CCCG: Clique Conversion Ratio Driven Clustering of Graphs

by

Prathyush Sambaturu, Kamalakar Karlapalem

in

Pacific-Asia Conference on Knowledge Discovery and Data Mining
(*PAKDD 2017*)

Jeju, South Korea

Report No: IIIT/TR/2017/-1

Centre for Data Engineering
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2017

CCCG: Clique Conversion Ratio Driven Clustering of Graphs

Prathyush Sambaturu(✉) and Kamalakar Karlapalem

Data Science and Analytics Centre IIIT Hyderabad, Hyderabad 500 032, India
prathyush.sambaturu@research.iiit.ac.in, kamal@iiit.ac.in

Abstract. Networks have become ubiquitous in many real world applications and to cluster similar networks is an important problem. There are various properties of graphs such as clustering coefficient (CC), density, arboricity, etc. We introduce a measure, Clique Conversion Coefficient (CCC), which captures the clique forming tendency of nodes in an undirected graph. CCC could either be used as a weighted average of the values in a vector or as the vector itself. Our experiments show that CCC provides additional information about a graph in comparison to related measures like CC and density. We cluster the real world graphs using a combination of the features CCC, CC, and density and show that without CCC as one of the features, graphs with similar clique forming tendencies are not clustered together. The clustering with the use of CCC would have applications in the areas of Social Network Analysis, Protein-Protein Interaction Analysis, etc., where cliques have an important role. We perform the clustering of ego networks of the YOUTUBE network using values in CCC vector as features. The quality of the clustering is analyzed by contrasting the frequent subgraphs in each cluster. The results highlight the utility of CCC in clustering subgraphs of a large graph.

1 Introduction

In many real world applications, data is naturally organized in the form of networks such as social networks [5], road networks [7], collaboration networks [3], communication networks, protein-protein interaction networks [4] and web graphs. In graph theory various measures exist such as diameter, clustering coefficient, density, arboricity [6], k-core number [2] and betweenness centrality [13] which describe certain properties of graphs. An interesting problem corresponding to data mining is to find similarity [14,15] between graphs. One of the main challenges here is to list the properties of graphs that can be used as their features in finding similarity.

Problem Statement: To determine the "goodness" of features for graph similarity.

To list all the graph properties is near to impossible. Our focus in this paper is on−Clique Conversion Ratio−the ratio at which the cliques of a particular order expand into the cliques of a higher order. In this paper, we define Clique Conversion Coefficient (CCC) to be a measure of the aforementioned tendency

J. Kim et al. (Eds.): PAKDD 2017, Part II, LNAI 10235, pp. 762–773, 2017.
DOI: 10.1007/978-3-319-57529-2_59

in graphs. Also, we show that the existing measures such as *clustering coefficient* [1] and *density* do not capture the idea of the Clique Conversion Ratio as well as CCC. To motivate the need for a measure like CCC we provide the following example. Let A, B, and C be three people in a large Social Network G who are mutually connected to each other. Let D be another person who forms at least one triangle in G. The expansion of the clique $\{A, B, C\}$ into the clique $\{A, B, C, D\}$ would require the presence of the cliques $\{A, B, D\}$, $\{A, C, D\}$ and $\{B, C, D\}$ in G. Although D forms more than one triangle it need not be connected to any of A, B and C. The clique $\{A, B, C\}$ could expand into multiple such cliques of size four. CCC, in fact, measures the ratio of actual conversions of all cliques of a particular number of people into cliques having one more person to the maximum number of such conversions possible. We could construct dense graphs with many cliques of size three with only a few of those actually expanding into cliques of size four. The Clustering Coefficient and Density would not be able to differentiate such graphs from the ones that have high clique conversion ratio. CCC is also useful in applications where the existence and expansion of cliques plays an important role. For example, an advertiser might want to advertise his product in a Social Network which is not only dense and well-connected but also has higher clique forming tendency. This would help his advertisement to percolate to more people in comparison to as in those networks having lesser clique forming tendency. In this paper, we define CCC, explore its properties, further provide experimental evidence that CCC provides new information in comparison to the existing measures. The remainder of this paper is organized as follows: definition of CCC in Sect. 2, heuristics to compute CCC are given in Sect. 3, experimental results showing the utility of CCC are presented in Sects. 4 and 5 and the last section provides some conclusions.

