Regularity lemmas for clustering graphs

Fan Chung
University of California, San Diego

Abstract

For a graph G with a positive clustering coefficient C, it is proved that for any positive constant ϵ, the vertex set of G can be partitioned into finitely many parts, say $S_1, S_2, \ldots, S_m$, such that all but an ϵ fraction of the triangles in G are contained in the projections of tripartite subgraphs induced by (S_i, S_j, S_k) which are ϵ-Δ-regular, where the size m of the partition depends only on ϵ and C. The notion of ϵ-Δ-regular, which is a variation of ϵ-regular for the original regularity lemma, concerns triangle density instead of edge density. Several generalizations and variations of the regularity lemma for clustering graphs are derived.

1 Introduction

One of the celebrated results of Szemerédi [21] is the so-called regularity lemma which asserts that for any graph on n vertices, the vertex set can be partitioned into finitely many parts so that almost all but ϵn^2 edges are contained in the union of bipartite subgraphs between pairs of the parts that are random-like under the notion of *ϵ-regular*. A bipartite graph is said to be ϵ-regular, if the edge density on any induced sub-bipartite graph on at least ϵn vertices differs from the edge density of the bipartite graph by at most ϵ. The regularity lemma has been a powerful tool in graph theory with numerous applications [13, 16, 19] because any graph (with more than ϵn^2 edges) can be approximated by a finite graph in the sense that each vertex of the finite graph can be replaced by a subset of vertices and the bipartite subgraphs between any two subsets are quasirandom.

A major deficiency of the regularity lemma is the fact that it is useful only for graphs with a positive edge density since the error bound of approximation is of order ϵn^2. There have been numerous attempts for possible extensions of the regularity lemma to sparse graphs, mostly with either additional assumptions [15] or weakened conditions [9, 20].

In this paper, we give a regularity lemma for clustering graphs without any restriction on edge density. We note that many information networks and social network graphs contain a large number of triangles and thus have nontrivial clustering coefficients [18, 22]. Such a clustering effect is one of the main characteristics of the so-called "*small world phenomenon*" that appear in a variety of real world graphs [17]. There are many research papers concerning finding dense subgraphs [2, 3] or partitioning into dense clique-like subgraphs [14] for such small-world graphs.

In this paper, we focus on graphs with nontrivial clustering coefficients (or triangle density). Let t_G denote the number of triangles in G and p_G denote the number of paths of two edges. The *clustering coefficient* C_G is defined to be (see [18])

$$C_G = \frac{3\ t_G}{p_G}. \tag{1}$$

If $p_G = 0$, we define $C_G = 0$. We say G is a *clustering graph* if its clustering coefficient C_G is a positive constant independent of the number of vertices of G.

Theorem 1 *For any $\epsilon > 0$ and any graph G with clustering coefficient C, the vertex set of G can be partitioned into $S_1, S_2, \ldots, S_m$ for some m depending only on ϵ and C, such that all but $\epsilon\ t_G$ triangles in G are contained in the projections of tripartite subgraphs with vertex set (S_i, S_j, S_k) that are ϵ-Δ-regular.*

The detailed definitions of various terms above will be given in Section 2. The proof of the regularity lemma for clustering graphs are quite similar to the previous proofs for the original regularity lemma [4, 16, 21] except for using an index function involving clustering coefficients. In Section 3 we give a proof of the regularity lemma for tripartite graphs with nontrivial clustering coefficient. The proof is self-contained and relatively short. In Section 4 we then consider a strong version of ϵ-Δ-regular for tripartite graphs. In Section 5 we give a proof of Theorem 1 and a weighted version of the regularity lemma both of which are straightforward applications of the regularity lemma for tripartite graphs with nontrivial clustering coefficients. In Section 6, we consider several generalizations of the regularity lemma. We will give a regularity for graphs which is dense in 4-cycles and, in general, graphs which contain a relatively large number of any specified graph (in comparison with its subgraphs). Some remarks and problems are mentioned in Section 7.

