Conditional Lower Bound for Subgraph Isomorphism with a Tree Pattern*

Robert Krauthgamer† Ohad Trabelsi‡
Weizmann Institute of Science

August 28, 2017

Abstract

The $k$Tree problem is a special case of Subgraph Isomorphism where the pattern graph is a tree, that is, the input is an $n$-node graph $G$ and a $k$-node tree $T$, and the goal is to determine whether $G$ has a subgraph isomorphic to $T$. We provide evidence that this problem cannot be computed significantly faster (i.e., by an exponential factor) than $O(2^k \text{poly}(n))$, which matches the fastest algorithm known for this problem [Koutis and Williams, ICALP 2009 and TALG 2016]. Specifically, we show that if $k$Tree can be solved in time $O(2^{(1-\varepsilon)k} \text{poly}(n))$ for some constant $\varepsilon > 0$, then Set Cover with $n'$ elements and $m'$ sets can be solved in time $2^{(1-\delta)\varepsilon n'} \text{poly}(m')$ for a constant $\delta(\varepsilon) > 0$, refuting the Set Cover Conjecture [Cygan et al., CCC 2012 and TALG 2016].

1 Introduction

The Subgraph Isomorphism problem was studied extensively in theoretical computer science. The most basic version of it asks whether a host graph $G$ contains a copy of a pattern graph $H$ as a subgraph. It is well known to be NP-hard since it generalizes hard problems such as Maximum Clique and Hamiltonicity [Kar72], but unlike many natural NP-hard problems, it requires $N^{\Omega(N)}$ time where $N = |V(G)| + |V(H)|$ is the total number of vertices, unless the exponential time hypothesis (ETH) fails [CFG+16]. Hence, most past research addressed its special cases that are in P, including the case where the pattern graph is of constant size [MP14], or when both graphs are trees [AVY15], biconnected outerplanar graphs [Lin89], two-connected series-parallel graphs [LP09], and more [DLP00, MT92].

We will focus on the version where the pattern is a tree $T$ on $k$ nodes, and the goal is to decide whether $G$ contains a copy of $T$ as a subgraph. For this special case, called $k$Tree, a couple of different techniques were used in order to design algorithms. The color-coding method, designed by Alon, Yuster, and Zwick [AYZ95], yields an algorithm with runtime $O^*((2e)^k)$, where throughout, $O^*(\cdot)$ hides polynomial factors in the instance size. Later, a new method that was developed utilizes $k$MLD (stands for $k$ Multilinear Monomial Detection – the problem of detecting multilinear monomials of degree $k$ in polynomials presented as circuits) to create a $k$Tree algorithm with runtime $O^*(2^k)$ [KW16]. We show that this runtime is actually optimal (up to exponential improvements)

---

*This work was partially supported by the Israel Science Foundation grant #897/13 and by a Minerva Foundation grant.

†Email: robert.krauthgamer@weizmann.ac.il
‡Email: ohad.trabelsi@weizmann.ac.il
Based on the Set Cover Conjecture, introduced by [CDL+16]. In the Set Cover problem, the input is a ground set \( \{1, \ldots, n\} \) and a collection of \( m \) sets, and the goal is to find the smallest sub-collection of sets whose union is the entire ground set. The Set Cover Conjecture implies that it cannot be solved significantly faster than \( O^*(2^n) \). The following is a formal statement of our main theorem.

**Theorem 1.1.** If for some fixed \( \varepsilon > 0 \), \( k\text{Tree} \) can be solved in time \( O^*(2^{(1-\varepsilon)k}) \), then for some \( \delta(\varepsilon) > 0 \), Set Cover on \( n \) elements and \( m \) sets can be solved in time \( O^*(2^{(1-\delta)\varepsilon m}) \).

In spite of extensive effort, the fastest algorithm for Set Cover is still essentially the folklore dynamic programming algorithm that runs in time \( O^*(2^n) \), with several improvements in special cases [Koi09, BHK09, Ned16, BHPK17]. The Set Cover Conjecture states that for every fixed \( \varepsilon > 0 \) there is an integer \( t(\varepsilon) > 0 \) such that Set Cover with sets of size at most \( t \) cannot be computed in time \( O^*(2^{(1-\varepsilon)t}) \). It clearly implies that for every fixed \( \varepsilon > 0 \), Set Cover cannot be solved in time \( O^*(2^{(1-\varepsilon)n}) \).

