Higher-order clustering in networks

CS224W Project Report

Hao Yin *

1 Introduction

Graph, or network, is a standard way to model complex system and data, where each node represents an entity and each edge represents some interaction or relationship between the corresponding two entities. Clustering is observed in various of real-world networks, and it is the basic hypothesis of any community detection research. The prototypical measurement for the extent to which the nodes of a network cluster is the *clustering coefficient*, or the frequency at which a node and two of its neighbors form a triangle. In this project, we generalize the this definition to the higher-order, which measures the density of clique in the graph.

Basic notations and definitions, including the classic clustering coefficient, are presented in Section 3. The definition of higher-order clustering coefficient is given in Section 4. In Section 4–6, I explored a couple of interesting properties of this generalization.

2 Basic definitions and notations

Let $G = (V, E)$ be an undirected, loop-less graph. We use $n = |V|$ to denote the number of vertices and $m = |E|$ to denote the number of edges. For any node u, we denote the degree of u by d_u and the set of its direct neighbors as $N_1(u)$.

For any integer $\ell \geq 2$, a set of ℓ nodes $\{u_1, u_2, \ldots, u_\ell\}$ forms an ℓ-clique if the induced graph on these ℓ nodes is a complete graph. Let K_ℓ be the set of ℓ-cliques in G, and $K_\ell(u)$ be the set of ℓ-cliques that vertex u is one of its endpoints.

2.1 Classic clustering coefficient

We introduce the definition and notation of the classic clustering coefficient here. My definition of higher-order clustering coefficient is seen in Section 3.

A *wedge* is an unordered pair of edges $\{(u, v), (u, w)\}$ that share exactly one common node, and the node u is called the *center* of the wedge. A wedge is called *closed* if there is an edge between

*Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. E–mail: `yinh@stanford.edu`. I would like to thank Austin Benson for his kind help and guidance. This project is part of my work as a research assistant in Prof. Leskovec's group, where I work with Austin Benson.

v and w, and is called *open* otherwise. Let W be the set of wedges in G, and $W(u)$ be the set of wedges centered at node u. Note that $|W(u)| = \binom{d_u}{2}$, and $|W| = \sum_{u \in V} |W(u)|$.

The *global clustering coefficient* C is defined as the proportion of closed wedges in G, i.e.,

$$C = \frac{3|K_3|}{|W|}, \tag{1}$$

where the coefficient 3 comes from the fact that each triangle closes three wedges. At any node u, the *local clustering coefficient* $C(u)$ is the proportion of closed wedges that centered at u, i.e.,

$$C(u) = \frac{|K_3(u)|}{|W(u)|}. \tag{2}$$

In comparison to global clustering coefficient, the *average clustering coefficient* $\bar{C}$ is the mean of the local clustering coefficient at all nodes, i.e.,

$$\bar{C} = \frac{1}{n} \sum_{u \in V} C(u). \tag{3}$$

2.2 Cut and conductance

For a given set of nodes, the *cut* is defined as the number of edges that has one endpoint in S and the other one in $\bar{S}$, where $\bar{S} = V - S$. We denote the cut by $\texttt{Cut}(S)$.

Conductance of a set of nodes S is defined as

$$\phi(S) = \frac{\texttt{Cut}(S)}{\min\{\texttt{Vol}(S), \texttt{Vol}(\bar{S})\}} \tag{4}$$

where the *volume* $\texttt{Vol}(S) = \sum_{v \in S} d_v$ is the sum of degree of all the nodes in S. Conductance is a commonly-used measure on how good this set of nodes is a community.

Benson et al. (2016) generalized the definition of cut and conductance to motif level. *Motif* is a high-order version (generalization) of edge which relates to more than two nodes. The most simple example of motif is triangle (3-clique) in an undirected graph. Given a motif M, the *motif cut* of a set S, denoted by $\texttt{Cut}_M(S)$, is the number of instances of motif M that has points in both S and $\bar{S}$. *Motif conductance* is defined as

$$\phi_M(S) = \frac{\texttt{Cut}_M(S)}{\min\{\texttt{Vol}_M(S), \texttt{Vol}_M(\bar{S})\}} \tag{5}$$

where *motif volume* $\texttt{Vol}_M(S)$ is the sum, over all the nodes in S, of the number of instances of M each node belongs to. For simplicity of notation, for any integer $\ell \geq 2$, we denote $\texttt{Cut}_\ell(S)$, $\texttt{Vol}_\ell(S)$, and $\phi_\ell(S)$ as the motif cut, volume, and conductance with motif being ℓ-clique, and we call them *ℓ-cut*, *ℓ-volume*, and *ℓ-conductance* respectively.

