6.5240 Sublinear Time Algorithms Lecture 18

Szemerédi's Regularity Lemma (SRL):
Testing Triangle-freeness

Lecturer: Ronitt Rubinfeld
Scribe: John Kuszmaul*

1 Szemerédi's Regularity Lemma (SRL)

1.1 Motivation: Graphs with "Random" Properties

Consider an example question: how many triangles are in a random tripartite graph with density η? Suppose the graph is tripartite on vertices $A, B, C \subset V$. Then for any $u \in A, v \in B, w \in C$, we have that

$$\Pr[u,\ v,\ w \text{ forms a triangle}] = \eta^3$$

and thus

$$\mathbb{E}[\#\text{ of triangles}] = |A||B||C|\eta^3.$$

A natural question then, is can we make weaker assumptions and still get reasonable bounds? To answer this question, we will turn to Szemerédi's Regularity Lemma. First, we will start with some preliminary definitions.

1.2 Definitions

For $A, B \subseteq V$ satisfying $A \cap B = \emptyset$, we define $e(A, B)$ to be the number of edges between A and B. The ***density*** is

$$d(A, B) = \frac{e(A, B)}{|A| \cdot |B|}.$$

We say (A, B) is ***γ-regular*** if for all $A' \subseteq A$, $B' \subseteq B$ such that $|A'| \geq \gamma|A|$ and $|B'| \geq \gamma|B|$, satisfy

$$|d(A', B') - d(A, B)| < \gamma.$$

Note that this latter definition is similar to saying that the graph "behaves" like a random graph.

1.3 Using Regularity to Count Triangles

Lemma 1. *For all $\eta \in (0, 1]$, there exists a γ, δ (Spoiler: $\gamma = \eta/2$ and $\delta = (1-\eta)\eta^3/8$) such that if A, B, C are disjoint subsets of V and each pair γ-regular with density $> \eta$, then G contains at least $\delta|A||B||C|$ distinct triangles with nodes in each of A, B, and C.*

Proof. Let A^* be the nodes in A that are ***typical***; we define them to be the nodes with at least $|\eta - \gamma||B|$ neighbors in B and at least $|\eta - \gamma||C|$ neighbors in C. By the claim below, we know $|A^*| \geq (1 - 2\gamma)|A|$.

For each $u \in A^*$, we count the number of triangles that contain u. We define B_u to be the neighbors of u in B and C_u to be the neighbors of u in C. In fact, we know these sets are not too small; since $\gamma = \eta/2$, we have that $|B_u| \geq (\eta - \gamma)|B| \geq \gamma|B|$ and $|C_u| \geq (\eta - \gamma)|C| \geq \gamma|C|$.

*MIT CSAIL. Email: john.kuszmaul@gmail.com

We observe that the number of edges between B_u and C_u is a lower bound on the number of triangles in which u participates. By the γ-regularity of the pair (B, C), we know

$$d(B_u, C_u) \geq \eta - \gamma.$$

Thus the number of edges between them is at least

$$\begin{aligned} e(B_u, C_u) &\geq (\eta - \gamma)|B_u||C_u| \\ &\geq (\eta - \gamma)^3|B||C|. \end{aligned}$$

Thus $(\eta - \gamma)^3|B||C|$ is a lower bound on the number of triangles in which $u \in A$ participates. Summing over all $u \in A$, we find that the total number of triangles is at least

$$(1 - 2\gamma)|A|(\eta - \gamma)^3|B||C| \geq (1 - \eta)(\eta/2)^3|A||B||C|,$$

where the inequality follows since $\gamma = \eta/2$. □

Claim 2. $|A^*| \geq (1 - 2\gamma)|A|$.

Proof of claim. Let A' be the bad nodes with regard to B, ($< |\eta - \gamma||B|$ neighbors in B) and let A'' be the bad nodes with regard to C (less than $|\eta - \gamma||C|$ neighbors in C). Then we argue that $|A'| \leq \gamma|A|$ and $|A''| \leq \gamma|A|$. Why is this the case? Consider the pair A', B. We have

$$\begin{aligned} d(A', B) &< \frac{|A'| \cdot |\eta - \gamma||B|}{|A'||B|} \\ &= \eta - \gamma. \end{aligned}$$

But $d(A, B) > \eta$, so $|d(A', B) - d(A, B)| > \gamma$.

We also know that $|B| \geq \gamma|B|$, so if $A' \geq \gamma|A|$, then (A, B) is not γ-regular. This is a contradiction. Thus $|A'| \leq \gamma|A|$. By a symmetric argument for A'', we have that $|A''| \leq \gamma|A|$. Let $A^* = A \setminus (A' \cup A'')$. Then we have

$$\begin{aligned} |A^*| &\geq |A| - |A'| - |A''| \\ &\geq |A| - 2\gamma|A| \\ &= (1 - 2\gamma)|A|. \end{aligned}$$

□

1.4 SRL Lemma

Do interesting graphs have regularity properties? Yes, in some sense all graphs do – they "can be approximated as a small collection of random graphs." We would like to be able to say something like: "we can always equipartition V into $V_1, \ldots, V_k$ (for constant k) such that all pairs (V_i, V_j) are ε-regular." (And actually at least $(1 - \varepsilon)$ fraction of the pairs being ε-regular would usually be good enough.)

