An Example of Performance Analysis for Network Community Detection

Luke Fostvedt, Dan Nordman, Alyson Wilson

fostvedt@iastate.edu, dnordman@iastate.edu, agw@iastate.edu

Iowa State University
OF SCIENCE AND TECHNOLOGY

Setting

Many algorithms analyzing clustering and neighborhoods within large networks incorporate the enumeration of the number of triangles. One method for listing triangles is by giving each node a "bucket" where each edge in the graph is placed into the bucket of its endpoint of lowest degree. We show the expected size of a "bucket" and the expected number of pairs in a bucket to be finite provided the degree distribution of the nodes has a finite 4/3 moment.

Random Graph Generation

- Let $\mathcal{G}_n$ be a graph with n nodes where the i^{th} node has degree D_i, $i = 1, \dots, n$.
- The D_i are iid from a discrete probability mass function

$$f(d) = P(D_1 = d), \qquad \text{integer } d \in [\ell, \infty). \qquad (1)$$

- Randomly increment some degree by 1 if $\sum_{i=1}^n D_i$ is odd.
- Randomly pair edges with every pairing equally likely (including self-loops and multi-edges).
- This is known as the configuration model (CM).

The Problem

- Nodes i and j are defined to be **neighbors** if they share at least one edge.
- Fix an arbitrary node i among the n nodes of the graph then we define a "**bucket**" as

$$\mathcal{B}_{i,n} = \{j : i \neq j, D_i \leq D_j, \text{node } i \text{ and node } j \text{ are neighbors}\},$$

  which corresponds to the set of all neighbors of node i having degree at least as great as that of node i.
- Let $N_{i,n} = |\mathcal{B}_{i,n}|$ be the size of the bucket for node i in a size n graph.
- The number of possible node pairs that can be formed from nodes in the bucket $\mathcal{B}_{i,n}$ is

$$\binom{N_{i,n}}{2} = \frac{N_{i,n}(N_{i,n}-1)}{2}.$$

- We are interested in the expected value of $\binom{N_{i,n}}{2}$ as the number of nodes $n \to \infty$ in the CM.

Results

Theorem 1: Under the CM, suppose $\ell \geq 1$ in (1) and $\mathrm{E}D_1 = \sum_{d=\ell}^{\infty} df(d) > 0$.

(i) If $\mathrm{E}D_1 < \infty$, then as $n \to \infty$,

$$\mathrm{E}N_{i,n} \to \sum_{d_1=\ell}^{\infty} \sum_{d_2=d_1}^{\infty} f(d_1) f(d_2) \frac{d_1 d_2}{\mathrm{E}D_1} < \infty$$

(ii) If $\mathrm{E}D_1^{4/3} < \infty$, then as $n \to \infty$,

$$\mathrm{E}\binom{N_{i,n}}{2} \to \frac{1}{2} \sum_{d_1=\ell}^{\infty} \sum_{d_2=d_1}^{\infty} \sum_{d_3=d_1}^{\infty} f(d_1) f(d_2) f(d_3) \frac{d_1 d_2}{\mathrm{E}D_1} \frac{(d_1-1)d_3}{\mathrm{E}D_1} < \infty$$

- Given $D_1 = d_1$ and $D_2 = d_2$, node 1 is neighbor of node 2 in $\mathcal{G}_n$ with probability $\approx \frac{d_1 d_2}{n\mathrm{E}D_1}$ for large n.

Power Law Degree Distribution

Under the CM, suppose the degree distribution is a power law

$$f(d) = P(D_1 = d) \propto d^{-\alpha} \qquad \text{integer } d \geq \ell \geq 1$$

Then the following table summarizes moments and limits as finite (F) or infinite (∞), where the values of finite limits are as in Theorem 1 (F-Th1).

	$\mathrm{E}D_1$	$\lim_{n\to\infty} \mathrm{E}N_{i,n}$	$\mathrm{E}D_1^{4/3}$	$\lim_{n\to\infty} \mathrm{E}\binom{N_{i,n}}{2}$	$\mathrm{E}D_1^2$
$\alpha \leq 2$	∞	∞	∞	∞	∞
$\alpha \in (2, 2\frac{1}{3}]$	F	F-Th1	∞	∞	∞
$\alpha \in (2\frac{1}{3}, 3]$	F	F-Th1	F	F-Th1	∞
$\alpha > 3$	F	F-Th1	F	F-Th1	F

Extensions

- Erased Configuration Model (ECM)
  (i) Generate graph with n nodes, $\mathcal{G}_n$, using CM.
  (ii) Erase self-loops and multi-edges.
- If $\mathrm{E}D_1 < \infty$,
  (i) The distribution of the erased degrees in the ECM converges to the degree distribution from the CM.
- The results for the CM in Theorem 1 also hold for the ECM if $\mathrm{E}D_1^2 < \infty$ or if initial degrees are truncated to at most $\sqrt{n}$.

