Sandia National Laboratories

SAND2013-1658C

The Inherent Community Structure in Real-World Graphs

Ali Pinar, C. Seshadhri, and Tamara G. Kolda

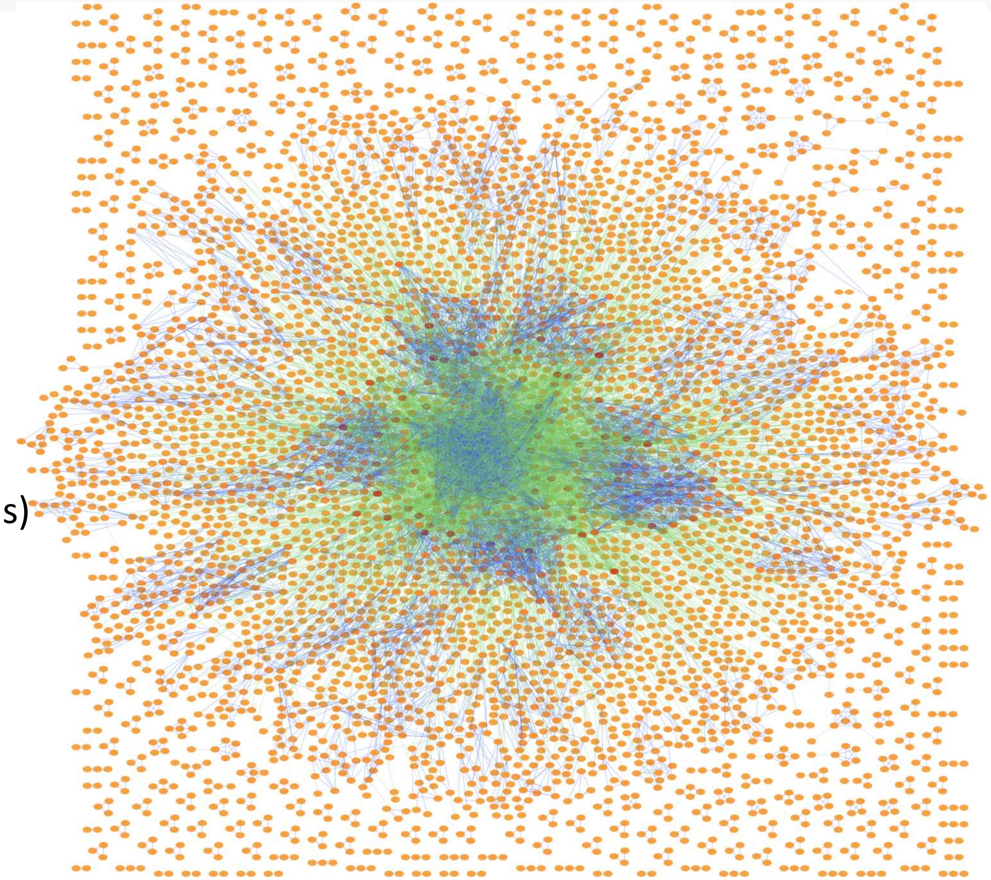
Sandia National Labs

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

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.

Why generate random graphs?

- Enable sharing of surrogate data
  - Computer network traffic
  - Social networks
  - Financial transactions
- Statistical analysis
  - Sample from a specified space
- Testing graph algorithms
  - Scalability
  - Versatility (e.g., vary degree distributions)
  - Characterizing algorithm performance
- Insight into…
  - Generative process
  - Community structure
  - Comparison
  - Evolution
  - Uncertainty



Block Two-Level Erdös-Rényi (BTER) graph; image courtesy of Nurcan Durak.

Markov Chains: common method to generate random graphs

- For this talk, a Markov chain is a graph whose nodes are realizations of a graph.
- Framework:
  - Find an arbitrary node of an MC
  - Take a loooong random walk
  - You will arrive at a uniform random walk
    - if certain conditions are satisfied.
- Challenges
  - Generating a graph with given properties
  - Rewiring a graph to preserve desired features
  - Patience

