Maximum Motif Problem in Vertex-Colored Graphs

Riccardo Dondi, Guillaume Fertin, Stéphane Vialette

Maximum Motif Problem in Vertex-Colored Graphs *

Riccardo Dondi[1], Guillaume Fertin[2], and Stéphane Vialette[3]

[1] Dipartimento di Scienze dei Linguaggi, della Comunicazione e degli Studi Culturali
Università degli Studi di Bergamo, Piazza Vecchia 8, 24129 Bergamo - Italy
riccardo.dondi@unimib.it
[2] Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR CNRS 6241
Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 3 - France
guillaume.fertin@univ-nantes.fr
[3] IGM-LabInfo, CNRS UMR 8049, Université Paris-Est,
5 Bd Descartes 77454 Marne-la-Vallée, France
vialette@univ-mlv.fr

Abstract. Searching for motifs in graphs has become a crucial problem in the analysis of biological networks. In this context, different graph motif problems have been considered [12, 6, 4]. Pursuing a line of research pioneered by Lacroix *et al.* [12], we introduce in this paper a new graph motif problem: given a vertex colored graph G and a motif $\mathcal{M}$, where a motif is a multiset of colors, find a maximum cardinality submotif $\mathcal{M}' \subseteq \mathcal{M}$ that occurs as a connected motif in G. We prove that the problem is **APX**-hard even in the case where the target graph is a tree of maximum degree 3, the motif is actually a set and each color occurs at most twice in the tree. Next, we strengthen this result by proving that the problem is not approximable within factor $2^{\log^{\delta} n}$ unless $\mathbf{NP} \subseteq \mathbf{DTIME}(2^{\text{poly}\log n})$. We complement these results by presenting two fixed-parameter algorithms for the problem, where the parameter is the size of the solution. Finally, we give exact efficient exponential-time algorithms for the problem.

1 Introduction

Searching for motifs in graphs has become a crucial problem in the analysis of biological networks (*e.g.* protein-protein interaction, regulatory and metabolic networks). Roughly speaking, there exist two different views of graph motifs. Topological motifs (patterns occurring in the network) are the classical view [10, 16, 17, 15, 11] and computationally reduce to graph isomorphism, in the broad meaning of that term. These motifs have recently been identified as basic modules of molecular information processing. By way of contrast, functional motifs, introduced recently by Lacroix *et al.* [12], do not rely on the key concept of topology conservation but focus on connectedness of the network vertices sought. This latter approach has been considered in subsequent papers [6, 4, 2]. Formally, searching for a functional motif reduces to the following graph problem (referred hereafter as Graph Motif) [12]: Given a target vertex-colored graph $G = (V, E)$ and a multiset of colors $\mathcal{M}$ of size k, find a subset $V' \subseteq V$, $|V'| = k$ $(= |\mathcal{M}|)$ such that (i) the vertex induced subgraph $G[V']$ is connected and (ii) there exists a color-preserving bijective mapping from $\mathcal{M}$ to V'.

Graph Motif is **NP**-complete even if G is a tree with maximum degree 3 and $\mathcal{M}$ is actually a set [6]. **NP**-completeness has also been showed in case G is a bipartite graph with maximum degree 4 and $\mathcal{M}$ is built over two colors only [6]. The seemingly intractability of Graph Motif has naturally led to parameterized complexity considerations [5]. Graph Motif can be solved in $\mathcal{O}(4.32^k\, k^2\, m)$ randomized time [2], where m is the number of edges in G, and in $\mathcal{O}(n^{2c\omega+2})$ time [6], where ω is the tree-width of G and c is the number of distinct colors in $\mathcal{M}$. When the number

* Supported by the Italian-French PAI Galileo Project 08484VH

of distinct colors in the motif is taken as a parameter, Graph Motif is, however, **W**[**1**]-hard even in case G is a tree.

