IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 8, 2020, accepted May 10, 2020, date of publication May 29, 2020, date of current version June 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2998744

Clustering and Closure Coefficient Based on $k - CT$ Components

PETR PROKOP, VÁCLAV SNÁŠEL, (Senior Member, IEEE), PAVLA DRÁŽDILOVÁ, (Member, IEEE), AND JAN PLATOŠ, (Member, IEEE)

Department of Computer Science, Faculty of Electrical Engineering and Computer Science, VSB—Technical University of Ostrava, 70800 Ostrava, Czech Republic

Corresponding author: Jan Platoš (jan.platos@vsb.cz)

This work was supported by Student Grant Competition (SGS), VŠB—Technical University of Ostrava, Czech Republic, through the Parallel Processing of Big Data VII, under Grant SP2020/108.

ABSTRACT Real-world networks contain many cliques since they are usually built from them. The analysis that goes behind the cliques is fundamental because it discovers the real structure of the network. This article proposed new high-order closed trail clustering and closure coefficients for evaluation of the network structure. These coefficients are able to describe the inner structure of the network concerning its randomized or organized behavior. Moreover, the coefficients can cluster networks with similar structures together. The experiments show that the coefficients are useful in both the local and global context.

INDEX TERMS Closed trail distance, clustering coefficient, closure coefficient, cyclic structure, higher-order structure.

I. INTRODUCTION

The networks that are built in real life have many standard features. The most important feature is related to their evolution and the process through which they are created. Co-authorship networks depict the fact that authors co-authored a book or a paper. If the book has three authors, a clique on three vertices is added to the network, similarly for more authors. Social networks suggest a connection between members according to the current relationship, regardless of whether it is on the internet or in real life. Therefore, the analysis that goes behind the cliques is fundamental.

The global clustering coefficient (or transitivity [1]) is a standard approach to characterizing networks and the tendency of vertices to clustering. In the article [2], a higher-order clustering coefficient as a natural generalization of the traditional clustering coefficient is defined. Higher-order cliques beyond triangles are crucial to an understanding of complex networks and the clustering behavior of their vertices concerning the standard metric on network structures.

The local closure coefficient [3] is defined in a similar way to the standard local clustering coefficient. It is a metric quantifying head-node-based edge clustering, and it is defined as the fraction of length-2 paths starting from the head node that induces a triangle. This small difference in definition, leads to different properties than the traditional clustering coefficient has. Benson *et al.* [4] developed a generalized framework for clustering networks based on higher-order connectivity patterns. Their results show that networks exhibit rich higher-order organizational structures detected by clustering based on higher-order connectivity patterns. The article [5] continues with the idea of higher-order graph clustering, and the authors present a class of local graph clustering methods that incorporate higher-order network information captured by network motifs. The higher-order structure is also the focus of the article [6]. The authors found that tie strength and edge density are the competing positive indicators of higher-order organization. These trends are consistent across interactions that involve a different number of nodes.

The measuring of the distances between two nodes in a graph is a frequent task. The standard measure for this distance is the shortest path ($d_{SP}(u, v)$) between two nodes $u,\ v$ in a graph [7], [8]. Another way that can be used is the expected lengths of the commuting time distance [9]. Variants of node distances are described in detail in [10]–[13].

The closed trail distance ($d_{CT}(u, v)$) as a metric in the graph is based on the definition of a biconnected component. The distance between two vertices in the graph is defined as the length of the shortest closed trail that contains these

The associate editor coordinating the review of this manuscript and approving it for publication was Tu Ngoc Nguyen.

two vertices. A $k - CT$ component is maximal subgraph that contains those vertices for which the closed trail distance among the vertices is less than or equal to k.

The $k - CT$ components that are detected highlight locally and cyclic connected subgraphs. Moreover, these components are not based on the biconnectivity property and, therefore, they can easily partition densely connected biconnected components. These components are more difficult to partition and detect the structure of communities. A list of the largest biconnected component in the selected networks was published by Leskovec *et al.* [14].

Local clustering and closure coefficients measure the tendency of vertices to be in a cluster. Both are based on the expansion of the clique. The higher-order clustering and closure coefficients are based on higher-order (bigger) cliques. In the graph, we can detect a dense subgraph, which is not a clique, but it is very close to a clique. This subgraph can be composed of numerous smaller cliques, and they create the $k - CT$ component. The new approach to clustering and closure coefficient is based on the expansion of $k - CT$ components to a $(k + 1) - CT$ subgraph. Higher-order clustering and closure coefficients are integrated into the clustering and closure $k - CT$ coefficients because all $3 - CT$ components are cliques. Sparser subgraphs, which can contain structural holes of the graph or chains of $k - CT$ components with a smaller k, are detectable via $k - CT$ components with a higher k.

The organization of the article is as follows. First, the terminology and the notation, which is used in the article, are introduced. In the next section, the closed trail distance in connected undirected graphs without bridges is defined. Moreover, the new coefficients based on $k - CT$ components are introduced. These coefficients extend the clustering and closure coefficient and characterize the tendency of vertices to participate in some $(k+1) - CT$ subgraph. Section IV contains the experimental results of selected real networks and two types of generated networks. In conclusion, the advantages and limitations of the coefficients that are defined are discussed.