2 Preliminaries

We consider a tripartite graph $\mathcal{H}$ with the vertex set as the disjoint union $V_1 \sqcup V_2 \sqcup V_3$. Any triangle in $\mathcal{H}$ has one vertex in each V_i for $i = 1, 2, 3$. Let $t_{\mathcal{H}}$ denote the number of triangles in $\mathcal{H}$. Let $p_{\mathcal{H}}$ denote the number of triples (v_1, v_2, v_3) with $v_i \in V_i$ and $\{v_1, v_2\}, \{v_2, v_3\}$ are edges in $\mathcal{H}$. The clustering coefficient of a tripartite graph is defined to be

$$c_{\mathcal{H}} = \frac{t_{\mathcal{H}}}{p_{\mathcal{H}}} \tag{2}$$

For a graph $G = (V, E)$, it is helpful to consider the associated tripartite graph $\mathcal{G}$ which has vertex set as the disjoint union $V_1 \sqcup V_2 \sqcup V_3$ where V_i is a copy of V. Each vertex v in G is associated with three vertices, say, v_1, v_2, v_3. If $\{u, v\}$ is an edge in G, then $\{u_i, v_j\}$ is an edge in $\mathcal{G}$ if $i \neq j$. Thus, $|V(\mathcal{G})| = 3|V(G)|$ and $|E(\mathcal{G})| = 6|E(G)|$. Each triangle in G is associated with six triangles in $\mathcal{G}$. Namely, $t_{\mathcal{G}}$ can be viewed as the number of embeddings or homomorphism that maps a triple (u_1, v_2, w_3) where all three pairs are edges in $\mathcal{G}$.

In a tripartite graph $\mathcal{H}$ with $T_1 \subseteq V_1, T_2 \subseteq V_2, T_3 \subseteq V_3$, let $t(T_1, T_2, T_3)$ denote the number of triangles (u_1, v_2, w_3) in the induced tripartite subgraph with vertex set $T_1 \sqcup T_2 \sqcup T_3$. Similarly, let $p(T_1, T_2, T_3)$ denote the number of paths on 3 vertices u_1, v_2, w_3 with the first vertex in T_1, the middle vertex in T_2 and the third vertex in T_3 and $u \neq w$. The clustering coefficient of the induced tripartite subgraph with vertex set $T_1 \sqcup T_2 \sqcup T_3$ is denoted by $c(T_1, T_2, T_3)$.

Lemma 1 *For a graph G with the associated tripartite graph $\mathcal{G}$, we have $c_{\mathcal{G}} = C_G$.*

Proof: For the associated tripartite graph $\mathcal{G}$ associated with G, there are six ways to embed a triangle but there are just two days to embed a path of length two. We have $t(V_1, V_2, V_3) = 6t_G$ and $p(V_1, V_2, V_3) = 2p_G$. From the definitions in (1) and (2), we have $c_{\mathcal{G}} = c(V_1, V_2, V_3) = C_G$. □

We say a tripartite graph with vertex set $T_1 \sqcup T_2 \sqcup T_3$ is ϵ-Δ-regular if for any $S_i \subset T_i$ for $i = 1, 2, 3$, with $p(S_1, S_2, S_3) \geq \epsilon \, p(T_1, T_2, T_3)$, we have

$$|c(S_1, S_2, S_3) - c(T_1, T_2, T_3)| \leq \epsilon \tag{3}$$

We say a graph G is ϵ-Δ-regular if the associated tripartite graph $\mathcal{G}$ is ϵ-Δ-regular. Examples of ϵ-Δ-regular graphs include the complete graph with $c(K_n) = 1$ and random graphs $G_{n,p}$ for $p \geq \epsilon$ while $c(G_{n,p}) \sim p$.

The following inequality will be used later in the proofs (see [4]).

Fact: For positive values α_i and x_i, $i = 1, \ldots, N$, with $\sum_{i=1}^N \alpha_i = 1$, if for some $M < N$, we have

$$\left| \sum_{i=1}^M \alpha_i x_i - \Big(\sum_{i=1}^M \alpha_i \Big) \Big(\sum_{i=1}^N \alpha_i x_i \Big) \right| \geq \gamma,$$

$$\text{then} \qquad \sum_{i=1}^N \alpha_i x_i^2 \geq \Big(\sum_{i=1}^N \alpha_i x_i \Big)^2 + \gamma^2. \tag{4}$$

3 A regularity lemma for tripartite graphs

We first prove the following version of the regularity lemma for tripartite clustering graphs.

