On Reshaping of Clustering Coefficients in Degree-based Topology Generators

Xiafeng Li, Derek Leonard, and Dmitri Loguinov
Texas A&M University

Presented by Derek Leonard

Agenda

- Motivation
- Statement of the Problem
- Our Algorithm
- Sampling Technique
- Questions

Simulating the Internet

- Important for:
  - Understanding its evolution
  - Testing new protocols
  - Verifying the performance of network applications
  - Etc.
- The key is to create a random graph that has the same fundamental properties as the Internet.

Properties of Internet Graph

- The Internet is a small-world graph
  - Small diameter
    - Bu *et al*. found that the average shortest path between pairs of nodes is small.
  - High clustering coefficient
    - Measured by Bu *et al.* in 2002.
- Power-law degree distribution
  - First measured by Faloutsos *et al.* in 1999.

Properties of Internet Graph II

- The clustering of node v is

$$\gamma_v = \frac{T_v}{d_v(d_v - 1)/2},$$

  where d_v is the degree of node v and T_v is the number of triangles containing node v.

- The clustering of a graph is the average of γ_v for all $d_v \geq 2$.

Properties of Internet Graph III

$$\gamma_A = \frac{1}{2(2-1)/2} = 1,$$

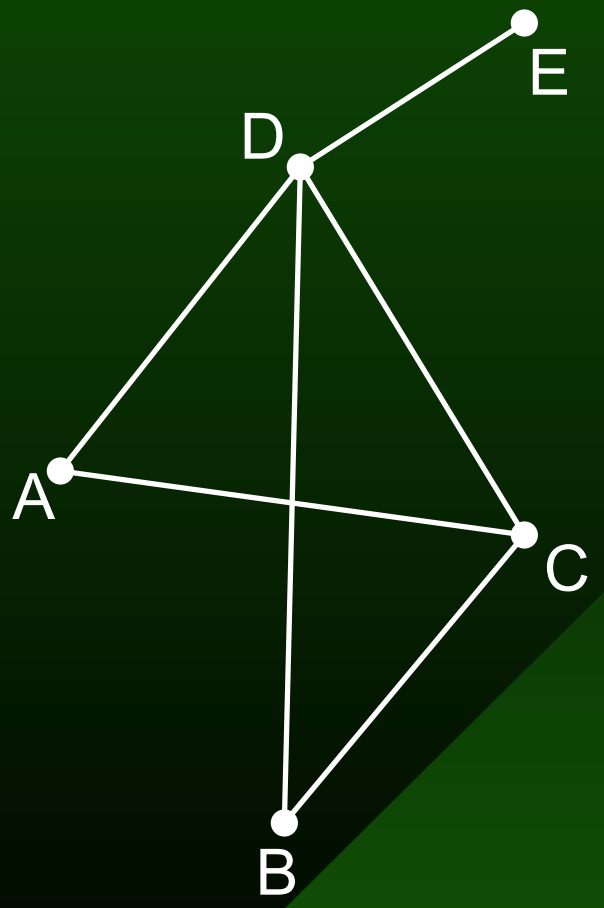
$$\gamma_B = \frac{1}{2(2-1)/2} = 1,$$

$$\gamma_C = \frac{2}{3(3-1)/2} = \frac{2}{3},$$

$$\gamma_D = \frac{2}{4(4-1)/2} = \frac{1}{3}.$$

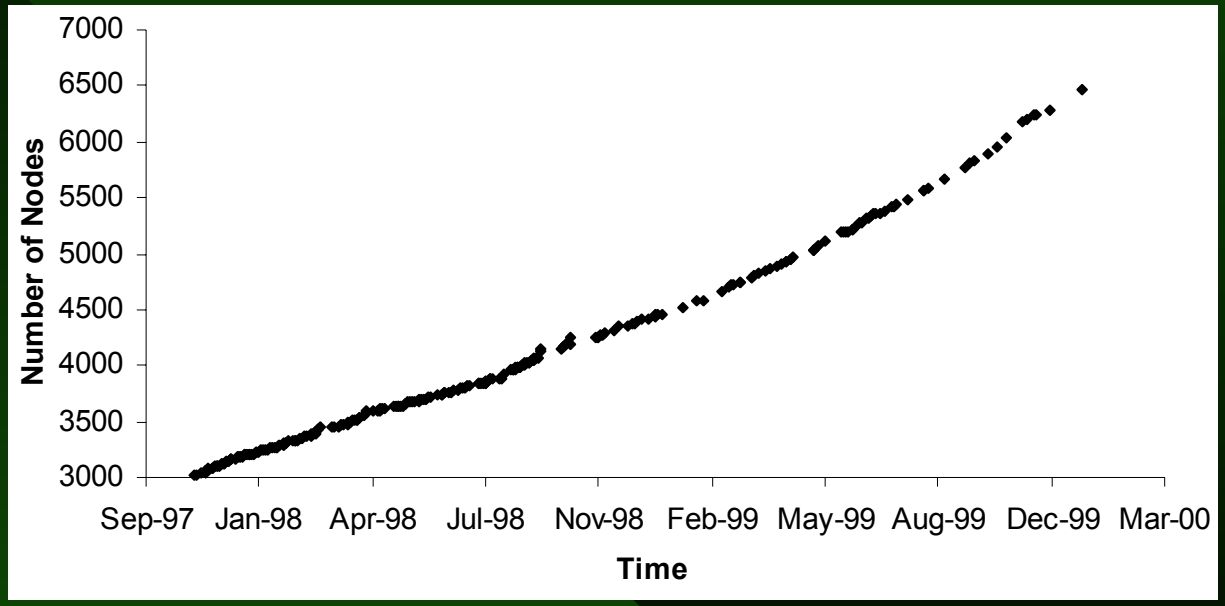
$$\gamma(G) = \frac{\gamma_A + \gamma_B + \gamma_C + \gamma_D}{4}$$

