# Triangle Enumeration, Graph Motifs, and Geometric Bounds

*Notes by Kanat Tangwongsan* (kanat.tan@mahidol.edu)

## 1 Background and Motivation

In an undirected graph $G = (V, E)$, a *triangle* is any triplet $\{a, b, c\} \subseteq V$ with edges between the vertices. This aligns with our idea of what a triangle looks like.

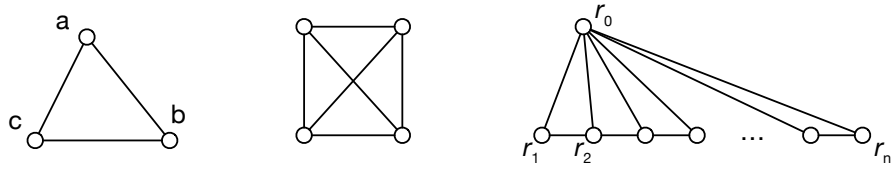


Figure 1: (left) a connected triplet, aka. a triangle; (center) $K_4$; (right) a graph with a high-degree node.

I will tell you in a bit why this problem is of interest to computer science. For now, say you have a big graph and wish to enumerate all triangles in this graph. Because the graph is big, you want this process to be efficient—taking time proportional to the number of triangles you can possibly have on a graph of that size. *How would you list all triangles in a graph?*

One natural idea is to look at all possible triplets, checking to see for each one whether there are edges between the vertices:

**foreach** $\{a, b, c\} \in \binom{V}{3}$ **do**
  **if** $\{a, b\}, \{b, c\}, \{c, a\}$ *are edges* **then** `emitTriangle`$(a, b, c)$ ;
**end**

*How does this algorithm perform?* If the input graph is $K_n$ (the complete graph on $n$ nodes), there are $\binom{n}{3}$ triangles—because if you pick any triplet, there are edges between them and hence that is a triangle. The algorithm above loops through $\binom{n}{3} = O(n^3)$ times, which is the best we can hope for.

Throughout this talk, let $n = |V|$ and $m = |E|$.

However, if we're to use this algorithm on a graph such as the right of Figure 1, the $O(n^3)$ running time is not at all efficient. The number of triangles is only $O(n)$. The following algorithm examines each vertex in turn, treating the high-degree vs. low-degree case differently.

**foreach** $a \in V$ **do**
  **if** $|\mathsf{nbrs}(a)|^2 \leqslant m$ **then**
    **foreach** $\{b, c\} \in \binom{\mathsf{nbrs}(a)}{2}$ **do**
      **if** $\{b, c\}$ *is an edge* **then** `emitTriangle`$(a, b, c)$;
    **end**
  **else**
    **foreach** $\{b, c\} \in E$ **do**
      **if** $\{a, b\}$ *and* $\{a, c\}$ *are edges* **then** `emitTriangle`$(a, b, c)$;
    **end**
  **end**
**end**

This algorithm is simple but handles skewed graphs pretty well. For any vertex $a$, the inner loops promises that the cost will be at most $\min\{\deg^2(a), m\}$. On the example graph, the cost will be

$$\min\{\deg^2(r_0), m\} + \sum_{i=1}^{n} \min\{\deg^2(r_1), m\} = m + \sum_{i=1}^{n} \deg^2(r_1) \leqslant m + 10n = O(m).$$

This is linear! The algorithm also indirectly gives us an upper bound on the total number of triangles:

$$\#\text{ triangles} \leqslant \sum_{a \in V} \min\{\deg^2(a), m\} \leqslant \sum_{a \in V} \sqrt{\deg^2(a) \cdot m} \leqslant \sqrt{m} \sum_{a \in V} \sqrt{\deg^2(a)} = 2m^{3/2}$$

### Why should we care?

For some time now, social graphs (Facebook friends, Twitter followers, etc.) are the largest graphs available on the planet. Computer scientists and sociologists alike are fascinated by them and want to study them. Because friends of friends tend to be friends themselves, triangles emerge as a pattern that they look for in measuring the strength of a community. Over time, people in areas such as graph mining (think: data mining on graphs) have found that there is a "right" amount of triangles per community. Beyond this threshold, it's usually not an organically-created community.

Another motivation came from database query processing. Database joins (very useful in data processing) are deeply connected to graph-motif enumeration.

## 2 How many times can a graph pattern occur?

The main character of this talk is a lemma that has been used time and again:

**Lemma 1 (Shearer)** *Let $X_1, X_2, \ldots, X_n$ be random variables. For each subset $F \subseteq [n]$, let $\boldsymbol{X}_F = (X_i)_{i \in F}$ and $\boldsymbol{X} = \boldsymbol{X}_{[n]}$. Let $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ be a hypergraph and*

$$P = \left\{ \boldsymbol{x} = (x_F)_{F \in \mathcal{E}} \mid \sum_{F : v \in F} x_F \geqslant 1, \forall v \in \mathcal{V}, \boldsymbol{x} \geqslant \boldsymbol{0} \right\}$$

*be a polytope of any fractional edge cover of the hypergraph. Then, for $\boldsymbol{x} \in P$,*

$$H[\boldsymbol{X}] \leqslant \sum_{F \in \mathcal{E}} x_F \cdot H[\boldsymbol{X}_F].$$

We will prove this lemma soon but to do that, we will need a few basic facts about discrete entropy.