Theorem 2 *For any $\epsilon > 0$ and any tripartite graph $\mathcal{H}$ with clustering coefficient c, the vertex set $V_1 \sqcup V_2 \sqcup V_3$ of $\mathcal{H}$ can be partitioned into $S_1, S_2, \ldots, S_m$ for some m depending only on ϵ and c, such that all but $\epsilon t_{\mathcal{H}}$ triangles in $\mathcal{H}$ are contained in the ϵ-Δ-regular tripartite subgraphs with vertex set $S_i \sqcup S_j \sqcup S_k$.*

Proof: For a partition $\mathcal{P}$ consisting of partitions P_i of V_i, for $i = 1, 2, 3$, we define the index function $I(\mathcal{P})$:

$$I(\mathcal{P}) = I(P_1, P_2, P_3) = \sum_{\substack{T_1 \in P_1 \\ T_2 \in P_2 \\ T_3 \in P_3}} \frac{p(T_1, T_2, T_3)}{p(V_1, V_2, V_3)} c(T_1, T_2, T_3)^2$$

Initially, we consider the partition $\mathcal{P}^{(0)} = (\{V_1\}, \{V_2\}, \{V_3\})$. Clearly, we have $I(\mathcal{P}^{(0)}) = c_{\mathcal{G}}^2 = c^2$. It follows from the Cauchy-Schwarz inequality that if $\mathcal{P}'$ is a refinement of $\mathcal{P}$, then $I(\mathcal{P}') \geq I(\mathcal{P})$.

A partition $\mathcal{P} = (P_1, P_2, P_3)$ is said to be a *good* partition (or a ϵ-Δ-regular partition) if

$$\sum_{\substack{T_1 \in P_1, T_2 \in P_2, T_3 \in P_3 \\ (T_1, T_2, T_3)\ \epsilon\text{-}\Delta\text{-regular}}} \frac{p(T_1, T_2, T_3)}{p(V_1, V_2, V_3)} \geq 1 - \epsilon.$$

We will proceed by an iterative process. We start with $\mathcal{P}^{(0)}$. If we arrive at a good partition $\mathcal{P}^{(j)}$, stop. If the partition $\mathcal{P}^{(j)} = (P_1^{(j)}, P_2^{(j)}, P_3^{(j)})$ is not good, for each tripartite ϵ-Δ-irregular subgraph on (T_i, T_j, T_k), where $T_1 \in P_i, T_2 \in P_j, T_3 \in P_k$, we identify subsets $T_i' \subset T_i, T_j' \subset T_j, T_k' \subset T_k$

with $p(T_i', T_j', T_k') \ge \epsilon\, p(T_i, T_j, T_k)$ and $|c(T_i', T_j', T_k') - c(T_i, T_j, T_k)| \ge \epsilon$. Let $\mathcal{P}^{(j+1)} = (P_1^{(j+1)}, P_2^{(j+1)}, P_3^{(j+1)})$ denote the refinement of $\mathcal{P}^{(j)}$ using all (T_i', T_j', T_k')'s in the irregular (T_1, T_2, T_3)'s. We consider

$$\begin{aligned} I(\mathcal{P}^{(j+1)}) &= \sum_{\substack{S_1 \in P_1^{(j+1)} \\ S_2 \in P_2^{(j+1)} \\ S_3 \in P_3^{(j+1)}}} \frac{p(S_1, S_2, S_3)}{p(V_1, V_2, V_3)} c(S_1, S_2, S_3)^2 \\ &\ge \sum_{\substack{T_1 \in P_1^{(j)} \\ T_2 \in P_2^{(j)} \\ T_3 \in P_3^{(j)}}} \frac{p(T_1, T_2, T_3)}{p(V_1, V_2, V_3)} \sum_{\substack{S_1 \subset T_1, S_1 \in P_1^{(j+1)} \\ S_2 \subset T_2, S_2 \in P_2^{(j+1)} \\ S_3 \subset T_3, S_3 \in P_3^{(j+1)}}} \frac{p(S_1, S_2, S_3)}{p(T_1, T_2, T_3)} c(S_1, S_2, S_3)^2 \end{aligned}$$

