COMS 6998: Algebraic Techniques in TCS (Fall'21) Sep 21, 2021

Lecture 2: Algebraic graph algorithms

Instructor: *Josh Alman* Scribe notes by: *Xuan Zhao, Shida Jing*

Disclaimer: This draft may be incomplete or have errors. Consult the course webpage for the most up-to-date version.

1 Induced Subgraphs

Recap from last lecture. We consider the following problem: Given an n-node input graph G, does G contain H as an induced subgraph? (see Figure 1)

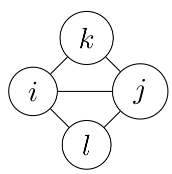


Figure 1: The target induced subgraph H.

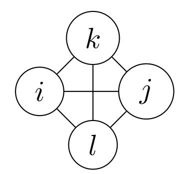


Figure 2: The 4-clique K_4.

In the last lecture we proposed the following approach: Let $A \in \{0,1\}^{n\times n}$ be the adjacency matrix of G:

$$A[i,j] = \begin{cases} 1 & \text{if } (i,j) \in E(G), \\ 0 & \text{otherwise.} \end{cases}$$

The algorithm first computes $A^2 \in \mathbb{R}^{n\times n}$ in $O(n^{2.373})$ time, then computes the value $\sum_{(i,j)\in E(G)} \binom{A^2[i,j]}{2}$ in $O(n^2)$ time. This value counts the number of our target graph in G, but also counts 4-cliques in G:

$$\begin{aligned} \sum_{(i,j)\in E(G)} \binom{A^2[i,j]}{2} &= \sum_{(i,j)\in E(G)} \#\text{ of pairs } (k,l) \text{ s.t. } (i,k),(k,j),(i,l),(l,j) \in E(G) \\ &= (\#\text{ of } H) + 6 \times (\#\text{ of } K_4). \end{aligned}$$

Each copy of K_4 is counted 6 times since it's counted once in the summand corresponding to (i,j) being each of its 6 edges.

A randomized algorithm. Denote $R(G) := \sum_{(i,j)\in E(G)} \binom{A^2[i,j]}{2}$.

- If $R(G)$ is not a multiple of 6, then we conclude that G must contain H as an induced subgraph.
- If $R(G)$ is a nonzero multiple of 6, we use the following randomized algorithm:

  Remove every node of G independently with probability $1/2$ to obtain a random subgraph G'. We will show that if H is an induced subgraph of G, then with probability $\geq 1/16$, # of H in G' is not a multiple of 6 (Lemma 1), and hence $R(G')$ is not a multiple of 6.

We repeat this sampling process for 100 times. If for at least one sampled subgraph G', $R(G')$ is not a multiple of 6, then we conclude that the original graph G contains H as an induced subgraph. The failure probability is $\Pr[\text{failure}] \le (15/16)^{100} < 1/100$.

Now it remains to prove Lemma 1.

Lemma 1 (Correctness of the randomized algorithm)**.** *If the graph G contains at least one induced copy of the subgraph H, then if we remove every node of G independently with probability $1/2$, the resulting graph will have (# of H) $\not\equiv 0 \pmod 6$ with probability at least $1/16$.*

Proof. Let

$$P(x_1, x_2, \cdots, x_n) = \sum_{0<i<j<k<l\le n, \text{ s.t. nodes } i,j,k,l \text{ form } H \text{ in } G} x_i x_j x_k x_l.$$

Then the number of H in G is $P(1, 1, \cdots, 1)$.

Plug into P the vector $x \in \{0,1\}^n$ defined as

$$x_i = \begin{cases} 1 & \text{if we randomly keep node } i, \\ 0 & \text{if we randomly remove node } i. \end{cases}$$

Then $P(x)$ is the number of H in the random subgraph.

It remains to bound the probability that $P(x) \not\equiv 0 \pmod 6$ when x is drawn uniformly at random from $\{0,1\}^n$. Using Lemma 2 (to be proved below), since the degree of P is 4, we have

$$\Pr_{x_1, x_2, \cdots, x_n \sim \{0,1\}}[P(x) \not\equiv 0 \pmod 6] \ge \frac{1}{16}.$$

This finishes the proof. □

The only missing piece in the previous proof is to show that for a uniformly random Boolean vector x, the probability that $P(x) \not\equiv 0 \pmod 6$ is at least $1/16$. We will prove a slightly more general lemma where the degree of P is any integer d, and the modulus is any integer $m > 1$.

Lemma 2. *If $P(x_1, x_2, \cdots, x_n)$ is a multilinear polynomial with integer coefficients which are not all divisible by m, and P has degree at most d, then*

$$\Pr_{x_1, x_2, \cdots, x_n \sim \{0,1\}}[P(x) \not\equiv 0 \pmod m] \ge \frac{1}{2^d}.$$

Before proving Lemma 2, we first prove the following warm-up lemma, which will be used in the proof of Lemma 2.