II. TERMINOLOGY AND NOTATION

In this section, knowledge of graph theory will be required. The definitions of the following terms were taken from [15]:

A *walk* on a graph is an alternating series of vertices and edges

$$W(v_0, v_k) = v_0 e_1 v_1 e_2 \dots v_{k-1} e_k v_k,$$

such that for $j = 1, \dots, k$ the vertices v_{j-1} and v_j are the endpoints of the edge e_j. A *closed walk* is a walk in which the initial vertex is also the final vertex. The *length of a walk* is the number of edges. We will denote the length of a walk as $|W(u, v)|$. A *trail* is a walk in which no edge occurs more than once. A *closed trail* is a closed walk with no repeated edges. We will denote a closed trail which contains the vertices u, v as

$$CT(u, v) = ue_1\, v_1 e_2 \dots v \dots e_k u.$$

A *path* is a walk in which no edge or internal vertex occurs more than once (a trail in which all the internal vertices are distinct). We will denote a path with an initial vertex u and a final vertex v as $P(u, v)$. A *circuit* is a closed trail. A *cycle* is a closed path with a length at least one and an induced cycle of length four or more is a *hole*. A *clique* is a subgraph in which each vertex is adjacent to every other vertex. We will denote the clique with k vertices as Q_k. A *diameter of graph* is the maximum of distances between any pair of vertices in the graph.

A *connected* graph is a graph such that between every pair of vertices, there exists a walk. A *biconnected* graph is a connected and "nonseparable" graph, meaning that if any vertex were to be removed, the graph would remain connected. A *component* of a graph is a maximal connected subgraph. An edge e is a *bridge* (cut-edge) of the connected graph G if $\{e\}$ is a disconnecting edge-set of G. An *articulation* is a vertex of a graph which removal increases the number of components. Therefore, a biconnected graph has no articulation vertices. A *biconnected component* is a maximal biconnected subgraph.

III. COEFFICIENTS BASED ON CLOSED TRAIL DISTANCE

The closed trail distance is a metric between vertices in a connected graph without bridges and loops. It is useful for the detection of subgraphs with a specified $CT-$distance among the vertices.

A. CLOSED TRAIL DISTANCE IN AN UNDIRECTED GRAPH

Definition 1: A graph is a k-closed trail connected graph $(k - CT)$ if every two vertices lie on the closed trail (circuit) with a length $\leq k$. The $k - CT$ component of the graph is a maximal $k - CT$ subgraph.

A maximal $k - CT$ subgraph is a $k - CT$ subgraph that cannot be extended by including one more adjacent vertex.

Definition 2: Let $G = (V, E)$ be a graph. Let $d_{CT} : V \times V \to R_0^+$ be defined by the equation

$$d_{CT}(u, v) = min_{CT(u,v) \in G} |CT(u, v)|,$$

where $CT(u, v)$ is a closed trail that contains the vertices $u,\ v$. Then the function d_{CT} is called the closed trail distance (CT-distance).

Theorem 1: The $CT-$distance is a metric on the set V.

The theorem was proven in the article [16].

Lemma 1: Every $3 - CT$ component is a clique.

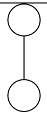
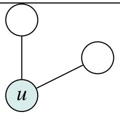
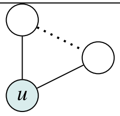
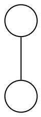
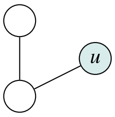
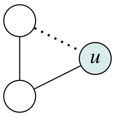
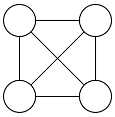
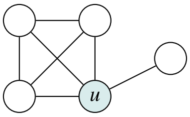
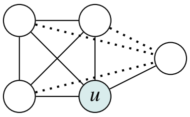
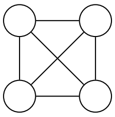
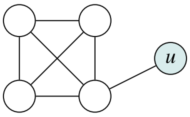
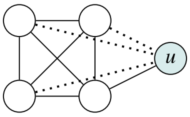
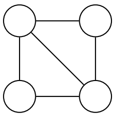
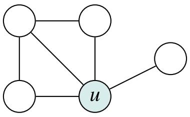
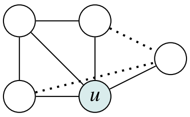
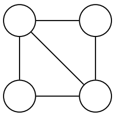
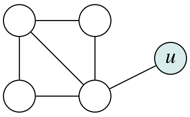
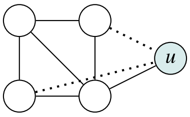
The lemma was proven in the article [16].

Lemma 2: $d_{CT}(u, v)$ is a metric in any connected graph without bridges and defines the distance between two nodes u and v.

We can define the $CT-$distance for a disconnected or connected graph with bridges in this way:

Definition 3: The $CT-$distance between the vertices u and v is equal to ∞ ($d_{CT}(u, v) = \infty$) if not closed trail containing these vertices exists.

TABLE 1. The list of the coefficients. The first column contains the formulas of the coefficients, and the other columns are examples of starting subgraphs (cliques or $k-CT$ components), starting subgraphs with an added edge, that is centered or headed in the vertex u and checked subgraphs (cliques or $(k+1)-CT$ subgraphs.