What we actually get is Szemerédi's Regularity Lemma:

Lemma 3 (Szemerédi's Regularity Lemma)**.** *For all $m, \varepsilon > 0$, there exists a constant $T = T(m, \varepsilon)$ such that for any graph $G = (V, E)$ satisfying $|V| > T$ and any equipartition $\mathcal{A}$ of V, there exists an equipartition $\mathcal{B}$ of V into k sets refining $\mathcal{A}$ such that $m \leq k \leq T$ and at most $\varepsilon\binom{k}{2}$ set pairs are not ε-regular.*

Remark 4. *SRL was first studied to prove that sequences of integers have long arithmetic progressions.*

2 An Application of SRL to Property Testing

Given a graph G in adjacency matrix format, we would like to answer the following question: is G triangle-free? The desired behavior of our algorithm is: if G is triangle-free, output PASS. If G is ε-far from being triangle-free, then output FAIL with probability at least 2/3. Notice this is one-sided error; all graphs that are triangle-free must pass deterministically. By ε-far from triangle-free, we mean that at least εn^2 edges must be deleted to make the graph triangle-free.

Our algorithm does the following $O(1/\delta)$ times. First it uniformly samples $v_1, v_2, v_3 \in_r V$. If they form a triangle, the algorithm rejects and halts. After completing this $O(1/\delta)$ times, the algorithm accepts.

2.1 Algorithm Analysis

Theorem 5. *For all ε, there exists a δ such that for any graph G whose number of nodes n is sufficiently large, if G is ε-far from triangle-free, then G has at least $\delta\binom{n}{3}$ distinct triangles.*

Corollary 6. *Our algorithm achieves the desired behavior.*

Proof. Why does it achieve the desired behavior?

- If G is triangle-free, then the algorithm will never reject.
- If G is ε-far from being triangle-free, then by the theorem there are at least $\delta\binom{n}{3}$ triangles in G. Thus each loops passes with probability at most $1-\delta$. So the probability that none of the $O(1/\delta)$ loops see a triangle is at most $(1-\delta)^{c/\delta} \leq e^{-c} < 1/3$ for an appropiate choice of the constant c in the $O(1/\delta)$. So the algorithm rejects with probability at least $2/3$.

□

Proof of Theorem. We start with an arbitrary equipartition $\mathcal{A}$ of G of into $5/\varepsilon$ sets. Let $\varepsilon' = \frac{\varepsilon}{5}$. If $n \geq T(\frac{5}{\varepsilon}, \varepsilon')$, we can use SRL to get an equipartition $\{V_1, \ldots, V_k\}$ such that

$$\frac{5}{\varepsilon} \leq k \leq T(\frac{5}{\varepsilon}, \varepsilon').$$

Or equivalently, $\varepsilon n/5 \geq n/k \geq n/T(5/\varepsilon, \varepsilon')$. SRL says that at most $\varepsilon'\binom{k}{2}$ pairs are not ε'-regular. Note that we need at least $5/\varepsilon$ sets in the partition so that no set has more than a $\varepsilon/5$ fraction of the nodes.

We assume n/k is an integer, and we will "clean up" G. We define G' to be the resulting graph when we take G and

1. For all i, delete V_i's internal edges. If $|V_i|$ is small, there will be few such edges. The number of such edges is at most

$$\frac{n}{k} \cdot n \leq \frac{\varepsilon n^2}{5}$$

   where k is the degree within V_i.

2. Delete edges between non-regular pairs. How many such edges are there? At most

$$\begin{aligned} \varepsilon'\binom{k}{2}\left(\frac{n}{k}\right)^2 &\leq \frac{\varepsilon}{5}\frac{k^2}{2}\frac{n^2}{k^2} \\ &= \frac{\varepsilon n^2}{10}. \end{aligned}$$

3. Delete all edges between low-density ($< \frac{\varepsilon}{5}$) pairs. The number of such edges is at most

$$\sum_{\text{low density}} \frac{\varepsilon}{5}\left(\frac{n}{k}\right)^2 \leq \frac{\varepsilon}{5}\binom{n}{2} \approx \frac{\varepsilon}{10}n^2$$

Thus the total number of deleted edges during this process is less than εn^2. But recall that G is ε-far from being triangle free. So G' therefore must still have a triangle, since at least εn^2 edges must be deleted to make such a graph triangle-free.

The main point, is that in this cleaned-up graph, the existence of one triangle implies the existence of many triangles. Any triangle in G' remaining must simultaneously connect:

1. Nodes in 3 distinct V_i, V_j, and V_k.
2. ε'-regular pairs.

3. High density ($\geq \frac{\varepsilon}{5}$) pairs.

Applying the triangle-counting lemma to V_i, V_j, and V_k, we find that G' has at least

$$\frac{\varepsilon}{10}(\frac{\varepsilon}{5})^3|V_i||V_j||V_k|$$

triangles, which is at least $\delta'\binom{n}{3}$ triangles in G' for

$$\delta' = \frac{\varepsilon^4}{1250}\frac{1}{6}\left[T\left(\frac{5}{\varepsilon},\varepsilon'\right)\right]^{-3}.$$

□

The runtime of the tester is $O(1/\delta)$, which is $O(\varepsilon^{-4}T(5/\varepsilon,\varepsilon')^3)$. This is roughly a "tower" of $\log(1/\varepsilon)$. This is a powerful technique! Further extensions allow:

- A similar lemma like triangle-counting can be used to test all constant-sized subgraphs.
- almost as is, we can use the same method to test all "first order" graph properties

2.2 Brief Discussion of Further Results

In general, we are interested in which kinds of graph properties are testable in dense graphs.

- 1-sided error in constant time can be done for hereditary graph properties (closed under vertex removal, chordal, perfect, interval).
- 2-sided error in constant time can be done for any property that can be reduced to testing if the graph satisfies one of a finite number of Szemerédi partitions.

2.3 Future Questions

Are there faster testers (in terms of ε) for specific graph properties? Maybe the reason that the dependence on ε is so bad, is that the technique is too "general purpose"? Still, we will see in the next lecture that triangle-freeness cannot be tested in time poly$(1/\varepsilon)$.