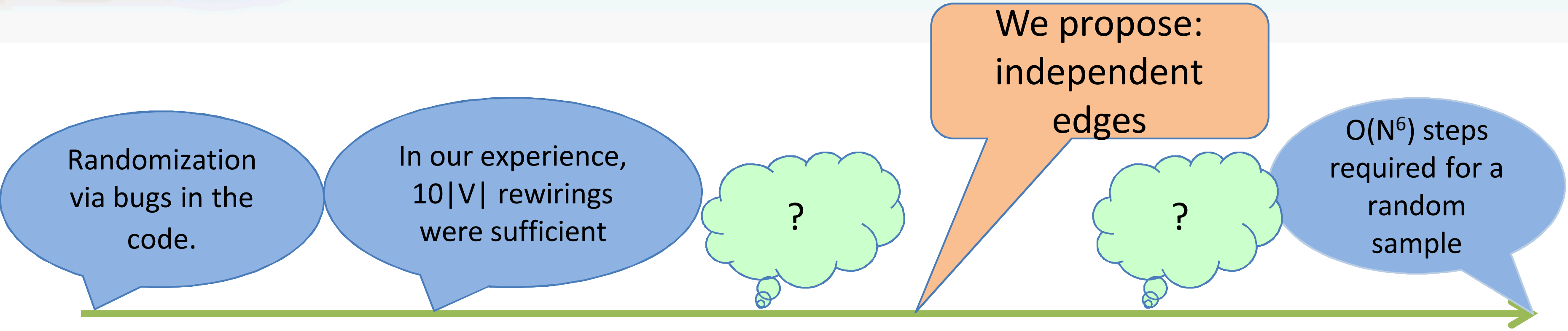
Convergence is a problem

[Source: http://metsmerizedonline.com/wp-content/uploads/2013/02/Are-We-There-Yet.jpg](http://metsmerizedonline.com/wp-content/uploads/2013/02/Are-We-There-Yet.jpg)

Can we find principled and practical metrics for convergence?

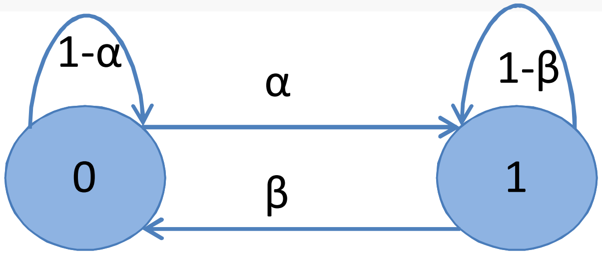
- In theory, we need to prove the stationary distribution of the MC is uniform.
- In practice, bounds for convergence may be impractical or nonexistent.
- Practitioners use unprincipled methods.
  - e.g., 10,000 rewiring operatiions
- Interpretations of statistical tools may be hard.
  - What does Gelman Rubin test mean from a graphs perspective?

Can we find principled and practical metrics for convergence?



- What is a mathematically sound definition of random enough?
- Goals: practical, sound, and interpretable.
- An imperfect analogy:
  - To solve $Ax=b$, we do not compute $A^{-1}b$, we compute an x, that yields a small residual, $Ax-b$.
  - We have done quite well living with this.

Testing independence of edges



$$T = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

α: probability that the edge will be inserted

β: probability that the edge will be deleted

T: transition matrix of the edge

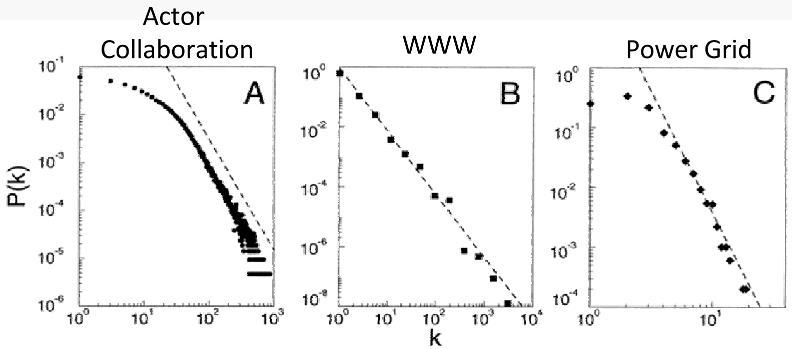
- Assume the addition/deletion of an edge can be approximated as a Markov process.
- The full Markov chain (MC) can be approximated as a collection of smaller Markov chains.
- Convergence of smaller MCs is a necessary condition for convergence of the full MC.