Note that if (T_1, T_2, T_3) is ϵ-Δ-regular, we have

$$\begin{aligned} &\sum_{\substack{S_1 \subset T_1, S_1 \in P_1^{(j+1)} \\ S_2 \subset T_2, S_2 \in P_2^{(j+1)} \\ S_3 \subset T_3, S_3 \in P_3^{(j+1)}}} \frac{p(S_1, S_2, S_3)}{p(T_1, T_2, T_3)} c(S_1, S_2, S_3)^2 \\ &\ge \Big(\sum_{\substack{S_1 \subset T_1, S_1 \in P_1^{(j+1)} \\ S_2 \subset T_2, S_2 \in P_2^{(j+1)} \\ S_3 \subset T_3, S_3 \in P_3^{(j+1)}}} \frac{p(S_1, S_2, S_3)}{p(T_1, T_2, T_3)} c(S_1, S_2, S_3) \Big)^2 = c(T_1, T_2, T_3)^2 \end{aligned}$$

However, if (T_1, T_2, T_3) is ϵ-Δ-irregular, (4) implies

$$\begin{aligned} &\sum_{\substack{S_1 \subset T_1, S_1 \in P_1^{(j+1)} \\ S_2 \subset T_2, S_2 \in P_2^{(j+1)} \\ S_3 \subset T_3, S_3 \in P_3^{(j+1)}}} \frac{p(S_1, S_2, S_3)}{p(T_1, T_2, T_3)} c(S_1, S_2, S_3)^2 \\ &\ge \Big(\sum_{\substack{S_1 \subset T_1, S_1 \in P_1^{(j+1)} \\ S_2 \subset T_2, S_2 \in P_2^{(j+1)} \\ S_3 \subset T_3, S_3 \in P_3^{(j+1)}}} \frac{p(S_1, S_2, S_3)}{p(T_1, T_2, T_3)} c(S_1, S_2, S_3) \Big)^2 + \epsilon^4 \\ &= c(T_1, T_2, T_3)^2 + \epsilon^4 \end{aligned}$$

since there are $T_i' \subset T_i$, for $i = 1, 2, 3$, such that

$$\frac{p(T_1', T_2', T_3')}{p(T_1, T_2, T_3)} |c(T_1', T_2', T_3') - c(T_1, T_2, T_3)| \ge \epsilon^2.$$

Therefore, we have

$$I(\mathcal{P}^{(j+1)}) \geq \sum_{\substack{T_1 \in P_1^{(j)} \\ T_2 \in P_2^{(j)} \\ T_3 \in P_3^{(j)}}} \frac{p(T_1,T_2,T_3)}{p(V_1,V_2,V_3)} c(T_1,T_2,T_3)^2 + \epsilon^4 \sum_{\substack{T_1 \in P_1^{(j)} \\ T_2 \in P_2^{(j)} \\ T_3 \in P_3^{(j)} \\ \epsilon\text{-}\Delta\text{-irregular}}} \frac{p(T_1,T_2,T_3)}{p(V_1,V_2,V_3)}$$

$$\geq I(P^{(j)}) + \epsilon^5.$$

Since the clustering coefficient is no more than 1, we have $I(\mathcal{P}) \leq 1$. The iterated process must stop after at most $(1-C^2)/\epsilon^5$ rounds. A crude upper bound for the size $|\mathcal{P}^{(j+1)}|$ is $2^{|\mathcal{P}^{(j)}|^5}$. The final partition must be good and has finitely many parts. This completes the proof of Theorem 2. □

We remark that by using a slight different definition for ϵ-regular, with more careful treatment as in [8] and by using a slight different definition of regularity, the term ϵ^5 in the above proof can be further reduced to ϵ^2 with a more careful treatment as in [8]. and the tower upper bound for the size of the partition can be improved accordingly.

4 A strong regularity lemma for tripartite graphs

For a tripartite graph $\mathcal{H}$ with vertex set $T_1 \sqcup T_2 \sqcup T_3$, we consider some variations of clustering coefficient. Recall that

$$c_{\mathcal{H}} = c^{(1)}_{\mathcal{H}} = c(T_1,T_2,T_3) = \frac{t(T_1,T_2,T_3)}{p(T_1,T_2,T_3)}.$$

$$\text{We define } \; p^{(2)}_{\mathcal{H}} = p(T_2,T_3,T_1), \; c^{(2)}_{\mathcal{H}} = c(T_2,T_3,T_1),$$

$$\text{and } \; p^{(3)}_{\mathcal{H}} = p(T_3,T_1,T_2), \; c^{(3)}_{\mathcal{H}} = c(T_3,T_1,T_2).$$