$$= \frac{3}{4}.$$



Properties of Internet Graph IV

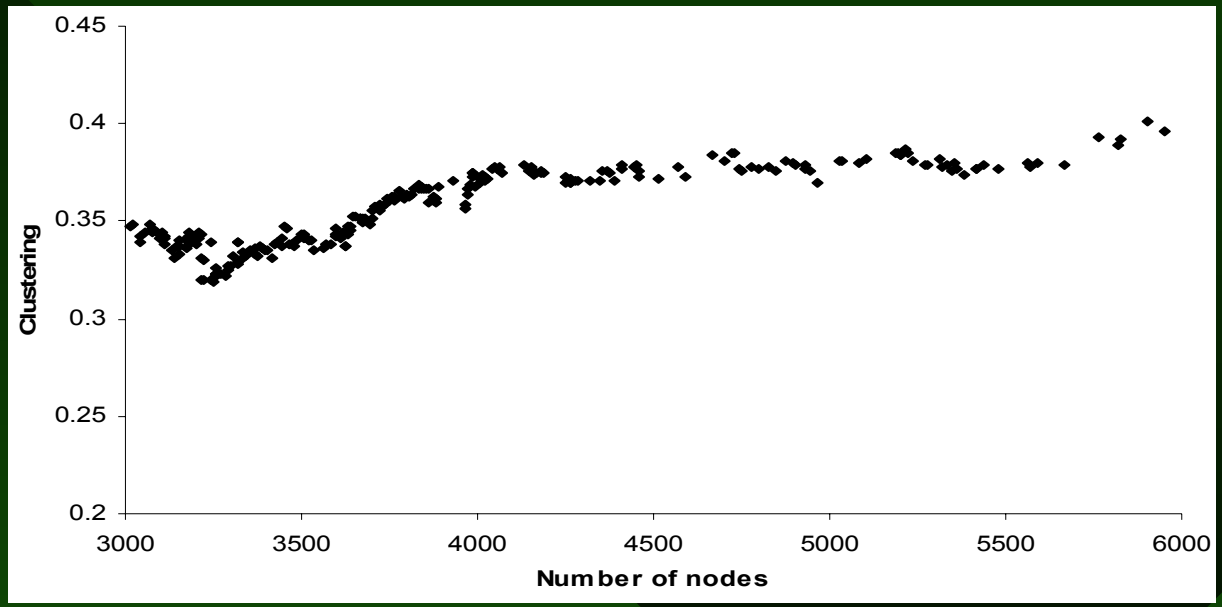
- The Internet graph evolves over time



- Small-world graph properties are maintained regardless of the size of the Internet!

Properties of Internet Graph V

- Clustering coefficient as the Internet evolves.



- The clustering coefficient increases over time!

Internet Graph Generators I

- Existing generators can be categorized into two groups
  - Non-evolving: Produces a graph with a fixed number of nodes n which cannot evolve.
    - Given expected degree (GED)
    - Power-law random graph (PLRG)
  - Evolving: Allows for a variable number of nodes and maintains properties throughout evolution
    - Barabasi-Albert (BA)

Internet Graph Generators II

- Given expected degree (GED)
  - N pre-assigned weights (w_1, ..., w_n) are drawn from a power-law distribution and distributed to n nodes in the graph
  - Edge (i,j) exists with probability:

$$p_{ij} = \frac{w_i w_j}{\sum_{k=1}^{n} w_k}.$$

Internet Graph Generators III

- Power-law random graph (PLRG)
  - Weight w_i drawn from a power-law distribution is assigned to node i for all n nodes
  - Then w_i virtual copies of node i are produced and edges are randomly created among all virtual copies of the nodes
  - Merge the edges created by all virtual copies into the original n nodes
  - Gives results similar to GED

