# Challenges of Modeling Directed Networks

Tamara G. Kolda, Ali Pinar, C. "Sesh" Seshadhri

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.

# UNDIRECTED GRAPHS

# A Good Graph Model…

- In an ideal world, encapsulates underlying driving principals
  - "Physics"
- Captures some measurable characteristics of real-world data
  - Degree distributions
  - Clustering coefficients
  - Community structure
  - Largest connected component size
  - Connectedness, Diameter
  - Eigenvalues
- Calibrates to specific data sets
  - Quantitative vs. qualitative
  - Surrogate for real data
  - Easy to share, reproduce results
- Ultimately, yields understanding
  - Serve as null model
  - Predictive capabilities

Today's assumptions:
unweighted, no loops, no multi-edges

Chung-Lu (aka Configuration) Model

$$\bar{d}_i = \text{desired degree of node } i$$

$$\bar{m} = \tfrac{1}{2}\sum_i \bar{d}_i$$

$$\text{Prob}\left((i,j) \in E\right) = \bar{d}_i \cdot \bar{d}_j / 4\bar{m}$$

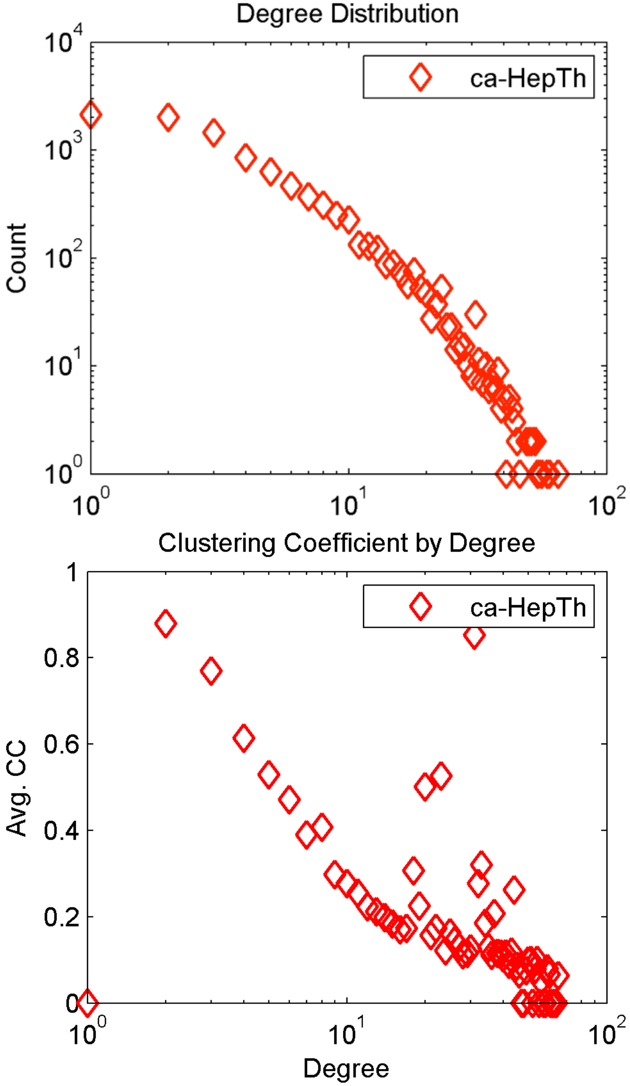
"Fast" Chung-Lu Model

$$\text{Prob}\left(i_k = i \mid e_k = (i_k, j_k)\right) = \bar{d}_i / 2\bar{m}$$

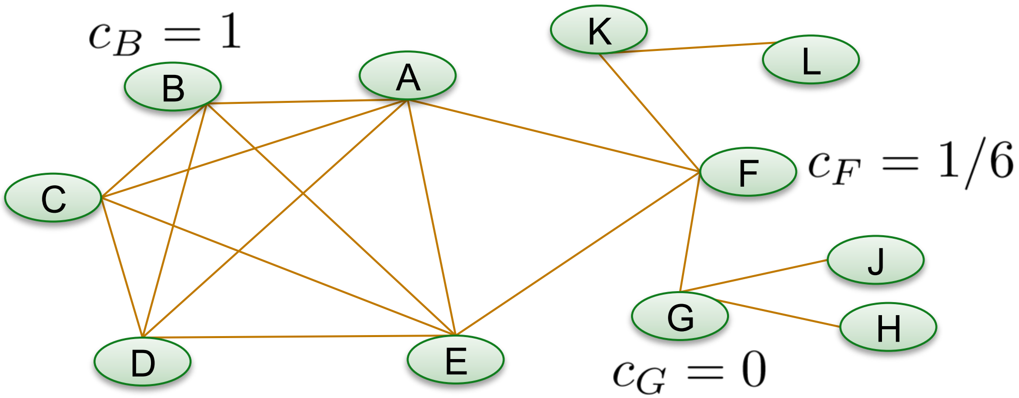
$$\text{Prob}\left(j_k = j \mid e_k = (i_k, j_k)\right) = \bar{d}_j / 2\bar{m}$$

# Goal for Undirected Graph: Match Degree Dist. & Clustering Coeffs. by Degree

Sandia National Laboratories




Recall: ***Clustering coefficient*** *measures rate of wedge closure*




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

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

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

# The Physics of Graphs

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

$$\Downarrow$$

*Thm:* This graph *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)

**Chung-Lu (aka Configuration) Model**

$$\bar{d}_i = \text{desired degree of node } i$$

$$\bar{m} = \tfrac{1}{2} \textstyle\sum_i \bar{d}_i$$

$$\text{Prob}\left((i,j) \in E\right) = \bar{d}_i \cdot \bar{d}_j / 4\bar{m}$$

**Global Clustering Coefficient**

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

**Dense Erdös-Rényi Subgraph**

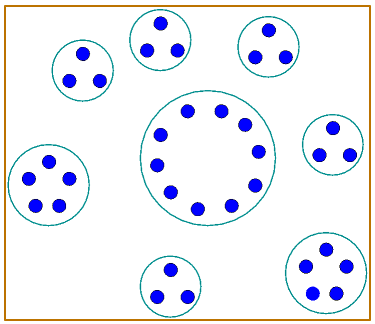
$$\bar{V} \subset V, \bar{E} \subset E$$

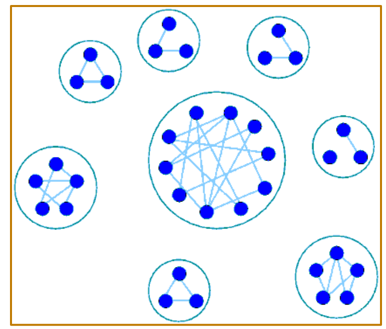
$$\text{Prob}\left((i,j) \in \bar{E} \mid i,j \in \bar{V}\right) \propto \text{constant}$$

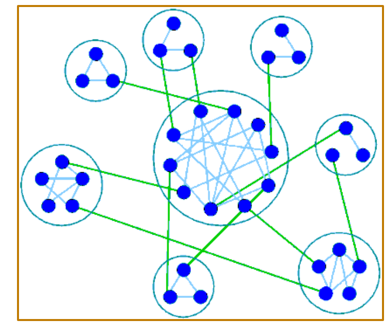
Seshadhri, Kolda, Pinar, *Phys. Rev. E,* 2012

