# Clustering coefficients of large networks*

Yusheng Li, Yilun Shang†, Yiting Yang
Department of Mathematics, Tongji University
Shanghai 200092, China

**Abstract**

Let $G$ be a network with $n$ nodes and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then $G$ is called an $(n, d, \lambda)$-network if it is $d$-regular and $\lambda = \max\{|\lambda_2|, |\lambda_3|, \ldots, |\lambda_n|\}$. It is shown that if $G$ is an $(n, d, \lambda)$-network and $\lambda = O(\sqrt{d})$, the average clustering coefficient $\overline{c}(G)$ of $G$ satisfies $\overline{c}(G) \sim d/n$ for large $d$. We show that this description also holds for strongly regular graphs and Erdős-Rényi graphs. Although most real-world networks are not theoretically constructed, we find that, interestingly, many of them have $\overline{c}(G)$ close to $\overline{d}/n$, and many close to $1 - \frac{\overline{\mu_2}(n-\overline{d}-1)}{\overline{d}(\overline{d}-1)}$, where $\overline{d}$ is the average degree of $G$, and $\overline{\mu_2}$ is the average of the numbers of common neighbors over all non-adjacent pairs of nodes.

*Key Words*: Clustering coefficient; theoretic graph; real-world network

## 1 Introduction

Complex systems from various fields, such as physics, biology, and sociology, can be systematically analyzed using their network representation. A network (also known as a graph) is composed of vertices (or nodes) and edges, where vertices represent the constitutes in the system and edges represent the relationships between these constitutes. Mathematically, we define $G = (V, E)$ as a simple graph with vertex set $V$ and edge set $E \subseteq V \times V$. Denote by $N(v) = \{u \in V : uv \in E\}$ the neighborhood of a vertex $v$, $d_v$ the degree of $v$, and $e_v = e(N(v))$ the number of edges in the subgraph of $G$ induced by $N(v)$, respectively.

An important measure of network topology, called *clustering coefficient*, assesses the triangular pattern as well as the connectivity in a vertex's neighborhood: a vertex has a high clustering coefficient if its neighbors tend to be directly connected with each other. The clustering coefficient $c_v$ of a vertex $v$ can be calculated as

$$c_v = \begin{cases} 0, & \text{if } d_v = 0, \\ \frac{e_v}{\binom{d_v}{2}}, & \text{if } d_v \geq 2. \end{cases}$$

*Research supported by the National Natural Science Foundation of China under Grants No. 11331003, No. 11101360, No. 11505127, the Shanghai Pujiang Program under Grant No. 15PJ1408300, and Outstanding Young Scholar Foundation of Tongji University under Grants No. 2013KJ031, No. 2014KJ036.

†Correspondence: shyl@tongji.edu.cn

For $d_v = 1$, it is a convention to define $c_v \in [0, 1]$ depending on the situation. Thus $0 \leq c_v \leq 1$. The clustering coefficient $c_v$ for $d_v \geq 2$ is the ratio of number of triangles and all possible triangles that share vertex $v$.

For a graph $G$ of order $n$ (i.e., $G$ contains $n$ vertices) and minimum degree $\delta(G) \geq 2$, its *average clustering coefficient* is defined as

$$\overline{c}(G) = \frac{1}{n}\sum_{v\in V} c_v = \frac{2}{n}\sum_{v\in V}\frac{e_v}{d_v(d_v-1)}.$$

Average clustering coefficient explains the clustering (triangulation) within a network by averaging the clustering coefficients of all its nodes. The idea of clustering coefficient is proposed (especially in the analysis of social networks) to measure the degree to which nodes in a graph tend to cluster together [31]. It is considered to be an effective measure of the local connectivity or "cliqueness" of a social network. For example, the clustering coefficient and community structure of a social network may reveal its collaborative relationship between agents [41]. If a network has a high average clustering coefficient and a small average distance, it is often called a "small-world" network [42]. High clustering is also associated with robustness of a network, that is, resilience against random damage [2, 24, 28, 38]. The clustering coefficient has also been extended to the settings of weighted networks [35, 43] and directed networks [15]. Network models with tunable clustering coefficients [20, 22, 37], degree-related clustering coefficients [29], and granular clustering coefficients [26] have been proposed. Moreover, the work [36] presents a fast sampling-based approximation algorithm for calculating the average clustering coefficient.

Some basic properties of clustering coefficient can be easily observed. Let $G$ be a graph with minimum degree $\delta(G) \geq 2$. Then $\overline{c}(G) = 0$ if and only if $G$ is triangle-free. Let $\delta(N(v))$ be the minimum degree of the subgraph of $G$ induced by $N(v)$. The following result gives us a nontrivial lower bound of $\overline{c}(G)$.

