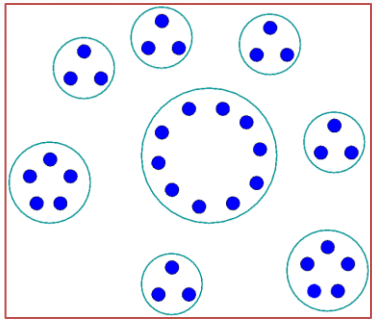
The math



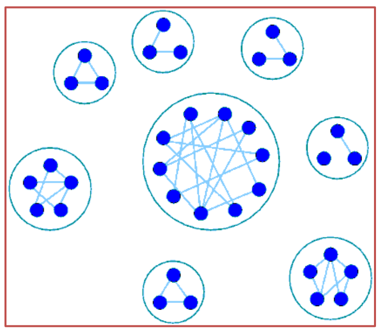
- Suppose a graph has a heavy tail, large CCs and contains communities
  - We can mathematically formalize all of this
- Then a constant fraction of its edges lie in disjoint dense Erdos-Renyi graphs, the sizes of which also form a heavy tail

BTER: Block Two-level Erdos-Renyi

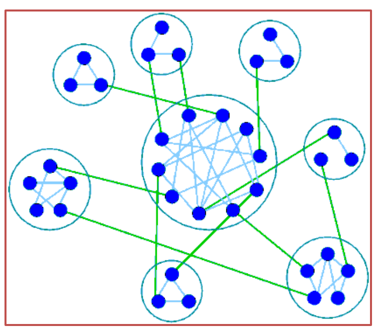
Preprocessing: Create explicit communities



Phase 1: Erdös-Rényi graphs in each community



Phase 2: CL model on “excess” degree



- **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, ρ_k
  - We use a function of the min degree in the community, d_k
- **Phase 2:** Generate CL graph on “excess” degree
  - $e(i) = d(i) - \rho_k \, d_k$ where vertex i is in community k