Lemma 3 (Warm-up lemma)**.** *If $P(x_1, x_2, \cdots, x_n)$ is a multilinear polynomial with integer coefficients which are not all divisible by m, then there exists an $x \in \{0,1\}^n$ such that*

$$P(x) \not\equiv 0 \pmod m.$$

Proof. The multilinear polynomial P can be written in the form

$$P(x_1, x_2, \cdots, x_n) = \sum_{S \subseteq \{1, \cdots, n\}} a_S \cdot \prod_{i \in S} x_i,$$

where the a_S coefficients are integers which are not all divisible by m. Pick a minimum size set $T \subseteq \{1, 2, \cdots, n\}$ such that $a_T \not\equiv 0 \pmod{m}$.
Define $y \in \{0,1\}^n$ by

$$y_i = \begin{cases} 1 & \text{if } i \in T, \\ 0 & \text{otherwise.} \end{cases}$$

Then for any set $S \subseteq \{1, 2, \cdots, n\}$, we have $\prod_{i \in S} y_i = 1$ only if $S \subseteq T$. And since T is the minimum size set where $a_T \not\equiv 0 \pmod{m}$, we have $a_S \equiv 0 \pmod{m}$ for any strict subset $S \subsetneq T$. It follows that

$$P(y) = \sum_{S \subseteq T} a_S = a_T \not\equiv 0 \pmod{m}.$$

□

Now we are ready to prove Lemma 2.

Proof of Lemma 2. Again write the multilinear polynomial P in the following form:

$$P(x_1, x_2, \cdots, x_n) = \sum_{S \subseteq \{1, \cdots, n\}} a_S \cdot \prod_{i \in S} x_i.$$

Pick any set $T \subseteq \{1, 2, \cdots, n\}$ of maximum size such that $a_T \not\equiv 0 \pmod{m}$. Let $d' = |T|$. Since P has degree d, we know that $d' \leq d$. For convenience we re-order the subscripts and denote this set as $T = \{1, 2, \ldots d'\}$.

Consider any $0/1$ assignment to $x_{d'+1}, \ldots, x_n$. Let $Q : \mathbb{Z}^{d'} \to \mathbb{Z}$ be the resulting polynomial from this partial assignment to P, where Q has d' variables $x_1, \ldots, x_{d'}$. Q is nonzero (mod m) since the monomial $a_T \cdot \prod_{i \in T} x_i$ is still in Q. By Lemma 3, there exists an assignment to $x_1, \ldots, x_{d'}$ such that $Q(x_1, \ldots, x_{d'}) \not\equiv 0 \pmod{m}$.

Thus, for any $0/1$ assignment to $x_{d'+1}, \ldots, x_n$, there is at least one $0/1$ assignment to $x_1, \ldots, x_{d'}$ such that $P(x_1, \cdots, x_n) \not\equiv 0 \pmod{m}$. Hence, as desired,

$$\Pr_{x_1, \cdots, x_n \sim \{0,1\}} [P(x_1, \cdots, x_n) \not\equiv 0 \pmod{m}] \geq \frac{1}{2^{d'}} \geq \frac{1}{2^d}.$$

□

2 Polynomial Identity Testing

Suppose we are given an expression for a polynomial and we want to test whether all its terms cancel out resulting in the zero polynomial. For instance, we know that

$$P(x, y) = x^2 - y^2 - (x + y)(x - y)$$

is the zero polynomial, but for larger examples such as

$$Q(x,y) = x^6 - y^6 - (x^2 - y^2)(x^2 + xy + y^2)(x^2 - xy + y^2)$$

or

$$R(x,y,z) = (x+y)^{100} + (x+z)^{100} + (y+z)^{100} - 2(x+y+z)^{100},$$

it may be unclear. Expanding out all the terms of the polynomial by brute force can be slow. An efficient test is instead to evaluate the polynomial on a few random points, and check whether the result is always zero. If the polynomial is zero, then it will always output 0, but otherwise, the next lemma shows that it will output a nonzero value with decent probability.

Lemma 4 (Schwartz-Zippel)**.** *Let $\mathbb{F}$ be any field, and let $S \subseteq \mathbb{F}$ be a finite subset. Let $P(x_1, \cdots, x_n)$ be a nonzero polynomial over $\mathbb{F}$ with degree $\leq d$. Then*

$$\Pr_{x \sim S^n}[P(x) = 0] \leq \frac{d}{|S|}.$$

Proof. We proceed by strong induction on n.