# BTER: Block Two-Level Erdös-Rényi

Sandia National Laboratories







### 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

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

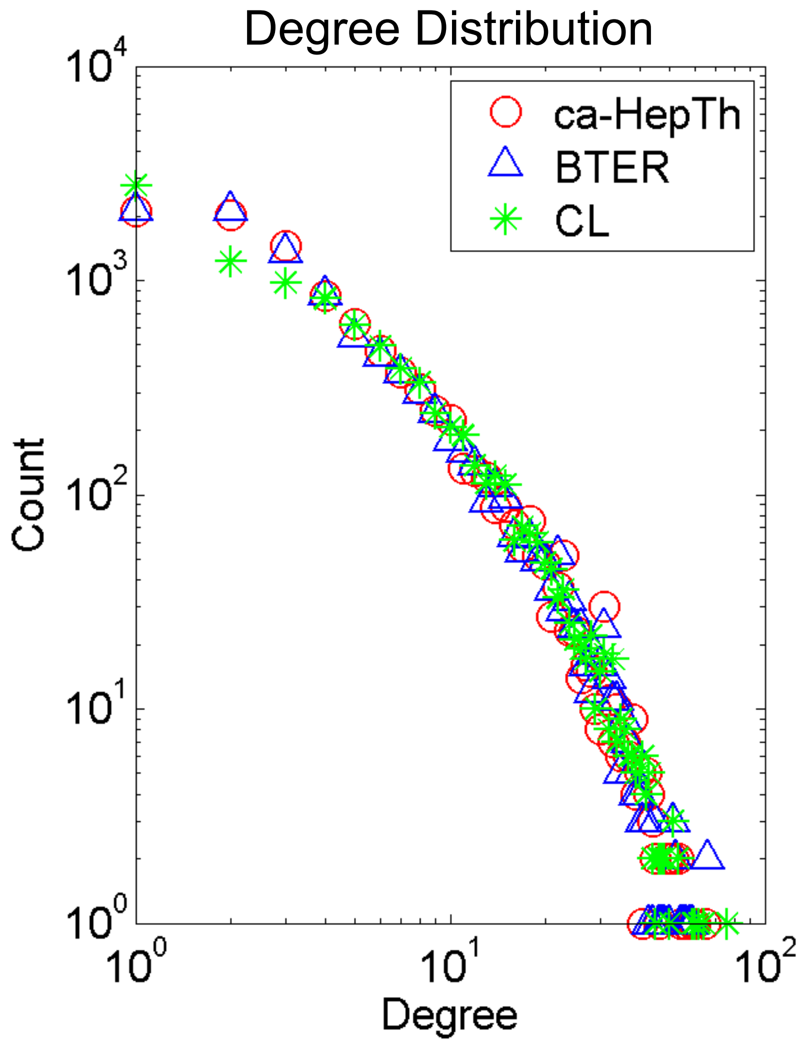
### Phase 2

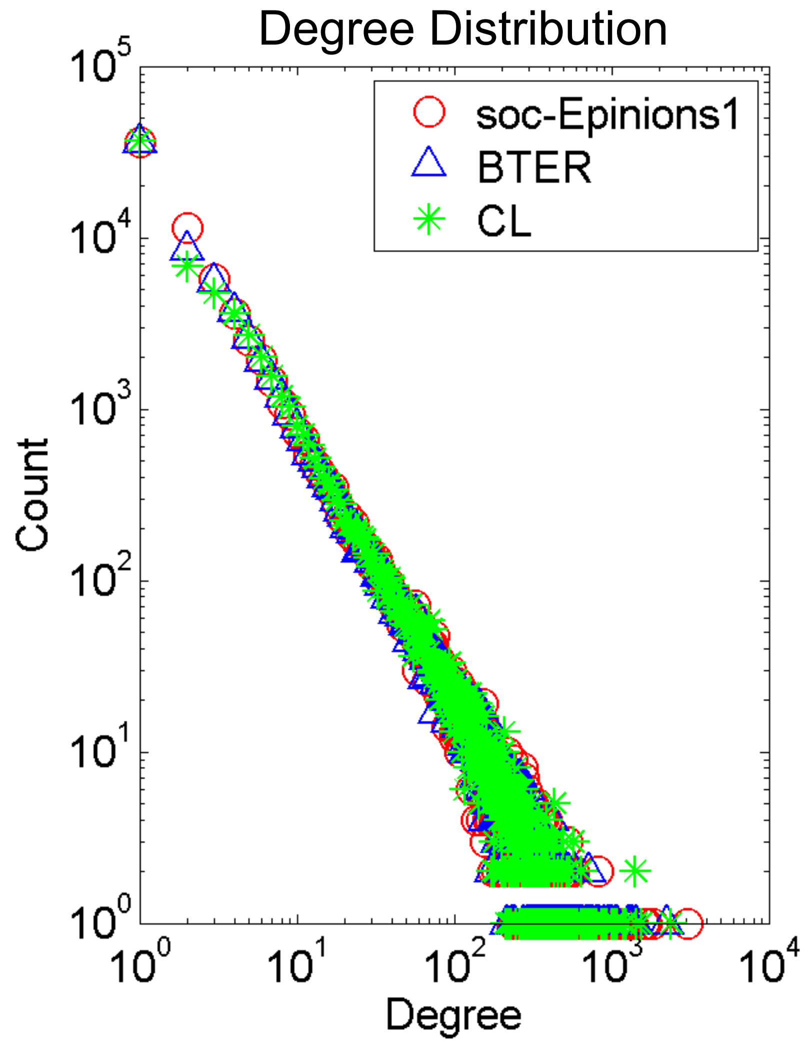
- CL model on excess degree (a sort of weighted Erdös-Rényi)
- Creates connections across blocks

*Occurring independently*

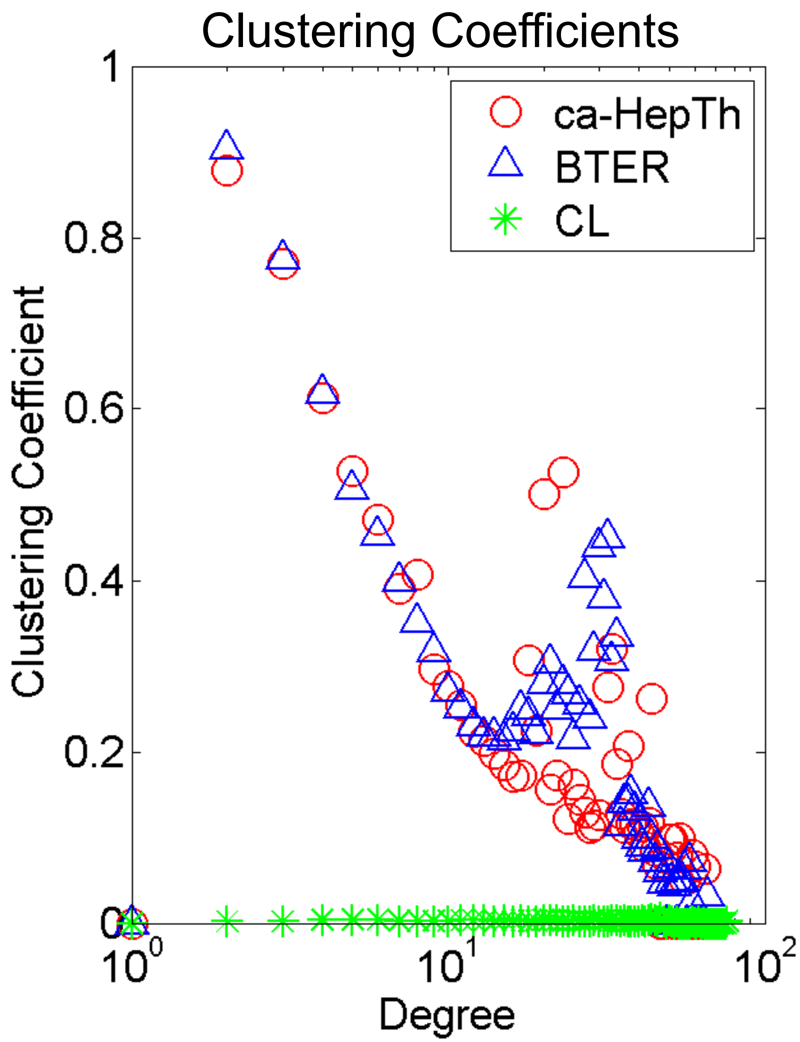
Seshadhri, Kolda, Pinar, *Phys. Rev. E,* 2012
Kolda, Plantenga, Pinar, Seshadhri, arXiv:1302.6636, Feb. 2013

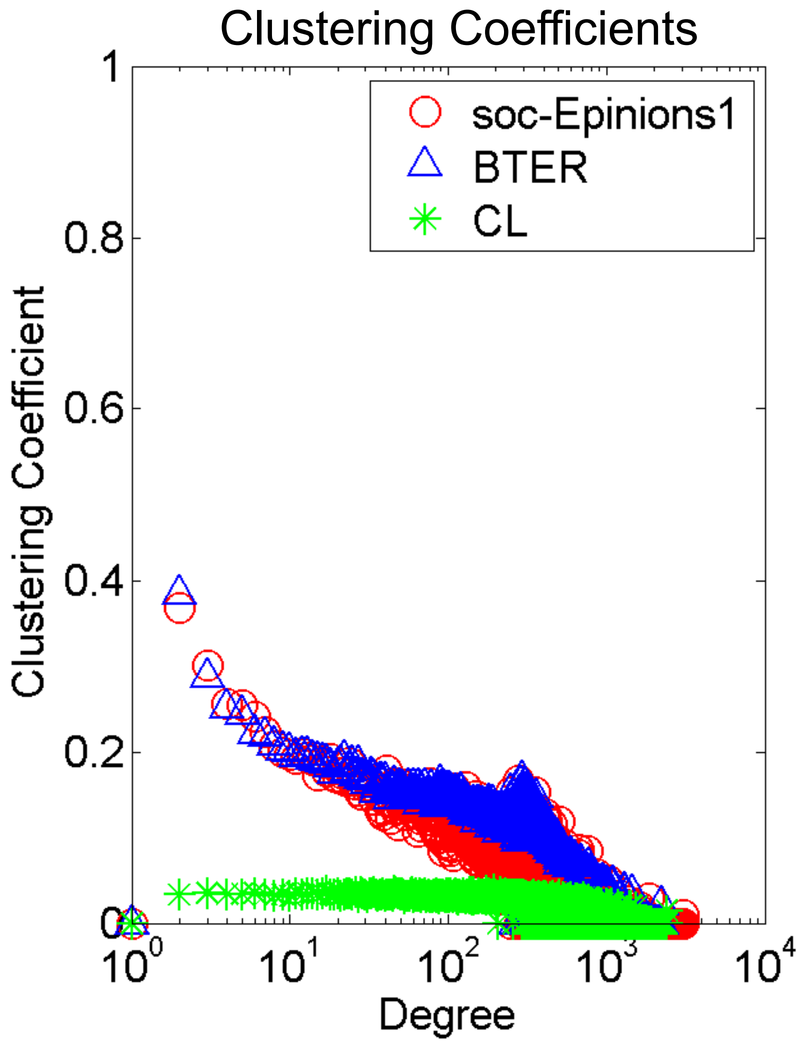
# Degree Distributions Captured by Both CL and BTER