3 Generalization of clustering coefficient

We first give an alternative way to interpret the classic clustering coefficient from which the higher-order clustering coefficient can be generalized naturally. First consider a 2-clique (i.e., an edge) in the network and another edge that share exactly one node with this 2-clique. We call this 2-clique – edge pair a *2-wedge*, and the common node the *center* of this 2-wedge. Now we say a 2-wedge is *closed* if the related 3 nodes form a $(2+1)$-clique, and open otherwise. Now the *global 2-clustering coefficient* C_2 is defined as the fraction of closed 2-wedges in the graph, i.e.,

$$C_2 = \frac{6|K_3|}{|W_2|} \tag{6}$$

where W_2 is the set of 2-wedges, and the coefficient 6 comes from the fact that each 3-clique closes six 2-wedges. Analogously, for any node u, the *local 2-clustering coefficient* $C_2(u)$ is the fraction of closed 2-wedges centered at this node, i.e.,

$$C_2(u) = \frac{6|K_3(u)|}{|W_2(u)|}, \tag{7}$$

and the *average 2-clustering coefficient* $\bar{C}_2$ is the mean of the local 2-clustering coefficient at all nodes, i.e.,

$$\bar{C}_2 = \frac{1}{n}\sum_{u\in V} C_2(u). \tag{8}$$

Note that each classic wedge is an unordered pair of edges, thus it corresponds to two 2-wedges as is defined here. Therefore, we have $|W(u)| = 2|W_2(u)|$ and $|W| = 2|W_2|$, and comparing Equation (1), (2) with Equation (6), (7), we have $C = C_2$, $C(u) = C_2(u)$, and thus $\bar{C} = \bar{C}_2$. To sum up, the classic definition of clustering coefficient can be interpreted as the probability of a 2-clique and an adjacent edge form a 3-clique.

With this viewpoint, the definition of higher-order clustering coefficients(HOCC) is straightforward. For any integer $\ell \geq 2$, we say that an ℓ-clique and an adjacent edge forms an *ℓ-wedge*; an *ℓ-wedge* is *closed* if the related $\ell+1$ nodes forms an $(\ell+1)$-clique, and is *open* otherwise. Now the definition of the ℓ-order clustering coefficients, globally, locally, and averagely, are defined as the following:

$$\begin{aligned}
C_\ell &= \frac{\binom{\ell+1}{\ell}\binom{\ell}{1}|K_{\ell+1}|}{|W_\ell|}, && (9)\\
C_\ell(u) &= \frac{\binom{\ell}{\ell-1}|K_{\ell+1}(u)|}{|W_\ell(u)|}, && (10)\\
\bar{C}_\ell &= \frac{1}{n}\sum_{u\in V} C_\ell(u). && (11)
\end{aligned}$$

When $\ell = 2$, our previous discussion shows that the higher-order clustering coefficient becomes equivalent to the classic definition of clustering coefficient.

4 Bound for HOCC

We now study the relationships between local higher-order clustering coefficients of different orders. In particular, we give an upper and lower bound for $C_\ell(u)$ based on $C_{\ell-1}(u)$. We have the following result for $\ell = 3$.

Theorem 1 *For any graph and any node u therein, we have $0 \leq C_3(u) \leq \sqrt{C_2(u)}$, and bounds on both directions are tight even if $C_2(u)$ is constant.*

Proof: The lower bound is trivial, and now we show its tightness. Denote G_u as the subgraph induced by all the 1-hop neighborhood of u (excluding u), then we have $C_3(u) = 0$ if G_u is bipartite, and in this case $C_2(u)$ can be any number between 0 and 0.5.