Configuration Model Simulations

Power Law with exponent $\alpha = 2.4$.

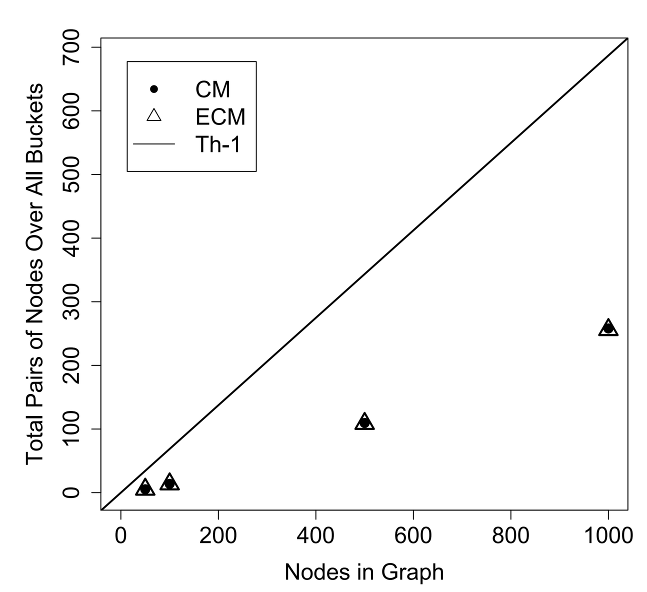
	Erased Configuration Model			Configuration Model		
n	$\mathrm{E}N_{i,n}$	$\mathrm{E}\binom{N_{i,n}}{2}$	$\mathrm{E}D_{i,n}$	$\mathrm{E}N_{i,n}$	$\mathrm{E}\binom{N_{i,n}}{2}$	$\mathrm{E}D_{i,n}$
50	1.034	0.106	1.720	1.029	0.103	2.219
100	1.071	0.138	1.807	1.068	0.135	2.291
500	1.139	0.219	1.966	1.138	0.217	2.223
1000	1.162	0.258	2.018	1.161	0.256	2.215
5000	1.202	0.346	2.107	1.202	0.345	2.233
∞	1.257	0.687	2.221	1.257	0.687	2.221

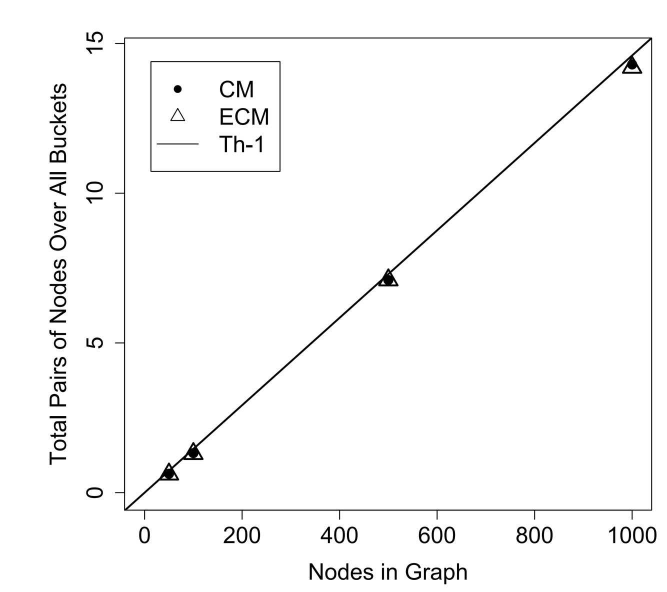
Power Law with exponent $\alpha = 3.2$.

	Erased Configuration Model			Configuration Model		
n	$\mathrm{E}N_{i,n}$	$\mathrm{E}\binom{N_{i,n}}{2}$	$\mathrm{E}D_{i,n}$	$\mathrm{E}N_{i,n}$	$\mathrm{E}\binom{N_{i,n}}{2}$	$\mathrm{E}D_{i,n}$
50	0.932	0.0126	1.252	0.931	0.0123	1.283
100	0.936	0.0132	1.261	0.935	0.0130	1.281
500	0.941	0.0142	1.274	0.941	0.0142	1.280
1000	0.942	0.0143	1.274	0.942	0.0142	1.277
5000	0.943	0.0145	1.277	0.943	0.0145	1.278
∞	0.943	0.0146	1.277	0.943	0.0146	1.277

Above ∞ denotes theoretical limits from Theorem 1.

Graphical Illustrations





The number of bucket pairs $\sum_{i=1}^n \binom{N_{i,n}}{2}$ over the entire graph grows at most linearly with n, as shown above for a Power Law with $\alpha = 2.4$ (Left) and a Power Law with $\alpha = 3.2$ (Right). Values are close under the CM and the ECM.

Funding and Acknowledgements

Research funded by Sandia National Laboratories' Network Grand Challenge. Thanks to research collaborators Cindy Phillips and Jon Berry.