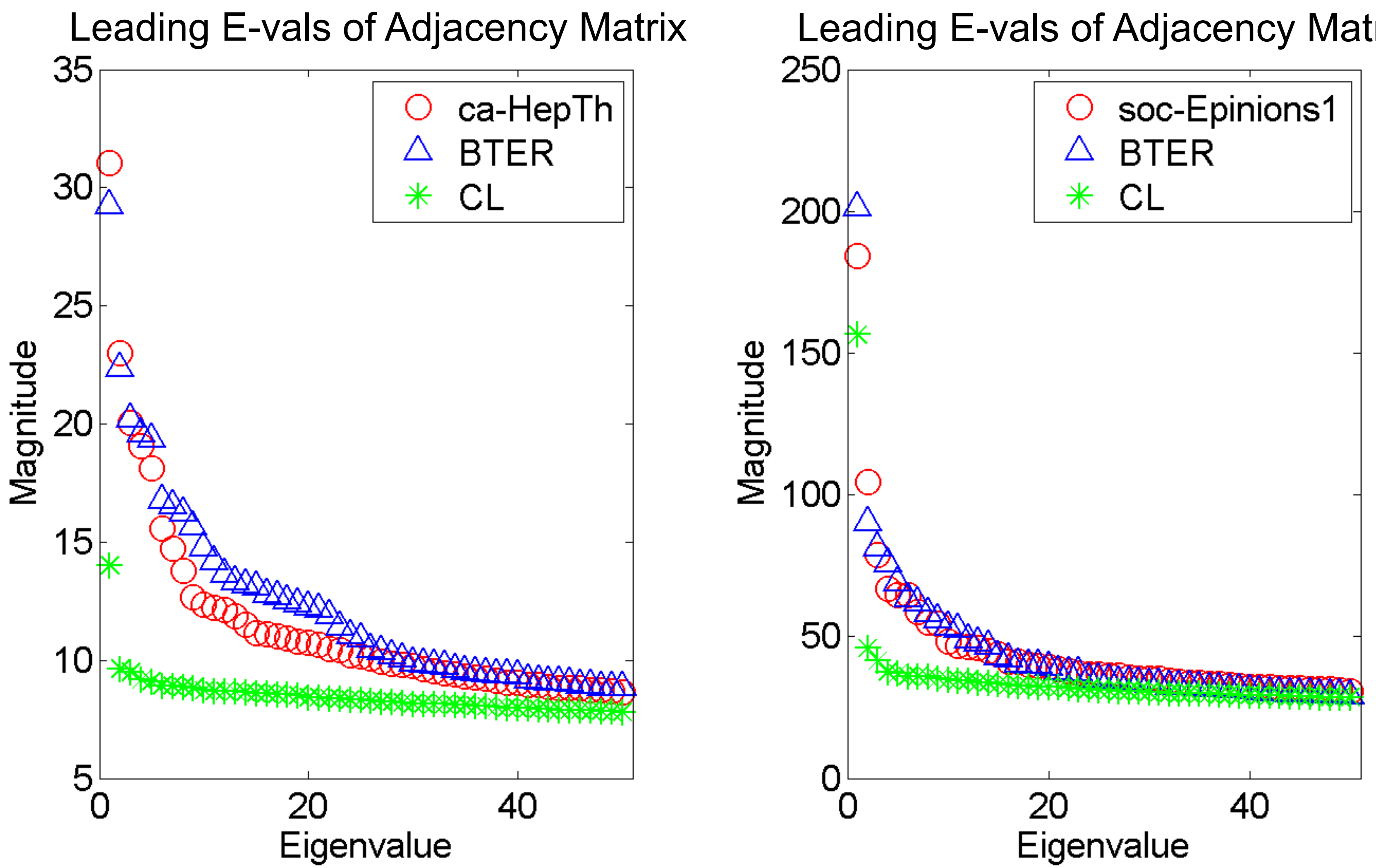



# Clustering Coefficients Captured by BTER, but not by CL

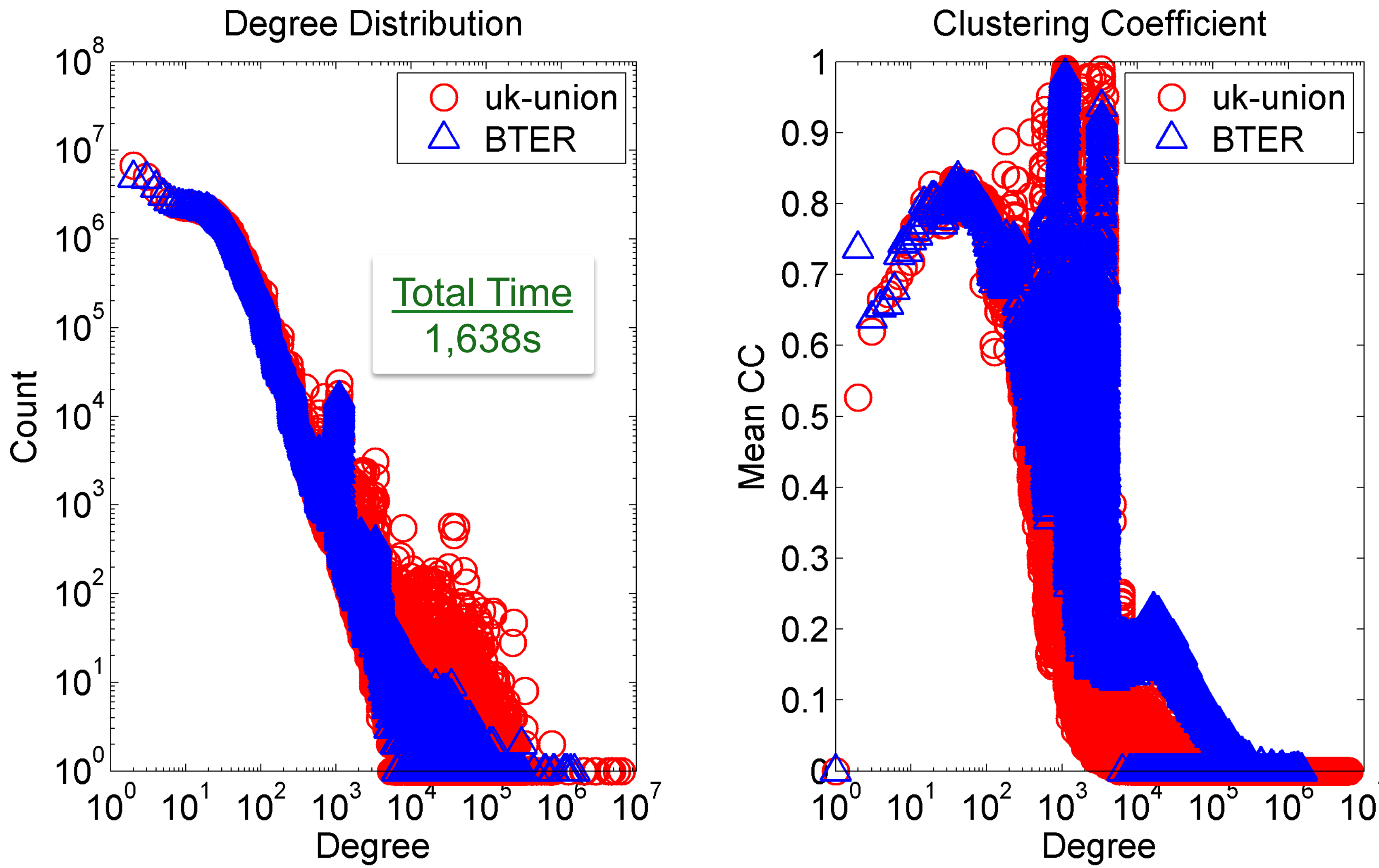


# Community Structure of BTER Improves Eigenvalue Fit versus CL

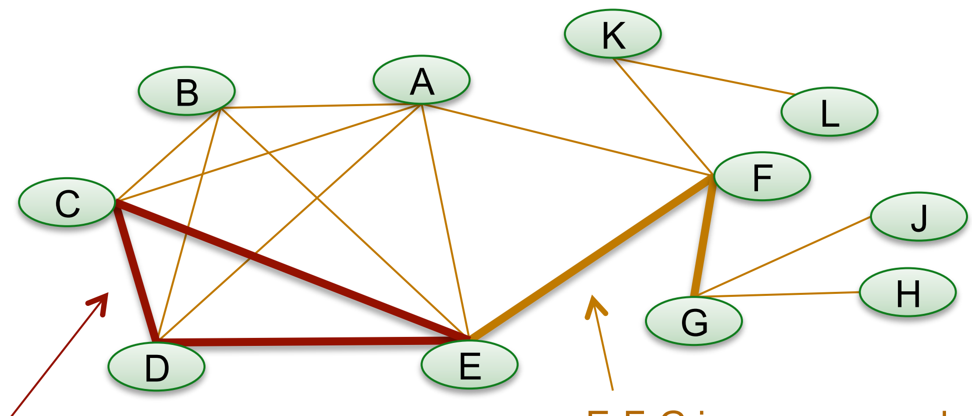


# FWIW, BTER Scales uk-union (4.6B edges)



# Clustering Coefficients for BIG Graphs

c = fraction of wedges that are closed




D-C-E is a closed wedge

E-F-G is an open wedge

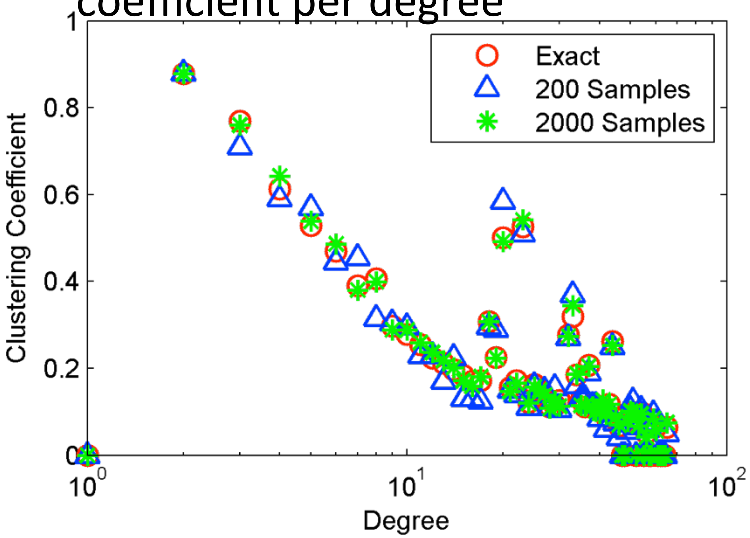
Enumeration: Find *every* wedge. Check if each is closed.
c = # closed wedges / # wedges

Sampling: Sample a few wedges (uniformly). Check if each is closed.
c ¼ # closed sampled wedges / # sampled wedges

Seshadhri, Pinar, Kolda, *Proc. 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 (² = 0.011)
- Faster than edge sampling (Doulion) and less variance
- Can also compute clustering coefficient per degree



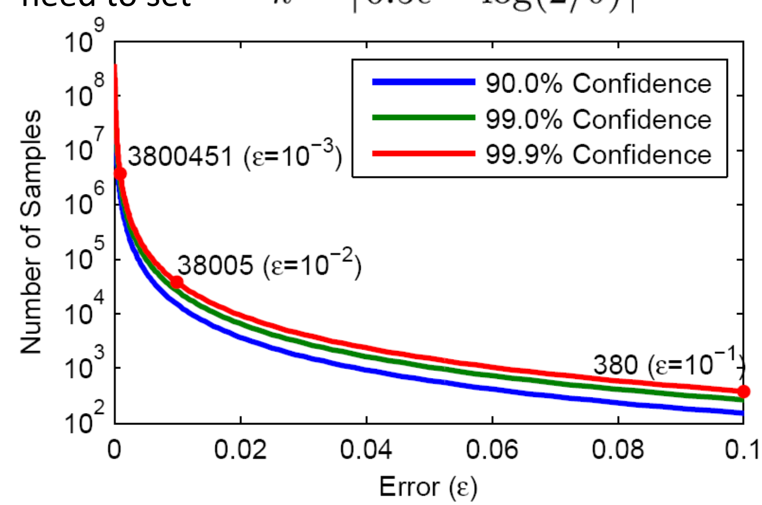

### Bounded Error: Hoeffding's Inequality

Theorem: (Hoeffding 1963) Let $X_1, X_2, \ldots, X_k$ 2 [0,1] be independent random variables. Define the sample mean: $\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$

Let ¹ be the true mean. Then for ² 2 (0,1-¹),

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

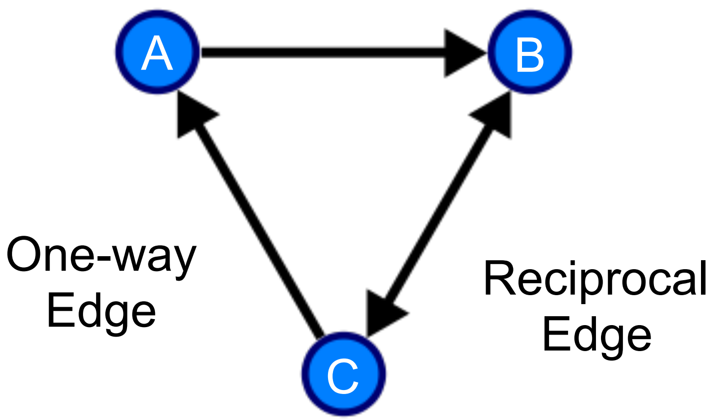
