

Notes on Applying the Regularity Lemma

We have seen how to analyze a simple algorithm for testing bipartiteness. There are also analyses of algorithms for k -colorability, having a clique of a certain size, and in general for the family of all *partition properties*. Namely these properties are defined by upper and lower bounds on the sizes of some constant number k of parts and upper and lower bounds on the edge-densities between these parts. The number of queries performed is polynomial in $1/\epsilon$ and exponential in k . The time-complexity is exponential in $1/\epsilon$, but this is inevitable (assuming $P \neq NP$) since partition problems include NP -hard problems.

There are other general results. One of the most general results is that all hereditary properties can be tested using a number of queries that depends only on $1/\epsilon$. A property is hereditary if it is closed under taking induced subgraphs. That is: if G has the property then every induced subgraph of G has the property (e.g. k -colorability, subgraph freeness). Actually, (roughly) these properties characterize what is testable with one-sided error (using a number of queries that depends only on $1/\epsilon$). This doesn't cover properties for which there are two-sided error algorithms (as is the case for some of the partition properties), but there is an even more general result for them.

All these results use variants of an important tool: Szemerédi's regularity lemma. Here we shall show how it can be used for one particular property: *triangle-freeness*.

In order to state the lemma, we need some definitions and notation. For any two non-empty disjoint sets of vertices, A and B , we let $E(A, B)$ be the set of edges between A and B , and we let $e(A, B) = |E(A, B)|$. The *edge-density* of the pair is defined as $d(A, B) = \frac{e(A, B)}{|A| \cdot |B|}$. We say that a pair A, B is γ -regular for some $\gamma \in [0, 1]$ if for every two subsets $A' \subseteq A$ and $B' \subseteq B$ satisfying $|A'| \geq \gamma|A|$ and $|B'| \geq \gamma|B|$ we have that $|d(A', B') - d(A, B)| < \gamma$. Note that if we consider a random bipartite graph between A and B (where there is an edge between each pair of vertices $v \in A$ and $u \in B$ with constant probability p), then it will be regular w.h.p. for some constant γ .

Lemma 1 *For every integer ℓ_0 and for every $\gamma \in (0, 1]$, there exists a number $u_0 = u_0(\ell_0, \gamma)$ with the following property: Every graph $G = (V, E)$ with $n \geq u_0$ vertices has an equipartition $\mathcal{A} = \{V_1, \dots, V_k\}$ of V where $\ell_0 \leq k \leq u_0$ for which all pairs (V_i, V_j) but at most $\gamma \cdot \binom{k}{2}$ of them are γ -regular.*

We won't prove this lemma and only use it to prove the correctness of the "naive" testing algorithm for triangle-freeness: It takes a sample of size $m = m(\epsilon)$ (which will be set later), queries all pairs of vertices in the sample to obtain the induced subgraph, and accepts or rejects depending on whether it sees a triangle in the induced subgraph. Clearly, if the graph is triangle-free then it always accepts. It remains to prove that if the graph is ϵ -far from triangle-free then (for sufficiently large $m = m(\epsilon)$), the sample will contain a triangle with high constant probability. An important note is in place concerning the size of m . It can be shown that (as opposed to bipartiteness and

other partition problems) it does **not** suffice to take m that is polynomial in $1/\epsilon$. That is, there exist graph that are ϵ -far from triangle-free but for which a $\text{poly}(1/\epsilon)$ -size sample will not show any triangle. This is a bit counterintuitive, but true. (The size m that we can show suffices for our needs, will be higher by quite a bit from the lower bound so there is quite a big gap between upper and lower bound.)

Suppose we apply the regularity lemma with $\ell_0 = 8/\epsilon$ and $\gamma = \epsilon/8$. Our first observation is that for this setting, the total number of edges in G that are between pairs of vertices that belong to the same part V_i of the partition is at most

$$k \cdot \left(\frac{n}{k}\right)^2 = \frac{1}{k} \cdot n^2 \leq \frac{1}{\ell_0} \cdot n^2 = \frac{\epsilon}{8} n^2$$

It follows that if we define G_1 as the graph that is the same as G except that we remove all edges within the part, the G_1 is at least $(7/8)\epsilon$ -far from triangle-free.

