**CSCI 1951-W Sublinear Algorithms for Big Data** **Fall 2020**

# Lecture 8: Dense Graph – Triangle Freeness

*Lecturer: Jasper Lee* *Scribe: Jerry Chu*

## 1 Overview

This lecture focuses on the analysis of a simple sublinear algorithm that tests for triangle-freeness, a special case of subgraph-freeness testing (Lecture 6) in dense graphs. To analyse the testing algorithm, we will introduce a deep graph-theoretic result, which is the Szemeredi Regularity Lemma. The algorithm we see in this lecture can be generalized to subgraph-freeness testing but its query complexity highly depends on whether the subgraph is bipartite.

## 2 Problem Setup and Algorithm

**Definition 8.1** A graph $G = (V, E)$ is triangle-free if and only if no triplet forms a triangle, i.e. $\{(u, v), (v, w), (w, u)\} \not\subseteq E$ for all $\{u, v, w\} \subseteq V$.

**Algorithm 8.2** Given a dense graph $G = (V, E)$, repeat the following steps $O\left(\frac{1}{C(\varepsilon)}\right)$ times ($C$ to be determined):

1. Pick $\{u, v, w\} \subseteq V$ uniformly at random.
2. If $u, v, w$ form a triangle, reject.

Otherwise accept.

**Proposition 8.3** *Algorithm 8.2 is complete. It accepts all triangle-free graphs. This is trivial because there doesn't exist any triangle in a triangle-free graph for step 2 to reject.*

**Theorem 8.4** *Algorithm 8.2 is sound. If $G$ is $\varepsilon$-far from triangle-freeness, then there exists $O(C(\varepsilon)n^3)$ triplets that form triangles.*

The contrapositive of Theorem 8.4 is known as the triangle removal lemma. The fact that it has a name speaks volume. The rest of the lecture focuses on the proof of it.

## 3 Strategy and Motivation

We want to lower-bound the number of triangle-forming triplets in graphs $\varepsilon$-far from triangle-freeness. The challenge mainly comes from the discrete nature of combinatorics. It turns out that random behaviors in graphs are actually easier to analyze, such as in the case of the Erdos-Renyi model.

Suppose $X, Y, Z$ are disjoint sets of vertices and tripartite edges between sets are Bernoulli($\eta$). Each triplet of vertices $(x, y, z) \in X \times Y \times Z$ form a triangle with probability $\eta^3$, so the expected number of triangles in $X \cup Y \cup Z$ is just $\eta^3|X||Y||Z|$.

If we could partition the vertices to create random-like behavior analogous to the Erdos-Renyi model, we might find a similar lower bound on the number of triangles.

Here is the strategy:

1. Use the Szemeredi Regularity Lemma to partition the graph vertices into subsets, where most of the pairs of subsets have edges between them that "look like" a random bipartite graph. We will later define formally what it means to "look like" a random bipartite graph, through the notion of a *regular* pair.

2. Use $\varepsilon$-farness to find a triplet of subsets that are densely connected in some sense.

3. Prove the Triangle Counting Lemma, which lower bounds the number of traingles in the dense triplet.

# 4 Analysis

The Szemeredi Regularity Lemma requires the definition of edge density and $\gamma$-regularity. For the following definitions, suppose $A, B \subseteq V$ are disjoint vertex sets in graph $G = (V, E)$.

**Definition 8.5** The edge density of the pair $(A, B)$ is defined as

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

$E(A, B)$ denotes the set of edges between $A$ and $B$, i.e. $\{(u, v) \in E \mid u \in A, v \in B\}$.

**Definition 8.6** The pair $(A, B)$ is $\gamma$-regular if and only if

$$\forall A' \subseteq A \text{ and } B' \subseteq B \text{ s.t. } \frac{|A'|}{|A|} \geqslant \gamma \text{ and } \frac{|B'|}{|B|} \geqslant \gamma, |d(A, B) - d(A', B')| < \gamma$$

**Remark 8.7** If a random Erdos-Renyi bipartite graph between vertex sets $A$ and $B$ have Bernoulli($\eta$) edges ($\eta$ is a constant), then with high probability, the graph is $\gamma$-regular for some constant $\gamma$ only depending on $\eta$.