Hence, for a given error ² and confidence 1-δ, we just need to set $k = \lceil 0.5\epsilon^{-2}\log(2/\delta)\rceil$




Seshadri, Pinar, Kolda, *Proc. SIAM Intl. Conf. Data Mining*, 2013
Kolda, Pinar, Plantenga, Seshadri, Task, *arXiv:1301.5886*, Jan. 2013

# DIRECTED GRAPHS

# Two Ways of Measuring the Degree Distribution of a Digraph




| Node | Total In | Total Out |
|---|---|---|
| A | 1 | 1 |
| B | 2 | 1 |
| C | 1 | 2 |

$$d_i^{\Leftarrow} = d_i^{\leftarrow} + d_i^{\leftrightarrow} = \text{total in-degree}$$
$$d_i^{\Rightarrow} = d_i^{\rightarrow} + d_i^{\leftrightarrow} = \text{total out-degree}$$

| Node | In | Out | Recip. |
|---|---|---|---|
| A | 1 | 1 | 0 |
| B | 1 | 0 | 1 |
| C | 0 | 1 | 1 |

$$d_i^{\leftrightarrow} = \text{reciprocal degree}$$
$$d_i^{\leftarrow} = \text{in-degree}$$
$$d_i^{\rightarrow} = \text{out-degree}$$

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

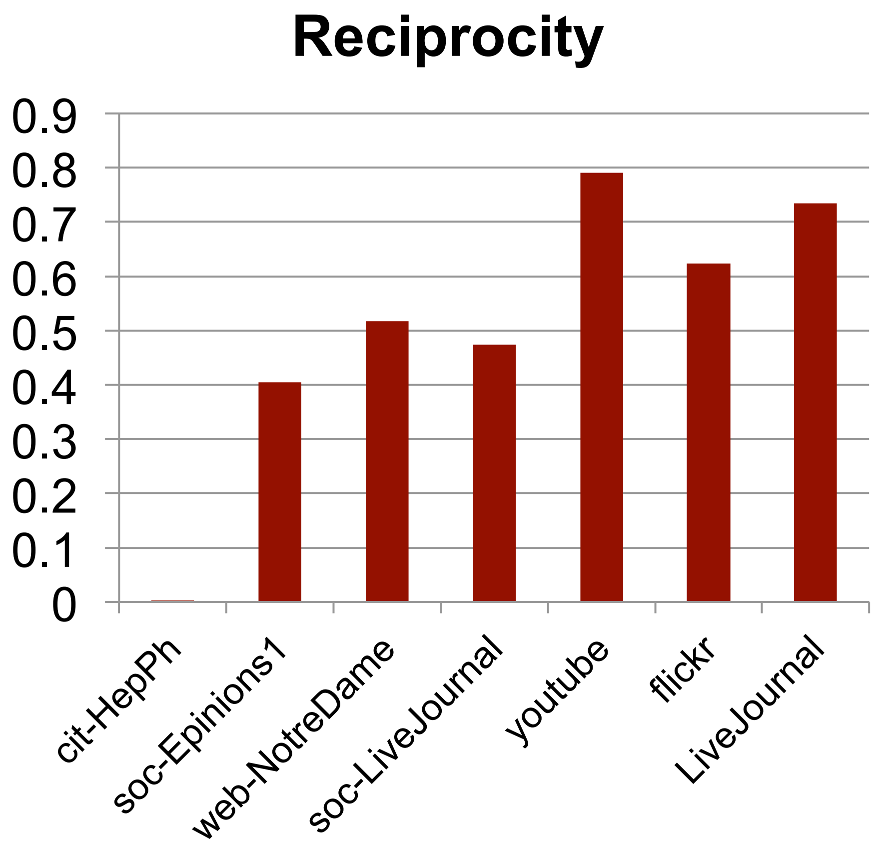
# Reciprocity is a Significant Behavior

$$n_d^{\leftrightarrow} = \# \text{ of nodes with reciprocal-degree } d$$
$$n_d^{\leftarrow} = \# \text{ of nodes with in-degree } d$$
$$n_d^{\rightarrow} = \# \text{ of nodes with out-degree } d$$

$$m = \sum_d d \cdot n_d^{\leftarrow} + d \cdot n_d^{\leftrightarrow} = \sum_d d \cdot n_d^{\rightarrow} + d \cdot n_d^{\leftrightarrow}$$