Local Coefficient	Starting subgraph	k-wedge	Checked subgraph
$C_2(u)=\frac{2\lvert K_3(u)\rvert}{\lvert W_2^c(u)\rvert}$			
$H_2(u)=\frac{2\lvert K_3(u)\rvert}{\lvert W_2^h(u)\rvert}$			
$C_4(u)=\frac{4\lvert K_5(u)\rvert}{\lvert W_4^c(u)\rvert}$			
$H_4(u)=\frac{4\lvert K_5(u)\rvert}{\lvert W_4^h(u)\rvert}$			
$C_4^{CT}(u)=\frac{\lvert 5-CT(u)\rvert}{\lvert W_{4-CT}^c(u)\rvert}$			
$H_4^{CT}(u)=\frac{\lvert 5-CT(u)\rvert}{\lvert W_{4-CT}^h(u)\rvert}$			

B. HIGHER-ORDER CLUSTERING AND CLOSURE COEFFICIENTS

The local clustering coefficient [17] of a vertex u in the network $G=(V,E)$ is the fraction of wedges centered at the vertex u that are closed. The wedge W_2^c is a subgraph composed of a clique Q_2 and an edge which are connected in the vertex u (see Table 1 coefficient C_2-2-wedge). The local higher-order clustering coefficient for the vertex u is defined in [2] as:

$$C_k(u)=\frac{k|K_{k+1}(u)|}{|W_k^c(u)|}=\frac{k|K_{k+1}(u)|}{(d_u-k+1)|K_k(u)|},$$

where $K_k(u)$ is the set of k-cliques containing u, $W_k^c(u)$ is the set of k-wedges (see Table 1 coefficient C_4-4-wedge) with its center in the vertex u and d_u is the degree of the vertex u. If $|W_k^c(u)|=0$, then $C_k(u)$ is undefined. The average kth-order clustering coefficient $\overline{C}_k$ is the mean of the local kth-order clustering coefficients,

$$\overline{C}_k=\frac{1}{|V_k|}\sum_{u\in V_k}C_k(u),$$

where V_k is the set of nodes in the network in which the local kth-order clustering coefficient is defined.

The global higher-order clustering coefficient of the network $G=(V,E)$ is defined in [2] as:

$$C_k=\frac{(k^2+k)|K_{k+1}|}{|W_k^c|},$$

where K_{k+1} is the set of $(k+1)$-cliques in G and W_k^c is the set of k-wedges, where a k-wedge is composed of a clique with k vertices and an edge. They are connected in the vertex u which is common for the clique and the edge.

A closure coefficient [3] is defined in a similar way.

The local closure coefficient of a vertex u in the network $G=(V,E)$ is the fraction of the wedges headed at the vertex u that are closed. The wedge W_2^h is a subgraph composed of a clique Q_2 and an edge which are connected in the vertex v and the vertex u is the head of the edge (see Table 1 coefficient H_2-2-wedge). The local higher-order closure coefficient for the vertex u is defined as:

$$H_k(u)=\frac{k|K_{k+1}(u)|}{|W_k^h(u)|},$$

where $K_k(u)$ is the set of k-cliques containing u and $W_k^h(u)$ is the set of k-wedges (see Table 1 coefficient H_4-4-wedge) headed in the vertex u. If $|W_k^h(u)|=0$, then $H_k(u)$ is undefined. The average kth-order closure coefficient $\overline{H}_k$ is the mean of the local kth-order closure coefficients,

$$\overline{H}_k=\frac{1}{|V_k|}\sum_{u\in V_k}H_k(u),$$

where V_k is the set of nodes in the network in which the local kth-order closure coefficient is defined.

C. CLUSTERING AND CLOSURE COEFFICIENTS BASED ON k − CT COMPONENTS

We denote the set of all $k-CT$ subgraphs containing the vertex u as $k-CT(u)$. The set of all the $k-CT$ components which contain the vertex u is denoted as $M_k(u)$. From the

TABLE 2. Overview of networks parameters used for experiments.

Networks group	Nodes	Edges	Density	Transitivity
Networks generated using BA model	100 - 200	196 - 7500	0.019 - 0.505	0.032 - 0.619
Networks generated using WS model	100 - 200	500 - 1400	0.067 - 0.101	0.057 - 0.692
Real world networks	33 - 18131	77 - 790261	0.001 - 0.146	0.047 - 0.661

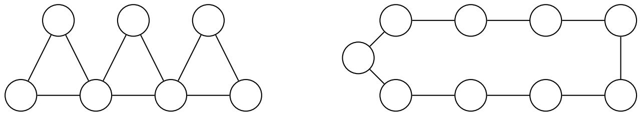


FIGURE 1. Example of $9-CT$ components where the first component (chain of triangles) has – $dim_{SP}=3$, $dim_{CT}=9$, length of the largest cycle is 3 and the second component (cycle) has – $dim_{SP}=4$, $dim_{CT}=9$, length of the largest cycle is 9.

definition 1 it follows that the set $M_k(u)$ contains all the maximal $k-CT$ subgraphs with the vertex u. We define an $k-CT$ wedge centered at u as consisting of an $k-CT$ component and an adjacent edge for $k\geq 3$ (see Fig. 1, row C_4^{CT}). A centered $k-CT$ wedge is denoted as W_{k-CT}^c. A $k-CT$ wedge headed at u is defined in a similar way (see Table 1, row H_4^{CT}) and it is denoted as W_{k-CT}^h.