**Lemma 8.8** (Szemeredi Regularity Lemma) *There exists a function $u : \mathbb{R}_+^2 \to \mathbb{N}$ with the following property. For any $l > 0$, $\gamma > 0$, and graph $G = (V, E)$ with $|V| = n \geqslant u(l, \gamma)$, there exists an equipartition of $V$ into $\mathcal{V} = \{V_1, ..., V_k\}$ s.t. $l \leqslant k \leqslant u$ and at most $\gamma\binom{k}{2}$ pairs in $\mathcal{V}$ fail to be $\gamma$-regular.*

This lemma has quite a few parameters. They each serves some purpose. A reasonably small $\gamma$ ensures behavior analogous to the Erdos-Renyi model. Conveniently, we get $\gamma$-regularity for almost all pairs $(V_i, V_j)$ in the partition $\mathcal{V}$.

What is the point of the parameter $\ell$ then? An application of the lemma equipartitions $V$ into at least $l$ different subsets. This turns out to have the effect of upper bounding the number of edges found entirely within a subset, which are not part of the regular/"random-like" behaviour of the graph. To see the upper bound, note that $|V_i| \leqslant \frac{n}{k}$, and so there are at most $\frac{n^2}{k^2}$ edges with both endpoints in $V_i$. Hence the total number of edges that are contained within any subset is upper bounded by $\frac{n^2}{k} \leqslant \frac{n^2}{l}$.

Also note that there's a small caveat if we try to make $l$ too big. Because $|\mathcal{V}|$ can be up to $u(l, \gamma)$, the vertex sets may become too small for the Erdos-Renyi-like bound (which involves the product of the sizes of all these subsets) to be meaningful. Thankfully, $|\mathcal{V}|$ has upper bound $u(l, \gamma)$, which is constant in $n$.

We won't prove the Szemeredi Regularity Lemma in this course. We will however use it to analyze Algorithm 8.2. The following lemma shows that, by the $\varepsilon$-farness of $G$ from being triangle-free, there must exists a triple of not-too-small subsets in $G$ that have pairwise not-too-small density. The consequence then, again, is that these three subsets behave essentially like a random tripartite graph, and we can easily lower bound the number of triangles in this triple (Lemma 8.10: Triangle Counting Lemma) to complete the analysis for Algorithm 8.2.

**Lemma 8.9** *Suppose $G = (V, E)$ is $\varepsilon$-far from triangle-freeness and $|V|$ is sufficiently large. Then there exists disjoint subsets $V_1, V_2, V_3 \subseteq V$ s.t.*

1. $\forall i \in \{1, 2, 3\}$,
$$|V_i| \in \Omega\left(\frac{|V|}{f(\varepsilon)}\right), \text{ where } f(\varepsilon) = u\left(\frac{8}{\varepsilon}, \frac{\varepsilon}{8}\right)$$
2. $(V_1, V_2), (V_2, V_3), (V_3, V_1)$ *are all $\frac{\varepsilon}{8}$-regular pairs with density at least $\frac{\varepsilon}{2}$.*

*Proof.* Suppose $G = (V, E)$ is $\varepsilon$-far from triangle-freeness and $n = |V|$ is sufficiently large. By applying the Szemeredi Regularity Lemma with $l = \frac{8}{\varepsilon}$ and $\gamma = \frac{\varepsilon}{8}$, we can equipartition $V$ into $\mathcal{V} = \{V_1, V_2, ..., V_k\}$ with
$$\frac{\varepsilon}{8} \leqslant k \leqslant u\left(\frac{8}{\varepsilon}, \frac{\varepsilon}{8}\right)$$
such that at most $\frac{\varepsilon}{8}\binom{k}{2}$ pairs in $\mathcal{V}$ aren't $\frac{\varepsilon}{8}$-regular. Because the size of each element in $\mathcal{V}$ satisfies the first requirement in Lemma 8.9, we just need to find three elements of $\mathcal{V}$ that satisfy the second requirement. To find them, it's helpful to focus exclusively on edges between $\frac{\varepsilon}{8}$-regular pairs with density at least $\frac{\varepsilon}{2}$, so we should try to exclude other edges:

- For each $i \in [k]$, the number of edges with endpoints in $V_i$ is upper bounded by
$$|V_i|^2 \leqslant \left(\frac{n}{k}\right)^2$$
Hence the total number of edges with endpoints in the same partition has upper bound
$$\frac{n^2}{k} \leqslant \frac{\varepsilon n^2}{8}$$
- For each $\frac{\varepsilon}{8}$-irregular pair $V_i, V_j$, the number of edges between them has upper bound
$$|V_i| \cdot |V_j| \leqslant \left(\frac{n}{k}\right)^2$$
Hence the total number of edges between $\frac{\varepsilon}{8}$-irregular pairs is upper bounded by
$$\frac{\varepsilon}{8}\binom{k}{2}\frac{n^2}{k^2} \leqslant \frac{\varepsilon n^2}{16}$$
- For each sparse pair $V_i, V_j$, i.e. $d(V_i, V_j) \leqslant \frac{\varepsilon}{2}$, the number of edges between them is upper bounded by
$$\frac{\varepsilon}{2}|V_i| \cdot |V_j| \leqslant \frac{\varepsilon}{2}\left(\frac{n}{k}\right)^2$$
Hence the total number of edges between sparse pairs is upper bounded by
$$\frac{\varepsilon}{2}\binom{k}{2}\frac{n^2}{k^2} \leqslant \frac{\varepsilon n^2}{4}$$

Except the three aforementioned types, (which account for no more than $\frac{\varepsilon n^2}{2}$ edges), all other edges are between $\frac{\varepsilon}{8}$-regular pairs of density at least $\frac{\varepsilon}{2}$.
Because $G$ is $\varepsilon$-far from triangle-freeness, we can still find a triangle-forming triplet $\{x, y, z\} \subseteq V$ using only edges between $\frac{\varepsilon}{8}$-regular pairs of density at least $\frac{\varepsilon}{2}$.
Due to the removal of edges within a single subset in $\mathcal{V}$, the subsets containing $x, y, z$ are distinct. Denote them $X, Y, Z$. The existence of triangle-forming triplet $x, y, z$ guarantees that $X, Y, Z$ are pairwise nonempty in terms of remaining edges. Hence $X, Y, Z$ must be pairwise $\frac{\varepsilon}{8}$-regular and $\frac{\varepsilon}{2}$-dense. □

Lastly, we show the Triangle Counting Lemma, which lower-bounds the number of triangle-forming triplets with a vertex in each special subset we just found with Lemma 8.9. After all, establishing a lower bound on the number of triangle-forming triplets is our ultimate goal.

**Lemma 8.10** (Triangle Counting Lemma) *Suppose $X, Y, Z$ are disjoint vertex sets that are pairwise $\gamma$-regular and at least $2\gamma$-dense for some small $\gamma$. Then there are at least $\Omega(\gamma^3)|X||Y||Z|$ many triplets $(x, y, z) \in X \times Y \times Z$ that form triangles.*

**Remark 8.11** The lower bound on the number of triangle-forming triplets is the same up to constant as what we expect for a random Erdos-Renyi tripartite graph with pairwise density $2\gamma$.

*Proof of Lemma 8.10.* Suppose $X, Y, Z$ are disjoint vertex sets that are pairwise $\gamma$-regular and $2\gamma$-dense for some small $\gamma$.
Let $X_Y = \{x \in X : |N(x) \cap Y| \geqslant \gamma|Y|\}$ and $X_Z = \{x \in X : |N(x) \cap Z| \geqslant \gamma|Z|\}$.