Base case ($n = 1$): Any single-variable polynomial over as field $\mathbb{F}$ with degree $\leq d$ has at most d roots. The probability that a random $x \in S$ is a root of $P(x)$ is hence at most $\frac{d}{|S|}$.

Induction step: Assume the lemma statement is true for $n-1$, and we want to prove it for n. Write the polynomial P in the following form:

$$P(x_1, \cdots, x_n) = \sum_{i=0}^{d} x_n^i \cdot P_i(x_1 \cdots, x_{n-1}),$$

where P_i is a polynomial over $n-1$ variables with degree $\leq d-i$.

Let k be the largest value such that $P_k \neq 0$. If we randomly pick $x_1, \cdots, x_{n-1} \in S$, then by the induction hypothesis,

$$\Pr[P_k(x_1, \cdots, x_{n-1}) = 0] \leq \frac{d-k}{|S|}.$$

Suppose that $x_1, \cdots, x_{n-1}$ satisfy $P_k(x_1, \cdots, x_{n-1}) \neq 0$. Then, after fixing $x_1, \cdots, x_{n-1}$, our polynomial P becomes a nonzero polynomial of degree $\leq k$ in the single variable x_n. As in the base case, it follows that

$$\Pr[P(x_n) = 0] \leq \frac{k}{|S|}.$$

By the union bound, we have

$$\Pr[P(x_1, \cdots, x_n) = 0] \leq \Pr[P_k(x_1, \ldots, x_{n-1}) = 0] + \Pr[P(x_1, \ldots, x_n) = 0 \mid P_k(x_1, \ldots, x_{n-1}) \neq 0],$$

which is at most $\frac{d-k}{|S|} + \frac{k}{|S|} = \frac{d}{|S|}$. $\square$

3 Maximum Matchings in Graphs

We next study the problem of determining the size of a maximum matching in an undirected graph. We will assume for now that we can compute the determinant of a $n \times n$ matrix in $O(n^{2.323})$ operations. This assumption will be discussed in following classes.

Definition 5 (Matching in a graph)**.** *For an undirected graph G with n nodes, a* matching *is a subset S of the edges of the graph G such that every node of G is incident to at most one edge of S. A* perfect matching *is a matching in which each node of G is incident to exactly one edge of S, or equivalently, a matching of size $n/2$.*

There is a simple reduction showing that if one can detect whether a graph has a perfect matching in time $T(n)$, then one can binary search for the maximum size of a matching in a graph in time $O(T(n)\log n)$. (The idea is that, for any k, to test whether G has a matching of size at least $(n-k)/2$, one can add k nodes to G which are adjacent to each other and every other node in G, and test whether the resulting graph has a perfect matching.)

Our goal is to find if a graph has a perfect matching in $O(n^{2.373})$ time, and we will use polynomial identity testing as discussed above. Our plan is to define a polynomial P such that

1. $P(x) \neq 0$ if and only if G has a perfect matching, and

2. $P(x)$ is easy to evaluate.

We will define our $P(x)$ as the determinant of the Tutte matrix of G.

Definition 6 (Tutte matrix)**.** *Let G be a graph with n nodes. Let $x_{i,j}$ where $i,j \in \{1,2,\cdots,n\}$ be n^2 variables. The Tutte matrix M is an $n \times n$ matrix:*

$$M[i,j] = \begin{cases} 0 & \text{if } (i,j) \notin E(G), \\ x_{i,j} & \text{if } (i,j) \in E(G) \text{ and } i<j, \\ -x_{j,i} & \text{if } (i,j) \in E(G) \text{ and } j<i. \end{cases}$$

The determinant of M, $\det(M)$, is the polynomial we are looking for with degree $\leq n$. We are going to prove that $\det(M) \neq 0$ if and only if G has a perfect matching.

Assuming this claim is true, our algorithm works as follows: pick a finite field $|\,\mathbb{F}_q\,| > 2n$, then evaluate $\det(M)$ on random points in $\mathbb{F}_q$. By the Schwartz-Zippel Lemma (Lemma 4), after trying a large constant number of random points, we learn with high probability whether $\det(M)$ is a zero polynomial.

Theorem 7. $\det(M) \neq 0$ *if and only if G has a perfect matching.*

Proof. Recall the determinant of the matrix M is given by

$$\det(M) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot \prod_{i=1}^{n} M[i,\sigma[i]],$$

where S_n is the set of all the permutations on $\{1,\cdots,n\}$. To make our discussion easier, we will use f_σ to denote $\prod_{i=1}^{n} M[i,\sigma[i]]$. So we have

$$\det(M) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) f_\sigma.$$

So what is f_σ anyways? Think about f_σ in terms of a "cycle cover" of G. A cycle cover of a graph is a cover where all the edges (directed) belong to a cycle or a loop. If you draw all the directed edges out in terms of the permutation σ (all edges have the form $(i, \sigma(i))$ with the arrow pointing towards $\sigma(i)$), then $f_\sigma = 0$ unless all of these edges are in G. This is because if one of the edges you draw is not in G, then the entry in M corresponding to your edge would have evaluated to be 0, based on how M is defined.