2 Clique Conversion Ratios and CCC

2.1 Notations

Let $G = (V, E)$ be an undirected graph where V is the set of nodes and E is the set of edges such that $|V| = n$ and $|E| = m$.

- C_p denotes the number of cliques of size p in G such that $2 \leq p \leq n$. C_2 is m.
- n_p denotes the number of nodes in G that participate in the formation of at least one clique of size p. We need at least p nodes to form at least one clique of size p. Hence $n_p \geq p$, if $C_p \geq 1$.
- r_{p+1} denotes the conversion ratio of cliques of size p to the cliques of size $p+1$ in G.

2.2 Conversion Ratios

We define r_{p+1} for $2 \leq p \leq n-1$ as

$$r_{p+1} = \begin{cases} \frac{C_{p+1}\ (p+1)}{C_p\ (n_p - p)} & \text{if } C_p \geq 1 \\ 0 & \text{if } C_p = 0 \end{cases} \tag{1}$$

It could be seen that a complete graph with k nodes has $r_{p+1} = 1, \forall p \in [2, k-1]$.

Combinatorial Justification of Ratios: In this section, we present a justification that the conversion ratio r_{p+1} correctly computes the ratio of conversion of cliques of size p to cliques of size $p+1$. As per our notation the number of cliques of size p are C_p and n_p nodes participate in at least formation of one clique of size p. For a clique $\{v_1, ..., v_p\}$ of size p there are at most $(n_p - p)$ nodes as choices to expand into a clique of size $p+1$. But for such an expansion to actually happen the node picked should have all p cliques of size p with each subset of $\{v_1, ..., v_p\}$ with size $(p-1)$. Therefore, C_p cliques each having $(n_p - p)$ nodes to expand into a maximum of $\frac{C_p(n_p-p)}{(p+1)}$ cliques of size $p+1$. The denominator comes from the observation that the node picked to expand a clique of size p could actually be forming all possible cliques of size p with subset of nodes in this clique. Therefore the resultant clique of size $(p+1)$ could at most be over counted $(p+1)$ times. We know that C_{p+1} is the number of cliques of size $p+1$. Therefore, the ratio of the number of cliques of size $p+1$ to the maximum number of possible cliques of size $p+1$ given C_p and n_p is our notion of conversion ratio, i.e.,

$$\frac{C_{p+1}}{\frac{C_p\ (n_p-p)}{p+1}} = \frac{C_{p+1}\ (p+1)}{C_p\ (n_p - p)} = r_{p+1} \tag{2}$$

Range of the Ratios: The definition of the conversion ratio can be viewed as the conditional probability of the event, where the formation of certain cliques of size $p+1$ is the event which is conditioned on the occurrence of an event in which certain cliques of size p are already formed by a set of n_p vertices. Therefore, the conversion ratios naturally have a range $[0, 1]$.

2.3 Global, Local and Average CCC

The Global CCC denoted by CCC_g is defined as the weighted average of the conversion ratios in a given graph defined for the following parameters:

1. A graph $G = (V, E)$ where $|V| = n$ and $|E| = m$.
2. A vector $\alpha = <\alpha_3, ..., \alpha_n>$ where α_i corresponds to the weight of the respective r_i and $\forall i \in [3, n],\ \alpha_i \in [0, 1],\ \sum_{i=3}^{n} \alpha_i = 1$.

$$CCC_g(G, \alpha) = \sum_{i=3}^{n} \alpha_i r_i$$

We now present global CCC values for two special graphs. Let $\alpha = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, ..., 0\}$.

1. In a *Petersen Graph* [9], we have 10 vertices and 15 edges connected as shown in Fig. 1(a). Let us denote the Petersen graph with H. In spite of its good connectivity it has no Cliques, i.e., all conversion ratios equal to 0. Hence, $CCC_g(H, \alpha)$ value is 0.