For $j = 1,2,3$, we say a tripartite graph with vertex set $T_1 \sqcup T_2 \sqcup T_3$ is ϵ-$\Delta^{(j)}$-regular if for any $S_i \subset T_i$ for $i = 1,2,3$, with $p^{(j)}(S_1,S_2,S_3) \geq \epsilon\, p^{(j)}(T_1,T_2,T_3)$, we have

$$\left| c^{(j)}(S_1,S_2,S_3) - c^{(j)}(T_1,T_2,T_3) \right| \leq \epsilon.$$

We say a tripartite graph with vertex set $T_1 \sqcup T_2 \sqcup T_3$ is strongly ϵ-Δ-regular if it is ϵ-$\Delta^{(j)}$-regular for $j = 1,2,3$. A graph G is ϵ-Δ-regular if the associated tripartite graph $\mathcal{G}$ is strongly ϵ-Δ-regular.

Theorem 3 *For any $\epsilon > 0$ and any tripartite graph $\mathcal{H}$ with clustering coefficient c, the vertex set $V_1 \sqcup V_2 \sqcup V_3$ of $\mathcal{H}$ can be partitioned into $S_1, S_2, \ldots, S_m$ for some m depending only on ϵ and c, such that all but $\epsilon t_{\mathcal{H}}$ triangles in $\mathcal{H}$ are contained in the strongly ϵ-Δ-regular tripartite subgraphs with vertex set $S_i \sqcup S_j \sqcup S_k$.*

Proof: For a partition $\mathcal{P} = (P_1, P_2, P_3)$ where P_i is a partition of V_i, for $i = 1, 2, 3$, we define the index function $I^{(j)}(\mathcal{P})$ for $j = 1, 2, 3$,

$$I^{(j)}(\mathcal{P}) = I^{(j)}(P_1, P_2, P_3) = \sum_{T_1 \in P_1} \sum_{T_2 \in P_2} \sum_{T_3 \in P_3} \frac{p^{(j)}(T_1, T_2, T_3)}{p^{(j)}(V_1, V_2, V_3)} c^{(j)}(T_1, T_2, T_3)^2.$$

The remaining proof is quite similar to that of Theorem 2 except that we do three rounds. In the first round, we start with the trivial partition $\mathcal{P}^{(0)}$ and we use the index function $I^{(1)}(\mathcal{P})$ instead of $I(\mathcal{P})$. If P is not ϵ-$\Delta^{(1)}$-regular, we can get a refinement of P and increase the index function by ϵ^5, if If $\mathcal{P}$ is ϵ-$\Delta^{(1)}$-regular, we move to the second round. By the same argument as in the proof of Theorem 2, after a finite number of steps, we are in the second round. Then we switch to use the index function $I^{(2)}(\mathcal{P})$ and we proceed with further refinements until the partition is ϵ-$\Delta^{(2)}$-regular. Then we move to the third round and use the index function $I^{(3)}(\mathcal{P})$. At the end of the third round, the partition is ϵ-$\Delta^{(j)}$-regular for $j = 1, 2, 3$ as desired. □

5 Regularity lemmas for triangle-dense graphs

In a graph $G = (V, E)$, we consider the associated tripartite graph $\mathcal{G} = \mathcal{G}(V_1, V_2, V_3)$, where V_i's are copies of V. For any three subsets $S_1, S_2, S_3 \subseteq V$, not necessarily distinct, we consider the associated induced subgraph of $\mathcal{G}$, denoted by $\mathcal{G}(T_1, T_2, T_3)$, where T_i is the copy of S_i in V_i. For a triple (v_1, v_2, v_3) where $v_i \in S_i$, we note that v_1, v_2, v_3 form a triangle in G if and only if (v_1, v_2, v_3) forms a triangle in $\mathcal{G}(T_1, T_2, T_3)$. In other words, the set of triangles in $\mathcal{G}$ are in one-to-one correspondence with ordered (or labeled) triangles in G. In this section, all triangles in G are considered to be ordered triples that form triangles.

We are ready to prove Theorem 1.