Aiming at accurate models, several variants of Graph Motif have been considered. Dondi *et al.* [4] introduced the problem of minimizing the number of connected components in $G[V']$, *i.e.*, finding an occurrence of $\mathcal{M}$ in G that results in as few connected components as possible. This problem was referred as Min-CC. It turns out that Min-CC is **APX**-hard even in the extremal case where the motif is a set and the target graph is a path and is not approximable within ratio $c \log n$ for some constant $c > 0$, where n is the order of the target graph. From a parameterized point of view, Min-CC is fixed-parameter tractable when the parameter is the size of the motif but becomes **W**[**2**]-hard when the parameter is the number of connected components in the occurrence of the motif (the problem is, however, only known to be **W**[**1**]-hard for paths [2]). Betzler *et al.* [2] replaced connectedness demand by more robust requirements, and proved the problem of finding a biconnected occurrence of $\mathcal{M}$ in G to be **W**[**1**]-complete when the parameter is the size of the motif. This result is important as it sheds light on the fact that a seemingly small step towards motif topology results in parameterized intractability. In this paper, we consider the Maximum Motif problem which is a natural dual variant of Graph Motif. This problem is concerned with finding a maximum cardinality submotif $\mathcal{M}' \subseteq \mathcal{M}$ that occurs as a connected motif in G. Notice that the problem is an optimization problem whereas Graph Motif is a pure decision problem.

This paper is organized as follows. We recall basic definitions in Section 2. In Section 3, we present inapproximability results for Maximum Motif. In Section 4, we present two exact exponential algorithms for Maximum Motif, when the target graph is a tree. In Section 5, we give two fixed-parameter algorithms, parameterized by the size of the solution, when the target graph is a tree and a general graph. Due to space constraint, most proofs are omitted.

2 Preliminaries

We assume readers have basic knowledge about graph theory [3] and we shall only recall basic notations. Let $G = (V, E)$ be a graph. For any $V' \subseteq V$, we denote by $G[V']$ the *subgraph of G induced by V'*, that is $G[V'] = (V', E')$ and $\{u, v\} \in E'$ if and only if $u, v \in V'$ and $\{u, v\} \in \mathbf{E}(G)$. Let $v \in V$, we denote by $N(v)$, the set of vertices $u \in V$ such that $\{u, v\} \in E$. Let $V' \subseteq V$, we denote by $N(V')$, the set of vertices $u \in (V \setminus V')$ such that $\{u, v\} \in E$, for some $v \in V'$. A *coloring* of G is a mapping $\lambda : V \to \mathcal{C}$, where $\mathcal{C}$ is a set of colors. For any subset V' of V, we let $\mathcal{C}(V')$ stand for the multiset of colors assigned to the vertices in V'. A motif $\mathcal{M}$ is a multiset of colors built over a set of colors $\mathcal{C}$. In case $\mathcal{M}$ is actually a set, we call it a *colorful motif*. An *occurrence* of $\mathcal{M}$ in G is a subset $V' \subseteq V$ such that (i) $G[V']$ is connected, and (ii) $\mathcal{C}(V') = \mathcal{M}$. A tree where a root has been specified is called a *rooted tree*. The edges of a rooted tree are often treated as directed. In a rooted tree, every non-root node has exactly one edge that leads to the root. This edge can be thought of as connecting each node to its *parent*. Two vertices with the same parent are said to be *siblings*. Rooted trees can also be considered as directed in the sense that all edges connect parents to their *children*. Given this parent-child relationship, a *descendant* of a node in a directed tree is defined as any other node reachable from that node.

We can now define the Maximum Motif problem we are interested in. Maximum Motif asks for a connected component $G' = (V', E')$ of maximum cardinality in G such that $\mathcal{C}(V') \subseteq \mathcal{M}$ (taken the number of occurrences of each color into account).

MAXIMUM MOTIF

• **Input** : A target vertex colored graph G and colored motif $\mathcal{M}$.

• **Output** : A maximum cardinality connected component $G' = (V', E')$ of G such that $\mathcal{C}(V') \subseteq \mathcal{M}$.

Intuitively, MAXIMUM MOTIF thus asks for the largest submotif $\mathcal{M}' \subseteq \mathcal{M}$ that occurs in G (as a connected component). Being a mere restriction of GRAPH MOTIF, MAXIMUM MOTIF is **NP**-complete as well [12].

3 Hardness of approximation

We prove **APX**-hardness of MAXIMUM MOTIF. Recall that, given a graph $G = (V, E)$, the maximum independent set problem (INDEPENDENT SET) seeks for a maximum cardinality subset $V' \subseteq V$ such that no two vertices in V' are joined by an edge. INDEPENDENT SET is known to be **APX**-hard even when restricted to cubic graphs [14].