Reciprocity (Newman et al., 2002)

$$r = \frac{\# \text{ reciprocated edges}}{\# \text{ edges}} = \frac{\sum_{d=1}^{d_{\max}} d \cdot n_d^{\leftrightarrow}}{m}.$$



# Two Null Models

## Fast Directed (FD)

Each node is *randomly* assigned a desired total in- and out-degree.

$$\bar{d}_i^{\Leftarrow} = \text{desired total in-degree}$$
$$\bar{d}_i^{\Rightarrow} = \text{desired total out-degree}$$
$$\bar{m} = \sum_i \bar{d}_i^{\Leftarrow} = \#\text{ edges}$$

For $k = 1, \ldots, \bar{m}$

$\quad \text{Prob}\,(i_k = i \mid e_k = (i_k, j_k)) = \bar{d}_i^{\Rightarrow} / \bar{m}$

$\quad \text{Prob}\,(j_k = j \mid e_k = (i_k, j_k)) = \bar{d}_j^{\Leftarrow} / \bar{m}$

## Fast Reciprocal Directed (FRD)

Each node is *randomly* assigned a desired reciprocal-, in-, and out-degree.

$$\bar{d}_i^{\leftrightarrow} = \text{desired reciprocal degree}$$
$$\bar{d}_i^{\leftarrow} = \text{desired in-degree}$$
$$\bar{d}_i^{\rightarrow} = \text{desired out-degree}$$
$$\bar{m}' = \tfrac{1}{2} \sum_i \bar{d}_i^{\leftrightarrow} = \#\text{ recip. edges}$$
$$\bar{m}'' = \sum_i \bar{d}_i^{\leftarrow} = \#\text{ one-way edges}$$

For $k = 1, \ldots, \bar{m}'$

$\quad \text{Prob}\,(i_k = i \mid e_k = (i_k, j_k)) = \bar{d}_i^{\leftrightarrow} / \bar{m}'$

$\quad \text{Prob}\,(j_k = j \mid e_k = (i_k, j_k)) = \bar{d}_j^{\leftrightarrow} / \bar{m}'$

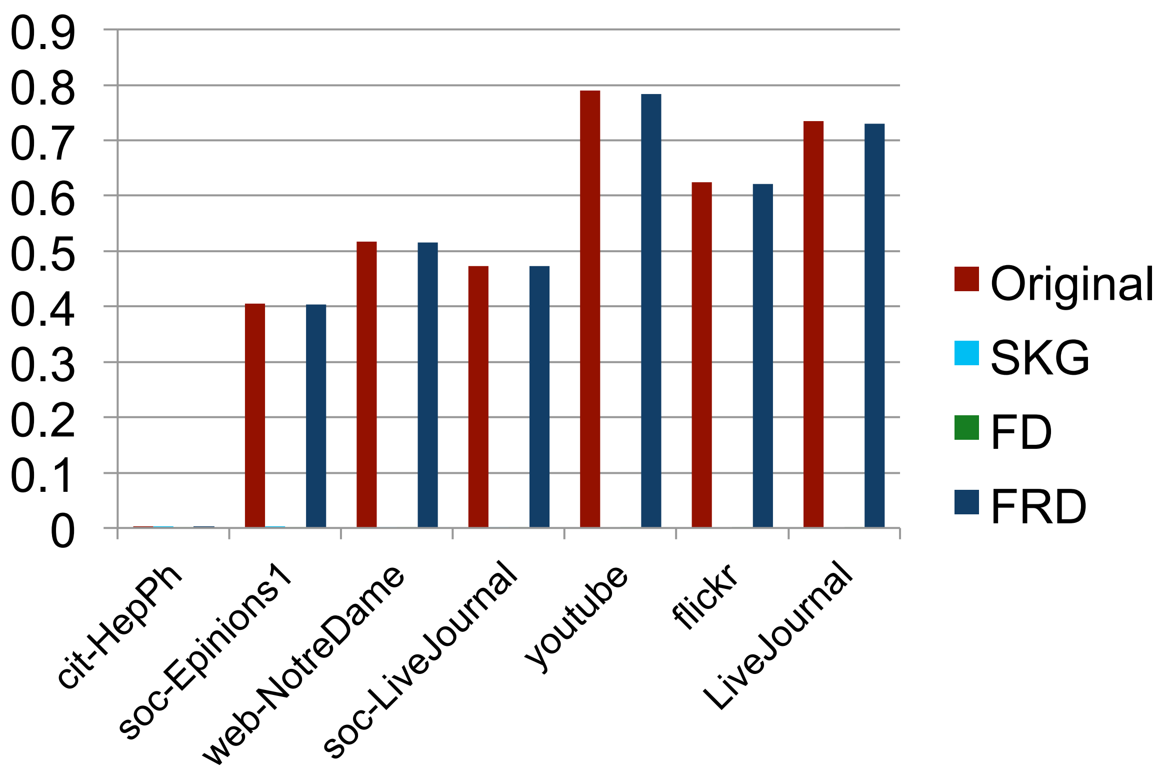
For $k = 1, \ldots, \bar{m}''$

$\quad \text{Prob}\,(i_k = i \mid e_k = (i_k, j_k)) = \bar{d}_i^{\rightarrow} / \bar{m}''$

$\quad \text{Prob}\,(j_k = j \mid e_k = (i_k, j_k)) = \bar{d}_j^{\leftarrow} / \bar{m}''$

# Models fail at capturing reciprocity

- SKG has very little reciprocation (Lescovec et al., JMLR, 2010)
- Forest Fire (FF) has <u>no</u> reciprocation (Leskovec, Kleinberg, Faloutsos, KDD'05)
- FD corresponds to a fast implementations of CL
- FRD takes reciprocal edges into account