**Lemma 1** *Let $G$ be a graph with $\delta(G) \geq 2$. If $m = \min_{v\in V}\delta(N(v))$, then*

$$\overline{c}(G) \geq \frac{m}{d-1},$$

*where $d$ is the average degree of $G$.*

**Proof.** Let $n$ be the order of $G$. Note that $e_v = e(N(v)) \geq \frac{md_v}{2}$ for each vertex $v$, and thus

$$\overline{c}(G) = \frac{1}{n}\sum_{v\in V} c_v = \frac{1}{n}\sum_{v\in V}\frac{e_v}{\binom{d_v}{2}} \geq \frac{1}{n}\sum_{v\in V}\frac{m}{d_v-1} \geq \frac{m}{d-1},$$

where the last inequality holds as the function $\frac{1}{x-1}$ is convex for $x > 1$. □

The aim of this article is to investigate the average clustering coefficient of some theoretic and large-scale real networks. The remainder of this paper is organized as follows. In Section 2 we derive the exact expression for the average clustering coefficient of strongly regular graphs of any order. In Section 3, we show that the average clustering coefficients of $(n, d, \lambda)$-graphs with $\lambda = O(\sqrt{d})$ follow the asymptotic $d/n$ for large $d$. The same expression is recovered for an almost regular algebraic graph, Erdős-Rényi graph, in Section 4. We present some data on networks

taken from real-world applications in Section 5. Interestingly, the average clustering coefficient of the Florentine families graph is shown to be close to the corresponding estimate obtained for strongly regular graphs (Theorem 1). Our finding highlights the fact that the asymptotic $d/n$ is not only a characteristic of random graphs and some theoretic graphs, but also observed in various real systems distinct from small-world networks. Finally, the paper is concluded in Section 6.

## 2 Clustering coefficients of strongly regular graphs

A graph $G$ of order $n$ is said to be a *strongly regular graph* with parameters $n, d, \mu_1, \mu_2$, denoted by $srg(n, d, \mu_1, \mu_2)$ in short, if it is $d$-regular, and any pair of vertices have $\mu_1$ common neighbors if they are adjacent, and $\mu_2$ common neighbors otherwise [19, p. 218]. For instance, $C_5$ is an $srg(5, 2, 0, 1)$. Another typical example of strong regular graphs is the *Paley graph* [19, p. 221] which is defined as follows:

Let $q \equiv 1 \,(\mathrm{mod}\ 4)$ be a prime power and $F_q$ the finite field of order $q$. The vertex set of Paley graph $P_q$ is $F_q$, and distinct vertices $x$ and $y$ are adjacent if and only if $x - y$ is non-zero quadratic. Clearly, Paley graph $P_q$ is an $srg(q, \frac{q-1}{2}, \frac{q-5}{4}, \frac{q-1}{4})$.

For an $srg(n, d, \mu_1, \mu_2)$, the number $m$ in Lemma 1 is exactly $\mu_1$. The next theorem says that Lemma 1 is sharp for strongly regular graphs.

**Theorem 1** *Let $G$ be an $srg(n, d, \mu_1, \mu_2)$ with $d \geq 2$. Then*

$$\overline{c}(G) = \frac{\mu_1}{d-1}.$$

**Proof.** For a vertex $v$ of $G$, we have $e_v = \frac{d\mu_1}{2}$ and thus $c_v = \frac{e_v}{\binom{d}{2}} = \frac{\mu_1}{d-1}$, hence the average of $c_v$ is also $\frac{\mu_1}{d-1}$ as claimed. □

**Corollary 1** *Let $P_q$ be a Paley graph. Then*

$$\overline{c}(P_q) = \frac{q-5}{2(q-3)}.$$

**Proof.** Since $P_q$ is an $srg(q, \frac{q-1}{2}, \frac{q-5}{4}, \frac{q-1}{4})$,

$$\overline{c}(P_q) = \frac{\mu_1}{d-1} = \frac{\frac{q-5}{4}}{\frac{q-1}{2}-1} = \frac{q-5}{2(q-3)}.$$

□

The eigenvalues of strongly regular graphs can be easily computed, which are summarized in the following lemma [19, p. 219].

**Lemma 2** *Let $G$ be a connected $srg(n, d, \mu_1, \mu_2)$ with $n \geq 3$. Assume that $G$ is neither complete nor empty. Then $\lambda_1 = d$ is an eigenvalue with multiplicity $m_1 = 1$, and any eigenvalue $\lambda \neq \lambda_1$ satisfies*

$$\lambda^2 + (\mu_2 - \mu_1)\lambda + (\mu_2 - d) = 0.$$

*The equation has two distinct solutions $\lambda_2 > \lambda_3$ with $\lambda_2 > 0 > \lambda_3$, and $\lambda_3$ is an eigenvalue. If $d + (n-1)\lambda_3 \neq 0$, then $\lambda_2$ is also an eigenvalue. Their multiplicities $m_2$ and $m_3$ can be determined by*

$$m_2 + m_3 = n - 1, \quad \text{and} \quad d + m_2\lambda_2 + m_3\lambda_3 = 0.$$

In the next section, we will consider a more general class of graphs, in which the average clustering coefficient can be asymptotically estimated under some condition regarding their second largest eigenvalues.

## 3 Clustering coefficients of $(n, d, \lambda)$-graphs

Let us label the vertices of $G$ of order $n$ as $v_1, v_2, \ldots, v_n$. Recall that $A = (a_{ij})_{n\times n}$ is the adjacency matrix of $G$, where

$$a_{ij} = \begin{cases} 1, & \text{if } v_iv_j \in E, \\ 0, & \text{otherwise.} \end{cases}$$

**Lemma 3** *Let $G$ be a graph of order $n$ with degree sequence $d_1, d_2, \ldots, d_n$, and let $A$ be the adjacency matrix. If $\delta(G) \geq 2$ and the diagonal elements of $A^3$ are $\gamma_1, \gamma_2, \ldots, \gamma_n$ in order, then*

$$\overline{c}(G) = \frac{1}{n}\sum_{i=1}^{n} \frac{\gamma_i}{d_i(d_i - 1)}.$$

*Particularly, if $G$ is $d$-regular, and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of $A$, then*

$$\overline{c}(G) = \frac{1}{nd(d-1)}\sum_{i=1}^{n} \lambda_i^3.$$

**Proof.** As $A$ is symmetric, it is diagonalizable. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of $A$. Then the eigenvalues of $A^k$ are $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$. Note that the $(i, j)$ element of $A^k$ is the number of walks from vertex $v_i$ to vertex $v_j$ [19, 23], and a closed walk of length 3 is a triangle. Thus the $ith$ diagonal element of $A^3$ is exactly $2e_{v_i}$, and the claimed equalities follow. □

We also call the eigenvalues of $A$ as eigenvalues of $G$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be eigenvalues of $G$ in a non-increasing order.