Proposition 1. MAXIMUM MOTIF *is* **APX***-hard even if the motif is colorful and the target graph is a tree with maximum degree* 3*.*

Proof. The proof is by reduction from INDEPENDENT SET for cubic graphs. Let $G = (V, E)$ be an instance of INDEPENDENT SET for cubic graphs. Write $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$. For each $v_i \in V$, let us denote by $E(v_i)$ the three edges of E that are incident to v_i. Furthermore, denote by $e(v_i, j)$ the j-th edge of $E(v_i)$, $1 \leq j \leq 3$, the order is arbitrary. We show how to construct the corresponding instance of MAXIMUM MOTIF. This instance consists in a vertex-colored tree $T = (V_T, E_T)$ of maximum degree 3 and a colorful motif $\mathcal{M}$. The tree T is defined as follows: $V_T = \{a_i, b_i, x_{i,I}, x_{i,C}, l_i : 1 \leq i \leq n\} \cup \{d_{i,j}, f_{i,j}, e_{i,j} : 1 \leq i \leq n \ \wedge \ 1 \leq j \leq 3\}$ and $E_T = \{\{a_i, b_i\}, \{b_i, x_{i,I}\}, \{b_i, x_{i,C}\}, \{x_{i,C}, d_{i,1}\}, \{x_{i,I}, f_{i,1}\} : 1 \leq i \leq n\} \cup \{\{a_i, a_{i+1}\} : 1 \leq i < n\} \cup \{\{d_{i,j}, d_{i,j+1}\}, \{f_{i,j}, f_{i,j+1}\} : 1 \leq i \leq n \ \wedge \ 1 \leq j < 3\} \cup \{\{d_{i,j}, e_{i,j}\} : 1 \leq i \leq n \ \wedge \ 1 \leq j \leq 3\}\{\{f_{i,3}, l_i\} : 1 \leq i \leq n\}$. Refer to Figure 1 for a schematic representation of the tree T. Vertex a_i, $1 \leq i \leq n$, is colored $c(a_i)$, vertex b_i, $1 \leq i \leq n$, is colored $c(b_i)$, the two vertices $x_{i,C}$ and $x_{i,I}$, $1 \leq i \leq n$, are colored $c(x_i)$, vertex l_i, $1 \leq i \leq n$, is colored $c(l_i)$, the two vertices $d_{i,j}$ and $f_{i,j}$, $1 \leq i \leq n$ and $1 \leq j \leq 3$, are colored $c(i, j)$, and vertex $e_{i,j}$, $1 \leq i \leq n$ and $1 \leq j \leq 3$, is colored $c(e_k)$, where $e_k = e(v_i, j)$. Write $\mathcal{C}$ for the set of all colors that occur in T (notice that each color in $\mathcal{C}$ occurs at most twice in T). The motif $\mathcal{M}$ is defined by $\mathcal{M} = \mathcal{C}$, and is hence colorful.

Suppose there exists an independent set V' of size k in G. For each $e = \{v_i, v_j\} \in E$, define $\min(e)$ to be

$$\min(e) = \begin{cases} v_i & \text{if } (v_j \in V') \ \vee \ (v_i \notin V' \ \wedge \ v_j \notin V' \ \wedge \ i < j), \\ v_j & \text{otherwise.} \end{cases}$$

Consider the subset $V'_T \subseteq V_T$ defined by $V'_T = \{a_i, b_i : 1 \leq i \leq n\} \cup \{x_{i,I}, f_{i,1}, f_{i,2}, f_{i,3}, l_i : v_i \in V'\} \cup \{x_{i,C}, d_{i,1}, d_{i,2}, d_{i,3} : v_i \notin V'\} \cup \{e_{i,j} : e \in E \ \wedge \ \min(e) = e(v_i, j)\}$. Observe that V'_T induces a connected component in T. Furthermore, $\mathcal{C}(V'_T) = \mathcal{M}' \subseteq \mathcal{M}$, contains all colors from $\mathcal{M}$ but those $c(l_i)$ with $v_i \notin V'$.

Conversely, suppose that there exists a motif $\mathcal{M}' \subset \mathcal{M}$, $|\mathcal{M}'| \geq 7$, that occurs in T. Fix one occurrence of $\mathcal{M}'$ in T and write $V'_T \subseteq V_T$ for the vertices of T involved in this occurrence. Without