It is not hard to see that if σ has a fixed point ($\sigma(i) = i$), then f_σ would be 0. Because if σ has a fixed point, when you draw out the edges defined by σ, you'll get a self-loop, which is impossible to be in G.

Now, define $O_n \subseteq S_n$ such that O_n are permutations that have an odd cycle.

Permutations with odd cycles. We claim that

$$\sum_{\sigma \in O_n} \operatorname{sgn}(\sigma) f_\sigma = 0.$$

In other words, the terms that corresponds to the permutations in O_n will cancel each other out in the calculation of the determinant. Next we prove this claim.

Fix a $\sigma \in O_n$. If there is a 1-cycle (a fixed point) in σ, then $f_\sigma = 0$, as discussed above. If there is no 1-cycle in σ, then we can pick the odd cycle in σ that contains the smallest index. Let σ' be σ but with that cycle reversed. (For instance, if σ mapped $1 \to 2 \to 3 \to 1$, then σ' instead maps $1 \leftarrow 2 \leftarrow 3 \leftarrow 1$, but is equal to σ on all other inputs.)

Since the permutations σ' and σ only differ in an odd cycle, by the definitions of the sign of permutations, we have $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma')$. It is not hard to see that $f_\sigma = -f_{\sigma'}$ due to the fact that M satisfies $M^T = -M$: We reversed an odd cycle, so that an odd number of signs are reversed. Thus we have

$$\operatorname{sgn}(\sigma) f_\sigma + \operatorname{sgn}(\sigma') f_{\sigma'} = 0.$$

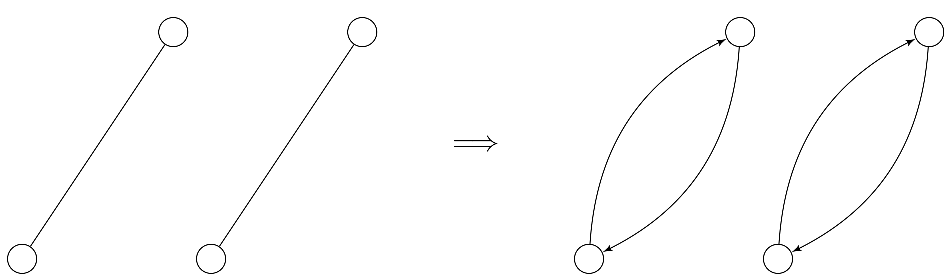
Therefore, the permutations in O_n cancel out in pairs, so $\sum_{\sigma \in O_n} \operatorname{sgn}(\sigma) f_\sigma = 0$.

Permutations without odd cycles. Now we will show that for all permutations that are not in O_n, we have

$$\sum_{\sigma \in S_n \setminus O_n} \operatorname{sgn}(\sigma) f_\sigma \neq 0 \text{ if and only if } G \text{ has a perfect matching.}$$

For the forward direction, if the formula is not equal to zero, then we have at least one σ with $f_\sigma \neq 0$. This means there is an even cycle cover of G, which in turn implies a perfect matching. Indeed, if we pick every other edge in each even cycle, or just the single edge in the case of a cycle of length 2, the result will be a perfect matching for G.

For the backward direction, if G has a perfect matching, then each node is connected by exactly one edge in the matching. We define a permutation $\sigma \in S_n$ by turning each edge into a cycle formed by two edges, as shown below:



By doing this, we convert a perfect matching into an even-cycle cover σ. Therefore:

$$\begin{aligned}
f_\sigma &= \prod_{i=1}^{n} M[i,\sigma(i)] \\
&= \prod_{(i,j)\ \in \text{ matching}} M[i,j]\cdot M[j,i], \quad \text{because all edges are part of a 2-cycle} \\
&= \prod_{(i,j)\ \in \text{ matching}} -x_{ij}^2.
\end{aligned}$$

There is no other $\sigma' \in S_n$ such that f_σ and $f_{\sigma'}$ use the same set of variables since σ uses all the copies of every variable it uses in M. Therefore, $\sum_{\sigma\in S_n\setminus O_n} \operatorname{sgn}(\sigma) f_\sigma \neq 0$, since the term corresponding to our σ can't get canceled out. This completes the proof. $\square$