Random graph has been proved to be one of the most important tools in modern graph theory and network research. Their tremendous triumph raised the question: what are the essential properties and how can we tell whether a given graph behaves like a random graph? A cornerstone contribution of Chung, Graham and Wilson [12] gave several equivalent properties to measure the similarity between a given graph and a random graph $G(n, p)$ with fixed $p$. This similarity is also called the quasi-randomness of the given graph. One of those properties is characterized by the magnitude of $\lambda = \lambda(G)$, where

$$\lambda = \lambda(G) = \max\{|\lambda_i| : \ 2 \le i \le n\}.$$

As called by Alon, see [4], a graph $G$ is an $(n, d, \lambda)$-graph if $G$ is $d$-regular with $n$ vertices and $\lambda = \lambda(G)$. Note that a $d$-regular connected graph satisfies that $\lambda_1 = d$. For sparse graphs with edge density $p = o(1)$, Chung and Graham [11] also gave some equivalent properties for quasi-randomness under certain conditions. One of the properties is that $\lambda_1 \sim pn$ and $\lambda = o(\lambda_1)$.

For an $(n, d, \lambda)$-graph, the spectral gap between $d$ and $\lambda$ is a measure for its quasi-random property. The smaller the value of $\lambda$ compared to $d$, the closer is the edge distribution to the ideal uniform distribution (i.e., it becomes a random graph). A natural question in order is "how small can $\lambda$ be?"

**Lemma 4** *Let $G$ be an $(n, d, \lambda)$-graph and let $\epsilon > 0$. If $d \le (1-\epsilon)n$, then*

$$\lambda \ge \sqrt{\epsilon d}.$$

**Proof.** Let $A$ be the adjacency matrix of $G$. Then

$$\begin{aligned} nd &= 2e(G) = tr(A^2) = \sum_{i=1}^{n} \lambda_i^2 \\ &\le d^2 + (n-1)\lambda^2 \le (1-\epsilon)nd + n\lambda^2, \end{aligned}$$

which concludes the proof. □

Based on this estimate, we may say, not precisely, that an $(n, d, \lambda)$-graph with $\lambda = O(\sqrt{d})$ has good quasi-randomness. Generally, this is a weak condition as most random graphs are such graphs, see [7]. Moreover, Lemma 2 suggests that for an $srg(n, d, \mu_1, \mu_2)$, when $|\mu_1 - \mu_2|$ is small compared to $d$, $\lambda$ is close to $\sqrt{d}$ and $G$ has good quasi-randomness.

It is known that, for any fixed $d \ge 4$, a random $d$-regular graph $G$ of order $n$ has $\lambda = (2 + o(1))\sqrt{d-1}$ and for any fixed $\epsilon > 0$, the probability

$$\Pr\left((1-\epsilon)\frac{d-1}{n} \le \overline{c}(G) \le (1+\epsilon)\frac{d-1}{n}\right) \to 1$$

as $n \to \infty$ [8, 17]. The following theorem gives an analogous estimate of average clustering coefficient for $(n, d, \lambda)$-graph with $\lambda = O(\sqrt{d})$.

**Theorem 2** *Let $G$ be an $(n, d, \lambda)$-graph. If $\lambda = O(\sqrt{d})$ as $d \to \infty$, then*

$$\overline{c}(G) \sim \frac{d}{n}.$$

**Proof.** Let $n$ be the order of $G$ whose eigenvalues are $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$. Since $G$ is $d$-regular, we have $\lambda_1 = d$. From Lemma 3 we obtain

$$\overline{c}(G) = \frac{1}{nd(d-1)}\left(d^3 + \sum_{i=2}^{n} \lambda_i^3\right).$$

The assumption $\lambda = O(\sqrt{d})$ implies that

$$\frac{\left|\sum_{i=2}^{n} \lambda_i^3\right|}{nd(d-1)} \leq \frac{n\lambda^3}{nd(d-1)} = \frac{O(d^{3/2})}{d^2} \to 0.$$

Thus

$$\bar{c}(G) \sim \frac{d^2}{n(d-1)} \sim \frac{d}{n} \tag{1}$$

for large $d$. □

**Remark 1.** If $G$ is an $srg(n, d, \mu_1, \mu_2)$, we have seen from Lemma 2 that $\lambda = O(\sqrt{d})$ provided $\mu_1 \approx \mu_2$. Note that there is a well-known equality relating the four parameters in a strongly regular graph, namely, $d(d-\mu_1-1) = \mu_2(n-d-1)$. Hence, by dividing $n$ on both sides of the equality and using the estimate $\mu_1 \approx \mu_2$, we derive that $\frac{d^2}{n} \approx \mu_2$. Thus, Theorem 1 implies that $\bar{c}(G) = \frac{\mu_1}{d-1} \approx \frac{\mu_2}{d} \approx \frac{d}{n}$, which agrees with Theorem 2.