Convergence of smaller Markov chains

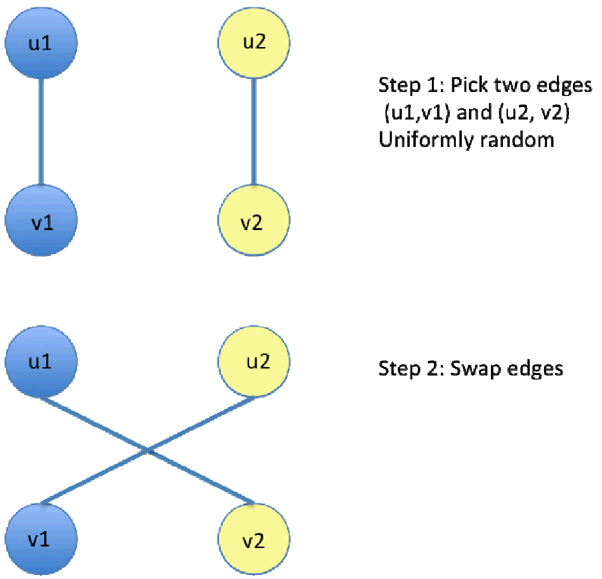
- Eigenvalues of *T* are *1* and *1 –(α+β)*.
  - Eigenvalues form a basis, so initial state *v* can be written as $v=c_1e_1+c_2e_2$.
  - After N iterations, we have $p = T^N v = c_1 e_1 + c_2 (1-(\alpha+\beta))^N e_2$
  - The second term decays and p converges to c_1e_1.
  - This vector c_1e_1 indicates the probability the edge is present/absent in a random graph.
  - For tolerance *ε,* the number of iterations required, *N*, is

$$N = \ln(1/\varepsilon)/(\alpha+\beta)$$

Preserving the degree distributions

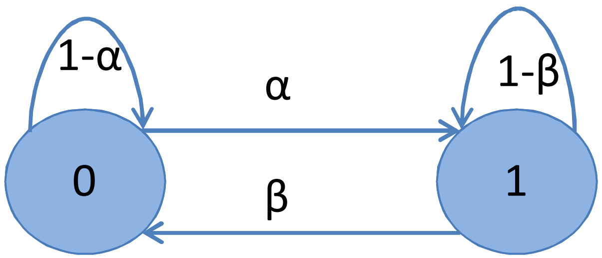


A.-L. Barabasi and R. Albert. Emergence of scaling
in random networks. Science, 286(5349):509-512, 1999.



- Degree distribution is like a histogram of degrees.
- It is one of the critical features that distinguish real graphs from arbitrary sparse graphs.
- Rewiring scheme has long been used to perturb graphs while preserving the degree distribution.
  - Converges in $O(|E|^6)$-time.
- Havel and Hakimi described the first algorithm to construct a graph with a given degree distribution.

Transition matrix for preserving degree distribution



d_u: degree of vertex u

m: total number of edges

α: probability that the edge will be inserted

β: probability that the edge will be deleted

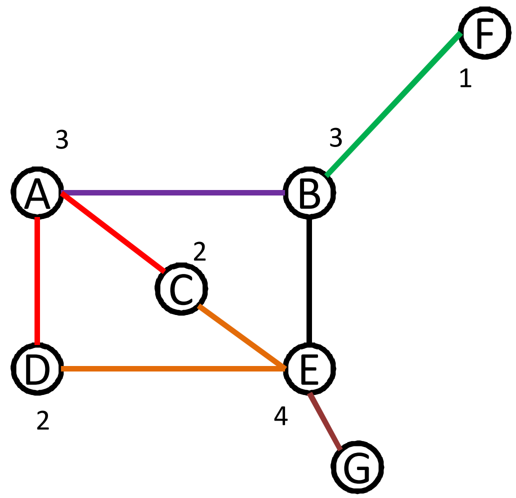
$$\alpha = \frac{d_u d_v}{2m^2} \qquad \beta = 1 - \left(1 - \frac{1}{m}\right)^2$$

$$N = \ln(1/\varepsilon)/(\alpha + \beta)$$

$$\alpha + \beta \cong \frac{2}{m}$$

To generate a graph with independent edges with a specified degree distribution we need $N = \frac{m}{2} \ln \frac{1}{\varepsilon}$