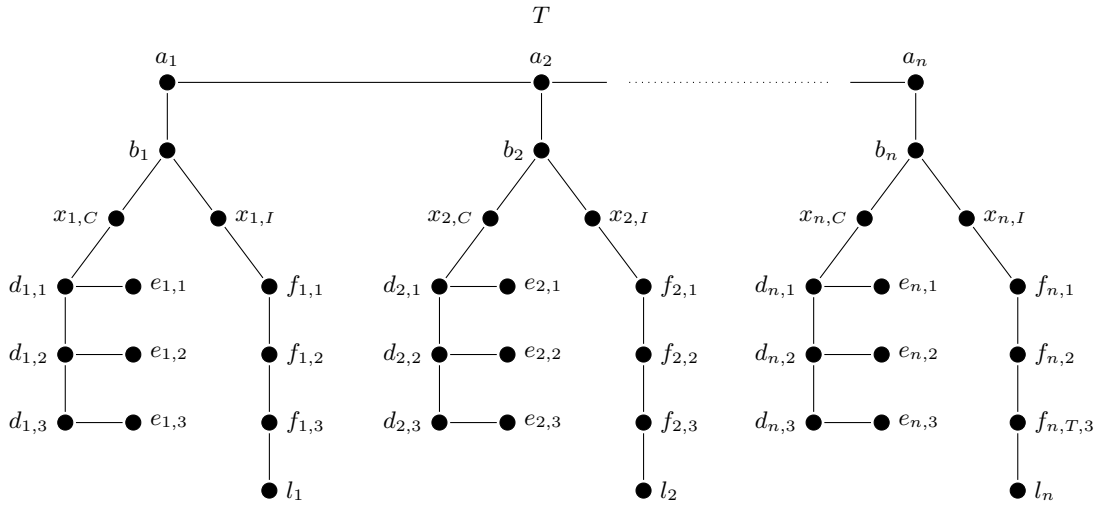

Fig. 1. Schematic representation of the tree T described in the proof of Proposition 1.

loss of generality, suppose that T' is maximal for inclusion (adding any adjacent vertex to T' results in a subtree that is not an occurrence of a submotif of $\mathcal{M}$). Observe first that $a_i, b_i \in V'_T$, $1 \leq i \leq n$, since adding any of these missing vertices would result in a larger connected component T'' of T, such that $\mathcal{C}(T'') \subseteq \mathcal{M}$, thereby contradicting the maximality of T'. Then it follows that $c(a_i), c(b_i) \in \mathcal{M}'$, $1 \leq i \leq n$. Moreover, since $\mathcal{M}$ is colorful, V'_T contains at most one of $x_{i,C}$ and $x_{i,I}$, $1 \leq i \leq n$; they indeed both have the same color. Therefore, by maximality of T', V'_T contains exactly one of $x_{i,C}$ and $x_{i,I}$, $1 \leq i \leq n$, and hence $\mathcal{M}'$ contains color $c(x_i)$, $1 \leq i \leq n$. Pursuing our maximality argument, if $x_{i,C} \in V'_T$ then V'_T also contains the three vertices $d_{i,j}$, $1 \leq i \leq n$ and $1 \leq j \leq 3$, and if $x_{i,I} \in V'_T$ then V'_T also contains the three vertices $f_{i,j}$, $1 \leq i \leq n$ and $1 \leq j \leq 3$. Therefore, $\mathcal{M}'$ contains colors $c(i,j)$, $1 \leq i \leq n$ and $1 \leq j \leq 3$. In case $x_{i,I}, f_{i,1}, f_{i,2}, f_{i,3} \in V'_T$, $1 \leq i \leq n$, $l_i \in V'_T$, and hence $\mathcal{M}'$ contains in addition color $c(l_i)$, $1 \leq i \leq n$. We now claim that we may assume that $c(e) \in \mathcal{M}'$ for all $e \in E$, *i.e.*, submotif $\mathcal{M}'$ contains the color associated with each edge of G. Indeed, suppose that for some color $c(e) \in \mathcal{M}$, say $e = \{v_i, v_j\}$, T' has no vertex colored $c(e)$, *i.e.*, $c(e) \notin \mathcal{M}'$. Then, by maximality of T' (and $\mathcal{M}'$), it follows that $\{x_{i,I}, f_{i,1}, f_{i,2}, f_{i,3}, l_i\} \subseteq V'_T$ and $\{x_{j,I}, f_{j,1}, f_{j,2}, f_{j,3}, l_j\} \subseteq V'_T$, and hence that $\{x_{i,C}, d_{i,1}, d_{i,2}, d_{i,3}\} \cap V'_T = \emptyset$ and $\{x_{j,C}, d_{j,1}, d_{j,2}, d_{j,3}\} \cap V'_T = \emptyset$. Therefore, $V''_T = (V'_T - \{x_{i,I}, f_{i,1}, f_{i,2}, f_{i,3}, l_i\}) \cup \{x_{i,C}, d_{i,1}, d_{i,2}, d_{i,3}\} \cup e_{i,p}$, with $c(e_{i,p}) = c(e)$, induces a subtree in T, and this subtree is an occurrence of $\mathcal{M}'' = (\mathcal{M}' - \{c(l_i)\}) \cup \{c(e)\}$. Applying the above procedure will eventually result in a submotif that contains the color associated with each edge of G. Then it follows that $\{v_i : x_{i,C} \in V'_T\}$ is a vertex cover of G, and hence $\{v_i : x_{i,I} \in V'_T\}$ is an independent set in G.