Several conditional lower bounds were based on this conjecture in the recent decade, including for Set Partitioning, Connected Vertex Cover, Steiner Tree, Subset Sum [CDL+16] (though for the last problem, it was later proved assuming instead Strong ETH (SETH) [ABHS17]), Maximum Graph Motif [BKK16], parity of the number of solutions to Set Cover with at most \( t \) sets [BHH15], Colorful Path (interestingly, this problem is a sub-routine in the aforementioned color-coding method, however, this lower bound says nothing about the \( k\text{Tree} \) problem) and Colorful Cycle [KL16], and the dynamic, general and connected versions of Dominating Set [KST17].

Note that our conditional lower bound is for the undirected version of \( k\text{Tree} \). The directed version of \( k\text{Tree} \) is defined similar to the undirected version, except that \( G \) is a directed graph, and \( T \) is a directed tree, that is the undirected version of \( T \) is a tree. This directed version of \( k\text{Tree} \) can only be harder - even when the directed tree is an arborescence, as one can reduce the undirected version to it with only a polynomial loss, as follows. Define the host graph \( G' \) to be \( G \) with edges in both directions, and direct the edges in \( T \) away from an arbitrary vertex \( v \in T \) to create the directed tree \( T' \), which is thus an arborescence. Clearly, the directed instance is a yes-instance if and only if the undirected instance also is.

**Prior Work.** As mentioned earlier, for the general version of Subgraph Isomorphism, no algorithm can decide whether a host \( N \)-vertex graph \( G \) contains a subgraph isomorphic to a pattern \( N \)-vertex graph \( H \) in time \( N^{o(N)} \), unless the ETH fails [CFG+16]. For the version where both the host and the pattern graphs are rooted trees of size \( N \), a tight lower bound of \( N^{2-o(1)} \) was proved [ABH+16] based on the Orthogonal Vectors Conjecture (and consequently on SETH). The \( k\text{Tree} \) problem cannot be solved in time \( 2^{o(k)} \) assuming the ETH, because directed Hamiltonicity is a special case of this problem and there is a simple reduction (with polynomial blowup) from 3SAT. If we care about the exact exponent in the running time, only a restricted lower bound is known. Williams and Koutis [KW16] used communication complexity to show that a faster algorithm for their intermediate problem \( k\text{MLD} \) in some settings is not possible, and so among a specific class of algorithms, their \( O^*(2^k) \) algorithm for \( k\text{Tree} \) is optimal.

## 2 Reduction to \( k\text{Tree} \)

In this section we prove Theorem 1.1. In order to make the proof simpler, we will have a couple of assumptions regarding the Set Cover instance. First, for a constant \( g > 0 \) to be determined later, we can assume that all the sets in the Set Cover instance are of size at most \( n/g^2 \), and that the
optimal solution has at least $g$, as these cases can already be solved significantly faster than $O^*(2^n)$, proving the theorem for them in a degenerate manner. We formalize it as follows.

**Assumption 2.1.** *All the sets in the Set Cover instance are of size at most $n/g^2$.***

To justify this assumption, notice that if some optimal solution for the Set Cover instance contains a set of size at least $n/g^2$, we can find such optimal solution by simply guessing one set of at least this size (using exhaustive search over at most $m$ choices) and then applying the known dynamic programming algorithm on the still uncovered elements (at most $n - n/g^2$ of them), and return the optimal solution in total time $O^*(2^{(1-1/g^2) n})$.

**Assumption 2.2.** *No solution has size less than $g$.***

The reason that this assumption can be made is that if some optimal solution for the Set Cover instance contains at most $g - 1$ sets, then it is easy to find it in polynomial time, as the number of possibilities is $O(m^g)$, and $g$ is a constant. We continue to the following lemma, which is the heart of the proof.

