Cliques and a New Measure of Clustering

Steve Lawford, Yll Mehmeti

CLIQUES AND A NEW MEASURE OF CLUSTERING

STEVE LAWFORD (JOINT WORK WITH YLL MEHMETI)

DATA, ECONOMICS AND INTERACTIVE VISUALIZATION (DEVI) GROUP

ENAC (UNIVERSITY OF TOULOUSE), FRANCE

1. INTRODUCTION

One widely used measure of clustering is the *overall clustering coefficient*, or "transitivity", on three nodes:

$$C(3) = \frac{3 \times \text{number of triangles in the network } G}{\text{number of connected triples of nodes in } G},$$

which quantifies the relative frequency with which two neighbours of a node are themselves neighbours.

Many real-world networks display **higher levels of clustering** than if those networks were random [1, 2].

Clustering related to cooperative social behaviour and beneficial information and reputation transfer [3].

Significant topological structures, **on more than three nodes**, can be found in real-world networks, and may perform precise specialized functions [4].

A generalized clustering coefficient could provide new insight into such higher-order network structure.

2. OBJECTIVES

a) Propose a **higher-order generalization** of $C(3)$, to any number of nodes, that nests standard clustering.

b) Develop and test a fast, practical, implementation based on **analytic subgraph enumeration** formulae.

REFERENCES

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] S.H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[3] M.O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[4] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[5] M. Agasse-Duval and S. Lawford. Subgraphs and motifs in a dynamic airline network. Technical Report arXiv:1807.02585, 2020.

[6] S. Lawford. Counting five-node subgraphs. Technical Report arXiv:2009.11318, 2020.

[7] H. Yin, A.R. Benson, and J. Leskovec. Higher-order clustering in networks. *Physical Review E*, 97:052306, 2018.

3. HIGHER-ORDER CLUSTERING

We define the *generalized clustering coefficient* as:

$$C(b) = \frac{a(b) \times \text{number of } b\text{-cliques } K_b \text{ in } G}{\text{number of } b\text{-spanning trees in } G}, \quad b \geq 3,$$

where *Cayley's formula* $a(b) = b^{b-2}$ gives the number of spanning trees in K_b, ensuring that $0 \leq C(b) \leq 1$.

We use analytic subgraph enumeration formulae to **count cliques and spanning trees** [5, 6]:

$$C(4) = \frac{16\,|K_4|}{|M_{11}^{(4)}| + |M_{13}^{(4)}|},$$

$$C(5) = \frac{125\,|K_5|}{|M_{75}^{(5)}| + |M_{77}^{(5)}| + |M_{86}^{(5)}|},$$

where $|M_a^{(b)}|$ is the count of subgraphs of "type" a on b nodes. For example, the 5-arrow subgraph count is:

$$|M_{77}^{(5)}| = \sum_{(i,j)^\star \in E} \binom{k_i - 1}{2}(k_j - 1) - 2|M_{15}^{(4)}|,$$

where edge $(i,j) \in E$ is summed in both directions, $k.$ is the node degree, and $|M_{15}^{(4)}|$ is the tadpole count.

An alternative measure was developed in 2018 by Yin-Benson-Leskovec (YBL), using clique expansion [7]:

$$C_{b-1} = \frac{(b^2 - b)\,|K_b|}{|L(b-1, 1)|}, \quad b \geq 4,$$

where $L(\cdot,\cdot)$ is the lollipop graph formed by joining a $(b-1)$-clique by a bridge to a single node.

Critical difference between $C(b)$ and C_{b-1} is in their definitions of the **"relative frequency" of cliques**.