The shortest closed trail which contains two vertices has to have a length greater than or equal to 3. It is the reason why coefficients are defined for $k\geq 3$.

The local $k-CT$ clustering coefficient of a vertex u in the network $G=(V,E)$ is the fraction of the CT wedges centered at the vertex u that are closed. The local higher-order $k-CT$ clustering coefficient for the vertex u is defined as:

$$C_k^{CT}(u)=\frac{|(k+1)-CT(u)|}{|W_{k-CT}^c(u)|}.$$

If $|W_{k-CT}^c(u)|=0$, then $C_k^{CT}(u)$ is undefined. The average kth-order clustering CT coefficient $\overline{C}_k^{CT}$ is the mean of the local kth-order clustering CT coefficients,

$$\overline{C}_k^{CT}=\frac{1}{|V_k|}\sum_{u\in V_k}C_k^{CT}(u),$$

where V_k is the set of nodes in the network where the local kth-order clustering CT coefficient is defined.

The interpretation of the local $k-CT$ clustering coefficient, is described as the expansion of $k-CT$ components to $(k+1)-CT$ subgraphs (see Table 1, row C_4^{CT}). The global $k-CT$ clustering coefficient C_k^{CT} is defined as the fraction of the $k-CT$ wedges centered at u that are closed, meaning that they induce a $(k+1)-CT$ subgraph in the network. We can formulate this as:

$$C_k^{CT}=\frac{1}{|M_k|}\sum_{k-CT\in M_k}\frac{|(k+1)-CT|}{|W_{k-CT}^c|}.$$

The local $k-CT$ closure coefficient of a vertex u in the network $G=(V,E)$ is the fraction of the CT wedges headed at the vertex u that are closed. The local higher-order $k-CT$ closure coefficient for the vertex u is defined as:

$$H_k^{CT}(u)=\frac{|(k+1)-CT(u)|}{|W_{k-CT}^h(u)|}.$$

If $|W_{k-CT}^h(u)|=0$, then $H_k^{CT}(u)$ is undefined. The average kth-order closure CT coefficient $\overline{H}_k^{CT}$ is the mean of the local kth-order closure CT coefficients,

$$\overline{H}_k^{CT}=\frac{1}{|V_k|}\sum_{u\in V_k}H_k^{CT}(u),$$

where V_k is the set of nodes in the network in which the local kth-order closure CT coefficient is defined.

D. METHODS FOR $k-CT$ COMPONENTS COMPUTATION

In the graph $G=(V,E)$ we need to detect all the maximal $k-CT$ subgraphs ($k-CT$ components) for the computation of the coefficients. $k-CT$ components are detected from the matrix of closed trail distances. We denote this full matrix as T and $T_{ij}=d_{ct}(i,j)$.

All the triangles and quadrangles in the graph are detected to fill the matrix T. the Chiba and Nishizeki algorithm [18] is used for these computations. The $CT-$ distances $d_{ct}(i,j)\geq 5$ are detected via the connection of the two shortest disjoint paths [19] between i and j, where the connection of these shortest paths creates a closed trail.

The $k-CT$ component in the graph $G=(V,E)$ is the maximal clique in the weighted graph $G_k=(V,E_k)$ where $\{i,j\}\in E_k \Leftrightarrow T_{i,j}\leq k$. Maximal cliques in G_k are detected with the Bron-Kerbosch algorithm [20].

IV. EXPERIMENTAL RESULTS

The experiments concentrate on comparing standard coefficients and $k-CT$ coefficients in selected real networks and two types of generated networks. Real networks were used for the experiments, as in the article [2]. Biological networks are represented by dataset C.elegans (a complete neural system) and Dros.-medulla (neural connections). Zachary Karate Club is a real small social network, and fb-Stanford and fb-Cornell are online friendship social networks on Facebook among students at universities since 2005. Two co-authorship networks are constructed from arxiv submission categories (arxiv-AstroPh and arxiv-HepPh). Human communication networks are created from emails (email-Enron-core, email-Eu-core) and Facebook-like messages among colleges (CollegeMsg). Oregon2-01052 is a technological network of an autonomous system.

A Barabási-Albert (BA) model [21] of a network was used for generating 14 networks with increasing numbers of edges $(2,3,5,7,\ldots,50)$ attached from a new node to existing nodes. The process of generating was repeated 15 times with 5 various random seeds, and the maximal number of vertices was $n\in\{100,150,200\}$. The result of the generating is 210 networks.