Internet Graph Generators IV

- Barabasi-Albert (BA)
  - BA evolves at discrete time steps
  - Initially, m_0 nodes are created in the graph
  - At each time step, a new node is added to the graph by attaching m edges to existing nodes chosen with probability:

$$\prod(d_i) = \frac{d_i(t)}{\sum_{k=1}^{t} d_k},$$

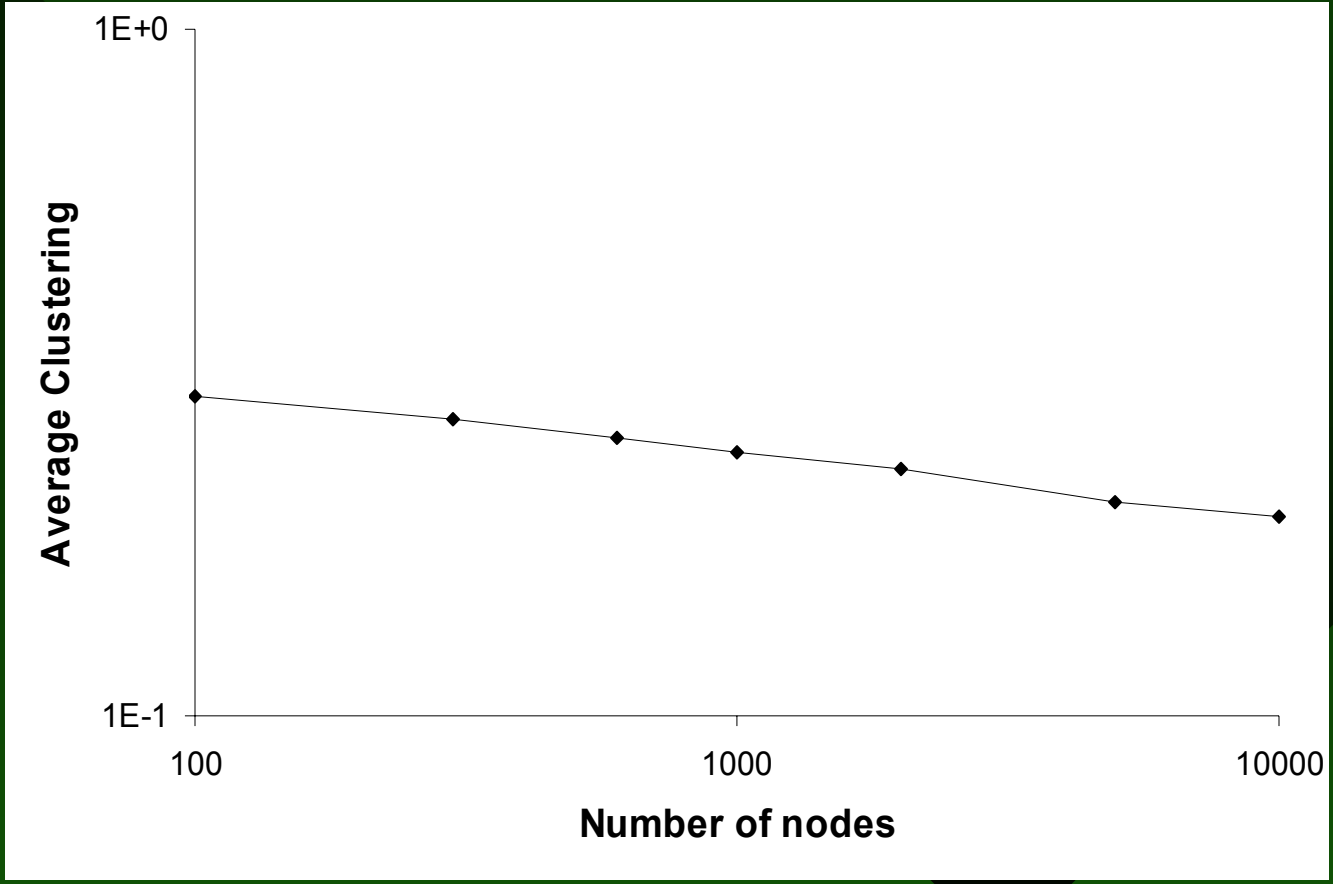
where $d_i(t)$ is the degree of node i at time t.