**Lemma 2.3.** *For every fixed $\varepsilon > 0$, Set Cover on a ground set $N = [n]$ and a collection $M$ of $m$ sets that satisfies assumptions 2.1 and 2.2, can be reduced to $2^{O(\sqrt{n})}$ instances of $k$Tree with $k = (1 + \varepsilon) n + O(1)$.***

**Proof of Theorem 1.1.** Assuming $k$Tree can be solved in time $O^*(2^{(1-\varepsilon) k})$, we reduce the Set Cover instance by applying Lemma 2.3 with $\varepsilon = \varepsilon'/2$, and then solve each of the $2^{\varepsilon n \sqrt{n}}$ instances of $k$Tree in the assumed time of $O^*(2^{(1-\varepsilon')(1+\varepsilon)n+c_2})$, where $c_1, c_2 > 0$ are the constants implicit in the terms $2^{O(\sqrt{n})}$ and $O(1)$ in the lemma, respectively. The total runtime is $O^*(2^{(1-\varepsilon')(1+\varepsilon)n+c_1 \sqrt{n}}) = O^*(2^{(1-\varepsilon'/2-\varepsilon^2/2)n+c_1 \sqrt{n}}) \leq O^*(2^{(1-\varepsilon'/2)n})$, which concludes the proof for $\delta(\varepsilon') = \varepsilon'/2$. \(\square\)

To outline the proof of Lemma 2.3, we will need the following definition. For an integer $a > 0$, let $p(a)$ be the set of all unordered partitions of $a$, where a partition of $a$ is a way of writing $a$ as a sum of positive integers, and unordered means that the order of the summands is insignificant. The asymptotic behaviour of $|p(a)|$ (as $a$ tends to infinity) is known [HR18] to be

$$e^{\pi \sqrt{2a/3}} / (4a^{\sqrt{3}}) = 2^{O(\sqrt{n})}.$$  

It is possible to enumerate all the partitions of $a$ with constant delay between two consecutive partitions, exclusive of the output [NW78, Chapter 9].

Now the intuition for our reduction of Set Cover to $k$Tree is that we first guess a partition of $n$ (the number of elements) that represents how an optimal solution covers the elements as follows — associate each element arbitrarily with one of the sets that contain it (so in effect, we assume each element is covered only once) and count how many elements are covered by each set in the optimal solution. This guessing is done by exhaustive search over $p(n) \leq 2^{O(\sqrt{n})}$ partitions of $n$. Then, we represent the Set Cover instance using a Subgraph Isomorphism instance, whose pattern tree $T$ succinctly reflects the guessed partition of $n$. The idea is that the tree is isomorphic to a subgraph of the Set Cover graph if and only if the Set Cover instance has a solution that agrees with our guess.

**Proof of Lemma 2.3.** Given a Set Cover instance on $n$ elements $N = \{n_i : i \in [n]\}$ and $m$ sets $M = \{S_i\}_{i \in [m]}$ and an $\varepsilon > 0$, we construct $2^{O(\sqrt{n})}$ instances of $k$Tree as follows. For a constant $g(\varepsilon)$ to be determined later, the host graph $G_g = (V_g, E_g)$ is the same for all the instances, and is built on the bipartite graph representation of the Set Cover instance, with some additions. This
is done in a way that a constructed tree will fit in \(G_g\) if and only if the Set Cover instance has a solution that corresponds to the structure of the tree, as follows (see Figure 1). The set of nodes is
\[V_g = N \cup M \cup M_g \cup R \cup \{r_g, r_1, r_2, r\},\]
where \(M_g = \{X \subseteq M : |X| = g\}\) and \(R = \{v_j^i : i \in [4], j \in \lfloor n/(g/2)\rfloor\}\). Intuitively, the role of \(M_g\) is to keep the size of the trees small by representing multiple vertices in \(M\) (multiple sets in Set Cover) at once, and the role of \(R\) and \(\{r_g, r_1, r_2, r\}\) is to enforce that the trees we construct will fit only in certain ways.

The set of edges is constructed as follows. Edges between \(N\) and \(M\) are the usual bipartite graph representation of Set Cover (i.e., connect vertices \(n_j \in N\) and \(S_i \in M\) whenever \(n_j \in S_i\)). We also connect vertex \(X \in M_g\) to vertex \(n_j \in N\) if at least one of the sets in \(X\) contains \(n_j\). Additionally, we add edges between \(r_g\) and every vertex in \(M_g\), and \(v_j^i \in R\) for \(j \in \lfloor n/(g/2)\rfloor\), between \(r_i\) and \(v_j^i\) for every \(i \in \{1, 2\}\) and \(j \in \lfloor n/(g/2)\rfloor\), and finally between \(r\) and every vertex \(v \in \{r_g, r_1, r_2\}\), \(S_i \in M\), and \(v_j^3 \in R\) for \(j \in \lfloor n/(g/2)\rfloor\).