4. THEORETICAL RESULTS ON RANDOM GRAPHS

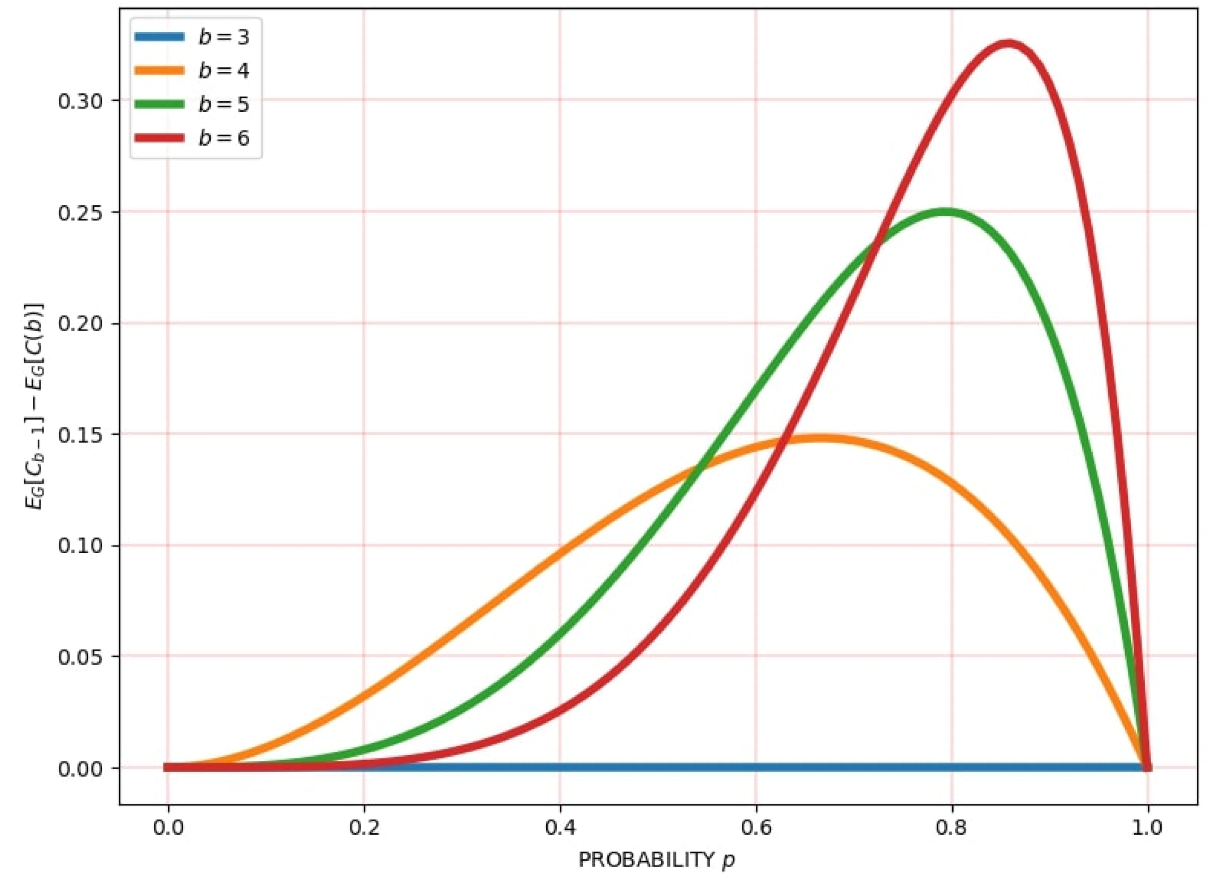

Figure 1: Theoretical difference in expectation for the **Erdős-Rényi** random graph $G(n, p)$ is $\mathbb{E}_G[C_{b-1}] - \mathbb{E}_G[C(b)] = p^{b-2}(1 - p^{(b-2)(b-3)/2})$, with edge-formation probability p.