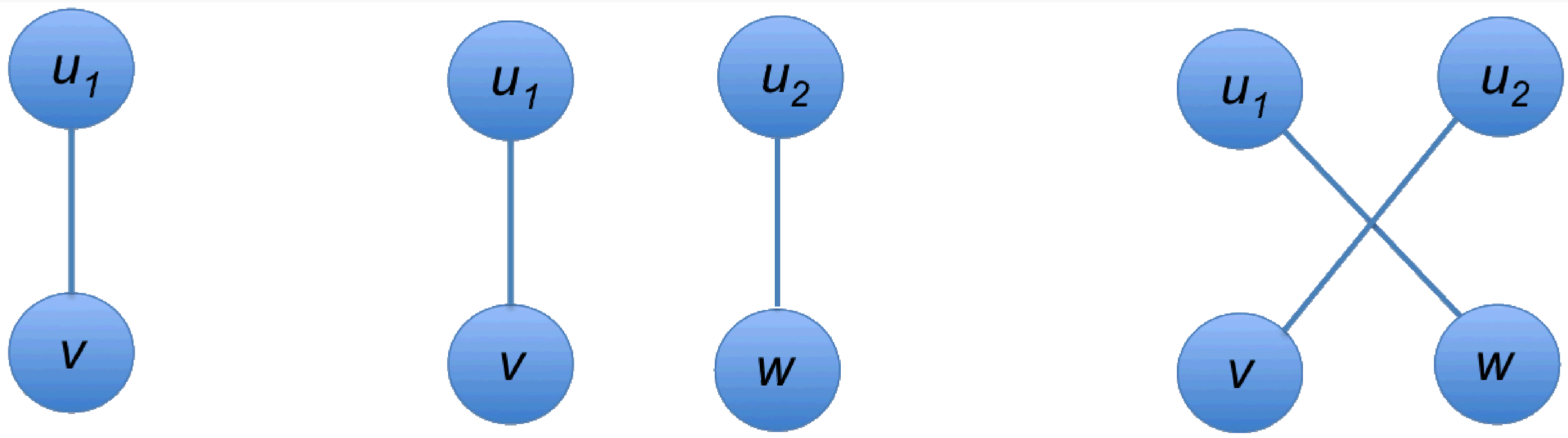
Joint Degree Distribution



Degree	1	2	3	4
1	0	0	1	1
2	0	0	2	2
3	1	2	1	1
4	1	2	1	0

- Joint Degree Distribution (JDD) specifies the number of *edges* between vertices of specified degrees.
- *JDD provides more information abot a graph.*
  - *The degree distribution is implicitly defined by JDD.*
- *Work on JDD is more recent and sparse.*

Preserving JDD



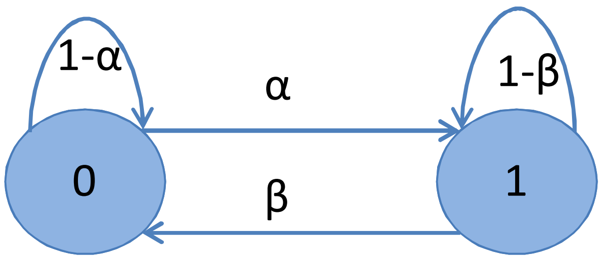
Step 1: Pick an edge (u_1,v), and pick one of its vertices, e.g., u_1

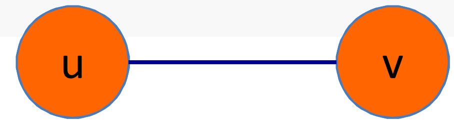
Step 2: Pick another edge (u,w), such that $d(u_1)=d(u_2)$ or $d(u_1)=d(w)$

Step 3: Swap edges

- This Markov chain can be used to construct uniformly random instances of a graph with a specified degree distribution.
- No theoretical bounds on convergence.
- A graph with a specified (feasible) joint degree distribution can be constructed in linear time.
- Stanton & P., *ACM J. Experimental Algorithmics*