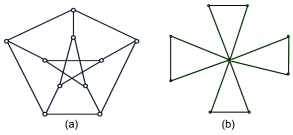

Fig. 1. (a) Petersen graph (b) F_4

2. A *Friendship Graph* [12], F_n has n triangles joined at a common vertex. Therefore, it has $2n+1$ vertices and $3n$ edges. This graph has no cliques of size greater than 3. Therefore, $CCC_g(F_n, \alpha) = \frac{1}{3(2n-1)}$ for $n \geq 2$. F_4 shown in Fig. 1(b), $CCC_g(F_4, \alpha) = \frac{1}{21}$.

   Let $u \in V$ and $N(u)$ denote the neighborhood of u such that every node in it is connected by an edge to u. The Local CCC denoted by CCC_l for u is calculated with respect to the subgraph H induced by $N(u)$. Assume H has p nodes. CCC_l is defined for the following parameters:
   1. A graph $G = (V, E)$ where $|V| = n$ and $|E| = m$.
   2. A node $u \in V$. Let us assume the neighborhood $N(u)$ induced subgraph H has p nodes.
   3. A vector $\alpha = <\alpha_2, ..., \alpha_p>$ where α_i corresponds to the weight of the respective r_i and $\forall i \in [2, p],\ \alpha_i \in [0, 1],\ \sum_{i=2}^{p} \alpha_i = 1$.

In local CCC, r_2 is also considered, because every edge in H corresponds to a clique of size three in G (as all nodes in H have an edge to u).

$$CCC_l(G, u, \alpha) = \sum_{i=2}^{p} \alpha_i r_i \tag{3}$$

The average CCC denoted by CCC_a is defined as an average on the local CCC of all nodes of G and a given vector $\alpha = <\alpha_2, ..., \alpha_n>$ where α_i corresponds to the weight of the respective r_i and $\forall i \in [2, n],\ \alpha_i \in [0, 1],\ \sum_{i=2}^{n} \alpha_i = 1$.

$$CCC_a(G, \alpha) = \frac{1}{n} \sum_{u \in V} CCC_l(G, u, \alpha) \tag{4}$$

The measure CCC in all its three variants has range $[0, 1]$. This follows from the fact that all the conversion ratios are defined to be in $[0, 1]$ and the choice that the sum of weights is to be 1.

2.4 Selection of Vector α

As per our experiments, different choices of vector α could strongly impact the value of CCC. One simple choice of α is to give each r_i equal weight. An example where this choice may not seem ideal for all requirements is as follows: Let us assume that a graph G consists of many triangles, only five cliques of size 4 and one clique of size 5. In this case, r_5 of G would be 1. This is because of all the Cliques of size 4 join to form a clique of size 5. In other words, the conversion of Cliques of size 4 if they exist to Cliques of size 5 is complete in G. If G has very low r_3 and r_4, r_5 would still skew the CCC value to be moderate to high. If such phenomenon is not wanted the weights in α could be set in decreasing order with increasing i values.

3 Computation of CCC

Computing CCC is a computationally intensive task. In all our experiments the heuristics presented in this section are used to compute conversion ratios and consequently CCC of any graph. The two heuristics to make the computation of CCC effective:

1. All vertices with degree less than $p-1$ can never form a clique of size p. Hence, these vertices can be pruned iteratively, so that the remaining vertices are selected as candidates that could possibly form cliques of size p.
2. Any vertex that can not form a clique of size $p-1$ also would not be able to form a clique of size i, when $i > p-1$. Therefore, only those vertices that form at least a clique of size $p-1$ are considered as candidates to check if they could form cliques of size p.

Let $G = (V, E), \alpha$ be the graph and the weight vector respectively which are the parameters required for computation of global CCC, $CCC_g(G, \alpha)$.

Complexity Analysis of the Algorithm Based on Heuristics: Let $\delta = \Delta(G)$ be the maximum degree in G. Let $\gamma = \delta + 1$ be the size of maximum clique possible in G. The upper bound on the running time of the algorithm is given by $\sum_{i=3}^{\gamma} \left(n_{i-1} \binom{\delta}{i-1} \right) = O(n\ \delta^{\delta})$, where n_{i-1} is the number of nodes that form at least one clique of size $i-1$. Nevertheless, in the worst case, where G is a complete graph, $\forall i,\ n_i = n$ and $\delta = n-1$. But on sparse graphs the heuristics presented however help to reduce the running time, when $\forall i,\ n_i << n$, and $\delta << n$.