Internet Graph Generators V

- These three generators provide a representative example of the state of Internet graph generators
- They have two desirable common properties
  - They each produce power-law degree sequences
  - They exhibit low diameter
- However, they fail to exhibit the high clustering property of the Internet graph
  - In fact, clustering tends to zero as n increases

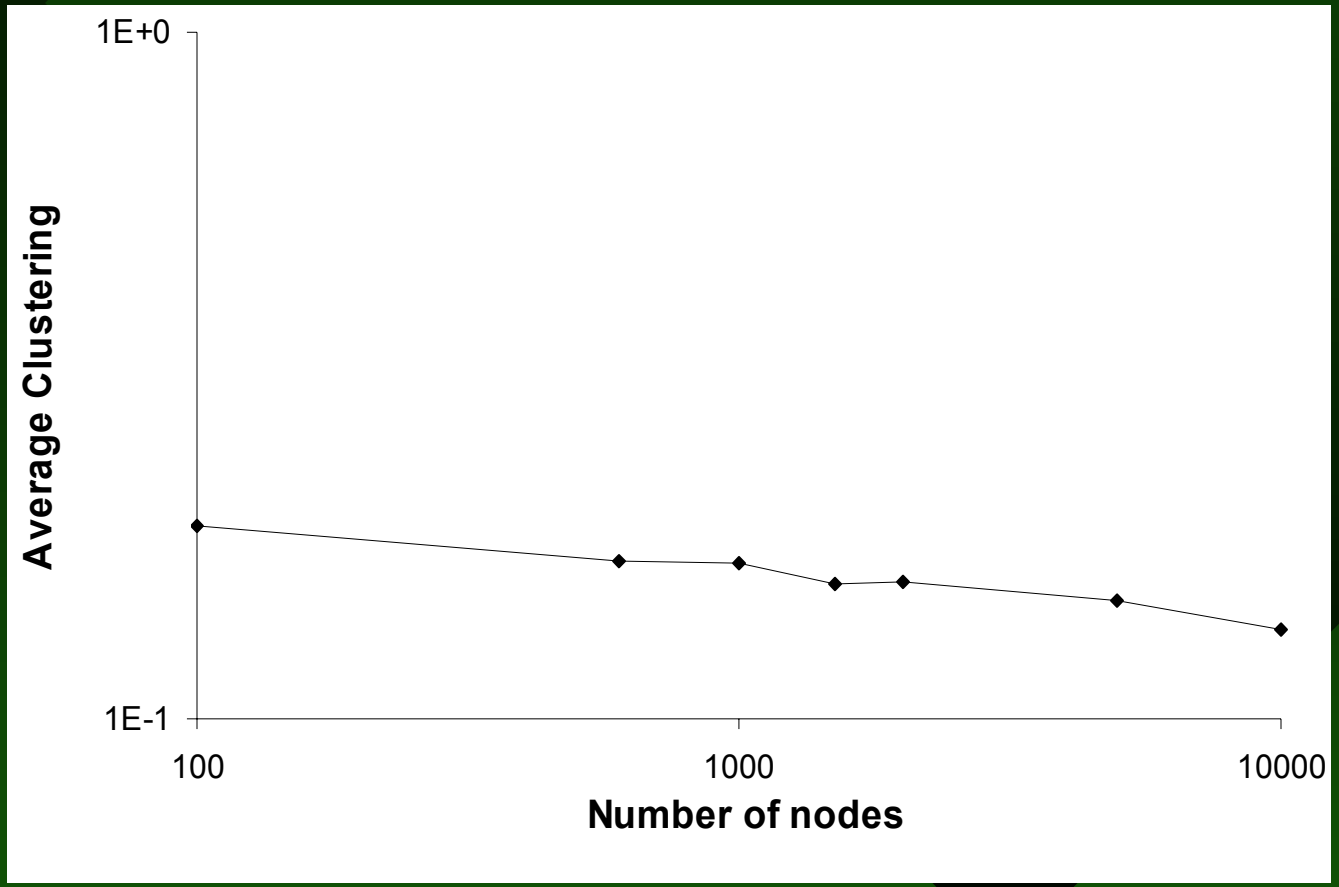
Internet Graph Generators VI

- Clustering in GED decreases as n increases



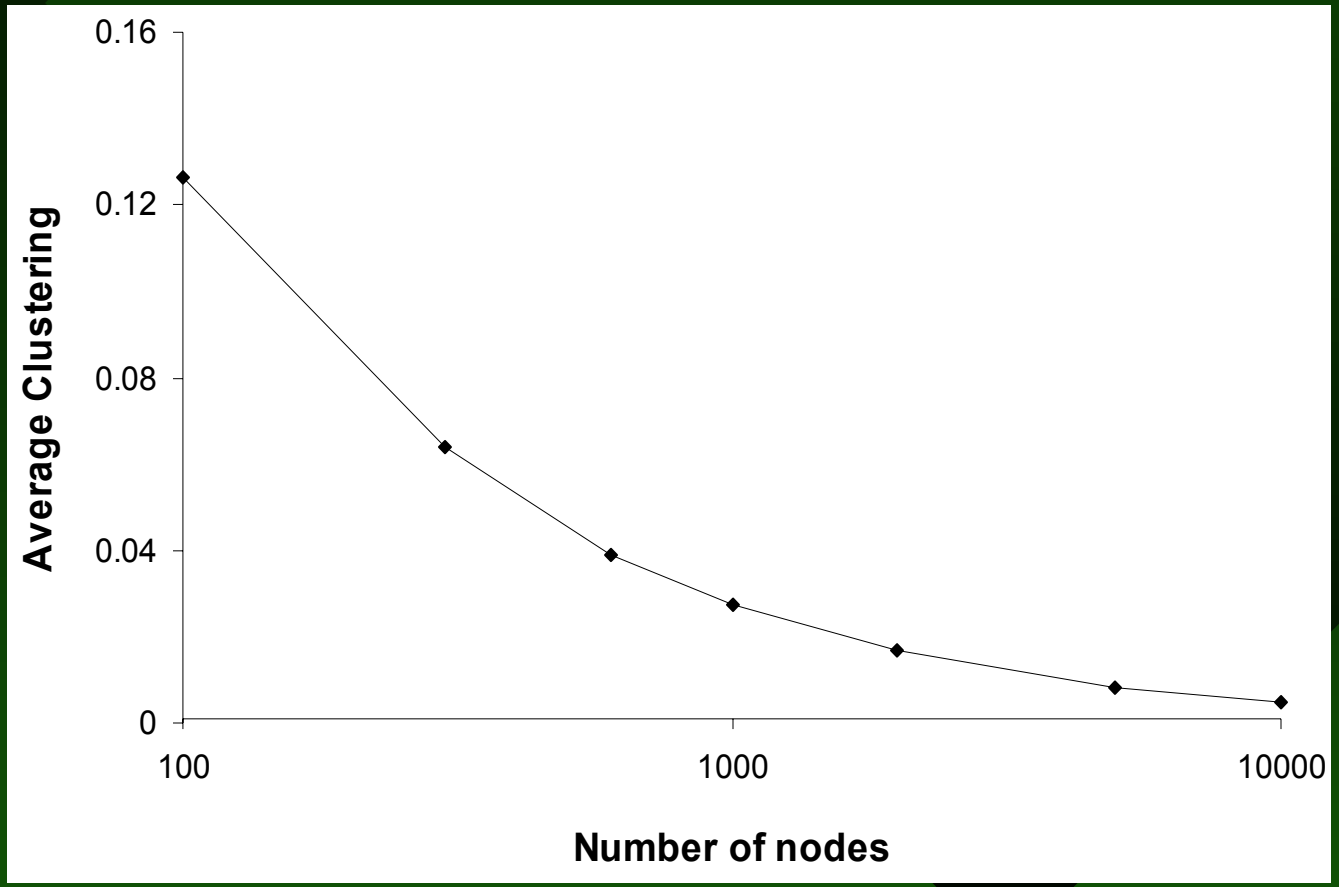
Internet Graph Generators VII

- Clustering in PLRG decreases as n increases



Internet Graph Generators VIII

- Clustering in BA decreases as n increases



Agenda

- Motivation
- Statement of the Problem
- Our Algorithm
- Sampling Technique
- Questions

Problem Statement

- Clustering should increase, not decrease!
- Given a connected graph G and a target clustering value γ_T, rewire G's edges and produce a new graph G′ satisfying
  - G′ is connected
  - The degree sequence in G′ is the same as that in G
  - G′ has low diameter
  - Clustering of G′ is larger than or equal to γ_T .

Agenda

- Motivation
- Statement of the Problem
- Our Algorithm
- Sampling Technique
- Questions