Proof of Theorem 1:
We apply Theorem 3 to obtain a vertex-partition $\mathcal{P} = P_1 \sqcup P_2 \sqcup P_3$ for the associated tripartite graph $\mathcal{G}$ of G and all but $\epsilon t_{\mathcal{H}}$ triangles in $\mathcal{H}$ are contained in the strongly ϵ-Δ-regular tripartite subgraphs with vertex set

formed from three sets in $\mathcal{P}$. Then the desired partition P for V is taken to be the refinement of P_1, P_2, P_3 and P has size at most $|P_1| \cdot |P_2| \cdot |P_3|$. $\square$

In the remainder of this paper, we consider weighted graphs. In previous sections and in Theorem 1, we concern graphs with edge weight 1. For a tripartite graph $\mathcal{H} = \mathcal{H}(X, Y, Z)$, we consider an associated weighted graph on the same vertex set $X \sqcup Y \sqcup Z$, the edge weight $\omega(x, z)$, for $x \in X, y \in Y$ is defined to be the number of triangles containing x, y in $\mathcal{H}$. Our goal here is to define a notion of being ϵ-regular for weighted bipartite subgraphs of $\mathcal{H}$. For $X' \subseteq X$, and $Z' \subseteq Z$, we define $\mu(X', Z')$ to be $p(X', Y, Z')$, the number of paths on three vertices with one of the end-vertices in X' and the other in Z'. We say the weighted bipartite graph with vertex set $X \sqcup Z$ is ϵ-Δ-regular if for any $X' \subset X, Z' \subset Z$ with $\mu(X', Z') \geq \epsilon\mu(X, Z)$, the weight $\omega(X', Z') = \sum_{x \in X', z \in Z'} \omega(x, z)$ satisfies

$$\left| \omega(X', Z') - \frac{\mu(X', Z')}{\mu(X, Z)} w(X', Z') \right| \leq \epsilon\mu(X', Z') \tag{5}$$

Note that (5) is equivalent to the definition previously given in (3). As an immediate consequence of (3) and Theorem 1, we have

Theorem 4 *For any $\epsilon > 0$ and any graph G with clustering coefficient C, the vertex set of G can be partitioned into $S_1, S_2, \ldots, S_m$ for some m depending only on ϵ and C, such that all but ϵt_G triangles in G are associated with the tripartite subgraphs with vertex set (S_i, S_j, S_k) in the associated tripartite graph $\mathcal{G}$ and the three weighted bipartite subgraphs with vertex set on $S_i \sqcup S_j, S_j \sqcup S_k, S_k \sqcup S_i$, respectively, are ϵ-Δ-regular.*

6 Several regularity lemmas for general clustering graphs

Many information networks are bipartite and therefore do not have non-trivial clustering coefficient as defined in (1). Nevertheless, some of these graphs contain a relatively large number of 4-cycles C_4. For a graph G, we can define the C_4-clustering coefficient of G, defined by

$$C(G; C_4) = \frac{4N(G; C_4)}{N(G; P_4)} \tag{6}$$

where $N(G, H)$ denotes the number of subgraph of G isomorphic to H. The usual clustering coefficient is just $C(G; C_3)$.

Before we define ϵ-C_4-regular, we consider the 4-partite graph with vertex set $V_1 \sqcup V_2 \sqcup V_3 \sqcup V_4$. For $T_i \subseteq V_i$, we define $p(T_1, T_2, T_3, T_4)$, to be the number of 4-paths (v_1, v_2, v_3, v_4) with $v_i \in T_i$ and $q(T_1, T_2, T_3, T_4)$ to be the number of 4-cycles (u_1, u_2, u_3, u_4) with $u_i \in T_i$. Finally we define $c(T_1, T_2, T_3, T_4) = q(T_1, T_2, T_3, T_4)/p(T_1, T_2, T_3, T_4)$.

We say a 4-partite graph with vertex set $T_1 \sqcup T_2 \sqcup T_3 \sqcup T_4$ is ϵ-C_4-regular if for any $S_i \subset T_i$ for $i = 1, 2, 3, 4$, with $p(S_1, S_2, S_3, S_4) \geq \epsilon\, p(T_1, T_2, T_3, T_4)$, we have

$$|c(S_1, S_2, S_3, S_4) - c(T_1, T_2, T_3, T_4)| \leq \epsilon \tag{7}$$

We say a graph G is ϵ-C_4-regular if the associated 4-partite graph $\mathcal{G}$ is ϵ-C_4-regular.

We have the following.