Transition matrix for preserving degree distribution





d_u: degree of vertex u m: # edges

$f(d_u)$: #vertices of degree d_u

$J(d_u, d_v)$: #edges between d_u and d_v

α: probability that the edge will be inserted

β: probability that the edge will be deleted

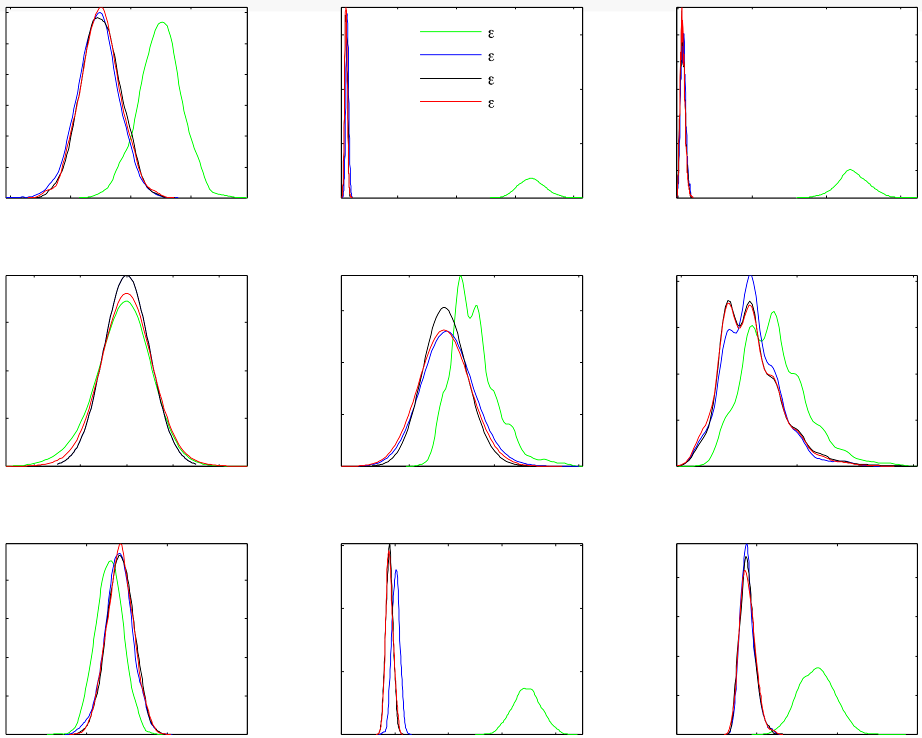
$$\beta = \frac{1}{m} + \frac{f(d_u)-1}{2mf(d_u)} + \frac{f(d_v)-1}{2mf(d_v)}$$

$$N = \ln(1/\varepsilon)/(\alpha+\beta)$$

$$\alpha \cong \frac{2J(d_u, d_v)}{mf(d_u)f(d_v)} \qquad \alpha + \beta \geq \frac{1}{m}$$

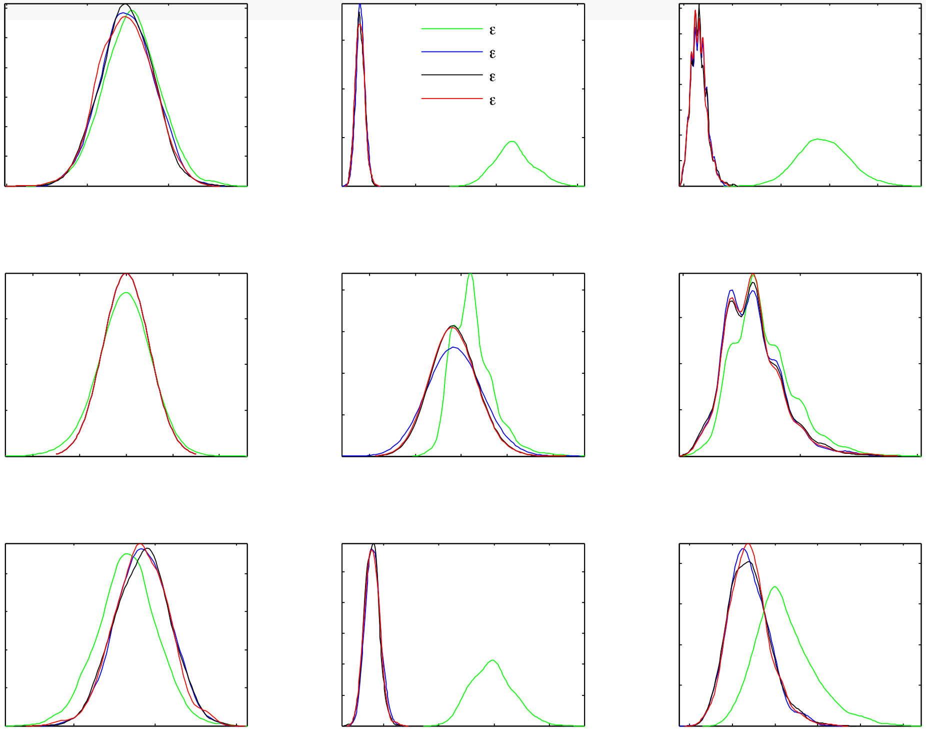
To generate a graph with independent edges with a specified degree distribution we need $N = m \ln \frac{1}{\varepsilon}$

How much error can we tolerate?