Now we prove the upper bound. Again consider G_u which contains d_u nodes. Denote m_u and t_u as the number of edges and triangles in G_u respectively. Now note the fact that $m_u = |K_3(u)|$ and $t_u = |K_4(u)|$, we have $C_2(u) = 2m_u/d_u(d_u - 1)$ and $C_3(u) = 3t_u/m_u(d_u - 2)$. Following the standard bounds in extremal graph theory (Rivin 2002), we have that

$$t_u \leq \frac{\sqrt{2}(d_v - 2)}{3\sqrt{d_v(d_v - 1)}} m_u^{3/2}. \tag{12}$$

Combine everything together, we have

$$\begin{aligned} C_3(u) &= \frac{3t_u}{m_u(d_u - 2)} \\ &\leq \frac{\sqrt{2}(d_v - 2) \cdot m_u^{3/2}}{\sqrt{d_v(d_v - 1)} \cdot m_u(d_u - 2)} \\ &= \sqrt{\frac{2m_v}{d_v(d_v - 1)}} = \sqrt{C_2(u)}. \end{aligned}$$

The upper bound is tight if G_u consists of a clique and isolated nodes. Suppose G_u contains a k-clique and $d_u - k$ isolated nodes, then we have

$$\begin{aligned} C_2(u) &= \frac{\binom{k}{2}}{\binom{d_u}{2}} = \frac{k(k-1)}{d_u(d_u - 1)}, \\ C_3(u) &= \frac{3 \cdot \binom{k}{3}}{\binom{k}{2} \cdot (d_u - 2)} = \frac{k-2}{d_u - 2}. \end{aligned}$$

Now for any constant $\alpha \in [0, 1]$, let $k = \sqrt{\alpha} \cdot d_u$ and as $d_u \to +\infty$, we have $C_2(u) \to \alpha$ and $C_3(u) \to \sqrt{\alpha}$. $\square$

For $\ell \geq 4$, we also have the trivial lower bound $C_\ell(u) \geq 0$ which is tight when G_u is ℓ-partite. For the upper bound, I conjecture that $C_\ell(u) < C_{\ell-1}(u)$ but not able to prove it.

5 HOCC in random graph models

In this section, we discuss the higher-order clustering coefficients in common random graph models. For Erdös-Rényi model, we have the following result:

Theorem 2 *For Erdös-Rényi model $G_{n,p}$, we have $C_\ell \sim p^{\ell-1}$, $C_\ell(u) \sim p^{\ell-1}$, and $\bar{C}_\ell \sim p^{\ell-1}$. Moreover, conditioning on lower-order clustering coefficients, we have $C_\ell(u) \sim (C_2(u))^{\ell-1}$.*

Proof: First note that any ℓ-wedge is closed if and only if the $\ell - 1$ possible edges between the ℓ-clique and outside nodes in the adjacent edge exist to form an $(\ell + 1)$-clique. In the classical Erdös-Rényi model $G_{n,p}$, each of the $\ell - 1$ edges exist independently with probability p(Erdös and Rényi 1959), thus $C_\ell = p^{\ell-1}$, and locally $C_\ell(u) = p^{\ell-1}$, and thus $\bar{C}_\ell = p^{\ell-1}$.

Note that $C_2(u)$ is the edge density in G_u, conditioning on $C_2(u)$, we have $C_\ell(u) = (C_2(u))^{\ell-1}$. □

A better model than Erdös-Rényi model that captures the clustering property of real-world network is the small-world model(Watts and Strogatz 1998). This model begins with a ring-like network where each node connects to its k nearest neighbors on both side. Then, for each node u and each of the k edges (u, v) with v following u "clockwise" in the ring, with rewiring probability p, the edge is "rewired" to (u, w) where w is chosen uniformly at random.

With no rewiring ($p = 0$) and $k \ll n$, we have $\bar{C} \sim 3/4$(Watts and Strogatz 1998). Here we generalize this result for higher-order clustering coefficients.

Theorem 3 *In the ring-based small-world model with rewiring probability $p = 0$, as $k \to +\infty$, we have*

$$|K_\ell(v)| = \frac{\ell}{(\ell-1)!}k^{\ell-1} + O(k^{\ell-2}); \tag{13}$$

$$C_\ell(u) \sim \frac{\ell+1}{2\ell}. \tag{14}$$

for any $\ell \geq 2$.