4 Clustering of Graphs Using CCC as a Feature

Experimental Setup: In our experiments, we use the large networks of the SNAP Dataset [8]. These graphs are selected from four different categories: road networks, networks with ground truth communities, collaboration networks, and peer-to-peer networks. We find the conversion ratios and global CCC values of each network with the parameter α set as

$$\alpha_i = \begin{cases} \frac{1}{3} & \text{if } i \in \{3, 4, 5\} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Table 1 presents the $CCC_g(., \alpha)$ for each network where n, m represent the number of nodes and edges in the network respectively. The Road Networks(9, 10, 11) are almost planar and hence have very low conversion ratios, consequently low global CCC values. This is also the case with Peer-to-peer Gnutella networks(5, 6, 8) except the fact that these are relatively very small networks and that they have a high conversion of edges into triangles in comparison to the road networks. Interestingly, in the Amazon network(7) the conversion ratios all have

Table 1. Table presenting values of conversion ratios and global CCC values of each network with CCC values in decreasing order

ID	Dataset	n	m	r_3	r_4	r_5	$CCC_g(.,\alpha)$
1	CA-GrQc	5242	14496	0.19×10^{-2}	0.71×10^{-2}	0.14×10^{-1}	0.77×10^{-2}
2	CA-HepTh	9877	25998	0.33×10^{-3}	0.12×10^{-2}	0.48×10^{-2}	0.21×10^{-2}
3	CA-CondMat	23133	93497	0.24×10^{-3}	0.33×10^{-3}	0.53×10^{-3}	0.37×10^{-3}
4	com-dblp	317080	1049866	0.20×10^{-4}	0.11×10^{-3}	0.40×10^{-3}	0.18×10^{-3}
5	P2p-Gnuetella05	8846	31839	0.12×10^{-4}	0.15×10^{-3}	0	0.53×10^{-4}
6	P2p-Gnutella06	8717	31525	0.12×10^{-4}	0.13×10^{-3}	0	0.46×10^{-4}
7	com-amazon	334863	925872	0.65×10^{-5}	0.62×10^{-5}	0.69×10^{-5}	0.65×10^{-5}
8	p2p-Gnuetella04	10876	39994	0.64×10^{-5}	0.74×10^{-5}	0	0.46×10^{-5}
9	RoadNet-PA	1088092	1541898	0.12×10^{-6}	0.70×10^{-8}	0	0.42×10^{-7}
10	RoadNet-TX	1379917	1921660	0.94×10^{-7}	0.69×10^{-8}	0	0.34×10^{-7}
11	RoadNet-CA	1965206	2766607	0.67×10^{-7}	0.44×10^{-8}	0	0.24×10^{-7}

similar values. Also, this phenomenon happens in CA-CondMat(3) adding to the observation that this network is dense and has higher clique conversion in comparison to the Amazon network. In both these cases, any weight vector α would give more or less a similar CCC_g value for each network respectively. The collaboration networks CA-GrQc(1) and CA-HepTh(2) along with the ground truth community com-dblp(4) have increasing conversion ratios r_p as the p increases. Hence, CCC_g for these networks is sensitive to the changes in the weight vector α. All these three networks are dense and have very high clique conversion ratios.

Comparison Between CCC, Clustering, and Density for Graphs in our Dataset: Table 2 presents the normalized scores of $CCC_g(.,\alpha)$, clustering

Table 2. Z-normalization scores [10] of values of CCC, CC, and Density for each graph