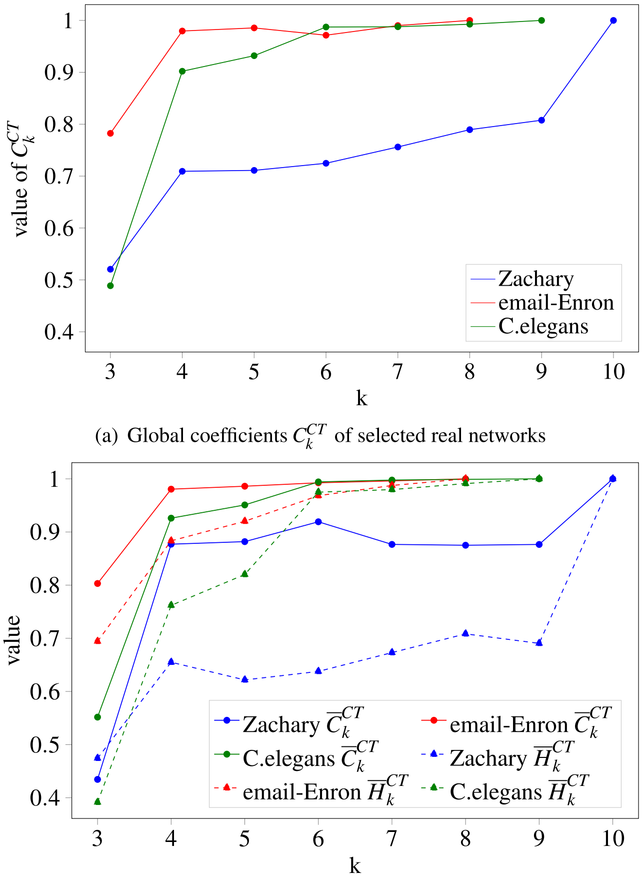
A Watts-Strogatz (WS) model [17] of the network is the second model which was used for generating 20 networks

TABLE 3. A short description of the largest connected components without bridges in selected real networks. The shortest path distance and $CT-$distance between vertices in the largest connected components without bridges were used for the computation of diameters and average distances.

Network name	Nodes	Edges	Density	dim_{SP}	Avg. d_{SP}	dim_{CT}	Avg. d_{CT}
C.elegans	282	2133	0.054	5	2.399	10	5.238
Dros.-medulla	1373	8613	0.009	5	2.586	11	5.996
Zachary	33	77	0.146	5	2.316	11	5.433
fb-Stanford	11081	567804	0.009	8	2.728	17	5.695
fb-Cornell	18131	790261	0.005	7	2.822	15	5.881
arxiv-HepPh	9945	116128	0.002	11	4.483	28	9.806
arxiv-AstroPh	16829	195893	0.001	14	4.073	33	8.738
email-Enron-core	143	1425	0.140	4	2.132	9	4.638
email-Eu-core	891	16575	0.042	5	2.435	11	5.136
CollageMsg	1498	13440	0.012	6	2.785	12	5.963
oregon2-010526	8110	29379	0.001	8	3.344	16	7.397

TABLE 4. List of selected real networks with number of $k - CT$ components.

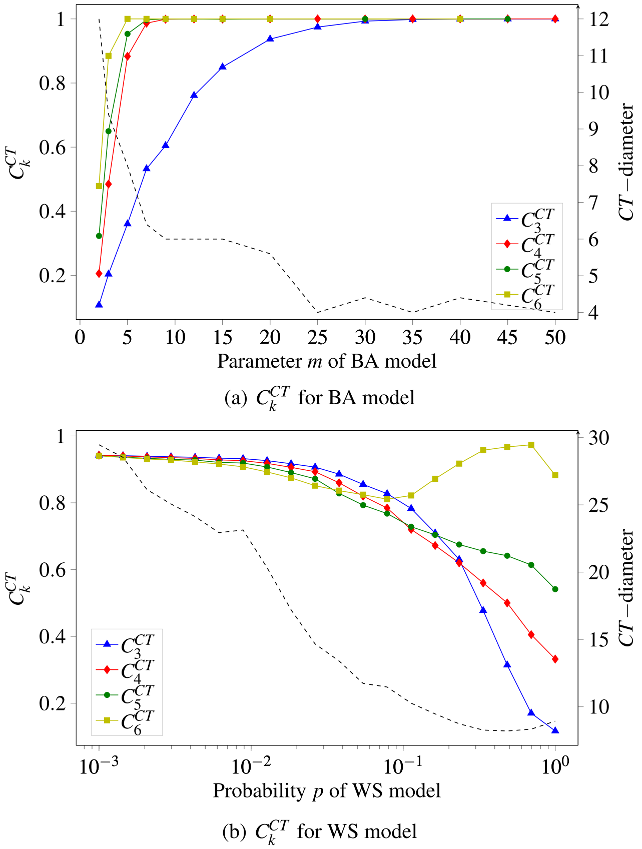
Network name	$3-CT$	$4-CT$	$5-CT$	$6-CT$	$7-CT$	$8-CT$	$9-CT$	$10-CT$	$11-CT$
Zachary	26	19	11	9	5	4	3	2	1
C.elegans	1 271	109 322	13 249 344	479 622	216 916	18	2	1	
Dros.-medulla	6 196	$> 10^9$	$> 10^9$	$> 10^9$	16 336 129	322	80	7	1
email-Enron-core	654	37 537	59 559	69	16	7	1		



(a) Global coefficients C_k^{CT} of selected real networks

(b) Average coefficients $\overline{C}_k^{CT}$ and $\overline{H}_k^{CT}$ of selected real networks

FIGURE 2. Coefficients C_k^{CT}, $\overline{C}_k^{CT}$ and $\overline{H}_k^{CT}$ of selected real networks.