We have thus shown that there is an independent set of size k in G if and only if there exists a submotif of size $6n+m+k$ that occurs in T. But G is a cubic graph, and hence $k \geq \frac{n}{4}$ and $m = \frac{3}{2}n$. Then it follows that the described reduction is indeed an L-reduction [14] from Independent Set for cubic graphs to Maximum Motif for trees, which proves the proposition. □

We now strengthen the inapproximability of Maximum Motif for trees and colorful motifs. More precisely, we show that there exists $\delta > 0$ such that Maximum Motif cannot be approximated within factor $2^{\log^{\delta} n}$ in polynomial-time unless $\mathbf{NP} \subseteq \mathbf{DTIME}[2^{\text{poly} \log n}]$. The proof is by the *self-*

improvement technique (see for example [7–9]). For the sake of clarity, let us introduce MAXIMUM LEVEL MOTIF which is the restriction of MAXIMUM MOTIF to colorful motifs and rooted trees in which two vertices can have the same color only if they are at the same level (distance to the root) in the target tree. It is easily seen that Proposition 1 can be modified to prove the following result.

Proposition 2. MAXIMUM LEVEL MOTIF *is* **APX**-*hard.*

The following easy lemma will prove useful in the sequel.

Lemma 1. *Let $I = (T, \mathcal{M})$ be an instance of* MAXIMUM LEVEL MOTIF *and T' be a solution for instance I. One can compute in polynomial-time a solution T'' for instance I such that (i) $|T''| \geq |T'|$ and (ii) T'' contains the root of T.*

Aiming at applying the self-improvement technique we need to precisely define the product of two instances I_1 and I_2 of MAXIMUM LEVEL MOTIF. Let $I_1 = (T_1, \mathcal{M}_1)$ and $I_2 = (T_2, \mathcal{M}_2)$ be two instances of MAXIMUM LEVEL MOTIF, where $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ are vertex-colored trees rooted at r_1 and r_2, respectively. The product $I_1 \times I_2$ is defined to be the instance $(T_{1,2}, \mathcal{M}_{1,2})$ where $T_{1,2} = (V_{1,2}, E_{1,2})$ is a rooted tree defined by $V_{1,2} = \{v_i(v_j) : v_i \in V_1 \ \wedge \ v_j \in V_2\}$ and $E_{1,2} = \{\{v_i(v_{j,1}), v_i(v_{j,2})\} : \{v_{j,1}, v_{j,2}\} \in E_2 \ \wedge \ v_i \in V_1\} \cup \{\{v_i(r_2), v_j(r_2)\} : \{v_i, v_j\} \in E_1\}$, and $\mathcal{M}_{1,2}$ is a motif defined by $\mathcal{M}_{1,2} = \{c_1(c_2) : c_1 \in \mathcal{M}_1 \wedge c_2 \in \mathcal{M}_2\}$. The tree $T_{1,2}$ is rooted at vertex $r_1(r_2)$. Informally, $T_{1,2}$ is obtained by replacing each vertex $v_i \in V_1$ by a copy of T_2, connecting these copies through their roots. As for the color of each vertex of $T_{1,2}$, if $v_i \in V_i$ is colored c_i and $v_j \in V_j$ is colored c_j then vertex $v_i(v_j) \in T_{1,2}$ is colored $c_i(c_j)$. Denote by $v_i[T_2]$ the subtree of $T_{1,2}$ isomorphic to T_2 rooted at $v_i(r_2)$. Write $V_{1,2,r} = \{v_i(r_2) : v_i \in V_1\}$. Observe that, by construction, the subtree of $T_{1,2}$ induced by $V_{1,2,r}$ is isomorphic to T_1.

Lemma 2. *Let $I_1 = (T_1, \mathcal{M}_1)$ and $I_2 = (T_2, \mathcal{M}_2)$ be two instances of* MAXIMUM LEVEL MOTIF. *Then $I_1 \times I_2$ is an instance of* MAXIMUM LEVEL MOTIF.