Next, since there are at most $\left(\frac{\epsilon}{8}\right) \cdot \binom{k}{2} < \frac{\epsilon}{16} k^2$ non-regular pairs, the total number of edges between non-regular pairs in the partition is at most

$$\frac{\epsilon}{16} k^2 \cdot \left(\frac{n}{k}\right)^2 = \frac{\epsilon}{16} n^2$$

Therefore, if we continue by removing all these edges, and let the resulting graph be denoted G_2 , then it is at least $(3/4)\epsilon$ -far from being triangle-free.

We'll perform one more step of this kind. Consider all pairs (V_i, V_j) such that $d(V_i, V_j) < \frac{\epsilon}{2}$. That is, $e(V_i, V_j) < \frac{\epsilon}{2} \cdot \left(\frac{n}{k}\right)^2$. Since there are at most $k^2/2$ such pairs, the total number of edges between such pairs is at most $\frac{\epsilon}{4} n^2$. Therefore, if we remove all these edges, and let the resulting graph be denoted G_3 , then G_3 is at least $(\epsilon/2)$ -far from being triangle-free. In particular this means that there exists at least one triplet (V_i, V_j, V_ℓ) such that all three edge densities, $d(V_i, V_j)$, $d(V_j, V_\ell)$ and $d(V_i, V_\ell)$ are at least $\epsilon/2$ in G_3 . (If no such triplet exists then G_3 would be triangle-free.) We shall show that since all three pairs are $(\epsilon/8)$ -regular, there are "many real triangles" $(u, v, w) \in V_i \times V_j \times V_\ell$, so that a sufficiently large sample will catch one.

For simplicity we denote the three subsets by V_1, V_2, V_3 . For each vertex $v \in V_1$, we let $\Gamma_2(v)$ denote the set of neighbors that v has in V_2 , and by $\Gamma_3(v)$ the set of neighbors that v has in V_3 . We shall say that v is *useful* if both $|\Gamma_2(v)| \geq \frac{\epsilon}{4} \left(\frac{n}{k}\right)$ and $|\Gamma_3(v)| \geq \frac{\epsilon}{4} \left(\frac{n}{k}\right)$. Since (V_2, V_3) are a regular pair,

$$e(\Gamma_2(v), \Gamma_3(v)) \geq (d(V_2, V_3) - \gamma) \left(\frac{\epsilon}{4}\right)^2 \left(\frac{n}{k}\right)^2 \geq \frac{\epsilon^3}{c \cdot k^2} n^2$$

for some constant c . It follows that if we get a useful vertex v from V_1 , and then we take an additional sample of $\Theta((u_0)^2/\epsilon)$ pairs of vertices (recall that $k \leq u_0$), then we will see a triangle. It remains to show that there are relatively many useful vertices in V_1 . Consider any vertex $z \in V_1$ that is *not* useful. We shall say that it is *unuseful of type 2* if $|\Gamma_2(v)| < \frac{\epsilon}{4} \left(\frac{n}{k}\right)$, and that it is *unuseful of type 3* if $|\Gamma_3(v)| < \frac{\epsilon}{4} \left(\frac{n}{k}\right)$. Without loss of generality, assume that there are more unuseful vertices of type 2. Suppose that at least half the vertices in V_1 are unuseful. Then at least a fourth are unuseful of type 2. Let V_1' consist of all these vertices, so that $|V_1'| > \gamma|V_1|$ (recall that $\gamma = \epsilon/8 \geq 1/8$). Let $V_2' = V_2$. But then,

$$d(V_1', V_2) \leq \frac{|V_1'| \cdot \frac{\epsilon}{4} \left(\frac{n}{k}\right)}{|V_1'| \cdot |V_2|} = \frac{\epsilon}{4} < d(V_1, V_2) - \gamma$$

and we have reached a contradiction to the regularity of V_1 and V_2 . Hence, at least a half of the vertices in V_1 are useful (that is $\Omega\left(\frac{n}{k}\right)$ vertices, and so a sample of size $\Theta(u_0)$ will contain a useful vertex with high probability.

Therefore, if we take a sample of size $\Theta(u_0) + \Theta((u_0)^2/\epsilon) = \Theta((u_0)^2/\epsilon)$, then we'll see a triangle with high probability.