Proof: We first count the number of ℓ-cliques node v. It can be easily verified that (13) holds for $\ell = 2$. Now for any $\ell \geq 3$, we first show that

$$|K_\ell(v)| = \sum_{s=\ell-2}^{k-1} (2k - 1 - s) \cdot \binom{s-1}{\ell-3}. \tag{15}$$

Node v has $2k$ neighbors lining up, and we can label them as $1, 2, \ldots, 2k$ sequentially from one end to the other. Now for any ℓ-clique at v, we define the *span* of this clique as the difference between the largest and smallest label of the $\ell - 1$ nodes in the clique other than v. Note that the span of any ℓ-clique, denoted by s, must satisfies $s \leq k - 1$ since any pair of neighbors that has an edge between them must have labels differ no greater than $k - 1$, and also $s \geq \ell - 2$ since there are $\ell - 1$ nodes in an ℓ-clique other than v. Now for each span s, we can find $2k - 1 - s$ pairs of (i, j) such that $1 \leq i, j \leq 2k$ and $j - i = s$. Also, for every such pair (i, j), there are $\binom{s-1}{\ell-3}$ choices of $\ell - 3$

nodes between i and j which will give us an ℓ-clique, together with v, i, and j. Therefore, we come up with (15).

Now starting from (15), we have

$$\begin{aligned} |K_\ell(v)| &= \sum_{s=\ell-2}^{k-1} (2k-1-s) \cdot \binom{s-1}{\ell-3} \\ &= \sum_{s=\ell-2}^{k-1} (2k-s-1) \cdot \frac{s^{\ell-3} + O(s^{\ell-2})}{(\ell-3)!} \\ &= \frac{1}{(\ell-3)!} \left(2k\frac{k^{\ell-2}}{\ell-2} - \frac{k^{\ell-1}}{\ell-1} + O(k^{\ell-2}) \right) \\ &= \frac{\ell}{(\ell-1)!} k^{\ell-1} + O(k^{\ell-2}). \end{aligned}$$

Now using (13), we have

$$\begin{aligned} C_\ell(u) &= \frac{\ell \cdot |K_v^{(\ell+1)}|}{|K_\ell(v)| \cdot (d_v - \ell + 1)} \\ &= \frac{\ell \cdot \frac{\ell+1}{l!} k^l + O(k^{\ell-1})}{\frac{\ell}{(\ell-1)!} k^{\ell-1} \cdot (2k - \ell + 1) + O(k^{\ell-1})} \\ &= \frac{(\ell+1) \cdot k + O(1)}{\ell \cdot (2k - \ell + 1) + O(1)} \\ &\sim \frac{\ell+1}{2\ell}, \end{aligned}$$

which proves (14). □

6 Connetion with neighborhood clique-cut

Clique-cut means motif cut with the motif being clique of some specified order. In this section, we show the connection between higher-order clustering coefficients and neighborhood clique-cut, which generalizes the result by Gleich and Seshadhri (2012). We first show the following simple and clean result.

Theorem 4 *For any integer $\ell \geq 2$, we have*

$$\sum_{v \in V} \boldsymbol{Cut}_\ell(N_1(v)) \leq (1 - C_\ell) \cdot |W_\ell|. \tag{16}$$

Proof: If an ℓ-clique $(u_1, \ldots u_\ell)$ gets cut by $N_1(v)$, then v must directly connects with one of $u_1, \ldots, u_\ell$, say u_1 without loss of generality. Now note that $((u_1, \ldots u_\ell), (u_1, v))$ forms an open $(\ell+1)$-wedge since v can not connect to all of $u_1, \ldots, u_\ell$. Therefore, we have built a map from any clique-cut on the left-hand side of (16) to open ℓ-wedge, and note that this map is injective. Therefore, we have $\sum_{v\in V}\texttt{Cut}_\ell(N_1(v))$ no greater than the number of open ℓ-wedges, which is exactly the number on the right-hand side of (16). □

We also have the following interesting result, with its corollary showing its connection with the neighborhood clique-cut.