Our Algorithm

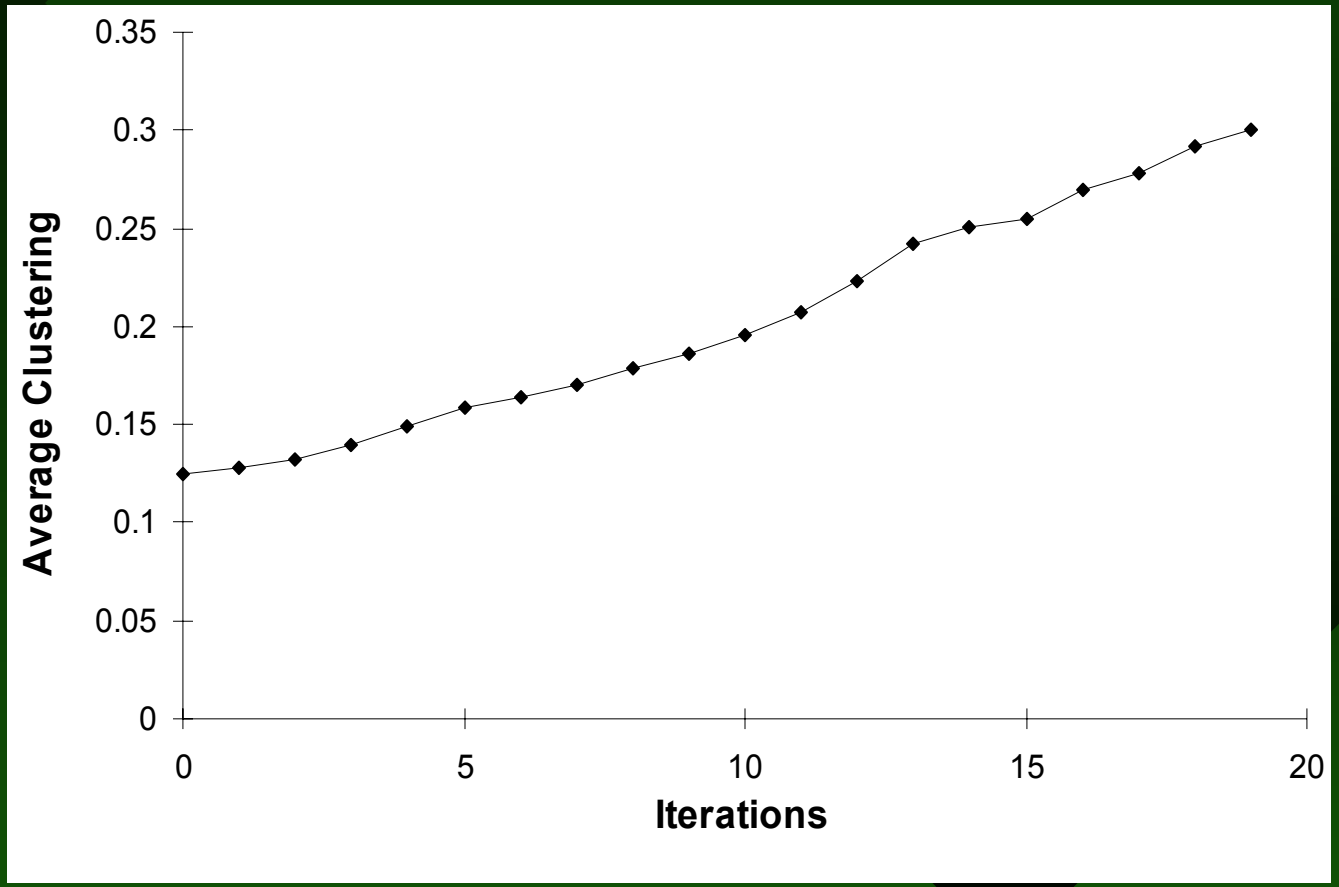
1. Mark all edges that are in at least one triangle.
2. Break an edge that is in some k-loop that is not marked. ($\forall$ k ≥ 4)
   - This maintains the connectivity of the graph
   - Degree of the two nodes incident on the edge will be reduced
3. Randomly connect two nodes if
   - Their degree is smaller than in the original graph G
   - Connecting them will produce at least one new triangle
4. Loop back to step 2 until the clustering of G′ reaches the target value γ_T

Simulation Results I

- Fine control of increase in clustering can be done by limiting the number of k-cycles broken per iteration
  - We break edges on 2% of all nodes each iteration
- In all simulations the shape parameter of the degree distribution is $\alpha = 1.2$ and scale parameter is $\beta = 0.63$
- The target clustering $\gamma_T = .3$ in all cases
- Our algorithm increases the clustering of the graphs created by existing generators in a small number of iterations