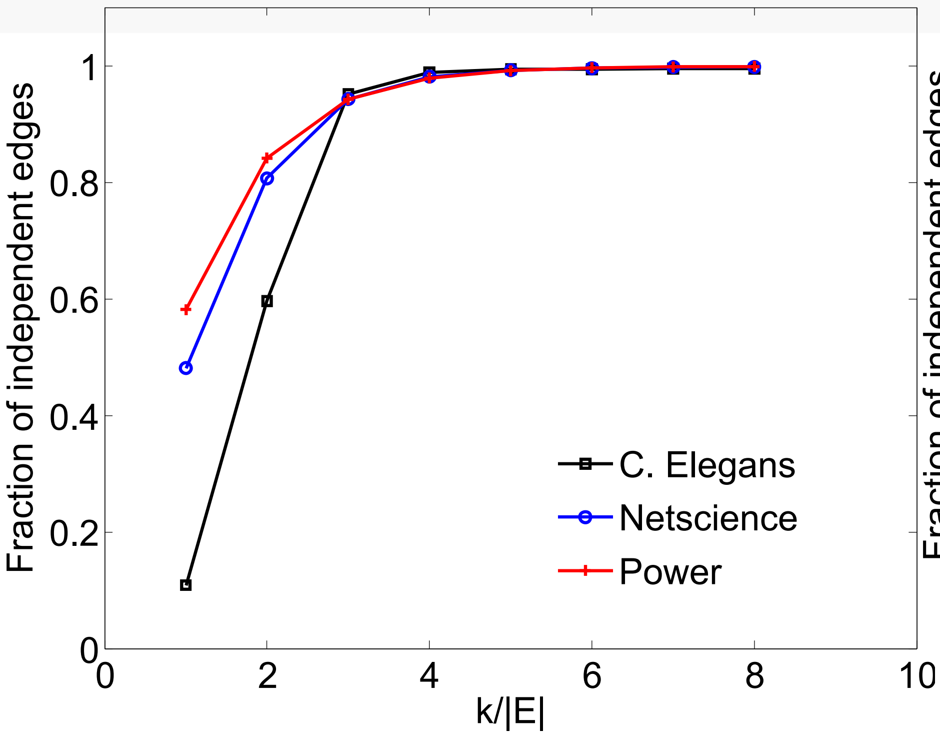
- Preserving degree distribution
- Errors correspond to 0.5, 2.5, 5, and 7.5|E| iterations.

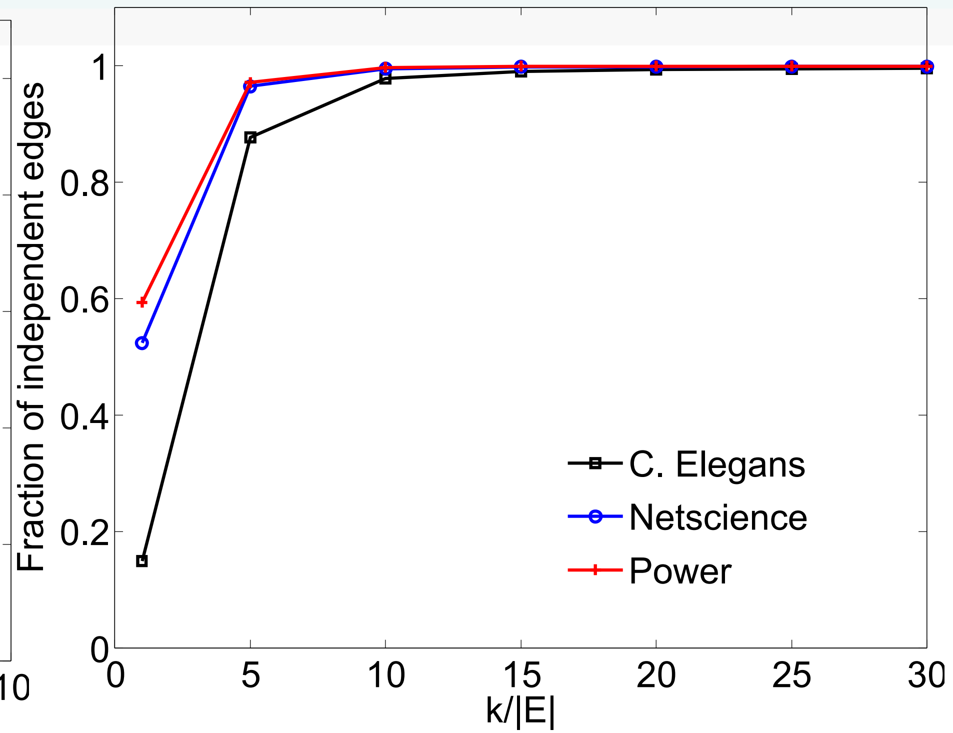
How much error can we tolerate?