### 2.1 An Information Theory Crash Course

**Definition 2 (Shannon Binary Entropy)** *Let $X$ be a discrete random variable. The **Shannon binary entropy function** $H[X]$ is a measure of the degree of uncertainty associated with $X$ and is given by*

$$H[X] \;=\; -\sum_{x} \mathbf{Pr}[X = x] \log_2 \mathbf{Pr}[X = x]\,.$$

A computer science-y interpretation of this definition is, Shannon entropy measures how many bits are needed to indicate what value $X$ takes on. The more uncertain we are, the higher number of bits we need. As a quick example, if we know $X$ is always going to be 42, we don't need any bit to describe it—it is always 42. But if there is a 50% chance it's going to be 42 and a 50% chance it's going to be 12, then you need more bits: $-(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$.

Notice that this definition naturally works on, say, a tuple $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. This is called *joint entropy*.

In the same way that we can define conditional probability, we can define a notion of *conditional entropy* $H[X \mid Y]$. This measures the degree of uncertainty about $X$ given that we know $Y$:

$$H[X \mid Y] = -\sum_{x,y} \mathbf{Pr}[X = y, Y = y] \log_2 \mathbf{Pr}[X = x \mid Y = y] .$$

One might expect then that given more information, there is less uncertainty. This is true and can be shown from first principles:

**Fact 3**

$$H[X \mid Y, Z] \leqslant H[X \mid Y],$$

Similarly, one can show that

$$H[(X_1, X_2)] = H[X_1 \mid X_2] + H(X_2),$$

or more generally

**Fact 4**

$$H[(X_1, X_2, \ldots, X_n)] = \sum_{j=1}^{n} H[X_j \mid X_1, X_2, \ldots, X_{j-1}].$$

It can be shown that the distribution that maximizes the entropy is the uniform distribution. After all, every outcome is equally likely. An easy way to see this is through Jensen's inequality:

**Fact 5** $H[X] \leqslant \log_2(\#\textit{support of } X)$.

## 2.2 Proof of Shearer's Lemma

We present a proof of Shearer's lemma using conditional entropy. The proof is due to Radhakrishnan.

*Proof of Lemma 1:* Let $F \in \mathcal{E}$ and $j \in F$. We will make three observations:

(i) (Fact 3) By knowing more, we are less uncertain. That is,

$$H[X_j \mid X_1, \ldots, X_{j-1}] \leqslant H[X_j \mid X_j \text{ where } i < j \wedge i \in F]$$

(ii) Because $\mathbf{x} \in P$ means $\sum_{F \in \mathcal{E} \text{ s.t } j \in F} x_F \geqslant 1$, we know that

$$\begin{aligned} H[X_j \mid X_1, \ldots, X_{j-1}] &= H[X_j \mid X_j \text{ where } i < j] && (1) \\ &\leqslant \sum_{F \in \mathcal{E} \text{ s.t } j \in F} x_F \cdot H[X_j \mid X_i \text{ where } i < j] && (2) \\ &\leqslant \sum_{F \in \mathcal{E} \text{ s.t } j \in F} x_F \cdot H[X_j \mid X_i \text{ where } i < j \wedge i \in F] && (3) \end{aligned}$$

(iii) Using Fact 4, we have

$$H[\mathbf{X}_F] = \sum_{j \in F} H[X_j \mid X_i \text{ where } i < j \wedge i \in F] \quad (4)$$

We proceed using Fact 4:

$$\begin{aligned}
H[\mathbf{X}] &= \sum_{j=1}^{n} H[X_j \mid X_1, \ldots, X_{j-1}] \\
&= \sum_{j=1}^{n} \sum_{F \in \mathcal{E} \text{ s.t } j \in F} x_F \cdot H[X_j \mid X_i \text{ where } i < j \wedge i \in F] && \text{[because of (3)]} \\
&= \sum_{F \in \mathcal{E}} \sum_{j \in F} x_F \cdot H[X_j \mid X_i \text{ where } i < j \wedge i \in F] && \text{[sum rewriting]} \\
&= \sum_{F \in \mathcal{E}} x_F \sum_{j \in F} H[X_j \mid X_i \text{ where } i < j \wedge i \in F] \\
&= \sum_{F \in \mathcal{E}} x_F \cdot H[\mathbf{X}_F] && \text{[because of (4)]}
\end{aligned}$$

which proves the lemma. ■

## 2.3 Bounding The Number of Triangles and Other Motifs

We will now apply Shearer's lemma to bound the number of triangles and other motifs:

**Lemma 6** *An undirected graph with m edges has at most $O(m^{3/2})$ triangles.*

*Proof:* Consider an undirected graph $G = (V, E)$ with $m = |E|$. Let $\mathcal{T} \subseteq V^3$ be the set of all triangles on $G$. The random process will use is the following: Pick $\mathbf{X} = (v_1, v_2, v_3)$ uniformly at random from $\mathcal{T}$. The entropy of $\mathbf{X}$ is $H(\mathbf{X}) = \log_2 |\mathcal{T}|$. Upon a closer look, a triangle is made up of three sides. Therefore, we can project $\mathbf{X}$ down to three edges as follows:

$$\mathbf{A} = \mathbf{X}_{\{1,2\}}, \quad \mathbf{B} = \mathbf{X}_{\{2,3\}}, \quad \text{and} \quad \mathbf{C} = \mathbf{X}_{\{3,1\}}$$

Each of these random variables is necessary an edge, i.e., $\mathbf{A}, \mathbf{B}, \mathbf{C} \in E$; however, their distributions may not be uniform over $E$. But we know that the entropy of a random variable is maximized when the distribution is uniform, so

$$H(\mathbf{A}) \leqslant \log_2 m, \quad H(\mathbf{B}) \leqslant \log_2 m, \quad \text{and} \quad H(\mathbf{C}) \leqslant \log_2 m$$

To use Shearer's lemma, we'll consider the following hypergraph

$$\mathcal{H} = (\mathcal{V} = \{1, 2, 3\}, \mathcal{E} = \{\{1, 2\}, \{2, 3\}, \{3, 1\}\}).$$

This is exactly the graph pattern we're trying to look for. Notice that every edge is incident on exactly 2 vertices, therefore setting $x_F = 1/2$ uniformly ensures that $\sum_{F \in \mathcal{E}: i \in F} x_F = 1$ for all $i = 1, 2, 3$. Hence, by Shearer's lemma, we have

$$\log_2 |\mathcal{T}| \leqslant \sum_{F \in \mathcal{F}} \mathbf{x}_F \cdot H(\mathbf{X}_F) \leqslant \frac{3}{2} \cdot \log_2 m \implies |\mathcal{T}| \leqslant m^{3/2}.$$

■

Reasoning in the same way, we can prove similar results for larger cliques $K_n$ and beyond. For example, the following can be shown for $K_n$:

**Lemma 7** *An undirected graph with m edges contains at most $O(m^{n/2})$ copies of $K_n$.*

## 2.4 Connections to Geometric Inequalities

In 1949, Loomis and Whitney proved an inequality allowing one to characterize the "size" of a $d$-dimensional set by the sizes of its $(d-1)$-dimensional projections. For the purpose of this talk, I'll state a discrete version of the LW theorem, which will be subsumed by what we are about to discuss.

**Theorem 8 (Discrete Loomis-Whitney (LW))** *Let $S \subset \mathbb{Z}^n$ be a finite set of n-dimensional grid points. For each dimension $i = 1, 2, \ldots, n$, let $S_{-i}$ denote the $(n-1)$-dimensional projection of $S$ onto the coordinates $[n] \backslash \{i\}$. Then,*

$$|S|^{n-1} \leqslant \prod_{i=1}^{n} |S_{-i}|.$$

In the context of discrete geometry, Bollobàs and Thomason later proved a generalization of LW (which clearly subsumes LW). Again, we'll focus on the discrete version:

**Theorem 9 (Discrete Bollobàs and Thomason (BT))** *Let $S \subset \mathbb{Z}^n$ be a finite set of n-dimensional grid points. Let $\mathcal{F}$ be a collection of subsets of $[n]$ in which every $i \in [n]$ occurs in exactly $d$ members of $\mathcal{F}$. Let $S_F$ be the set of projections $\mathbb{Z}^n \to \mathbb{Z}^F$ of points in $S$ onto the coordinates in $F$, specifically $S_F = \{\boldsymbol{x}_F \mid \boldsymbol{x} \in S\}$. Then,*

$$|S| \leqslant \prod_{F \in \mathcal{F}} |S_F|^{1/d}.$$

*Proof:* Consider the following random process. Pick $\mathbf{X}$ uniformly at random from $S$, so $H(\mathbf{X}) \leqslant \log_2 |S|$. Notice that for any $F \in \mathcal{F}$, $\mathbf{X}$ restricted to $F$, denoted by $\mathbf{X}_F$, is also a random variable—but it may not be uniform with respect to $S_F$. But we know that the entropy is maximized when the distribution is uniform, so $H(\mathbf{X}_F) \leqslant \log |S_F|$. To apply Shearer's lemma, we will work with the hypergraph $\mathcal{H} = ([n], \mathcal{F})$. Because every $i \in [n]$ occurs in exactly $d$ members of $F$, setting $x_F = 1/d$ ensures that for every $i \in [n]$, $\sum_{F \in \mathcal{F} : i \in F} x_F = 1$. Therefore, by Shearer's lemma, we have that

$$\log_2 |S| = H(\mathbf{X}) = \sum_{F \in \mathcal{F}} \frac{1}{d} \cdot H[\mathbf{X}_F] \leqslant \sum_{F \in \mathcal{F}} \frac{1}{d} \cdot \log_2 |S_F|,$$

or in other words, $|S| \leqslant \prod_{F \in \mathcal{F}} |S_F|^{1/d}$, proving the theorem. ∎