The condition that $d$ is large in Theorem 2 is for estimating $\sum_{i=2}^{n} \lambda_i^3$ in the proof. However, this is not necessary for some graphs which are described in the next section.

# 4 Clustering coefficients of Erdős-Rényi graphs

Erdős-Rényi graph $E_q$ [14] is one of the classical algebraic constructions in extremal combinatorics which is defined as follows. Let $F_q^* = F_q \setminus \{0\}$. Define an equivalence relation $\equiv$ on $F_q^3 \setminus \{(0,0,0)\}$ by letting $(a_1, a_2, a_3) \equiv (b_1, b_2, b_3)$ if there is $\gamma \in F_q^*$ such that $(a_1, a_2, a_3) = \gamma(b_1, b_2, b_3)$. Let $\langle a_1, a_2, a_3\rangle$ denote the equivalence class containing $(a_1, a_2, a_3)$, and let $V$ be the set of all equivalence classes. Then $|V| = q^2 + q + 1$.

Define a graph $E_q$ on vertex set $V$ by letting distinct vertices $\langle v_1, v_2, v_3\rangle$ and $\langle x_1, x_2, x_3\rangle$ be adjacent if and only if

$$v_1x_1 + v_2x_2 + v_3x_3 = 0.$$

For a vertex $v = \langle v_1, v_2, v_3\rangle$, since $v_1x_1 + v_2x_2 + v_3x_3 = 0$ has $q^2 - 1$ solutions $(x_1, x_2, x_3)$ forming $q + 1$ vertices,

$$d_v = \begin{cases} q, & \text{if } v_1^2 + v_2^2 + v_3^2 = 0, \\ q+1, & \text{otherwise.} \end{cases}$$

Unfortunately, $E_q$ is not regular and its spectrum is not known. If we add a loop for a vertex $v$ to $E_q$ when $v_1^2 + v_2^2 + v_3^2 = 0$, we obtain a $(q+1)$-regular graph, which is denoted by $E_q^o$. In the following, we will compute the spectrum of $E_q$ by using the spectrum of $E_q^0$.

Let $M$ be the adjacency matrix of $E_q^o$, where a diagonal element is equal to the number of loops incident to the vertex. From the eigenvalues of $M^2$ [3], one can find the spectrum of $M$ as follows.

**Lemma 5** *The spectrum of $E_q^o$ is as follows.*

| *eigenvalue* | $q+1$ | $\sqrt{q}$ | $-\sqrt{q}$ |
|---|---|---|---|
| *multiplicity* | 1 | $\frac{q(q+1)}{2}$ | $\frac{q(q+1)}{2}$ |

Return to the simple graph $E_q$ with maximum degree $\Delta(E_q) = q+1$ and minimum degree $\delta(E_q) = q$. Let $A$ be the adjacency matrix of $E_q$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ the eigenvalues of $A$, where $n = q^2+q+1$. We shall estimate these $\{\lambda_i\}_{i=1}^n$ by the following result, see [23, p. 181].

**Lemma 6** *Let $A$ and $B$ be real symmetric matrices of order $n$. Let the eigenvalues $\lambda_i(A)$ of $A$, $\lambda_i(B)$ of $B$ and $\lambda_i(A+B)$ of $A+B$ are labeled in non-increasing order, respectively. Then, for each $1 \leq i \leq n$,*

$$\lambda_i(A) + \lambda_1(B) \geq \lambda_i(A+B) \geq \lambda_i(A) + \lambda_n(B).$$

**Corollary 2** *Let $G$ be a simple graph of order $n$ with $\Delta = \Delta(G)$ and $\delta = \delta(G)$. Let $G'$ be a graph obtained from $G$ by attaching each vertex $v$ with $\Delta - d_v$ loops. Suppose that $G$ and $G'$ have eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\lambda'_1 \geq \lambda'_2 \geq \cdots \geq \lambda'_n$, respectively. Then, for each $1 \leq i \leq n$,*

$$\lambda_i + \Delta - \delta \geq \lambda'_i \geq \lambda_i.$$

**Proof.** It is easy to see that $A(G') = A(G) + D$, where $D$ is a diagonal matrix whose diagonal elements are $\Delta - d_v$ for each vertex $v$. Since $\lambda_1(D) = \Delta - \delta$ and $\lambda_n(D) = 0$, the assertion follows from Lemma 6. □

**Lemma 7** *Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be eigenvalues of $E_q$, where $n = q^2+q+1$. Then these eigenvalues can be bounded as follows.*

| *eigenvalue* | $\lambda_1$ | $\lambda_i;\ \lambda_i > 0, i \geq 2$ | $\lambda_i;\ \lambda_i < 0$ |
|---|---|---|---|
| *bounds* | $q \leq \lambda_1 \leq q+1$ | $\sqrt{q} \leq \lambda_i \leq \sqrt{q}+1$ | $-\sqrt{q} \leq \lambda_i \leq -\sqrt{q}+1$ |

Although $E_q$ is not regular, $\overline{c}(E_q)$ is still close to $d/n$, where $n = q^2+q+1$ and $d$ is the average degree of $E_q$.