- Preserving JDD
- Errors correspond to 1, 5,10, and 15|E| iterations.

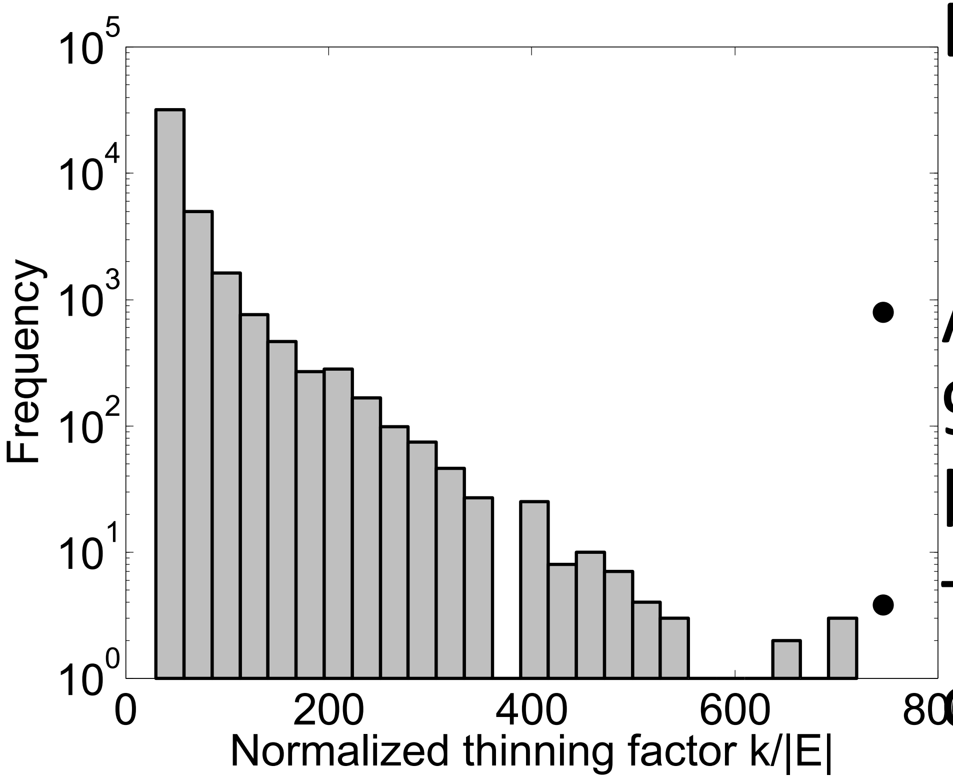
Edges become independent rapidly





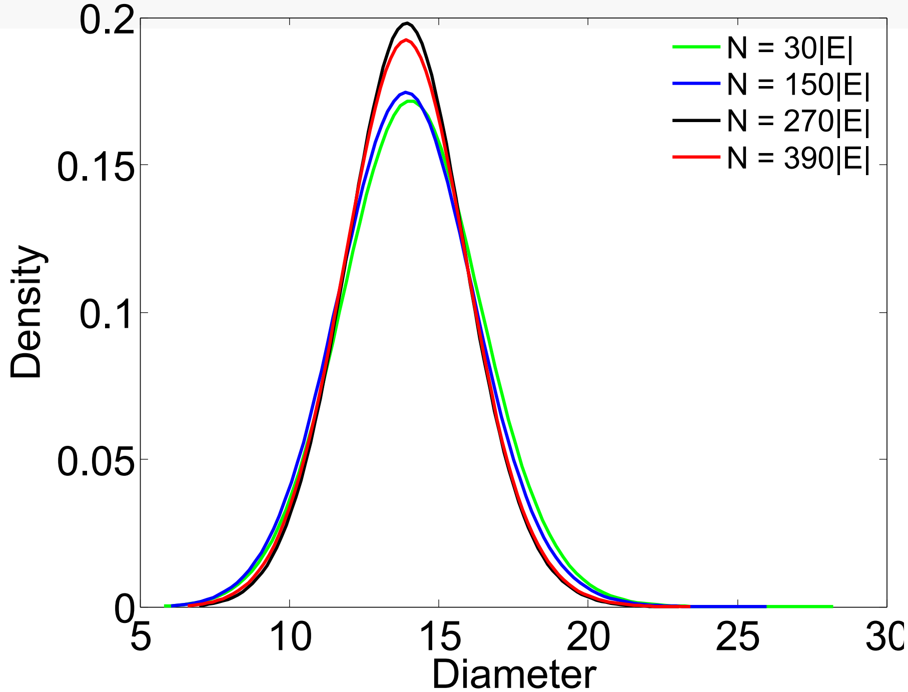
- Many edges become independent quickly.
- Only a few remain after 7.5|E| and 15|E| iterations for preserving DD and JDD, respectively.

