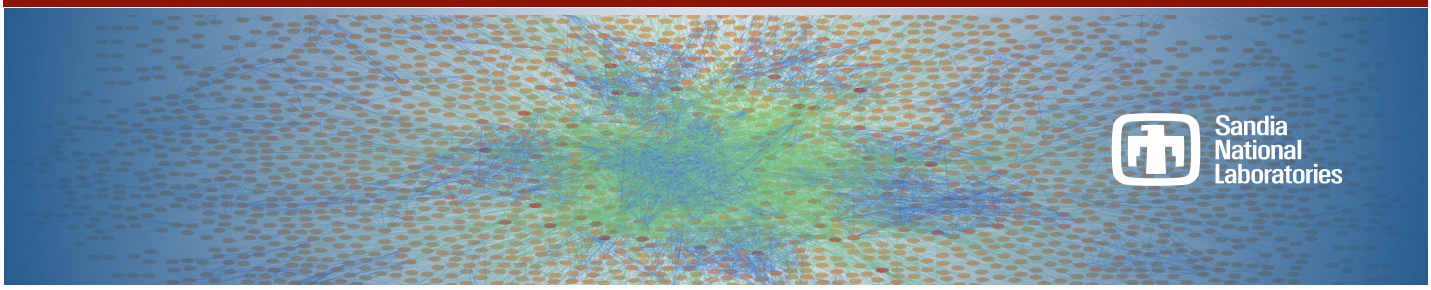
SAND2013-9482C



Scalable Models for Large Graphs

Why Model Graphs?

Enable sharing of surrogate data
- Computer network traffic
- Social networks
- Financial transactions

Testing graph algorithms
- Scalability
- Versatility
- Performance characterization
- Verification & validation

Insight into…
- Anomaly detection
- Generative process
- Community structure
- Comparison
- Evolution

Why Another Model?

Type	Deg. Dist.	Clust. Coeff.	Fitting	Scalable Generation?	Params
Incremental (PA, FF)	Qualitative	Qualitative	Expensive	No	Few
Markov Chain/ Rewiring (dK, 2.5K)	Exact	Near Exact for 2.5K	Compute DD/JDD and maybe CC	No	DD/JDD, plus CC for 2.5K
CL,EC	Near exact	No	Compute DD	Yes	DD
SKG/RMAT	No	No	Expensive	Yes	Few
BTER	Near exact	Near Exact	Compute DD & CC	Yes	DD & CC

Theory behind Block Two-level Erdös-Rényi (BTER) Model

Random graph:
(1) Formed according to CL Model
(2) "High" clustering coefficient

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

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.)

Chung-Lu (CL) Model

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

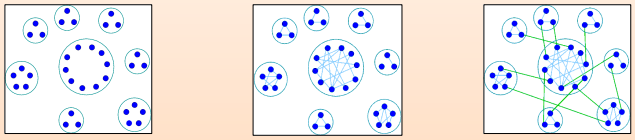
Global Clustering Coefficient

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

Dense Erdös-Rényi Subgraph

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

Theory describes the structure and enable generation



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

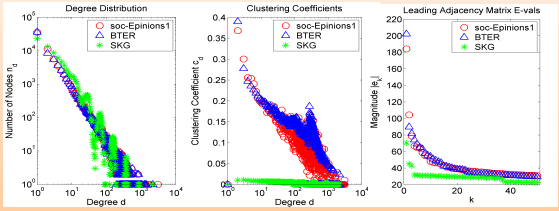
Phase 1
- Erdös-Rényi graphs in each block
- Need to insert extra links to insure enough *unique* links per block

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

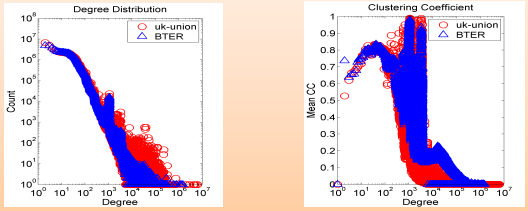
Phase 2
- CL model on excess degree (a sort of weighted Erdös-Rényi)
- Creates connections across blocks

Occurring independently

BTER can match properties of real world graphs



BTER is Scalable



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

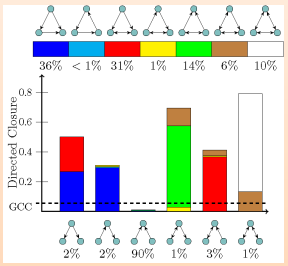
Modeling Directed Graphs

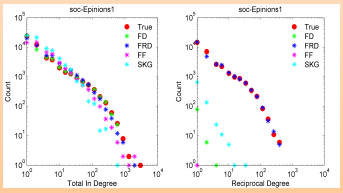
Degree Analysis

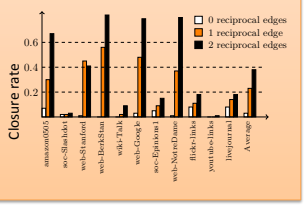
One-Way Edge Reciprocal Edge

Graph	# reciprocal edges
Soc-Epinions	0.405
Web-NotreDame	0.517
youtube	0.791
flickr	0.624
LiveJournal	0.735

Triadic Analysis







References:

- C. Seshadhri, A. Pinar, and T. G. Kolda. **An In-Depth Analysis of Stochastic Kronecker Graphs**, J. ACM, Vol. 60(2), pp:13:1--13:32, 2013.
- C. Seshadhri, T.G. Kolda, and A. Pinar. **Community structure and scale-free collections of Erdos-Renyi graphs**, Phys. Review E, Vol. 85(5), 2012.
- T. G. Kolda, A. Pinar, T. Plantenga, and C. Seshadhri, **A Scalable Generative Graph Model with Community Structure**, arXiv:1302.6636.
- C. Seshadhri, A. Pinar, N. Durak, and T.G. Kolda, **Directed closure measures for networks with reciprocity**, arXiv:1302:6220.
- N. Durak, T.G. Kolda, A. Pinar, and C. Seshadhri, **A Scalable Null Model to Match All Degree Distributions: In, Out, and Reciprocal,** Proc. IEEE Network Science, 2013.

U.S. DEPARTMENT OF ENERGY

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. SAND No. 2011-XXXXP