Theorem 5 *If a graph has global high-order clustering coefficient $C_\ell = 1$ for some integer $\ell \geq 2$, then each connected component of this graph is either complete or ℓ-clique free.*

Proof: We prove by contradiction. Suppose a connected component of this graph contains an ℓ-clique, then the maximum clique of this connected component is of size $j \geq \ell$. Now if this connected component is not a complete graph, there must be a node connecting to this maximum clique but not forming a bigger clique, thus we obtain an open j-wedge, which contains an open ℓ-wedge. This contradicts to the fact that any ℓ-wedge is closed in a graph with $C_\ell = 1$. □

Corollary 6 *If a graph has global high-order clustering coefficient $C_\ell = 1$ for some integer $\ell \geq 2$, then $\texttt{Cut}_\ell(N_1(u)) = 0$ for any node $u \in V$.*

Note that Corollary 6 can also be obtained from Theorem 5. It shows that, when $C_\ell = 1$, any neighborhood set is a perfect cut by the criterion of ℓ-conductance. Now we are going to generalize this result to the case with any $C_\ell \in [0, 1]$, which we will give an upper bound on the ℓ-conductance of neighborhood set.

We first define a probabilistic distribution on the nodes, $p_\ell(u) = |W_\ell(u)|/|W_\ell|$, which connects the global and local ℓ-th order clustering coefficient.

Lemma 7 $\sum_{u\in V} p_\ell(u)C_\ell(u) = C_\ell$.

Proof.

$$\begin{aligned}
\sum_{u\in V} p_\ell(u)C_\ell(u) &= \sum_{u\in V}\frac{|W_\ell(u)|}{|W_\ell|}\cdot\frac{\ell\cdot|K_{\ell+1}(u)|}{|W_\ell(u)|} \\
&= \frac{\ell}{|W_\ell|}\cdot\sum_{u\in V}|K_{\ell+1}(u)| \\
&= \frac{\ell}{|W_\ell|}\cdot(\ell+1)|K_{\ell+1}| = C_\ell.
\end{aligned}$$

where we use the fact that $\sum_{u\in V}|K_{\ell+1}(u)| = (\ell+1)|K_{\ell+1}|$. □

Lemma 8

$$\sum_{u\in V}\left(p_\ell(u)\frac{\texttt{Cut}_\ell(N_1(u))}{|W_\ell(u)|}\right) \leq 1 - C_\ell.$$

Proof:

$$\begin{aligned}
& \sum_{u\in V}\left(p_\ell(u)\frac{\mathtt{Cut}_\ell(N_1(u))}{|W_\ell(u)|}\right) \\
= & \frac{\sum_{u\in V}\mathtt{Cut}_\ell(N_1(u))}{W_\ell} \\
\leq & \frac{(1-C_\ell)\cdot|W_\ell|}{W_\ell} = 1-C_\ell.
\end{aligned}$$

where the inequality is due to Theorem 4. $\square$

The following theorem shows that large ℓ-clustering coefficient implies the existence of neighborhood cuts with low ℓ-conductance. Here we assume that any neighborhood set has smaller ℓ-volume than its complement, which is intuitively true in real-world large networks.

Theorem 9 *For any graph G of global ℓth-order clustering coefficient C_ℓ for some integer $\ell \geq 2$, then for any constant $a > 1$, there exists a node u such that*

$$\phi_\ell(N_1(u)) \leq \frac{1-C_\ell}{1-C_\ell+l(l+1)\cdot\frac{aC_\ell-1}{a(a-1)}}.$$

Proof: We prove the existence using probabilistic method. Suppose we choose a node u according to the probability distribution $p_\ell(u)$. Let

$$X = \frac{\mathtt{Cut}_\ell(N_1(u))}{|W_l(u)|}$$

which is a random variable, then $\mathbb{E}[X] = 1 - C_\ell$ according to Lemma 8. By Markov's inequality, we have $\mathbb{P}[X > a(1-\kappa)] < 1/a$.