**Theorem 3** *Let $q$ be a prime power and let $E_q$ be the Erdős-Rényi graph. Then $E_q$ has $n = q^2+q+1$ vertices and its average degree $d$ satisfies $q < d < q+1$, and $\overline{c}(E_q) \sim \frac{d}{n} \sim \frac{1}{q}$ for large $q$.*

**Proof.** It is easy to see that except the largest $\lambda_1$, the number of other positive eigenvalues is $q(q+1)/2$ and the number of negative eigenvalues is also $q(q+1)/2$. So we can estimate $\sum_i \lambda_i^3$ from above as

$$\sum_i \lambda_i^3 \leq (q+1)^3 + \frac{q(q+1)}{2}(\sqrt{q}+1) - \frac{q(q+1)}{2}(\sqrt{q}-1) = (q+1)^3 + q(q+1),$$

and thus by Lemma 3 we have

$$\overline{c}(E_q) \leq \frac{(q+1)^3 + q(q+1)}{(q^2+q+1)q(q-1)} \sim \frac{1}{q},$$

and the desired lower bound can be obtained similarly. □

## 5 Clustering coefficients of some real networks

In this section, we present some applications of the above theoretical results to realistic networks.

If a realistic network $G$ is shown to be close (roughly speaking) to a strongly regular graph $srg(n, d, \mu_1, \mu_2)$, say, by sampling a portion of the nodes, Theorem 1 can then be an estimate of its average clustering coefficient $\bar{c}(G)$. Since many realistic large-scale networks (such as social networks) are highly clustered, the calculation of parameter $\mu_1$ could create an enormous burden on the computation facility in practice. Alternatively, one can first calculate the parameter $\mu_2$ by using a heuristic algorithm, and then invoke the convenient relation $d(d-\mu_1-1) = \mu_2(n-d-1)$, which in turn yields an estimate of $\mu_1$. In other words, we have

$$\bar{c}(G) = \frac{\mu_1}{d-1} = 1 - \frac{\mu_2(n-d-1)}{d(d-1)}.$$

The advantage of evaluating $\mu_2$ over $\mu_1$ is shown in Appendix A.

Let $\overline{d}$ be the average degree of a network and $\overline{\mu_2}$ the average of the numbers of common neighbors of non-adjacent pairs of nodes. We shall estimate average clustering coefficients of some networks as follows:

$$\bar{c}(G) \approx 1 - \frac{\overline{\mu_2}(n-\overline{d}-1)}{\overline{d}(\overline{d}-1)}. \tag{2}$$

Algorithm 1 is a heuristic algorithm to calculate $\overline{\mu_2}$.

**Algorithm 1**

**Input**: $G$
**Output**: $\theta = \overline{\mu_2}$
let $a = b = 0$
let $Q_1$ and $Q_2$ be two empty queues
**for** $v$ in $G \backslash Q_1$
    add $v$ into $Q_1$ and $Q_2$
    add each vertex $u$ that is adjacent to $v$ into $Q_2$
    $b = b + |G \backslash Q_2|$, where $|\cdot|$ represents the length
    **for** $w$ in $G \backslash Q_2$
        let $c$ be the number of vertices in $Q_2$ that are adjacent to $w$
        $a = a + c$
    clear $Q_2$
$\theta = a/b$

**Algorithm 1.** A heuristic algorithm that calculates the average value $\overline{\mu_2}$ for a network $G$.

We worked out a concrete example by considering a social network $G$ of Florentine families [9]. This network describes the marriage relations among $n = 15$ families in fifteenth-century Florence with data collected by John Padgett from historical documents (see Fig. 1 for an illustration). It is direct to check that the average degree of $G$ is $\overline{d} = 2.667$ and the average

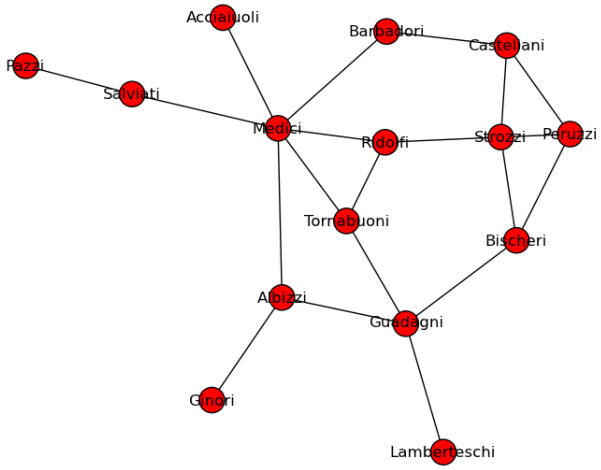



Fig. 1. The Florentine families graph $G$ on $n = 15$ vertices, where the family names are shown alongside the vertices.

clustering coefficient is $\bar{c}(G) = 0.16$. The Algorithm 1 gives $\overline{\mu_2} = 0.322$. Interestingly, by (2), $1 - \frac{\overline{\mu_2}(n-\overline{d}-1)}{\overline{d}(\overline{d}-1)} = 1 - \frac{3.650}{4.446} = 0.179$ gives a reasonably good approximation of the average clustering coefficient $\bar{c}(G)$. In addition, we studied two other real networks, i.e. the US top-500 airport network [13], in which two airports are connected if a flight was scheduled between them in 2002, and the autonomous system graph of the Internet [36]. The results are summarized in the following table, which illustrate the availability of the proposed method. It is worth mentioning that the sampling-based algorithm in [36] yields an estimation $\bar{c}(G) = 0.311$ for AS graph, which seems slightly worse as compared to the present method.