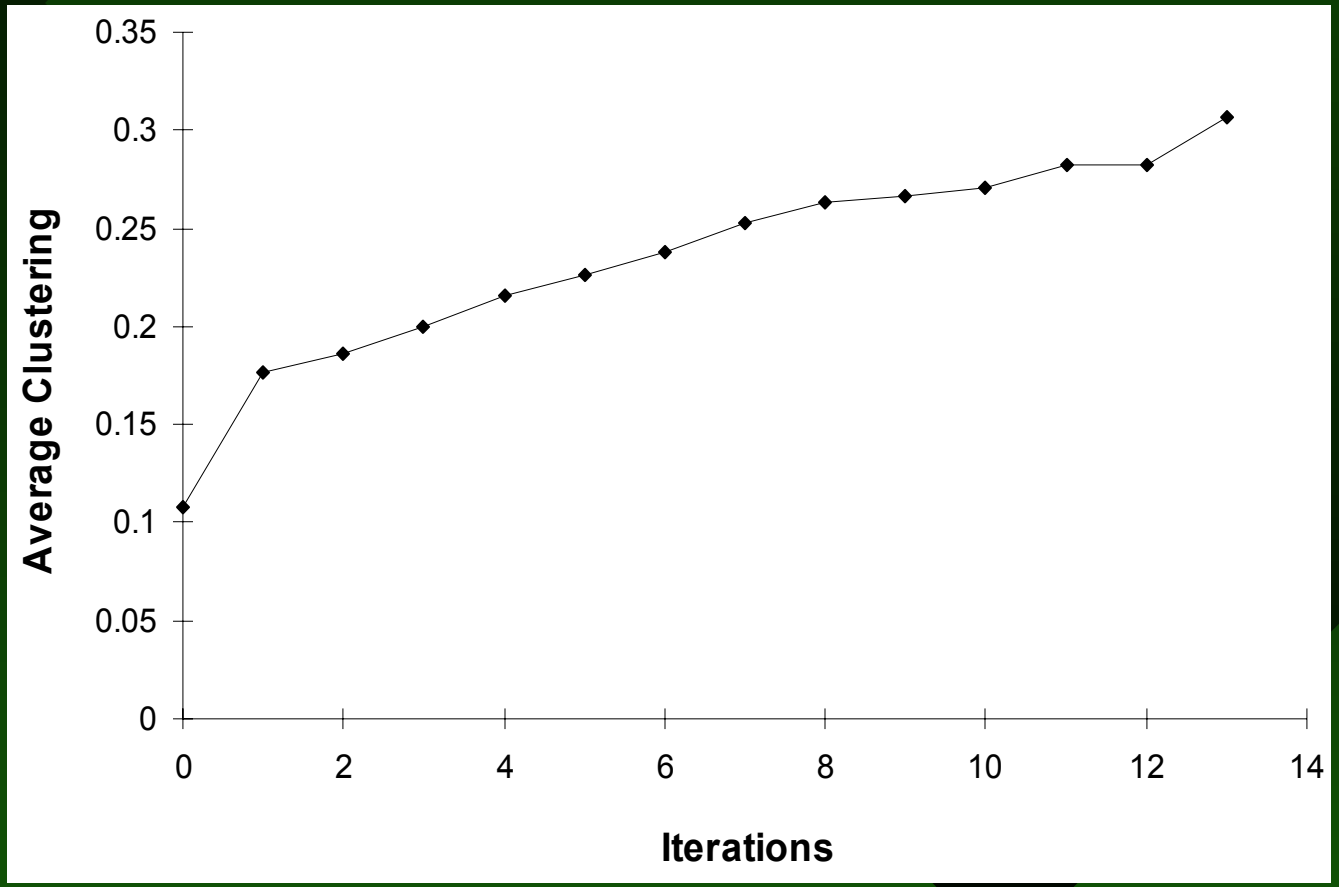
Simulations Results II

- Increase in clustering for a GED graph



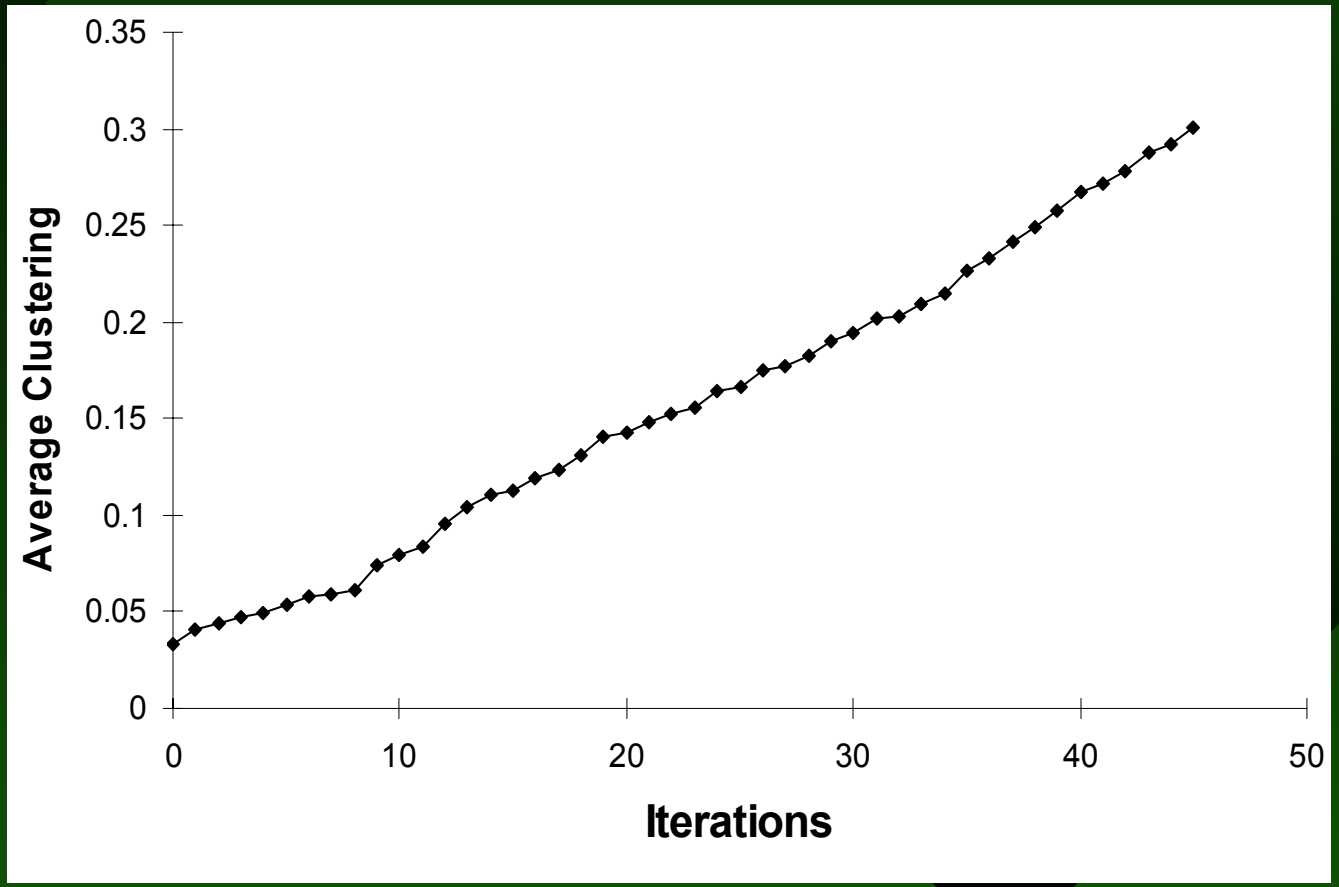
Simulations Results III

- Increase in clustering for a PLRG graph



Simulations Results IV

- Increase in clustering for a BA graph



Simulation Results V

- Average clustering is increased in all cases
- Distribution of node degree is not changed
  - The degree of each node remains relatively unchanged after our algorithm is run
- Average path length is still small
  - After the algorithm is run, the average path length increases by a very small percentage

Simulation Results VI

- Table of average path length before and after our algorithm is run

	G	G′
GED	3.68216	3.706753
PLRG	3.85151	3.633815
BA	4.10205	4.252274

- Average Path Length is not significantly changed

Agenda

- Motivation
- Statement of the Problem
- Our Algorithm
- Sampling Technique
- Questions

Sampling Technique I

- Calculating the clustering of a graph is not a trivial operation
  - Each edge must be examined at every node.
- It would be very inefficient to calculate the exact clustering coefficient at every iteration of our algorithm
- Make use of sampling to estimate clustering