Let $b = \frac{aC_\ell-1}{a-1}$, and $p = \mathbb{P}[C_\ell(u) < b]$. Now according to Lemma 7, we have

$$\begin{aligned}
C_\ell &= \sum_{u\in V} p_\ell(u)C_\ell(u) \\
&= \sum_{C_\ell(u)<b} p_\ell(u)C_\ell(u) + \sum_{C_\ell(u)\geq b} p_\ell(u)C_\ell(u) \\
&< b\cdot p + 1\cdot(1-p),
\end{aligned}$$

thus $p < \frac{1-C_\ell}{1-b} = 1 - \frac{1}{a}$.

Then by the union bound, the probability that $\frac{\mathtt{Cut}_\ell(N_1(u))}{|W_\ell(u)|} > a(1-C_\ell)$ or $C_\ell(u) < b$ is less than 1, thus there exists some vertex u such that $\mathtt{Cut}_\ell(N_1(u)) \leq a(1-C_\ell)\cdot|W_\ell(u)|$ and $C_\ell(u) \geq b$. Now we show that, for this u, we have

$$\phi_\ell(N_1(u)) \leq \frac{1-C_\ell}{1-C_\ell+\ell(\ell+1)\cdot\frac{aC_\ell-1}{a(a-1)}}.$$

We first find a lower bound on $\texttt{Vol}_\ell(N_1(u))$. First, each ℓ-clique cut would contribute at least one into $\texttt{Vol}_\ell(N_1(u))$. Second, for each closed ℓ-wedges centered at u, it is actually an $(\ell+1)$-clique, consists of $\ell+1$ unique ℓ-cliques, and each ℓ-clique would contribute ℓ in $\texttt{Vol}_\ell(N_1(u))$. Now note that there are $C_\ell(u)|W_\ell(u)| \geq b|W_\ell(u)|$ closed ℓ-wedges centered at u, which is contained in $N_1(u)$, we must have $\texttt{Vol}_\ell(N_1(u)) \geq \texttt{Cut}_\ell(N_1(u)) + \ell(\ell+1)b|W_\ell(u)|$.

Now combining that $\texttt{Cut}_\ell(N_1(u)) \leq a(1-C_\ell) \cdot |W_\ell(u)|$ and based on our assumption that $\texttt{Vol}_\ell(N_1(u)) \leq \texttt{Vol}_\ell(\overline{N_1(u)})$, we have

$$\begin{aligned}
\phi_\ell(N_1(u)) &= \frac{\texttt{Cut}_\ell(N_1(u))}{\texttt{Vol}_\ell(N_1(u))} \\
&\leq \frac{\texttt{Cut}_\ell(N_1(u))}{\texttt{Cut}_\ell(N_1(u)) + \ell(\ell+1)b|W_\ell(u)|} \\
&\leq \frac{a(1-C_\ell)\cdot|W_\ell(u)|}{a(1-C_\ell)|W_\ell(u)| + \ell(\ell+1)b|W_\ell(u)|} \\
&= \frac{1-C_\ell}{1-C_\ell+\ell(\ell+1)\cdot\frac{aC_\ell-1}{a(a-1)}}.
\end{aligned}$$

□

Note that for each constant a, as $C_\ell \to 1$, the upper bound will decrease to 0, which means that large global ℓ-order clustering coefficient implies the existence of neighborhood cuts with low ℓ-conductance. Moreover, the optimal choice of a for each C_ℓ is $a = \frac{1+\sqrt{1-C_\ell}}{C_\ell}$.

References

Andersen, Reid, Fan Chung, Kevin Lang. 2006. Local graph partitioning using pagerank vectors. *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE, 475–486.

Benson, Austin R, David F Gleich, Jure Leskovec. 2016. Higher-order organization of complex networks. *Science* **353**(6295) 163–166.

Erdös, Paul, Alfréd Rényi. 1959. On random graphs, i. *Publicationes Mathematicae (Debrecen)* **6** 290–297.

Fiedler, Miroslav. 1973. Algebraic connectivity of graphs. *Czechoslovak mathematical journal* **23**(2) 298–305.

Gleich, David F, C Seshadhri. 2012. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 597–605.

Rivin, Igor. 2002. Counting cycles and finite dimensional lp norms. *Advances in applied mathematics* **29**(4) 647–662.

Watts, Duncan J, Steven H Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* **393**(6684) 440–442.