ID	Dataset	$CCC_g(.,\alpha)$	CC	Density
1	CA-GrQc	2.9106	1.0074	1.3241
2	CA-HepTh	0.4851	0.7868	1.3241
3	CA-CondMat	−0.2511	1.3750	0.4712
4	com-dblp	−0.3319	1.3750	−0.9765
5	p2p-Gnuetella05	−0.3860	−0.9147	0.7058
6	p2p-Gnutella06	−0.3889	−0.9165	0.7484
7	com-amazon	−0.4057	0.5294	−0.9851
8	p2p-Gnuetella04	−0.4066	−0.9184	0.4286
9	RoadNet-PA	−0.4085	−0.7684	−1.0158
10	RoadNet-TX	−0.4085	−0.7684	−1.0171
11	RoadNet-CA	−0.4085	−0.7721	−1.0183

coefficient(CC) and density. Among the collaboration networks, CA-GrQc(1) and CA-HepTh(2) both have high density and a very different clique conversion nature. Also, CA-GrQc(1), CA-CondMat(3) and com-dblp(4) all have high CC values, but only CA-GrQc among them has high CCC. All the peer-to-peer networks(5, 6, 8) have moderate density values but very low CCC values. This table shows that networks could still be dissimilar with respect to their clique forming tendency in spite of having similar CC and density values.

Clustering Results: We perform three hierarchical clusterings with single linkage of the eleven networks. All values are taken from Table 2. The three clusterings are performed with: a) $CCC_g(.,\alpha)$ and CC as features, b) CC and $Density$ as features and c) $CCC_g(.,\alpha)$ and $Density$ as features. Euclidean distance is used to compute the similarity between any two networks in all clusterings. Figure 2 shows the clustering of the networks. The green dotted ovals represent the clusters, while the red dotted lines represent the level at which we break the dendrogram to obtain clusters. Table 3 gives an overview of the clustering results. Figure 2(a) shows the clustering obtained when $CCC_g(.,\alpha)$ and CC are used as features. The first cluster in the result is just CA-GrQc(1) which has both high CCC and CC. The second cluster consists of the networks CA-HepTh(2), CA-CondMat(3), com-dblp(4) and com-amazon(7) which have moderate to high values of CC but moderate CCC values. Only com-amazon(7) has a low CCC

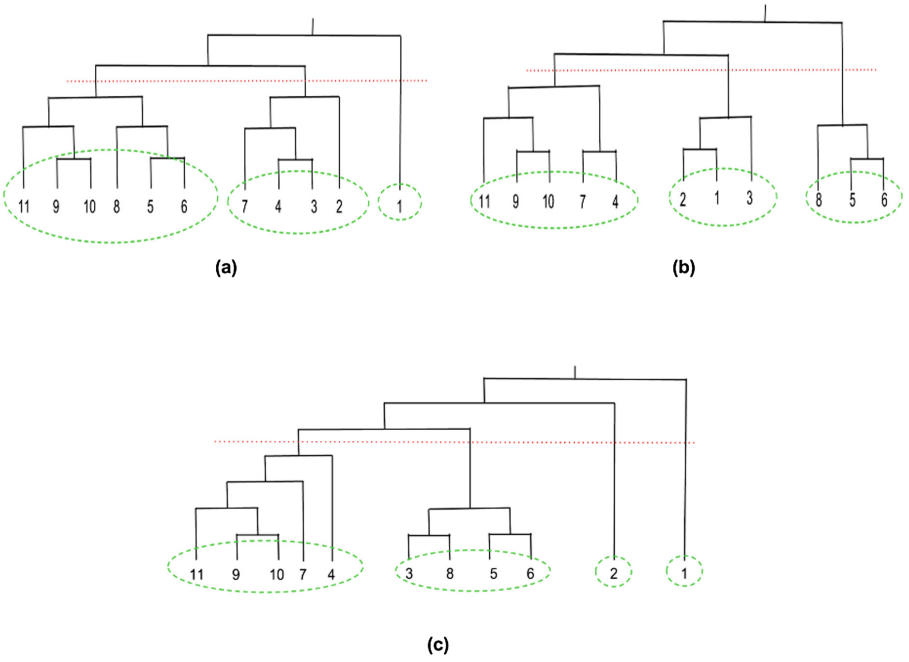

Fig. 2. The dendrogram for the three hierarchical clusterings of the networks. (Color figure online)