| Network $G$ | $n$ | $\overline{d}$ | $\overline{\mu_2}$ | $\bar{c}(G)$ | $\bar{c}(G)$ via eq.(2) |
|---|---|---|---|---|---|
| Florentine families graph | 15 | 2.667 | 0.322 | 0.160 | 0.179 |
| US airport network | 500 | 11.920 | 0.185 | 0.351 | 0.308 |
| AS graph | 13164 | 4.332 | 0.0005 | 0.458 | 0.544 |

**Table 1.** Statistics for three social and information networks.

Next, we collected statistics for some real-world complex networks in the following table. In this table, $n$ is the number of vertices; $m$ is the number of edges; $\overline{d}$ is the average degree; $\bar{c}(G)$ is the average clustering coefficient; $R$ is the relative error defined as $R = \frac{|\bar{c}(G) - \overline{d}/n|}{\bar{c}(G)}$.

| Network $G$ | $n$ | $m$ | $\bar{c}(G)$ | $\overline{d}/n$ | $\overline{d}$ | $R$ |
|---|---|---|---|---|---|---|
| p2p-Gnutella | 3234 | 13453 | 0.007 | 0.003 | 9.7 | 0.57 |
| German highway | 1168 | 1532 | 0.0013 | 0.0019 | 2.6 | 0.46 |
| cat brain | 55 | 454 | 0.55 | 0.3 | 16.5 | 0.45 |
| constraint programming | 240 | 6300 | 0.225 | 0.219 | 52.5 | 0.027 |
| passenger air traffic | 1148 | 16523 | 0.027 | 0.025 | 28.8 | 0.074 |
| protein interaction I | 97 | 2327 | 0.49 | 0.495 | 47.5 | 0.01 |
| protein interaction II | 109 | 2978 | 0.5 | 0.501 | 54.6 | 0.002 |
| protein interaction III | 54 | 1018 | 0.71 | 0.698 | 37.7 | 0.017 |

**Table 2.** Statistics for some real networks

In Table 2, p2p-Gnutella is a snapshot of the Gnutella peer-to-peer file sharing network on August 5, 2002 [27, 34], where nodes represent hosts in the Gnutella network and edges represent connections between the hosts. German highway is compiled from data of the "Autobahn-Informations-System" in July 2002 taking cities as nodes and highways as edges [25]. cat brain is a biological network depicting long-range cortical connectivity in the cat brains [21]. The dataset of constraint programming comes from monitoring of constraint-oriented programs on activity graphs [18]. passenger air traffic [5] is obtained by tracking the aerial routes among airports handling over 2 million passengers in 2000. protein interaction I, II, and III are three quasi-cliques for Ribosome biogenesis in protein-protein interaction networks of budding yeast [10], where proteins are nodes and interactions form edges.

We observe that these real networks have average clustering coefficients $\bar{c}(G)$ relatively close to $\overline{d}/n$. Although these real-world networks are not theoretically constructed, they obey the same asymptotic average clustering coefficient as those obtained in Sections 3 and 4. A clear trend can be discerned from Table 2: Network with large average degree $\overline{d}$ tends to have smaller $R$, namely, the approximation of $\overline{d}/n$ becomes better for larger $\overline{d}$ (compared to $n$). This phenomenon is consistent with our obtained theoretical results, although they are obtained under more restrictive assumptions (i.e., the $(n, d, \lambda)$-graph condition with $\lambda = O(\sqrt{d})$). This phenomenon may be explained by the fact that these considered networks are, in general, like edge-independent random graphs with large maximum degree. Let $G$ be a random graph on $n$ vertices, where $uv$ is an edge with probability $p_{uv}$ and each edge is independent of each other edge. It is shown that [30, 39] if $\max_u\{\sum_v p_{uv}\} \gg \ln^4 n$ then $\lambda = O\big(\sqrt{\overline{d}}\big)$ almost surely. The classical Erdős-Rényi random graph is a special case of such graphs. The passenger air traffic graph, for example, is shown to have asymptotically independent edges and a handful of hub nodes [5]. So, the result $\bar{c}$(passenger air traffic) $\approx \overline{d}/n$ presumably follows from Theorem 2.

Finally, we mention that there are many real networks whose average clustering coefficients $\bar{c}(G)$ are far from $\overline{d}/n$ as compared to those given in Table 2. In particular, networks with small-world properties usually have high clustering coefficients but low value of $\overline{d}/n$. In Table 3, we collected some real network data in which the values of $R$, namely, the relative errors, are typically much higher than those in Table 2. Most of these networks are proved to be small-world networks, and the edge structures therein are far from random. In movie actors, for example, an actor's choice of collaborating with one or another is highly correlated as the total potential number of collaborators is usually fixed. Distinct from the Erdős-Rényi like graphs, which by nature has $\bar{c}(G) \approx \overline{d}/n$, such small-world networks are prevalent in social, information,

and biological systems. Much effort on the study of clustering coefficients has been devoted to such networks with small-world phenomenon [31].

| Network $G$ | $n$ | $m$ | $\bar{c}(G)$ | $\overline{d}/n$ | $\overline{d}$ | $R$ |
|---|---|---|---|---|---|---|
| WWW [1] | 153127 | 2695801 | 0.1078 | 0.00023 | 35.21 | 0.998 |
| movie actors [42] | 225226 | 6869393 | 0.79 | 0.00027 | 61 | 0.999 |
| SPIRES co-authorship [32] | 56627 | 4898236 | 0.726 | 0.003 | 173 | 0.996 |
| Neurosci. co-authorship [6] | 209293 | 1203435 | 0.76 | 0.000005 | 11.5 | 0.999 |
| E. coli substrate [16] | 282 | 1036 | 0.32 | 0.026 | 7.35 | 0.919 |
| E. coli reaction [16] | 315 | 4475 | 0.59 | 0.09 | 28.3 | 0.847 |
| C. Elegans [42] | 282 | 1974 | 0.28 | 0.05 | 14 | 0.821 |