(a) C_k^{CT} for BA model

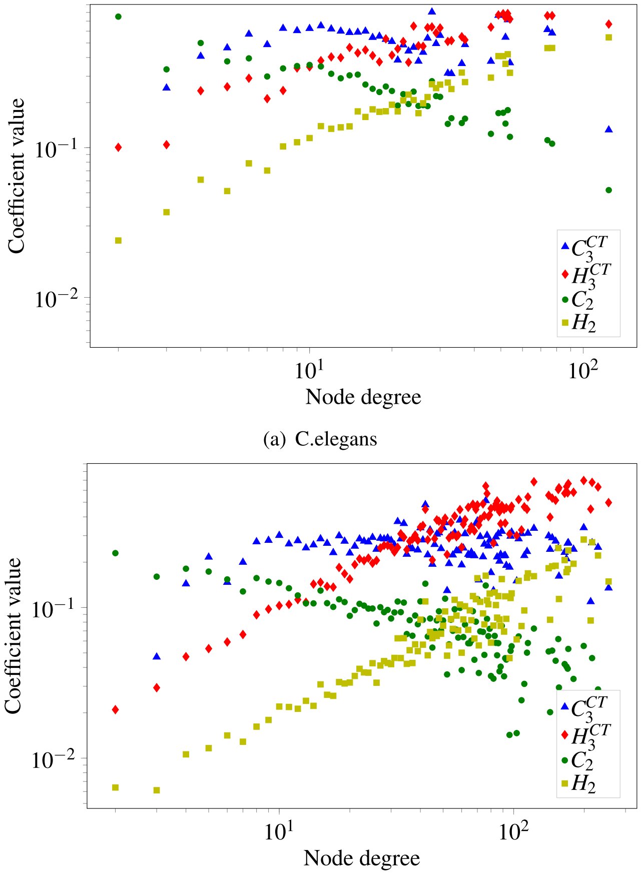
(b) C_k^{CT} for WS model

FIGURE 3. Graphs of global $k - CT$ clustering coefficients in the BA and WS network models; the dashed line represent the average $CT-$diameter of generated networks with specific parameters.

with an increasing rewiring probability p, the number of vertices is $n \in \{100, 150, 200\}$ and the number of neighbors in the ring topology is $k \in \{10, 15\}$. 5 various random seeds were used with a combination of the number of vertices and number of neighbors (100, 10), (150, 10), (200, 15). The result of the generating is 300 networks.

A brief description of network parameters for all types of networks are summarized in table 2.

Table 3 contains two specific networks (arxiv-HepPh and arxiv-AstroPh) that have the biggest average shortest path distances and their $CT-diameters$ (dim_{CT}) are not $(2*dim_{SP})$ or $(2*dim_{SP}+1)$. Figure 1 describes a situation in which the dim_{CT} is bigger than the $(2*dim_{SP}+1)$.



(a) C.elegans

(b) CollageMsg

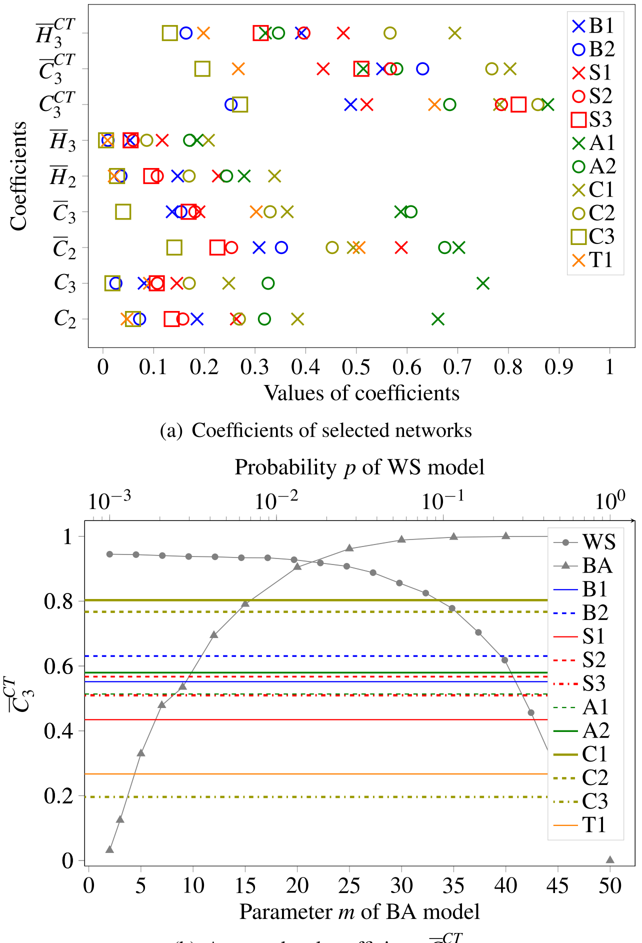
FIGURE 4. **Comparison of local clustering, closure, and $3 - CT$ clustering and $3 - CT$ closure coefficients ($C_2, H_2, C_3^{CT}, H_3^{CT}$) in terms of the dependency on the node degree.**

All the $k - CT$ coefficients are calculated for smaller networks. Bigger networks have a huge number of $k - CT$ components (see Table 4) which leads to more expensive computation of the $k - CT$ coefficients for $k \geq 4$. Higher-order clustering and closure coefficients use cliques of various sizes, as do the C_3^{CT} and H_3^{CT} coefficients.