Theorem 5 *For any $\epsilon > 0$ and any graph G with C_4-clustering coefficient $C(G; C_4)$, the vertex set of G can be partitioned into $S_1, S_2, \ldots, S_m$ for some m depending only on ϵ and $C(G; C_4)$, such that all but $\epsilon N(G; C_4)$ 4-cycles in G are contained in the 4-partite subgraphs with vertex set (S_i, S_j, S_k, S_l) in the associated 4-partite graph $\mathcal{G}$ which are strongly ϵ-C_4-regular.*

Proof: The proof is almost identical to those of Theorem 1 and 2 except for the definition of the index function for the 4-partite graph associated with G. Namely, for a partition $\mathcal{P} = (P_1, P_2, P_3, P_4)$ of the vertex set of a 4-partite graph with vertex set $V_1 \sqcup V_2 \sqcup V_3 \sqcup V_4$, define

$$I(\mathcal{P}) = I(P_1, P_2, P_3, P_4) = \sum_{\substack{T_1 \in P_1 \\ T_2 \in P_2 \\ T_3 \in P_3 \\ T_4 \in P_4}} \frac{p(T_1, T_2, T_3, T_4)}{p(V_1, V_2, V_3, V_4)} c(T_1, T_2, T_3, T_4)^2$$

where $p(T_1, T_2, T_3, T_4)$ is the number of 4-paths (v_1, v_2, v_3, v_4) with $v_i \in V_i$, $q(T_1, T_2, T_3, T_4)$ is the number of 4-cycles (u_1, u_2, u_3, u_4) with $u_i \in V_i$ and $c(T_1, T_2, T_3, T_4) = q((T_1, T_2, T_3, T_4)/p((T_1, T_2, T_3, T_4)$. □

Theorem 6 can easily be generalized to other specified graphs. Let F be a given graph (say, a Petersen graph). Let H denote a proper connected spanning subgraph of F's. In other words, $V(F) = V(H)$ and $E(H) \subset E(F)$. We define the (F, H)-clustering coefficient of G to be

$$C(G; F, H) = \frac{N(G; F)}{N(G; H)}. \tag{8}$$

A graph G is said to be (F, H)-clustering if its (F, H)-clustering coefficient $C(G; F, H)$ is an absolute positive constant. We can define ϵ-(F, H)-regular and strongly ϵ-(F, H)-regular in the same way as we have done for the case of F as a path on 3 vertices and H as a triangle.

Theorem 6 *Suppose F is a connected graph on k vertices and H is a proper connected spanning subgraph. For any $\epsilon > 0$ and any G with (F, H)-clustering coefficient $C(G; F, H)$, the vertex set of G can be partitioned into $S_1, S_2, \ldots, S_m$ for some m depending only on ϵ and $C(G; F, H)$, such that all but $\epsilon N(G; F)$ copies of F in G are contained in the associated k-partite subgraphs with vertex set $(S_{i_1}, S_{i_2}, \ldots, S_{i_k})$ which are strongly ϵ-(F, H)-regular.*

To avoid repetition, we skip the proof here.

7 Problems and remarks

A natural question is to derive a reasonable upper bound for the size of the ϵ-Δ-regular partition for clustering graphs. A crude upper bound as mentioned in the proof of Theorem 1 is of tower type, namely, a tower of 2's of height proportional to $1/\epsilon^5$ where C is the clustering coefficient and ϵ is the desired accuracy. For the original regularity lemma, Gowers [11] gave a lower bound for the size of the partition as a tower of 2's of height $1/\epsilon^{1/16}$. With a slightly different definition of regularity, Fox and Lovász [8] proved the tight bound as a tower of 2's of height $1/\epsilon^2$. It is of both theoretical and practical interest to see if the clustering property could be helpful for reducing the size of the partition.

A further direction for research is to further examine the ϵ-Δ-regular property. Note that the ϵ-regular property belongs to the family of quasi-random graph properties (see [5]) and therefore is equivalent to a large number of properties, shared by random graphs. It is desirable to develop a quasi-random theory for ϵ-Δ-regular graphs, perhaps in a similar fashion as for quasi-random hypergraphs [6] or sparse quasi-random graphs [7].

The above questions can be asked for the ϵ-C_4-regular partitions as described in Section 6. This also opens up many general questions for ϵ-(F, H)-regular graphs for any pair of connected graphs F and H where H is a proper connected spanning subgraph of F.