**Table 3.** Statistics for some real small-world networks, where the value $R$ is typically very high.

# 6 Conclusion

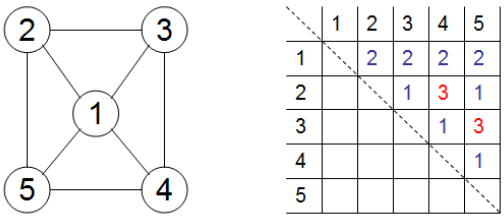
In this paper, we have shown analytically that the average clustering coefficient $\bar{c}(G)$ of an $(n, d, \lambda)$-network $G$ with $\lambda = O(\sqrt{d})$ satisfies $\bar{c}(G) \sim d/n$ for large $d$. This asymptotic expression also holds for strongly regular graphs and Erdős-Rényi graphs. In addition to the above theoretic graphs, we present numerical results based on real network data.

Our key finding, that a range of networks possess the asymptotic average clustering coefficient $\overline{d}/n$, where $\overline{d}$ is the empirical average degree of the network in question, as opposed to the small-world networks, suggests a new category of Erdős-Rényi like networks, which we hope could shed some lights on the network clustering phenomenon and stimulate further research efforts on the related topics.

# Appendix A

For a graph $G = (V, E)$, let $M_1 = \sum_{uv \in E} |\{w \in V : w \in N(u) \cap N(v)\}|$ be the total number of common neighbors of adjacent vertices. Similarly, let $M_2 = \sum_{uv \notin E} |\{w \in V : w \in N(u) \cap N(v)\}|$ be the total number of common neighbors of non-adjacent vertices. Apparently, the complexity of computing $\overline{\mu_i}$ is proportional to $M_i$ $(i = 1, 2)$. In Fig. 2 we show an illustrating example for a moderately clustered network, where $M_1 = 2M_2$, meaning that using (2) could halve the computational burden.

We show in Fig. 3 the ratio $M_2/M_1$ for general random networks created by the clustered random graph model following [33]. This model is basically the configuration random graph decorated by two sequences, i.e., $t_i$, the number of triangles in which the $i$-th vertex participates, and $s_i$, the number of single edges attached to the $i$-th vertex, other than those belonging to the triangles. We observe from Fig. 3 that $0 < M_2/M_1 < 0.8$ for networks of all orders considered, suggesting that our use of (2) could effectively reduce the computational burden for clustered networks.




| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | 2 | 2 | 2 | 2 |
| 2 | | | 1 | 3 | 1 |
| 3 | | | | 1 | 3 |
| 4 | | | | | 1 |
| 5 | | | | | |

Fig. 2. Left: a network with 5 vertices. Right: the associated table consisting of the numbers of common neighbors of each pair of vertices, where blue numbers indicate adjacent pairs of vertices while red numbers indicate non-adjacent pairs. Hence, $M_1 = 12$ and $M_2 = 6$.

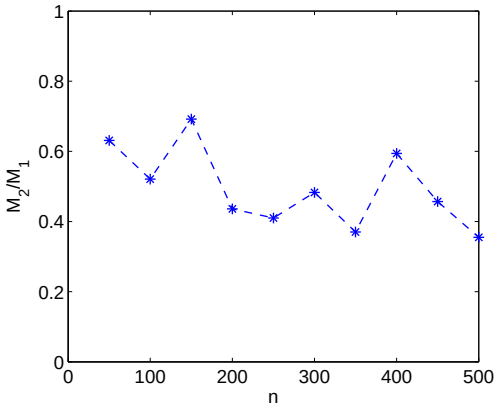



Fig. 3. The ratio $M_2/M_1$ versus $n$, the number of vertices, for clustered random graphs with $t_i, s_i \in \{1, 2, \cdots, 5\}$ uniformly at random. Each data point is based upon averaging over a sample of 30 independent realizations.

## Acknowledgement

The authors are very thankful to the editor and anonymous reviewers for their valuable comments and constructive suggestions that greatly help them improve the presentation of the manuscript.

## References

[1] L. A. Adamic and B. A. Huberman, Growth dynamics of the World Wide Web, *Nature*, **401** (1999), 131.

[2] S. Agreste, S. Catanese, P. D. Meo, E. Ferrara, G. Fiumara, Network structure and resilience of Mafia syndicates, *Inf. Sci.*, **351** (2016), 30-47.

[3] M. Aigner and G. Ziegler, *Proofs from THE BOOK* (3rd ed.), Springer, 2004.

[4] N. Alon and J. Spencer, *The Probabilistic Method, 3rd ed.*, Wiley-Interscience, New York, 2008.

[5] M. Amiel, G. Mélançon, and C. Rozenblat. Réseaux multi-niveaux : l'exemple des échanges aériens mondiaux de passagers. *M@ppemonde*, **79** (2005), 05302.

[6] A.-L. Barabasi, H. Jeonga, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, Evolution of the social network of scientific collaborations, *Physica A*, **311** (2002), 590-614.

[7] B. Bollobás, *Random Graphs (2nd ed.)*, Cambridge University Press, London-New York, 2001.