The global coefficient C_k^{CT} is a fraction of $k-CT$ centered wedges that are closed, meaning that they induce a $(k+1)-CT$ subgraph. In the situation when $k = (dim_{CT} - 1)$ then every $k-CT$ centered wedge has to be closed to the $(k+1)-CT$ subgraph because the dim_{CT} is the maximal value of the $CT-$distance between the vertices and the closed wedge has to have a maximal $CT-$distance between the vertices equal to dim_{CT} of the graph. Then the coefficient $C_{(dim_{CT}-1)}^{CT}$ has a value equal to one (see Figure 2(a)).

The average clustering and closure $k - CT$ coefficients $\overline{C}_k^{CT}, \overline{H}_k^{CT}$ have similar behavior for selected networks to the global coefficient C_k^{CT} (see Figure 2). Figure 2(b) shows the tendencies of average coefficients.

With an increase in the parameter m, networks created with the BA model (see Figure 3(a)) have higher density and then a smaller diameter of the network and an increasing value of C_k^{CT}, which goes to one, either faster or slowly. These networks with $m \geq 5$ have $dim_{CT} \leq 7$ and then the value of C_6^{CT} is equal to one.



(a) Coefficients of selected networks

(b) Average local coefficients $\overline{C}_3^{CT}$

FIGURE 5. **Comparison of coefficients' values (Figure 5(a)) and a detailed view of the average $3 - CT$ clustering coefficients (Figure 5(b)) of real networks compared with the BA and WS network models. The networks were grouped into categories according to the origin of the network and relabeled for better illustration of plots and better readability. The group of biological networks – C.elegans (B1), Dros.-medulla (B2), social networks – Zachary (S1), fb-Stanford (S2), fb-Cornell (S3), collaboration networks arxiv-HepPh (A1) and arxiv-AstroPh (A2), communication networks email-Enron-core (C1), email-Eu-core (C2), CollageMsg (C3), and finally the technological network oregon2-010526 (T1).**

Networks created with the WS model (see Figure 3(b)) started with a regular graph and an increasing rewiring probability p, causing a random graph with the same density and with a smaller $CT-$diameter. The coefficient C_6^{CT} increases in the range $\{5, \ldots, 11\}$ of the $CT-$diameter with increasing p. The coefficient C_3^{CT} for increasing p decreases to zero because the $CT-$diameter decreases to 8 and the density is the same as in the regular network.

The $3-CT$ clustering coefficient is calculated with cliques of all sizes and the resulting value is appropriate to the cumulative value of the higher-order clustering coefficients. The coefficients C_3^{CT} and C_2 have very often a similar tendency (see Figure 4) when they depend on the node degree. The same tendency is more significant for the closure coefficients H_3^{CT} and H_2 in the selected network (see Figure 4(b)).

The selected real networks have different values for their global and average local coefficients (see Figure 5(a)). The $3 - CT$ coefficients have mostly a greater range and a higher value. The calculation of the coefficients is not restricted only to cliques. The $3 - CT$ coefficients represent the fraction of $4-CT$ subgraphs to wedges based on cliques. The $4-CT$ subgraphs are still dense, but they are not as strict as the cliques. The extension to the $k-CT$ components allows the calculation with parts of the graph with $k - CT$ distance between the vertices.

V. CONCLUSION

This article suggests a new higher-order closed trail based clustering and closure coefficients that were designed for the discovery of the features of networks that are behind their clique-based structure. In many networks, cliques represent the way the network is built. The co-authorship networks contain cliques of co-authors connected with other cliques using common authors. Actor-Actor networks are build using the interconnection of cliques of actors. Therefore, the structure behind the cliques is the real structure of the networks. The coefficients C_3^{CT} and H_3^{CT}, as well as their averaged values $\overline{C}_3^{CT}$ and $\overline{H}_3^{CT}$, provide completely new knowledge about networks. The coefficients' values can identify the nature of the networks and consider their chaotic or organized behavior. Moreover, we demonstrated the relationship between the selected networks with the Barabási-Albert and Watts-Strogatz models. We computed coefficients for both models with different parameters, and any network may be compared to them, and the most similar parameter for each model may be chosen. Both parameters may be used as features for network similarity measurements because of the differences in the behavior of each model. The experiments were performed on the largest connected components without bridges of 11 well-known networks with hundreds of nodes up to seventeen thousand and up to eight hundred thousand edges. The results show that the coefficients are able to distinguish between different types of networks and cluster the networks across the source area.

REFERENCES

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, vol. 8. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[2] H. Yin, A. R. Benson, and J. Leskovec, "Higher-order clustering in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 97, no. 5, May 2018, Art. no. 052306.