in this cluster. The third cluster is the peer-to-peer networks and the road networks which have very low CCC and CC values. Figure 2(b) shows the clustering obtained when CC and Density are used as features. The first cluster consists of the three collaboration networks(1, 2, 3) which have high CC and Density values. The second cluster consists of all peer-to-peer networks(5, 6, 8) which have high density and low CC. The third cluster consists of the rest of networks(4, 7, 9, 10, 11) which have low density and low CC. The only exception to this is com-dblp(4) which has the highest CC and very low Density. Interestingly, the CCC value of com-dblp(4) is in the range of moderate to low which is because its conversion ratios have almost same values. Figure 2(c) shows the clustering obtained when $CCC_g(., \alpha)$ and Density are used as features. The first cluster consists of only CA-GrQC(1) which has very high density and CCC. The second cluster consists of just CA-HepTh(2) which has high density and positive CCC. The third cluster consists of networks(3, 5, 6, 8) which have positive density and moderate CCC. The fourth cluster consists of networks(4, 7, 9, 10, 11) which have low values for both CCC and density. The important observation is that when CCC is used as a feature, the clusters are formed taking into account the similarity based on clique conversion ratio of graphs. But, in the absence of CCC, the clusters do not take into account such similarity. For example, the highly clique forming tendency in CA-GrQc is not considered in the clustering obtained when CCC is not used as a feature. As CCC gives additional insights into the clique forming tendency of graphs, we propose the use of CCC as one of the features in clustering to get better clustering results.

Table 3. Clustering results

Features used	Clusters obtained
CCC, CC	{{1}, {2, 3, 4, 7}, {5, 6, 8, 9, 10, 11}}
CC, Density	{{1, 2, 3}, {5, 6, 8}, {4, 7, 9, 10, 11}}
CCC, Density	{{1}, {2}, {3, 5, 6, 8}, {4, 7, 9, 10, 11}}

5 Clustering of Subgraphs of YOUTUBE Network

Experimental Setup: The main idea in this section is to cluster the subgraphs of a large YOUTUBE graph using the Conversion Ratios as features. We also cluster the same subgraphs using CC and Density as features. We find the frequent subgraphs in each cluster of the clustering obtained when Conversion Ratios (resp. CC and Density) are used as features. The two clusterings are compared based on the frequent subgraphs in each cluster. The YOUTUBE graph [17] has 1157827 users (nodes) and 4945382 user-to-user links (edges). In our experiments, we generate 200 subgraphs such that each subgraph is induced by the neighbors of a particular node, say, the ego node, of the corresponding subgraph. All ego nodes are selected to have exactly 20 neighbors.

We find the local CC and Density of the 200 subgraphs. The clustering of the subgraphs is performed by using DBSCAN [19] algorithm in Weka [18].

The attributes of the clustering are CC and Density. The parameters of DBSCAN are set as follows: $\epsilon = 0.04$ and $minPoints = 5$. The result is two clusters of subgraphs as presented in Table 4. Around 10% of the subgraphs−the *Outliers*−are not assigned to any cluster. Cluster 1 has subgraphs with low CC and high density whereas the Cluster 2 has subgraphs with high CC and density. Cluster 2 is a collection of subgraphs with nodes having a high tendency to group with each other. From the cluster centers in Table 4 we can note that density does not play a major role in Clustering. Majority of subgraphs belong to the Cluster 1 and have nodes with less tendency to group with eachother inspite of some subgraphs having high density.

Table 4. Clustering of YOUTUBE subgraphs using CC and Density

Cluster ID (Size)	Average $\{CC, Density\}$ in cluster
1 (167)	$\{0.0814, 0.3125\}$
2 (23)	$\{0.2453, 0.3416\}$

Frequent Subgraphs in the Clusters: To analyze the quality of the clustering lets look at the frequent subgraphs in each cluster ignoring the outliers. Each cluster is viewed as a Graph Database and each subgraph in the cluster as a transaction in it. Finding the frequent subgraphs in a cluster is then similar to finding frequent itemsets in a set of transactions. All vertices in each subgraph are labeled same. The Support S of each frequent subgraph I in a cluster is defined to be the number of subgraphs in the cluster that consist of an isomorphic substructure to I. We use GASTON (a frequent subGrAph, Sequences, and Tree ExtractiON)[16] to find the frequent subgraphs in each cluster. The minimum support is set as 40%. We find frequent subgraphs with at most 6 vertices. GASTON outputs the frequent subgraph patterns of three types: (i) frequent cyclic graphs, (ii) real trees (iii) paths. The frequent cyclic graphs are the patterns which consist of at least one cycle. The real trees are the patterns which are tree kind of structures that are not just simple paths from a node to another via intermediate nodes. We are mostly interested in analyzing the frequent cyclic graphs as they correspond to clique forming tendency. The frequent subgraphs in cluster 1 are 21 cyclic graphs and 8 real trees. The interesting frequent cyclic graphs in cluster 1 are presented in Fig. 3 along with their support. We can note

