Cliques and a New Measure of Clustering
Steve Lawford, Yll Mehmeti

To cite this version:
Steve Lawford, Yll Mehmeti. Cliques and a New Measure of Clustering. CCS 2020, Dec 2020, Virtual event, France. ACM. hal-03142525

HAL Id: hal-03142525
https://enac.hal.science/hal-03142525
Submitted on 16 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 more 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), for any number of nodes, that nests standard clustering.

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

3. HIGHER-ORDER CLUSTERING

We define the generalized clustering coefficient as:
\[ C(b) = \frac{a(b) \times \text{number of } b\text{-cliques in } G}{\text{number of } b\text{-spanning trees in } G} \]
where Capley’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]:
\[ C(4) = \frac{16 |K_4|}{|M_4^{(1)}| + |M_4^{(2)}|} \]
\[ C(5) = \frac{125 |K_5|}{|M_5^{(0)}| + |M_5^{(1)}| + |M_5^{(4)}|} \]
where \( |M_{kl}^{(i)}| \) is the count of subgraphs of “type” \( a \) on \( b \) nodes. For example, the 5-arrow subgraph count is:
\[ |M_5^{(4)}| = \sum_{(i,j) \in E} \left( \frac{1}{2} (k_i - 1)(k_j - 2) |M_4^{(i)}| \right) \]
where edge \((i,j) \in E\) is summed in both directions, \( k_i \) is the node degree, and \( |M_{kl}^{(0)}| \) is the tadpole count.

An alternative measure was developed in 2018 by Yin-Benson-Leskovec (YBL), using clique expansion [7]:
\[ C_{YBL} = \frac{b^b - b|K_b|}{L(b - 1, 1)} \]
where \( L(\cdot, \cdot) \) is the laptop graph formed by joining a \( (b-1) \)-clique by a bridge to a single node.

Critical difference between \( C(b) \) and \( C_{YBL} \) 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 \( E_{ER}(C(6)) - E_{YBL}(C(6)) = p^6(1-p^3)(1-2p^2+p^4)(1-p) \), with edge-formation probability \( p \).

Figure 2: Simulated expected clustering \( E_{ER}(C(6)) \) 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.

<table>
<thead>
<tr>
<th>Carrier</th>
<th>Nodes</th>
<th>Edges</th>
<th>Density</th>
<th>BS</th>
<th>BM</th>
<th>WP</th>
<th>MS</th>
<th>10%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>71</td>
<td>153</td>
<td>0.06</td>
<td>1.04</td>
<td>1.02</td>
<td>0.125</td>
<td>0.061</td>
<td>0.018</td>
<td>0.000</td>
</tr>
<tr>
<td>AS</td>
<td>34</td>
<td>49</td>
<td>0.09</td>
<td>2.00</td>
<td>1.12</td>
<td>0.037</td>
<td>0.085</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>DL</td>
<td>85</td>
<td>221</td>
<td>0.09</td>
<td>1.58</td>
<td>2.84</td>
<td>0.146</td>
<td>0.061</td>
<td>0.021</td>
<td>0.000</td>
</tr>
<tr>
<td>FL</td>
<td>38</td>
<td>78</td>
<td>0.11</td>
<td>1.54</td>
<td>2.83</td>
<td>0.158</td>
<td>0.108</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>NK</td>
<td>29</td>
<td>47</td>
<td>0.25</td>
<td>1.95</td>
<td>1.96</td>
<td>0.279</td>
<td>0.223</td>
<td>0.097</td>
<td>0.011</td>
</tr>
<tr>
<td>UA</td>
<td>48</td>
<td>158</td>
<td>0.16</td>
<td>2.03</td>
<td>2.23</td>
<td>0.346</td>
<td>0.138</td>
<td>0.122</td>
<td>0.003</td>
</tr>
<tr>
<td>US</td>
<td>58</td>
<td>113</td>
<td>0.07</td>
<td>2.09</td>
<td>3.03</td>
<td>0.115</td>
<td>0.067</td>
<td>0.011</td>
<td>0.000</td>
</tr>
<tr>
<td>WN</td>
<td>89</td>
<td>522</td>
<td>0.14</td>
<td>1.69</td>
<td>2.04</td>
<td>0.235</td>
<td>0.136</td>
<td>0.106</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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 (aplrand). The clustering coefficients \( C(b) \) are typically higher than those from Erdős-Rényi random graphs \( C(b)_{ER}(\text{rand}) \). 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).

6. 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!