**Claim 8.12** $|X_Y \cap X_Z| \geqslant (1 - 2\gamma)|X|$

*Proof of Claim 8.12.* By the definition of $X_Y$ and $X_Z$,

$$d(X \setminus X_Y, Y) = \frac{|E(X \setminus X_Y, Y)|}{(|X| - |X_Y|)|Y|} \leqslant \gamma \text{ and similarly, } d(X \setminus X_Z, Z) \leqslant \gamma$$

Because $(X, Y)$ and $(X, Z)$ are $2\gamma$-dense pairs, we have $d(X, Y) - d(X \setminus X_Y, Y) \geqslant \gamma$ and $d(X, Z) - d(X \setminus X_Z, Z) \geqslant \gamma$.
Because we can't break the assumption that $X, Y, Z$ are pairwise $\gamma$-regular, we must have $|X \setminus X_Y|, |X \setminus X_Z| < \gamma|X|$. Hence $|X_Y|, |X_Z| \geqslant (1 - \gamma)|X|$, which by union bound implies $|X_Y \cap X_Z| \geqslant (1 - 2\gamma)|X|$. □

Now fix an arbitrary $x \in X_Y \cap X_Z$. By the definition of $N(x) \cap Y$ and $N(x) \cap Z$, $|N(x) \cap Y| \geqslant \gamma|Y|$ and $|N(x) \cap Z| \geqslant \gamma|Z|$.
Furthermore, by the $\gamma$-regularity of $(Y, Z)$, $|d(Y, Z) - d(N(x) \cap Y, N(x) \cap Z)| < \gamma$.
Because $d(Y, Z) \geqslant 2\gamma$, we must have $d(N(x) \cap Y, N(x) \cap Z) \geqslant \gamma$ which implies that there are at least $\gamma^3|Y||Z|$ edges between $N(x) \cap Y$ and $N(x) \cap Z$.
Hence there are at least $(1-2\gamma)\gamma^3|X||Y||Z|$ triangle-forming vertex triplets in $X \times Y \times Z$. □

Now we can actually prove the soundness of Algorithm 8.2.

*Proof of Theorem 8.4.* Suppose $G = (V, E)$ with $|V| = n$ is $\varepsilon$-far from triangle-freeness. By Lemma 8.9 (shown using the Szemeredi Regularity Lemma,) there exist disjoint vertex sets $X, Y, Z \subseteq V$ s.t.

1. $|X|, |Y|, |Z| \in \Omega(\frac{n}{f(\varepsilon)})$
2. $X, Y, Z$ are pairwise $\frac{\varepsilon}{8}$-regular and $\frac{\varepsilon}{2}$-dense.

Now apply Lemma 8.10 with $\gamma = \frac{\varepsilon}{8}$ (noting that $\frac{\varepsilon}{2} > 2\gamma$). We get that there are at least $\Omega(\varepsilon^3)|X||Y||Z|$ triplets $(x, y, z) \in X \times Y \times Z$ that form triangles. Hence the total number of triangles in $G$ is lower bounded by

$$\Omega\left(\frac{\varepsilon^3 n^3}{u^3\left(\frac{8}{\varepsilon}, \frac{\varepsilon}{8}\right)}\right)$$

Therefore, we can run Algorithm 8.2 with $O\left(\frac{1}{\varepsilon^3}u^3(\frac{8}{\varepsilon}, \frac{\varepsilon}{8})\right)$ iterations to achieve soundness. □

**Remark 8.13** Although the query complexity is independent of $n$, $u$ can be up to

$$2^{2^{2^{\cdot^{\cdot^{\cdot^{2}}}}}}$$

with $\Theta(\frac{1}{\varepsilon^2})$ 2s in this tower of exponentials. The humongous query complexity comes from the application of the Szemeredi Regularity Lemma, and this bound is tight for the Regularity Lemma [Fox and Lovasz 2017 `https://arxiv.org/pdf/1606.01230.pdf`].. Whether efficiency can be improved through an analysis circumventing the Szemeredi Regularity Lemma remains an open question. For one sided testers though, it has been shown that the query complexity is at least

$$\Omega\left(\exp\left(\text{polylog}\left(\frac{1}{\varepsilon}\right)\right)\right)$$

so there are no one-sided testers with polynomial query complexity.

Algorithm 8.2 can be extended to general subgraph-freeness testing, such as in [`https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.10056`].