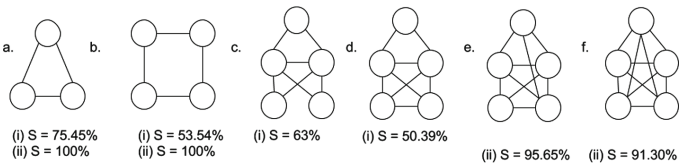

Fig. 3. Support for the presented frequent subgraphs in Cluster (i) and Cluster (ii) when CC and density are features.

that in spite of low CC values the support for cliques of size 3 (a. i) and four (b.i) in this cluster is over 75% and 50% respectively. More than 50% of the subgraphs in Cluster 1 have good clique forming tendencies in spite of low CC. The frequent subgraphs in cluster 2 are 23 cyclic graph and 8 real trees. The cliques of size 3 (a.ii) and 4 (b.ii) both have a support of 100% in Cluster 2. Also, a clique of size 5 (f.ii) has support over 90% in cluster 2. As cluster 2 has highly dense subgraphs with high CC this result is not surprising, but only 23 subgraphs are in this cluster out of the 200 subgraphs. The support for frequent subgraphs in the clusters shows us that both the clusters have many subgraphs that have high clique forming tendencies. To contrast, let us now use conversion ratios to find the clustering and analyze the frequent subgraphs in each cluster obtained. The conversion ratios r_2, r_3, r_4, r_5 of the 200 subgraphs are calculated and DBSCAN algorithm is used for the clustering with conversion ratios as the attributes of clustering. The parameters of DBSCAN are set as follows: $\epsilon = 0.07$ and $minPoints = 3$. DBSCAN outputs three Clusters with about 13% of the subgraphs in the 200 subgraphs being outliers. The details of the three Clusters obtained are presented in Table 5. Cluster 1 consists of 79 subgraphs with high clique forming tendency, while Cluster 2 consists of 91 subgraphs with low clique forming tendency. Cluster 3 consists of just four subgraphs with moderate values of r_2, but very high r_3, r_4 and r_5. Cluster 3 appeared because each of these subgraphs has fewer edges among the neighbors of the ego nodes in comparison to the subgraphs in Cluster 1. Most of these edges join to form at least a triangle. Also, most of those triangles join to form at least one clique of size 4. In other words, these subgraphs have nodes which once form at least an edge have high tendency to form Cliques of higher orders as well. This could be noticed from the Average Conversion Ratios values of the Cluster in Table 5.

Table 5. Clustering results of YOUTUBE subgraphs using the sequence of Clique Conversion Ratios

Cluster ID (Size)	Average $\{r_2, r_3, r_4, r_5\}$
1 (79)	{0.1744, 0.1545, 0.1422, 0.0875}
2 (91)	{0.0462, 0.0499, 0.0, 0.0}
3 (4)	{0.1158, 0.2679, 0.3929, 0.3611}

Frequent Subgraphs in the Clusters: Again, using GASTON we find the frequent subgraphs in each cluster obtained. We ignore the Cluster 3 (since it has only 4 subgraphs) and the Outliers from this discussion. The frequent subgraphs in Cluster 1 are 22 cyclic graphs and 8 real trees. Some interesting frequent cyclic graphs in Cluster 1 and Cluster 2 are presented in Fig. 4. In Cluster 1 the cliques of size 3, 4 and 5 (a.i, c.i and e.i) have support 100%, 94.94% and 41.77% respectively. This shows that the subgraphs in the cluster have higher clique forming tendency. The frequent subgraphs in Cluster 2 are 13 cyclic graphs and 8 real trees. In Cluster 2 the clique of size 3 (a.ii) has support over 54%. Cliques