[3] H. Yin, A. R. Benson, and J. Leskovec, "The local closure coefficient: A new perspective on network clustering," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*. New York, NY, USA: ACM, Jan. 2019, pp. 303–311.

[4] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, Jul. 2016.

[5] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: ACM, Aug. 2017, pp. 555–564.

[6] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 48, pp. E11221–E11230, Nov. 2018.

[7] A. R. Ashrafi and M. V. Diudea, *Distance, Symmetry, and Topology in Carbon Nanomaterials*, vol. 9. Cham, Switzerland: Springer, 2016.

[8] W. Goddard and O. R. Oellermann, "Distance in graphs," in *Structural Analysis of Complex Networks*. Boston, MA, USA: Birkhäuser, 2011, pp. 49–72.

[9] F. Göbel and A. Jagers, "Random walks on graphs," *Stochastic processes their Appl.*, vol. 2, no. 4, pp. 311–336, 1974.

[10] P. Chebotarev, "The walk distances in graphs," *Discrete Appl. Math.*, vol. 160, nos. 10–11, pp. 1484–1500, Jul. 2012.

[11] M. M. Deza and E. Deza, "Encyclopedia of distances," in *Encyclopedia of Distances*. Berlin, Germany: Springer-Verlag, 2013, pp. 1–583.

[12] E. Estrada, "The communicability distance in graphs," *Linear Algebra Appl.*, vol. 436, no. 11, pp. 4317–4328, Jun. 2012.

[13] U. V. Luxburg, A. Radl, and M. Hein, "Getting lost in space: Large sample analysis of the resistance distance," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2622–2630.

[14] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Math.*, vol. 6, no. 1, pp. 29–123, Jan. 2009.

[15] J. L. Gross and J. Yellen, *Handbook of Graph Theory*. Boca Raton, FL, USA: CRC Press, 2004. [Online]. Available: http://books.google.com/?id=mKkIGIea_BkC

[16] V. Snasel, P. Drazdilova, and J. Platos, "Closed trail distance in a biconnected graph," *PLoS ONE*, vol. 13, Aug. 2018, Art. no. e0202181.

[17] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, p. 440, 1998.

[18] N. Chiba and T. Nishizeki, "Arboricity and subgraph listing algorithms," *SIAM J. Comput.*, vol. 14, no. 1, pp. 210–223, Feb. 1985.

[19] J. W. Suurballe, "Disjoint paths in a network," *Networks*, vol. 4, no. 2, pp. 125–145, 1974.

[20] C. Bron and J. Kerbosch, "Algorithm 457: Finding all cliques of an undirected graph," *Commun. ACM*, vol. 16, no. 9, pp. 575–577, Sep. 1973.

[21] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

PETR PROKOP received the B.Sc. and M.Sc. degrees in computer science from the VSB—Technical University of Ostrava, Ostrava, Czech Republic, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree in computer science focused on data science and machine learning. His current research interests include big data and social network analysis.

VÁCLAV SNÁŠEL (Senior Member, IEEE) received the master's degree in numerical mathematics from the Faculty of Science, Palacky University, Olomouc, Czech Republic, in 1981, and the Ph.D. degree in algebra and number theory from Masaryk University, Brno, Czech Republic, in 1991. His research and development experience includes more than 30 years in the industry and academia. He is currently a Full Professor with the VSB—Technical University of Ostrava, Ostrava, Czech Republic. He also works in a multidisciplinary environment involving artificial intelligence, social networks, conceptual lattice, information retrieval, semantic web, knowledge management, data compression, machine intelligence, and nature and bio-inspired computing applied to various real-world problems. He has authored or coauthored several refereed journals/conference papers, books, and book chapters. He is the Chair of the IEEE International Conference on Systems, Man, and Cybernetics and the Czechoslovak Chapter. He also served as an Editor/Guest Editor for several journals, such as *Engineering Applications of Artificial Intelligence* (Elsevier), *Neurocomputing* (Elsevier), and *Journal of Applied Logic* (Elsevier).

PAVLA DRÁŽDILOVÁ (Member, IEEE) received the Ph.D. degree in computer science from the VSB—Technical University of Ostrava, Ostrava, Czech Republic, in 2012. She has coauthored of over 40 scientific articles published in proceedings and journals. Her citation report consists of 58 citations and H-index of five on *Web of Science*, 137 citations and H-index of seven on *Scopus*, and 300 citations and H-index of ten on Google Scholar. Her research interests include data mining, social and complex networks, and graph theory.

JAN PLATOŠ (Member, IEEE) received the Ph.D. degree in computer science from the VSB—Technical University of Ostrava, Ostrava, Czech Republic, in 2006.

He was an Associate Professor in computer science, in 2014. Since 2017, he has been the Head of the Department of Computer Science, Faculty of Electrical Engineering and Computer Science, VSB—Techincal University of Ostrava. He has coauthored more than 200 scientific articles published in proceedings and journals. His citation report consists of 338 citations and H-index of ten on *Web of Science*, 800 citations and H-index of 14 on *Scopus*, and 1213 citations and H-index of 17 on Google Scholar. His research interests include text processing, data compression, bio-inspired algorithms, information retrieval, data mining, data structures, and data prediction.