The set
\[
\alpha \leftarrow \left(\sum_{i=1}^{g} p_i, \sum_{i=g+1}^{2g} p_i, \ldots, \sum_{i=(g-1)\cdot\lfloor l/g \rfloor+1}^{g\lfloor l/g \rfloor} p_i, p_{g\lfloor l/g \rfloor+1}, \ldots, p_l\right)
\]

Next, we construct \(2^{O(\sqrt{n})}\) trees such that identifying those that are isomorphic to a subgraph of \(G_g\) will determine the optimum of the Set Cover instance.

For every partition \(\alpha = (p_1, p_2, \ldots, p_l) \in p(n)\) (with possible repetitions) where \(p(n)\) is as defined above, we construct a tree \(T^\alpha_g = (V^\alpha_g, E^\alpha_g)\). This tree has the same set of edges and vertices as \(G_g\), except for the vertices in \(M \cup M_g\) and the edges incident to them, which we replace by a set of new vertices \(M^\alpha \cup M^g_g\), and connect these new vertices to the rest in a way that the resulting graph is a tree. In more detail, \(V^\alpha_g = N' \cup M^\alpha \cup M^g_g \cup R' \cup \{r_g', r_1', r_2', r'\}\) where \(N', R', r_g', r_1', r_2', r'\) are tagged copies of the originals, and \(M^\alpha, M^g_g\) are initialized to be \(\emptyset\).

We define \(\alpha_g\) to be a partition of \(n\) which is also a shrinked representation of \(\alpha\) by partitioning \(\alpha\) into sums of \(g\) numbers for a total of \(\lfloor l/g \rfloor\) such sums, and a remaining of less than \(g\) numbers. Formally,
\[
\alpha_g = \left(\sum_{i=1}^{g} p_i, \sum_{i=g+1}^{2g} p_i, \ldots, \sum_{i=(g-1)\cdot\lfloor l/g \rfloor+1}^{g\lfloor l/g \rfloor} p_i, p_{g\lfloor l/g \rfloor+1}, \ldots, p_l\right)
\]

Note that all the numbers in \(\alpha_g\) are a sum of \(g\) numbers in \(\alpha\), except (maybe) for the last \(g' := \lfloor l/g \rfloor\).
$l - g(l/g) < g$ numbers in $\alpha_g$, a (multi)set which we denote $s(\alpha_g)$. For every $i \in \alpha_g$ (with possible repetitions) we add a star on $i + 1$ vertices to the constructed tree $T^\alpha_g$. If $i \in \alpha_g \setminus s(\alpha_g)$, we add the center vertex to $M^\alpha_g$, connect it to $r_g'$, and add the rest $i$ vertices to $N'$. Else, if $i \in s(\alpha_g)$ we add the center vertex to $M^\alpha_g$, connect it to $r'$, and again add the rest $i$ vertices to $N'$. We return the minimum cardinality of $\alpha$ for which $(G_g, T^\alpha_g)$ is a yes-instance. To see that this construction is small enough, note that the size of $G_g$ is at most $4 + 4 \cdot n/(g/2) + m^g + m + n$ which is polynomial in $m$, and the size of the tree $T^\alpha_g$ is at most

$$4 + 4 \cdot n/(g/2) + n/g + g + n = n \cdot (1 + 9/g) + O(1) = n \cdot (1 + \varepsilon) + O(1)$$

where the last equality holds for $g = 9/\varepsilon$, and so the size constraint follows.

We now prove that at least one of the trees $T^\alpha_g$ returns yes and satisfies $|\alpha| \leq d$, if and only if the Set Cover instance has a solution of size at most $d$. For the first direction, assume that the Set Cover instance has a solution $I$ with $|I| \leq d$. Consider a partition $\alpha_I \in p(n)$ of $n$ that corresponds to $I$ in the following way. Associate every element with exactly one of the sets in $I$, to $I$ with $|I| \leq d$. Consider a partition $\alpha_I \in p(n)$ of $n$ that corresponds to $I$ in the following way. Associate every element with exactly one of the sets in $I$, that contains it, and then consider the list of sizes of the sets in $I$ according to this association (eliminating zeroes). Clearly, $(G_g, T^\alpha_g)$ is a yes-instance and so we will return a number that is at most $|I|$.