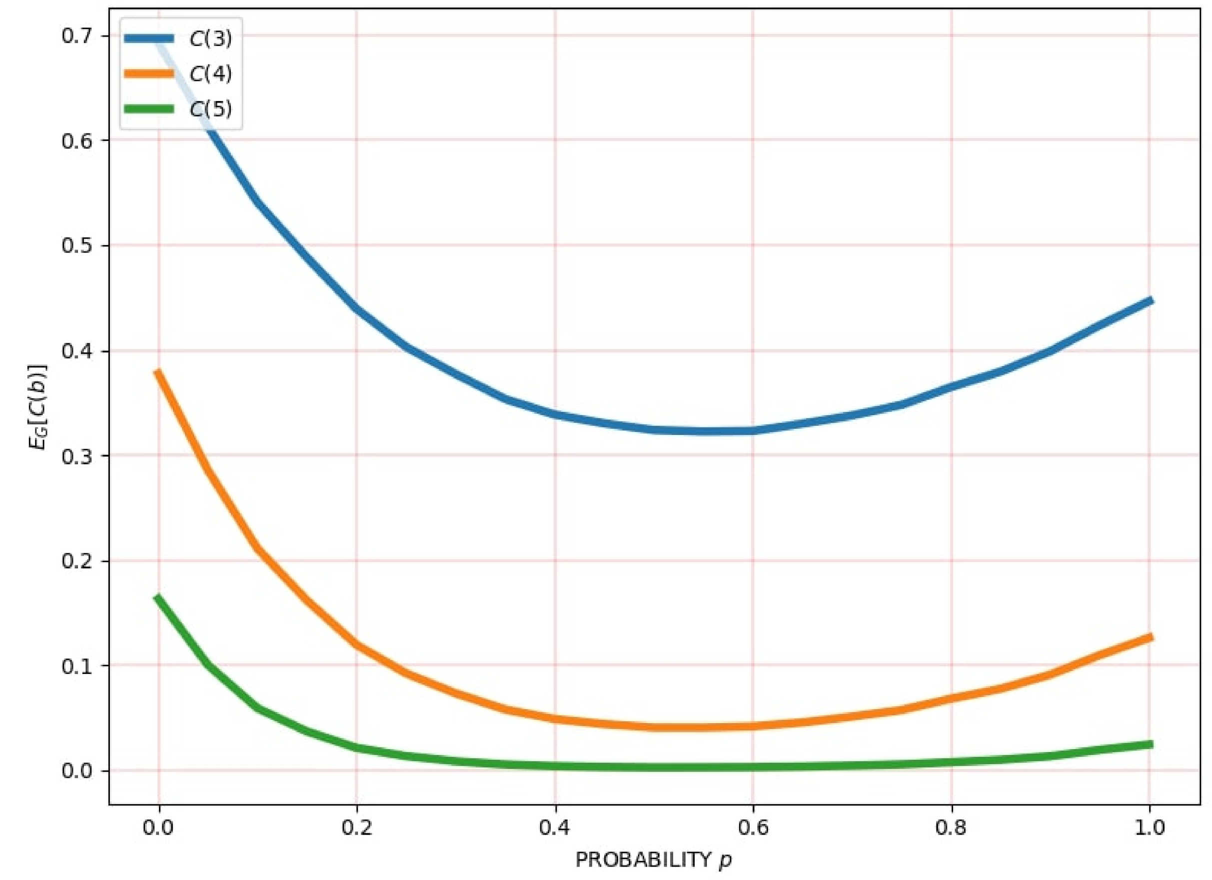

Figure 2: Simulated expected clustering $\mathbb{E}_G[C(b)]$ from 250 replications of a **small-world** graph with $n = 50$ nodes, each of which has degree 14, and edge-rewiring probability p.

5. EMPIRICAL RESULTS ON REAL-WORLD NETWORKS

Data on eight small, sparse, U.S. domestic airline **route networks**, 1999Q1 to 2013Q4 (most are *small-world*). Heterogeneity in clustering dynamics across carriers.

Small values of b capture much of the higher-order clustering present in these real-world networks. Some higher-order clustering left if control for lower-orders.

Carrier	Nodes	Edges	Density	apl	$\text{apl}_{\text{rand}}^{\text{conn}}$	$C(3)$	$C(3)_{\text{rand}}$	$C(4)$	$C(4)_{\text{rand}}$	$C(5)$	$C(5)_{\text{rand}}$	Connected %
AA	71	153	0.06	1.94	3.01	0.120	0.061	0.018	0.000	0.002	0.000	44.6
AS	34	49	0.09	2.00	3.12	0.037	0.085	0.000	0.001	0.000	0.000	18.2
DL	85	221	0.06	1.98	2.84	0.146	0.061	0.021	0.000	0.002	0.000	65.4
FL	38	78	0.11	1.94	2.63	0.154	0.108	0.008	0.001	0.000	0.000	61.7
NK	29	92	0.23	1.95	1.96	0.379	0.223	0.097	0.011	0.016	0.000	97.3
UA	48	158	0.14	2.03	2.23	0.346	0.138	0.122	0.003	0.034	0.000	96.0
US	58	113	0.07	2.09	3.03	0.115	0.067	0.011	0.000	0.000	0.000	35.3
WN	88	522	0.14	1.99	2.04	0.335	0.136	0.106	0.002	0.031	0.000	99.9

Figure 3: Descriptive statistics for 2013Q4. The average path lengths (apl) for real-world networks are close to those from connected Erdős-Rényi random graphs ($\text{apl}_{\text{rand}}^{\text{conn}}$). The clustering coefficients $C(b)$ are typically higher than those from Erdős-Rényi random graphs ($C(b)_{\text{rand}}$). Connected % gives the percentage of connected realizations across 1,000 replications.

6. DISCUSSION AND FUTURE DIRECTIONS

a) Our work complements YBL: (theory) with $C(b)$, we develop the other **natural generalization** of $C(3)$ to more nodes, (computational) we derive analytic higher-order clustering formulae, while YBL use numerical methods, (empirical) we apply $C(b)$ to airline networks, a classical example that is not covered by YBL. Our statistic **avoids several undesirable properties** of YBL's statistic, namely its invariance on some graphs with vanishing density, and its lack of applicability (when undefined) to even some connected graphs.

b) It is hard to derive analytic count formulae for subgraphs as b increases e.g. $C(8)$ has 23 denominator terms. There may be a role for **computer-assisted (or automated) theorem proving** in working towards this goal.

c) Airline carriers are increasingly developing small groups of highly-connected airports. The concept of a "hub" (or central) node in real-world networks can be extended to **"multi-node hubs"** (or central *groups* of nodes).

CONTACT INFORMATION

I will be happy to discuss problems, papers and projects in all areas of complex systems after CCS2020.

Email You can send me a message at: steve.lawford@enac.fr

Web For more about my research see: http://tinyurl.com/web-steve . . . or Google me and follow link #1

arXiv My complex systems papers are at: http://tinyurl.com/arxiv-steve

Scan the QR code (top-right) for the clustering paper!