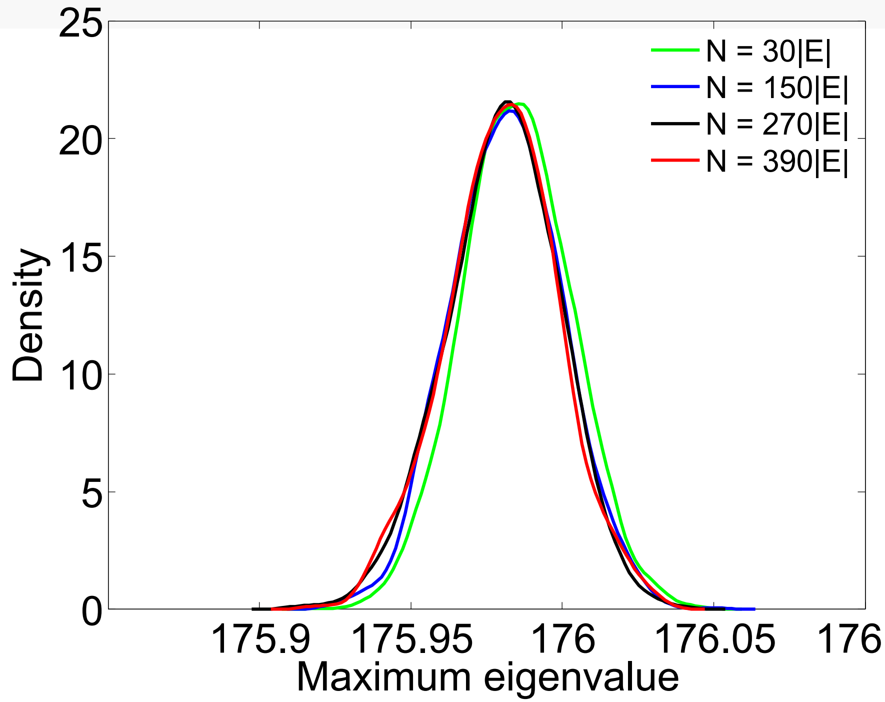
Some edges are tougher than others



- Preserving JDD on Soc-Epinions
  - Edges are sampled down to 10%.
- After 30|E| iterations 90% of the edges become independent.
- There are a few outliers.

Diminishing returns





- Preserving JDD on soc-epinions1

Conclusions

- Generating uniformly random instances of a graph with given properties is a fundamental problem in graph analysis.
- Markov chains are commonly used for this purpose, but guaranteeing/testing their convergence is a challenge.
- We proposed to use
  - edge independence as a practical metric for convergence.
  - Smaller Markov chains for presence/absence of edges as a guide.
- We showed how the method applies to DD and JDD preserving MCs.
- Empirical studies on several graphs validated the approach.
- We are not guaranteeing convergence of the chain, but providing a metric that quantifies what is satisfied.
  - Results should be interpreted accordingly.
- The same approach can be used to guarantee independence of a bigger structures.

Relevant Publications

- Generating a random graph
  - J. Ray, A. Pinar, and C. Sehadhri, Are we there yet? When to stop a Markov chain while generating random graphs," PROC. WAW 12. .
  - I. Stanton and A. Pinar, "Constructing and uniform sampling graphs with prescribed joint degree distribution using Markov Chains," ACM JEA.
  - I. Stanton and A. Pinar, "Sampling graphs with prescribed joint degree distribution using Markov Chains," ALENEX'11.