For the second direction, assume that every solution to the Set Cover instance is of size at least $d + 1$. We need to prove that for every tree $T^\alpha_g$ with $|\alpha| \leq d$, $(G_g, T^\alpha_g)$ is a no-instance. Assume for the contrary that there exists such $\alpha$ for which $(G_g, T^\alpha_g)$ is a yes-instance with the isomorphism function $f$ from $T^\alpha_g$ to $G_g$. We will show that the only way $f$ is feasible is if $f(r') = r$, $f(M^\alpha_g) \subseteq M$, $f(M^\alpha_g) \subseteq M_g$, and also $f(N') = N$, which together allows us to extract a corresponding solution for the Set Cover instance, leading to a contradiction. We start with the vertex $r' \in T^\alpha_g$. Since its degree is at least $n/(g/2) + 3$ and by Assumption 2.1 and the construction of $G_g$, it holds that $f(r') \notin \{r_1, r_2\} \cup R \cup M \cup M_g$. Moreover, if it was the case that $f(r') \in \{r_g\} \cup N$ then $\{f(r'_1), f(r'_2)\} \cap (M \cup M_g) \neq \emptyset$, however, the degree of $r'_1$ and $r'_2$ in $T^\alpha_g$ is $n/(g/2)$, and the degree of the vertices in $M \cup M_g$ in $G_g$ is at most $g \cdot n/g^2 = n/g$, so it must be that $f(r) = r$. Our next claim is that $f(r'_g) = r_g$. Observe that Assumption 2.2 implies that $M^\alpha_g \neq \emptyset$, and so $r'_g$ in the tree has vertices in distance 2 from it and away from $r'$, a structural constraint that cannot be satisfied by any vertex in $\{r_1, r_2\} \cup R$. Furthermore, the degree of $r'_g$ is at least $n/(g/2)$ and so again by Assumption 2.1 it is also impossible that $f(r'_g) \in M^\alpha_g$, and hence it must be that $f(r'_g) = r_g$. Finally, by the same Assumption and the degrees of $r_1$ and $r_2$, $f(r'_1)$ and $f(r'_2)$ must be in $\{r_1, r_2\}$. Altogether, it must be that $f(M^\alpha_g) \subseteq M_g$, $f(M^\alpha_g) \subseteq M$ and that $f(N') = N$, and therefore we can extract a feasible solution to the Set Cover instance that has at most $d$ sets in it, which is a contradiction.

\[\square\]

References

[ABH16] A. Abboud, A. Backurs, T. D. Hansen, V. V. Williams, and O. Zamir. Subtree isomorphism revisited. In Proceedings of the Twenty-seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA ’16, pages 1256–1271. SIAM, 2016. doi:10.1137/1.9781611974331.ch88.

[ABHS17] A. Abboud, K. Bringmann, D. Hermelin, and D. Shabtay. SETH-based lower bounds for subset sum and bicriteria path. CoRR, 2017. Available from: http://arxiv.org/abs/1704.04546.

[AVY15] A. Abboud, V. Vassilevska-Williams, and H. Yu. Matching triangles and basing hardness on an extremely popular conjecture. In Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing, STOC ’15, pages 41–50. ACM, 2015. doi:10.1145/2746539.2746594.

[AYZ95] N. Alon, R. Yuster, and U. Zwick. Color-coding. J. ACM, 42(4):844–856, July 1995. doi:10.1145/210332.210337.
[Ned16] J. Nederlof. Finding large set covers faster via the representation method. In 24th Annual European Symposium on Algorithms (ESA 2016), volume 57 of Leibniz International Proceedings in Informatics (LIPIcs), pages 69:1–69:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2016. doi:10.4230/LIPIcs.ESA.2016.69.

[NW78] A. Nijenhuis and H. S. Will. Combinatorial Algorithms: For Computers and Hard Calculators. Academic Press, Inc., 2nd edition, 1978.