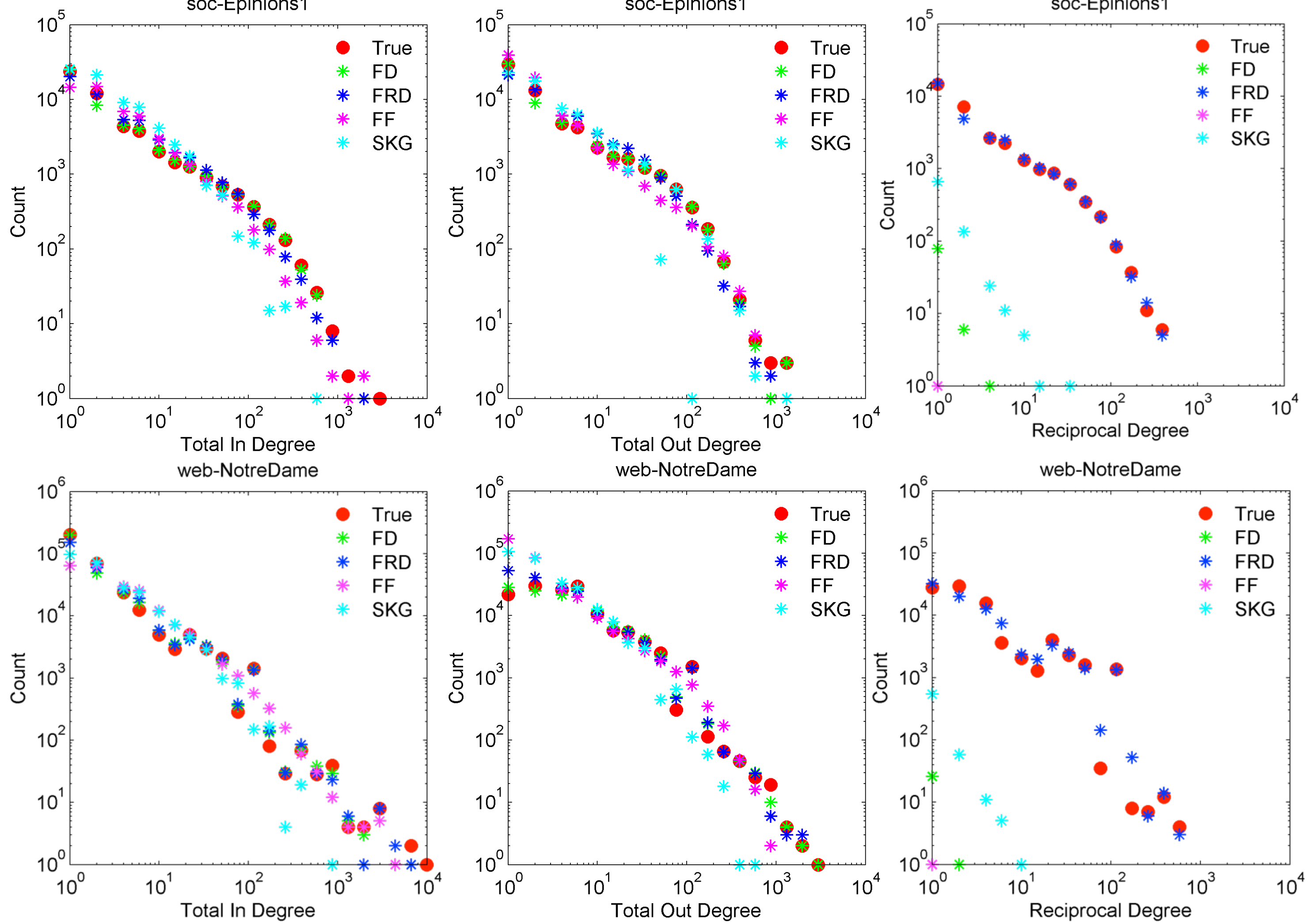


# Tough to Match Reciprocal Degree

Soc-Epinions
|V|= 76K
|E|=508K
r=0.405

Web-NotreDame
|V|=325K
|E|=1469K
r=0.517

**Total In-Degree** | **Total Out-Degree** | **Reciprocal Degree**

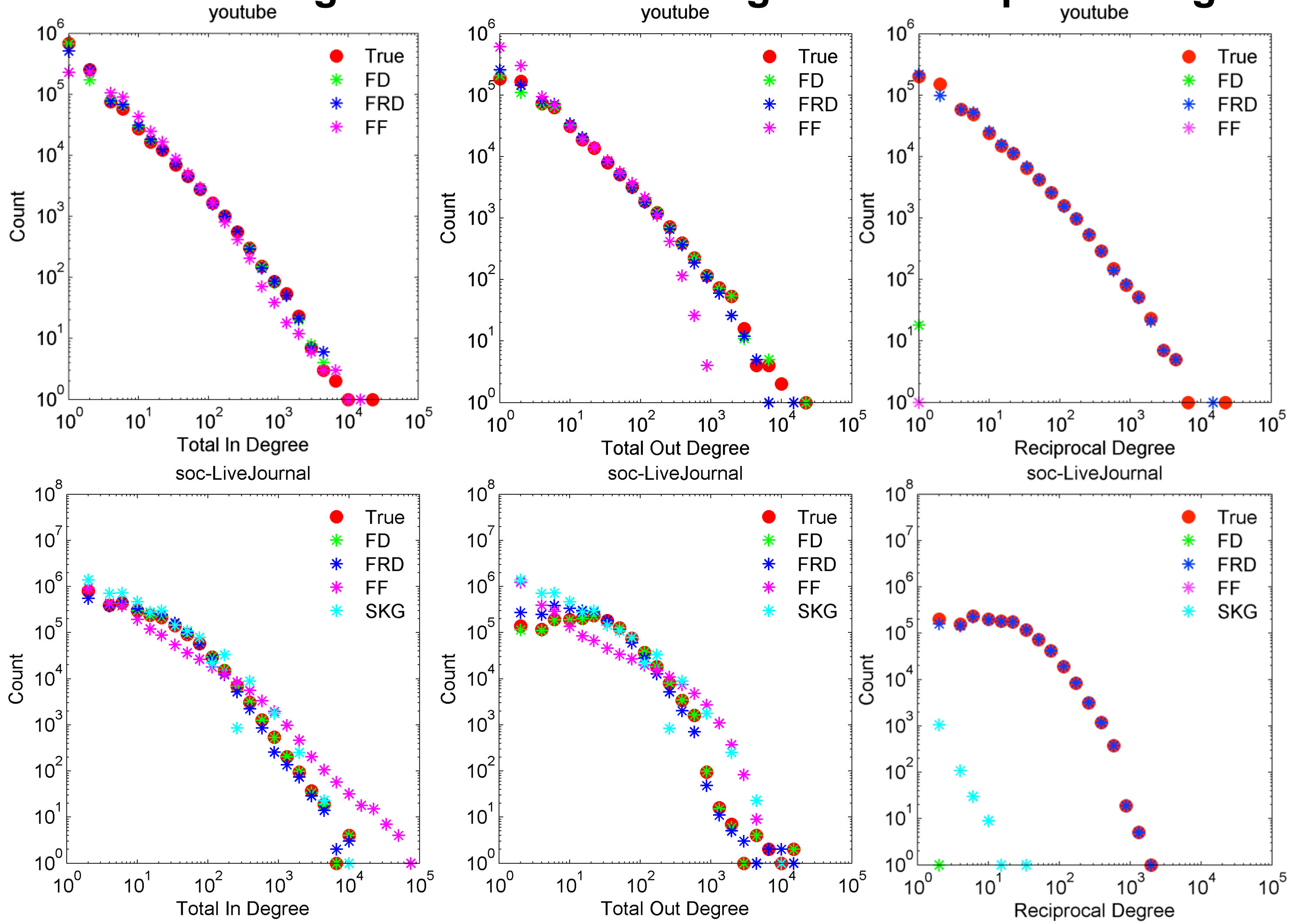


# Tough to Match Reciprocal Degree

youtube
|V|=1157K
|E|=4945K
r=0.791

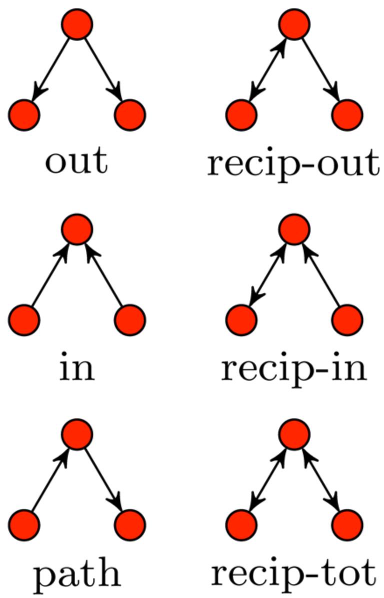
Soc-LiveJour
|V|=4847K
|E|=68475K
r=0.632