- In each iteration of the algorithm, this technique reduces the time complexity from $O(mn)$ to $O(m)$

Sampling Technique II

- Estimate actual clustering $\gamma_a(G)$ by sampled clustering $\gamma_s(G)$ such that:

$$P(|\gamma_a - \gamma_s| < E) = 1 - \rho$$

- To do this we use a well-known result from sampling theory to determine the sample size s:

$$s = \frac{Z^2_{\rho/2}}{(2E)^2} \geq \frac{\sigma^2 Z^2_{\rho/2}}{E^2}$$

Sampling Technique III

- This guarantees that $|\gamma_a(G) - \gamma_s(G)|$ does not exceed error E with probability $1\text{-}\rho$.
- Note that the sample size in the formula does not depend on the number of nodes in the graph.
- Example: for a graph with 100,000 nodes, margin of error $E=0.1$ and $\rho=0.05$, we get $Z_{\rho/2}=1.96$
  - We only need to sample $s=1.96^2/0.2^2 \approx 97$ nodes.

Conclusions and Future Work

- Our algorithm successfully increases the clustering of G without modifying its degree
  - GED, PLRG, and BA better match the Internet small-world properties after our algorithm is run
- Time complexity of the algorithm needs to be further analyzed
- Devise a new method for generating Internet graphs that inherently incorporates the properties of the Internet

Agenda

- Motivation
- Statement of the Problem
- Our Algorithm
- Sampling Technique
- Questions