Another direction are the algorithmic aspects of finding ϵ-Δ-regular partitions. The problem of finding ϵ-regular partitions of regularity lemma have been extensively investigated [1, 9]. Similar questions can be asked here for finding regularity partitions for clustering graphs that are rich in triangles, 4-cycles or general specified graphs.

References

[1] N. Alon, R. A. Duke, H. Lefmann, V. Rödl, and R. Yuster, The algorithmic aspects of the regularity lemma, *Journal of Algorithms*, **16**, (1994), 80–109.

[2] R. Andersen, A local algorithm for finding dense subgraphs, *Proc. 19th ACM-SIAM Symposium on Discrete Algorithms (SODA 2008)*, 1003–1009.

[3] M. Charikar, Greedy Approximation Algorithms for Finding Dense Components in a Graph, (K. Jansen, S. Khuller eds.), *Approximation Algorithms for Combinatorial Optimization, APPROX 2000, Lecture Notes in Computer Science*, **1913**. Springer, Berlin, (2000), 84–95.

[4] F. Chung, Regularity lemmas for hypergraphs and quasi-randomness *Random Structures & Algorithms*, **2**, (1991), 241–252.

[5] F. Chung, R. L. Graham and R. M. Wilson, Quasi-random graphs, *Combinatorica*, **9** (1989), 345–362.

[6] F. Chung and R. L. Graham, Quasi-random hypergraphs, *Random Structures and Algorithms*, **1**. (1990), 105–124.

[7] F. Chung and R. L. Graham, Sparse quasi-random graphs, *Combinatorica*, **22**, (2002), 217–244.

[8] J. Fox and L. M. Lov'asz, A tight lower bound for Szemer'edi's regularity lemma, *Combinatorics*, **37** (2017), 911–951.

[9] A. Frieze and R. Kannan, A simple algorithm for constructing Szemerdi's regularity partition, *The Electronic Journal of Combinatorics,* **6**, (1999), #R17, 7 pages.

[10] S. Gerke and A. Steger, The sparse regularity lemma and its applications, *Surveys in Combinatorics 2005*, 227–258, London Math. Soc. Lecture Notes Ser. 327, Cambridge Univ. Press, Cambridge.

[11] W. T. Gowers, Lower bounds of tower type for Szemerédi's uniformity lemma, *Geometric & Functional Analysis GAFA*, **7** (1997), 322–337.

[12] B. Green, A Szemerdi-type regularity lemma in abelian groups, with applications, *Geometric & Functional Analysis GAFA*, **15**, (2005), 340–376.

[13] B. Green and T. Tao, An arithmetic regularity lemma, an associated counting lemma, and applications *An Irregular Mind*, Springer, (2010), 261–334.

[14] R. Gupta, T. Roughgarden and C. Seshadhri, Decompositions of triangle-dense graphs, *SIAM J. Compt.*, **45** (2016), 197–215.

[15] Y. Kohayakawa, Szemerédi's regularity lemma for sparse graphs, *Foundations of Computational Mathematics*, Springer, Berlin, (1997), 216–230.

[16] J. Komlós and M. Simonovits, Szemerédi's regularity lemma and its applications in graph theory, *Combinatorics, Paul Erdős is Eighty*, Vol. 2 (Keszthely, 1993), 295–352, *Bolyai Soc. Math. Stud.*, 2, János Bolyai Math. Soc., Budapest, 1996.

[17] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Statistical properties of community structure in large social and information networks, *Proceedings of the 17th International Conference on World Wide Web (WWW2008)*, 695–704.

[18] R. D. Luce and A. D. Perry, A method of matrix analysis of group structure, *Psychometrika*, **14**, (1949), 95–116.

[19] V. Rödl and J. Skokan, Applications of the regularity lemma for uniform hypergraphs, *Random Structures & Algorithms*, **25**, (2006), 1–42.

[20] A. Scott, Szemerédi's regularity lemma for matrices and sparse graphs, *Combinatorics, Probability and Computing*, **20**, (2011), 455-466.

[21] E. Szemerédi, Regular partitions of graphs, Problèmes combinatoires et théorie des graphes, *Colloq. Internat. CNRS, Univ. Orsay*, Orsay, (1976), 399–401, *Colloq. Internat. CNRS*, 260, CNRS, Paris, 1978.

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