**Total In-Degree** | **Total Out-Degree** | **Reciprocal Degree**

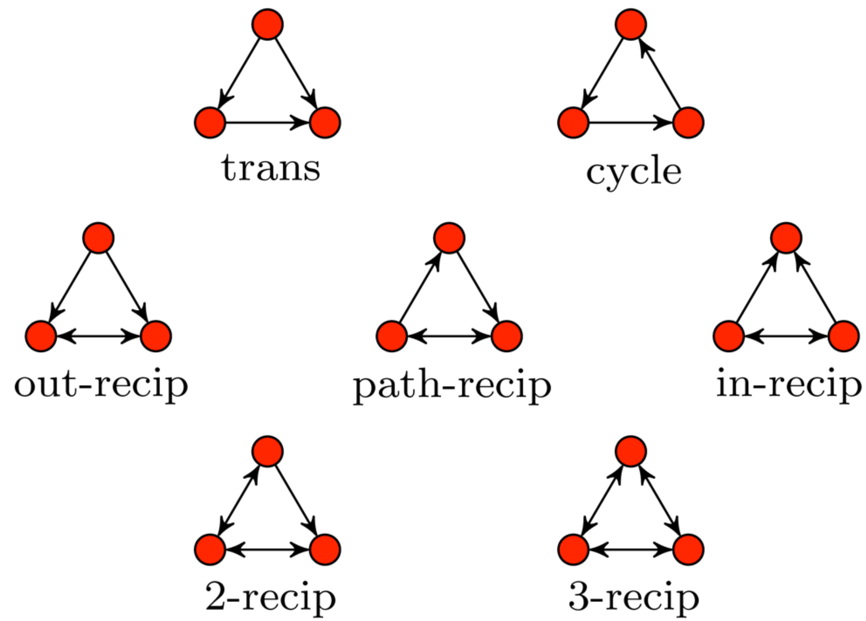


# Triangles in Directed Networks

## Directed Wedges



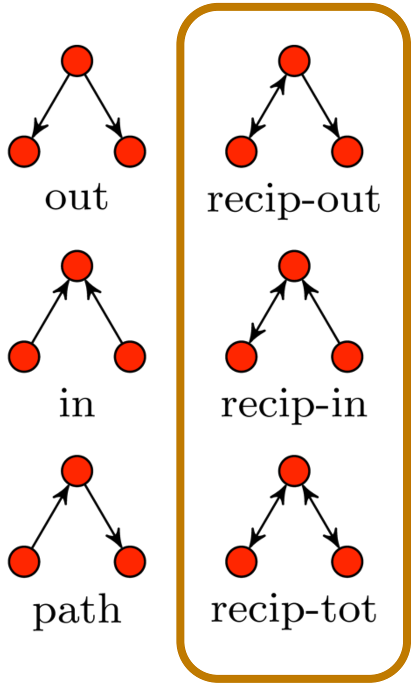

## Directed Triangles



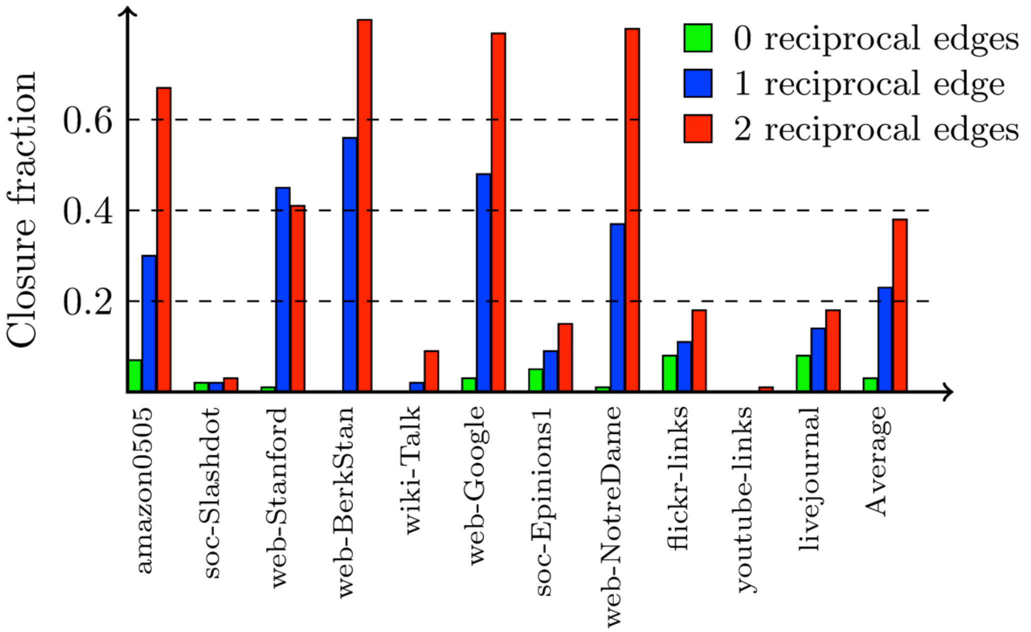

Seshadhri, Pinar, Durak, Kolda, arXiv:1302.6220, 2013

# Reciprocity and Wedge Closure

## Directed Wedges



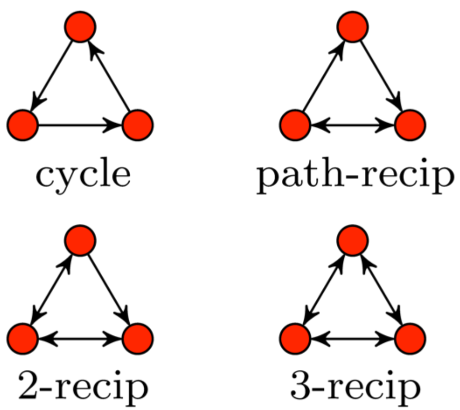

Wedges with reciprocal edges are much more likely to close in social and web networks.