[8] B. Bollobás and O. Riordan, Mathematical results on scale-free random graphs, in *Handbook of Graphs and Networks: From the Genome to the Internet* (eds. S. Bornholdt and H. G. Schuster), pp. 1-34, Wiley-VCH, Berlin, 2002.

[9] R. L. Breiger and P. E. Pattison, Cumulated social roles: the duality of persons and their algebras, *Social Networks*, **8** (1986), 215-256.

[10] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, Topological structure analysis of the protein-protein interaction network in budding yeast, *Nucleic Acids Research*, **31** (2003), 2443-2450.

[11] F. R. Chung and R. Graham, Sparse quasi-random graphs, *Combinatoria*, **22** (2002), 217-244.

[12] F. R. Chung, R. Graham, and R. Wilson, Quasi-random graphs, *Combinatoria*, **9** (1989), 345-362.

[13] V. Colizza, R. Pastor-Satorras, and A. Vespignani, Reaction-diffusion processes and metapopulation models in heterogeneous networks, *Nature Physics*, **3** (2007), 276-282.

[14] P. Erdős, and A. Réyi, On a problem in theory of graphs, *Publ. Math. Inst. Hungar. acad. Sci.*, **7A** (1962), 623-641.

[15] G. Fagiolo, Clustering in complex directed networks, *Phys. Rev. E*, **76** (2007), 026107.

[16] D. A. Fell and A. Wagner, The small world of metabolism, *Nat. Biotechnol*, **18** (2000), 1121-1122.

[17] J. Friedman, A proof of Alon's second eigenvalue conjecture and related problems, *Mem. Amer. Math. Soc.*, **195** (2008), #910.

[18] M. Ghoniem, Outils de visualisation et d'aide á la mise au point de programmes avec contraintes. Phd, Université de Nantes, 2005.

[19] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer, 2001.

[20] L. S. Heath, N. Parikh, Generating random graphs with tunable clustering coefficients, *Physica A*, **390** (2011), 4577-4587.

[21] C. Hilgetag, G. A. P. C. Burns, M. A. O'Neill, J. W. Scannell, and M. P. Young, Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat, *Philos. Trans. R. Soc. London, Ser. B*, **355** (2000), 91-110.

[22] P. Holme and B. J. Kim, Growing scale-free networks with tunable clustering, *Phys. Rev. E*, **65** (2002), 026107.

[23] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1986.

[24] S. Iyer, T. Killingback, B. Sundaram, and Z. Wang, Attack robustness and centrality of complex networks, *PLoS One*, **8** (2013), e59613.

[25] M. Kaiser and C. C. Hilgetag, Spatial growth of real-world networks, *Phys. Rev. E*, **69** (2004), 036103.

[26] S. Kundu, S. K. Pal, FGSN: fuzzy granular social networks - model and applications, *Inf. Sci.*, **314** (2015), 100-117.

[27] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: densification and shrinking diameters, *ACM Transactions on Knowledge Discovery from Data*, **1** (2007), 2.

[28] M. Li and C. O'Riordan, The effect of clustering coefficient and node degree on the robustness of cooperation, *2013 IEEE Congress on Evolutionary Computation*, Cancun, 2013, 2833-2839.

[29] Y. Liu, C. Zhao, X. Wang, Q. Huang, X. Zhang, D. Yi, The degree-related clustering coefficient and its application to link prediction, *Physica A*, **454** (2016), 24-33.

[30] L. Lu and X. Peng, Spectra of edge-independent random graphs, *Electron. J. Combin.*, **20** (2013), #P27.

[31] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, New York, 2010.

[32] M. E. J. Newman, Scientific collaboration networks. I. Network construction and fundamental results, *Phys. Rev. E*, **64** (2001), 016131.

[33] M. E. J. Newman, Random graphs with clustering, *Phys. Rev. E*, **103** (2009), 058701.

[34] M. Ripeanu, I. Foster, and A. Iamnitchi, Mapping the Gnutella network: properties of large-scale peer-to-peer systems and implications for system design, *IEEE Internet Computing Journal*, **6** (2002), 50-57.

[35] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész, Generalizations of the clustering coefficient to weighted complex networks, *Phys. Rev. E*, **75** (2007), 027105.

[36] T. Schank and D. Wagner, Approximating clustering coefficient and transitivity, *J. Graph Algor. Appl.*, **9** (2005), 265-275.

[37] Y. Shang, Distinct clusterings and characteristic path lengths in dynamic small-world networks with identical limit degree distribution, *J. Stat. Phys.*, **149** (2012), 505-518.

[38] Y. Shang, Unveiling robustness and heterogeneity through percolation triggered by random-link breakdown, *Phys. Rev. E*, **90** (2014) 032820.

[39] Y. Shang, Bounding extremal degrees of edge-independent random graphs using relative entropy, *Entropy*, **18** (2016), #53.

[40] T. Szabó, On the spectrum of projective norm-graphs, *Inform. Process. Lett.*, **86** (2003), 71-74.

[41] S. Wang, L. Huang, C.-H. Hsu, F. Yang, Collaboration reputation for trustworthy Web service selection in social networks, *J. Comput. Syst. Sci.*, **82** (2016), 130-143.

[42] D. J. Watts and S. Strogatz, Collective dynamics of "small-world" networks, *Nature*, **393** (1998), 440-442.

[43] B. Zhang and S. Horvath, A general framework for weighted gene co-expression network analysis, *Stat. Appl. Genet. Mol. Biol.*, **4** (2005), 17.