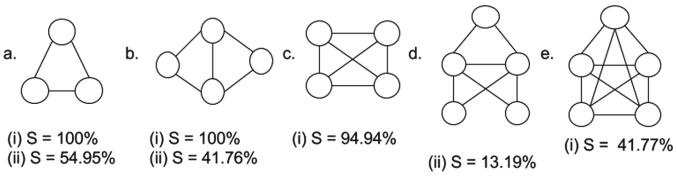

Fig. 4. Support for the presented frequent subgraphs in Cluster (i) and Cluster (ii) when conversion ratios are features

of size 4 and 5 are infrequent and have support less than 20%. Cluster 2 has subgraphs which do not have good clique conversions whereas Cluster 1 has subgraphs which have very good clique forming tendency.

6 Summary

The main contribution of this work is the measure CCC, which unlike existing measures, does not focus on the tendency of nodes to cluster alone, but also focuses on the tendency of nodes to form cliques. This measure gives new insights into graph properties, say, graphs with similar clustering coefficient or density, might have a lot of variation in their clique conversion ratios. We compare the clustering results of some real world graphs in the presence and absence of CCC. Our results show that clustering with CCC as a feature helps to Cluster the graphs with similar clique conversion ratios, while without CCC as a feature this is not always possible. This highlights the need for a measure like CCC. Also, we show the utility of CCC in clustering subgraphs of a large graph. The quality of the clusters obtained is verified using the frequent subgraph patterns in the clusters. This work could be further explored to find faster algorithms to compute Clique Conversion Ratios either deterministically or approximately if possible, to use the conversion ratios in the generation of synthetic graphs with desired clique forming tendencies. Also, the exploration of the uses of CCC in many applications areas has a good scope.

References

1. Luce, R.D., Perry, A.D.: A method of matrix analysis of group structure. Pschometrica **14**(1), 95–116 (1949)
2. Altaf-Ul-Amin, M., Nishikata, K., Koma, T., Miyasato, T., Shinbo, Y., Arifuzzaman, M., Wada, C., Maeda, M., Oshima, T.: Prediction of protein functions based on k-cores of protein-protein interaction networks and amino acid sequences. Genome Inf. **14**, 498–499 (2003)
3. Leskovec, J., Kleinberg, J.M., Faloutsos, C.: Graph evolution: densification and shrinking diameters. TKDD **1**(1) (2007)
4. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nat. Biotechnol. **18**(12), 1257–1261 (2000)

5. McAuley, J.J., Leskovec, J.: Discovering social circles in ego networks. TKDD **8**, 4:1–4:28 (2014)
6. Chen, B., Matsumoto, M., Wang, J., Zhang, Z., Zhang, J.: A short proof of Nash-Williams' theorem for the arboricity of a graph. Graphs Comb. **10**(1), 27–28 (1994)
7. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. Internet Math. **6**(1), 29–123 (2009)
8. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford Large Network Dataset Collection, June 2014. http://snap.stanford.edu/data
9. Holton, D.A., Sheehan, J.: The Petersen Graph. Cambridge University Press, Cambridge (1993). doi:10.2277/0521435943. ISBN 0-521-43594-3
10. Kreyszig, E.: Advanced Engineering Mathematics, 4th edn. Wiley, New York (1979)
11. Rosen, K.H.: Discrete Mathematics and Its Applications, 7th edn. McGraw-Hill (2011). p. 655
12. Erdos, P., Renyi, A., Sos, V.: On a problem of graph theory. Stud. Sci. Math. **1**, 215–235 (1966)
13. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977)
14. Awodey, S.: Isomorphisms. Oxford University Press, Category theory (2006)
15. Gao, X., Xiao, B., Tao, D., Li, X.: A survey of graph edit distance. Pattern Anal. Appl. **13**(1), 113–129 (2010)
16. Nijssen, S., Kok, J.: A quickstart in frequent structure mining can make a difference. In: Proceedings of the SIGKDD (2004). http://www.liacs.nl/home/snijssen/gaston
17. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC 2007), San Diego, CA, October 2007
18. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
19. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., Simoudis, E., Han, J., Fayyad, U.M. (eds.): A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231. AAAI Press (1996)