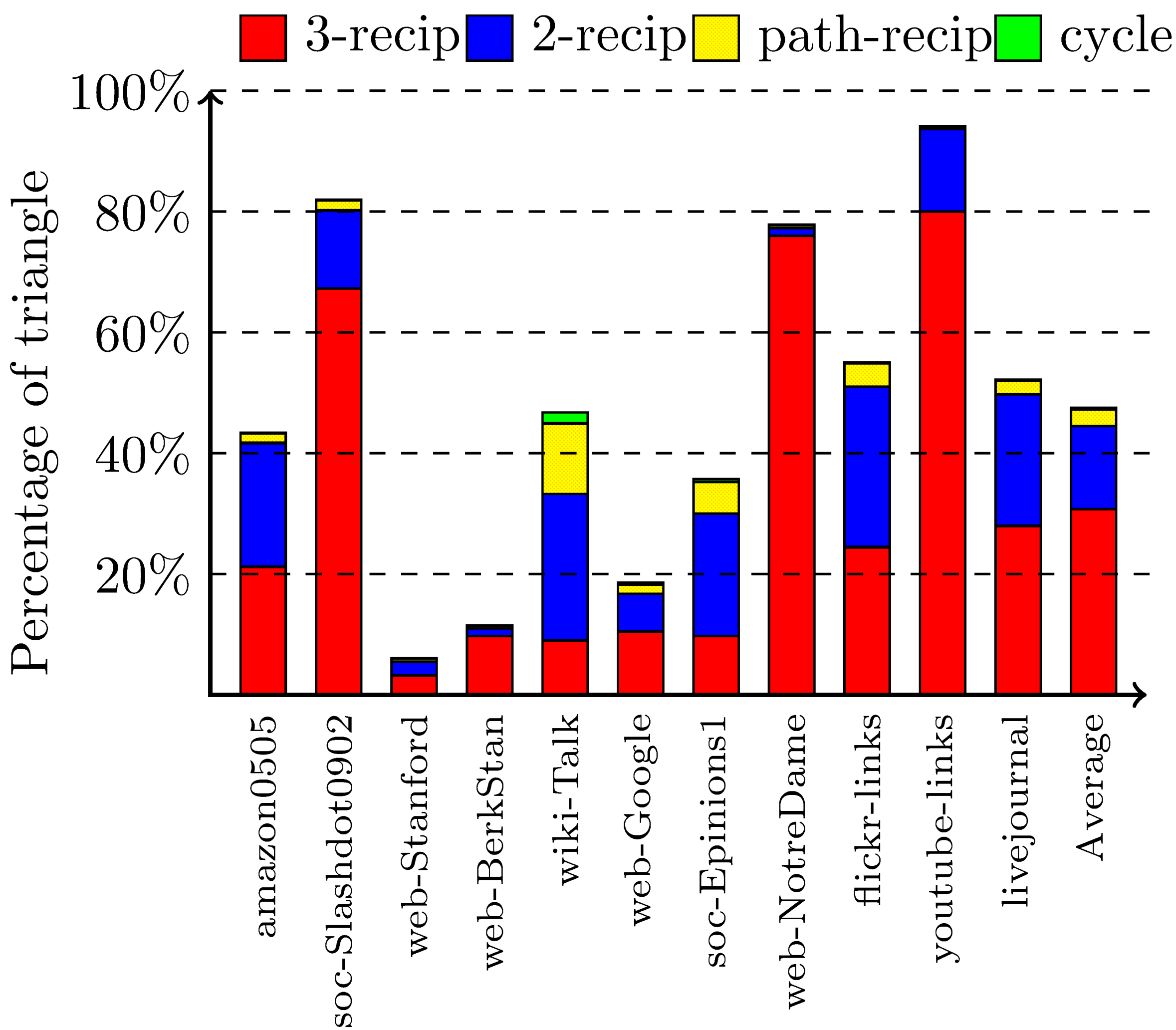

Seshadhri, Pinar, Durak, Kolda, arXiv:1302.6220, 2013

# Reciprocity and Cycles

## Triangles with Cycles



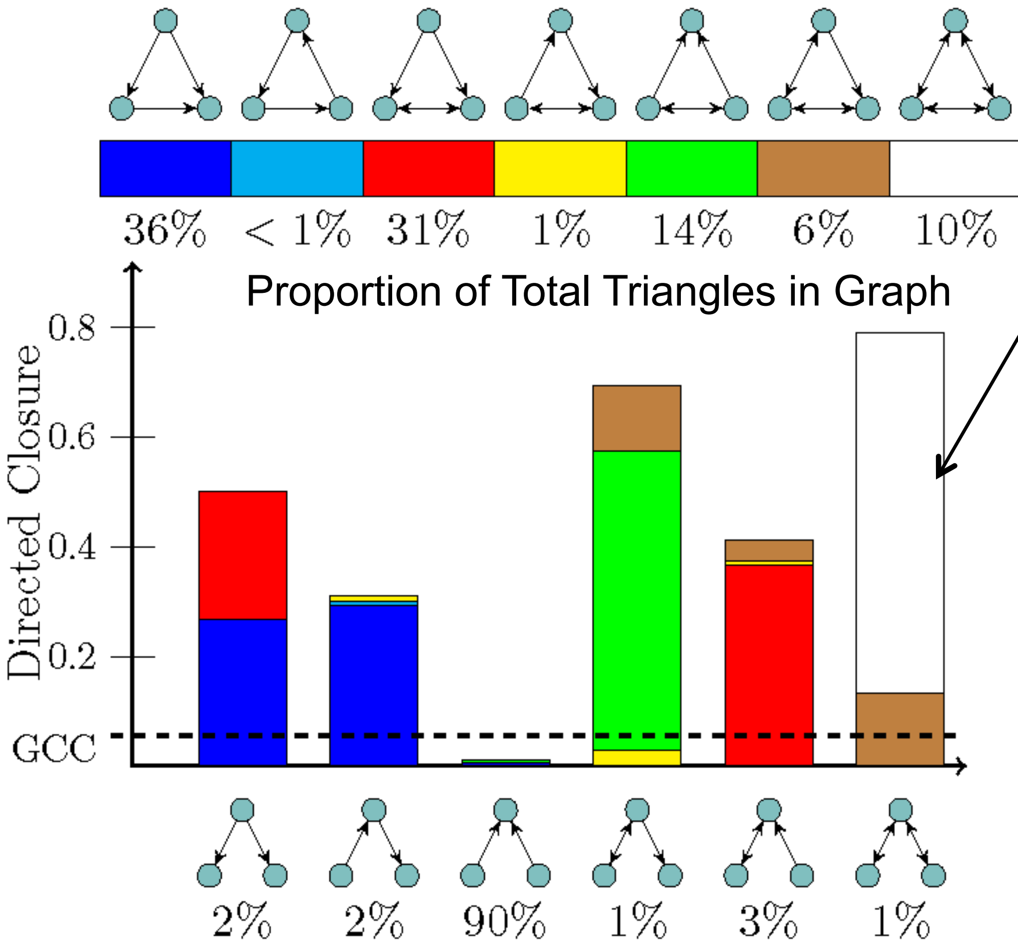

Cycles w/o reciprocation exceedingly rare




Seshadhri, Pinar, Durak, Kolda, arXiv:1302.6220, 2013

# Wedges and Triangles in Web Network: web-Google

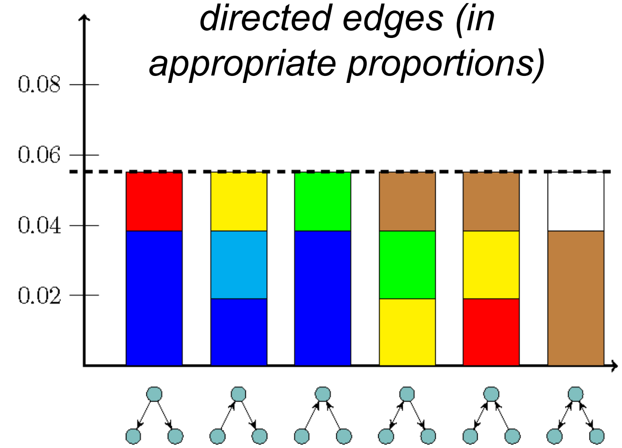
web-Google: 876K nodes, 5.1M edges, reciprocation = 31%, GCC=0.055




Proportion of Total Wedges in Graph

Proportion of these wedges that closed into that color triangle.

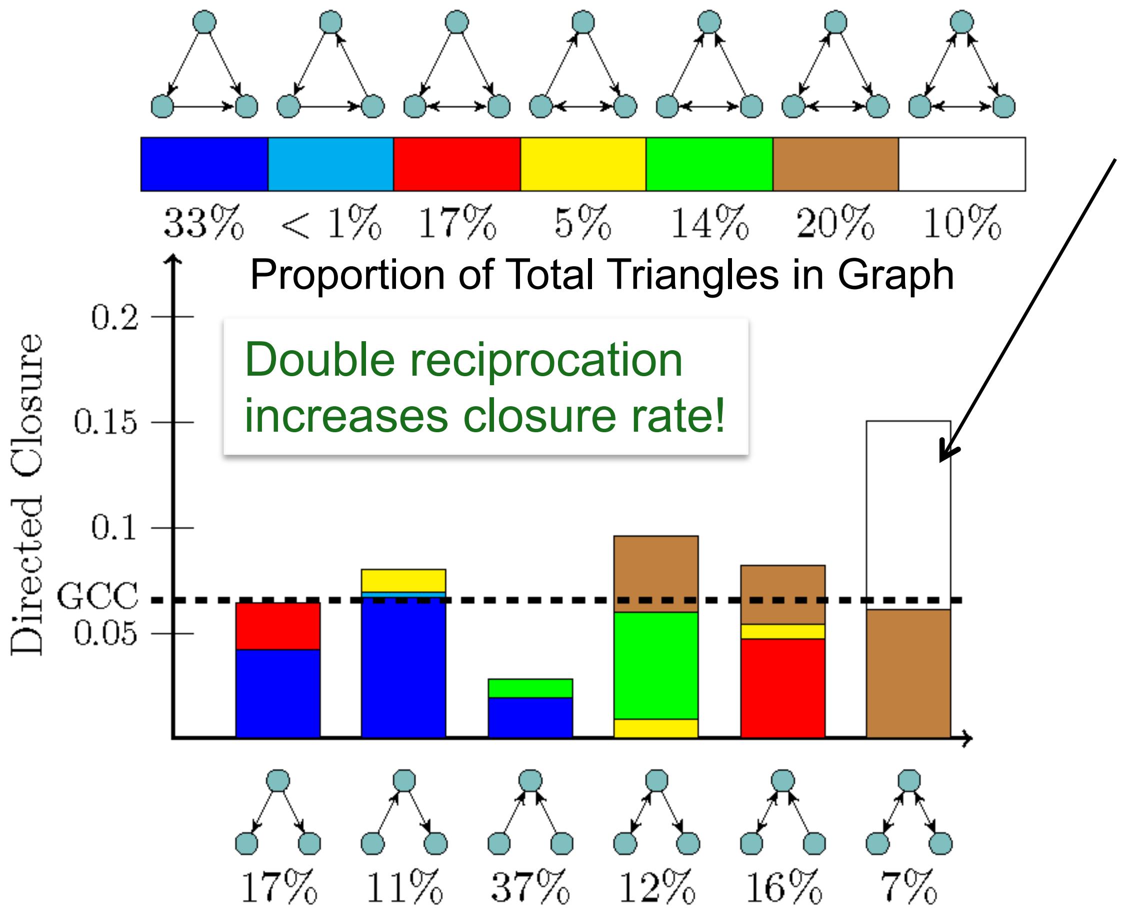
*Contrast to proportions for a graph with randomly directed edges (in appropriate proportions)*




Data from SNAP

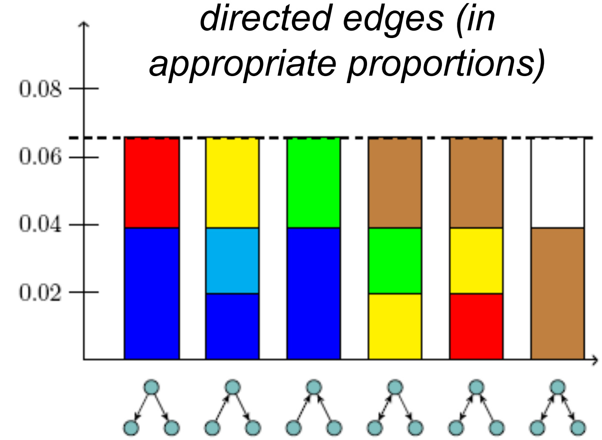
# Wedges and Triangles in Social Network: soc-Epinions-1

soc-Epinions1: 76K nodes, 509K edges, reciprocation = 41%, GCC=0.066




Proportion of these wedges that closed into that color triangle.

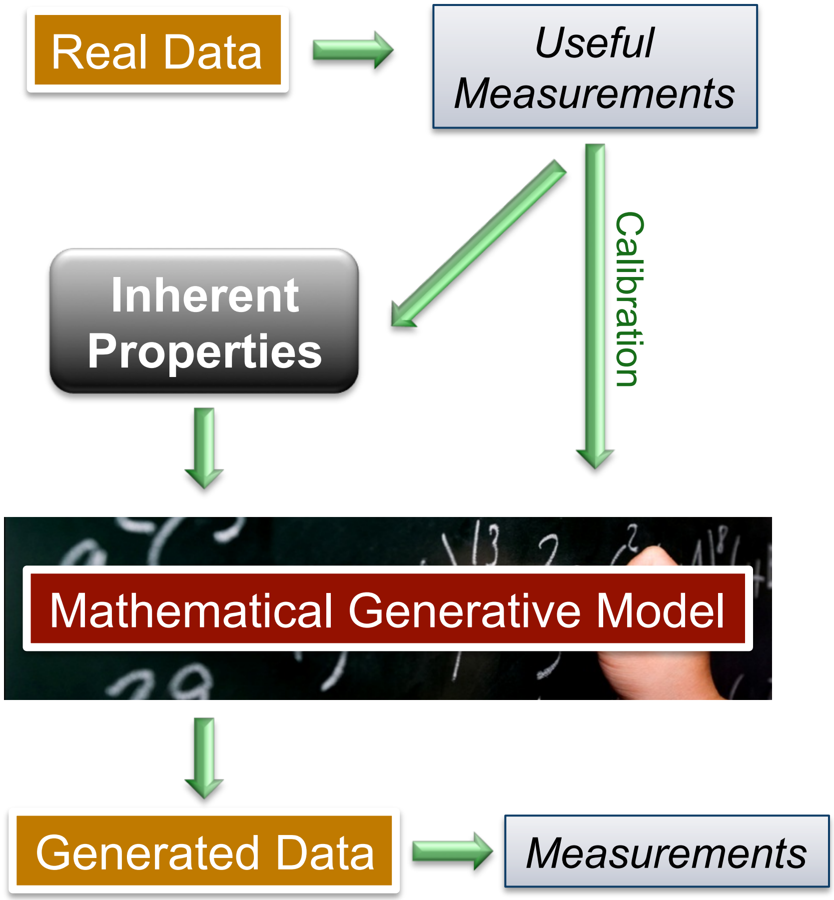
*Contrast to proportions for a graph with randomly directed edges (in appropriate proportions)*




Data from SNAP

# Conclusions

- Simple graphs
  - Useful measurements
    - Degree distribution
    - Clustering coefficients
  - Generative Models
    - CL matches degree distribution by not cluster coefficients
    - BTER matches degree distribution and clustering coefficients
- Digraphs
  - Useful measurements
    - Various degree distributions
    - Various directed clustering coefficients
  - Models
    - Most ignore reciprocation
    - FRD model matches degree distributions
    - No model yet matches triangle behavior



# References


- 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
- 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


- *For copies or information about job openings: Tammy Kolda, tgkolda@sandia.gov*