For any instance I of MAXIMUM LEVEL MOTIF, write $I^1 = I$ and $I^k = I \times I^{k-1}$ for all $k \geq 2$. According to Lemma 2, it follows by induction that I^k, $k \geq 1$, is an instance of MAXIMUM LEVEL MOTIF.

Lemma 3. *Let $I = (T, \mathcal{M})$ be an instance of* MAXIMUM LEVEL MOTIF *and let T_S be a solution for I. Then there exists a solution T_{S^k} for instance I^k such that $|T_{S^k}| \geq |T_S|^k$, for all $k \geq 1$.*

Lemma 4. *Let T_{S^k} be a solution of* MAXIMUM LEVEL MOTIF *for instance $I^k = (T^k, \mathcal{M}^k)$. Then, one can compute in polynomial-time a solution T_S for instance I such that $|T_S|^k \geq |T_{S^k}|$.*

Proof. We prove the lemma by induction on k. The result is certainly valid for $k = 1$. Let $k \geq 2$ and assume that the lemma holds for each $1 \leq k' \leq k-1$. Let $T_{S^k} = (V_{S^k}), E_{S^k}$ be a solution for MAXIMUM LEVEL MOTIF over instance I^k. According to Lemma 1, there is no loss of generality in assuming that the root of T^k is part of V_{S^k}. Then it follows that V_{S^k} contains vertices $x_1, \ldots x_p$ of T^k, with $p \leq |T|$, so that at least one vertex in subtree $x_i[T^{k-1}]$ isomorphic to T^{k-1} belongs to T_{S^k}. For each x_i, $1 \leq i \leq p$, denote by $x_i[T_S^{k-1}]$ the subtree of $x_i[T^{k-1}]$ part of T_{S^k}. Let $x_{\max}[T_S^{k-1}]$ be a subtree of maximum size among the subtrees $x_i[T_S^{k-1}]$, $1 \leq i \leq p$. Let T_S^{k-1} be a subtree of T^{k-1} isomorphic to $x_{\max}[T_S^{k-1}]$. Notice that T_S^{k-1} is a solution of MAXIMUM LEVEL MOTIF over instance I^{k-1}. By induction hypothesis, we can compute in polynomial time a solution $T_{S'}$ over instance

I such that $|T_{S'}|^{k-1} \geq |x_{\max}[T_S^{k-1}]|$. Denote now by T_p the subtree of T_{S^k} induced by $\{x_1 \dots x_p\}$. Now $|T_{S^k}| \leq |T_p||x_{\max}[T_S^{k-1}]| \leq |T_p||T_{S'}|^{k-1}$. If $|T_{S'}| \geq |T_p|$, then $T_S = T'_S$ and the lemma holds, since $|T_{S'}||T_{S'}|^{k-1} \geq |T_p||T_{S'}|^{k-1}| \geq |T_{S^k}|$. Otherwise, if $|T_{S'}| < |T_p|$, let T_S be the subtree of T isomorphic to T_p. It follows that $|T_p||T_p|^{k-1} > |T_p||T_{S'}|^{k-1} \geq |T_{S^k}|$.

Observe that T_S is a feasible solution of Maximum Level Motif over instance I. In the former case, when T_S is equal to $T_{S'}$, T_S is feasible by induction hypothesis. Consider the latter case, when T_S is equal to T_p. Let $x_1, \dots x_p$ be the vertices of T_p. Vertex x_i of T_p, with $1 \leq i \leq p$, is associated with color $c_i(c(r), c(r), \dots, c(r))$, where $c(r)$ is the color associated with the root of T and $c_i \in \mathcal{M}$. Observe that, since $\mathcal{M}^k$ is colorful, $c_i \neq c_j$, when $i \neq j$, hence the vertices of T_S have all distinct colors. □

We are now in position to state the main results of this section.

Theorem 1. *For any constant $\delta < 1$,* Maximum Level Motif *cannot be approximated within ratio $2^{\log^\delta n}$ in polynomial-time unless* **NP** $\subseteq$ **DTIME**$[2^{\text{poly}\log n}]$.

Proof. Assume that there exists a constant $\delta < 1$ such that Maximum Level Motif can be approximated within ratio $2^{\log^\delta n}$ in $\mathcal{O}(n^c)$ time, for some constant c. For any fixed $\varepsilon > 0$, let $k = \lceil (\frac{\log^\delta n}{\log(1+\varepsilon)})^{\frac{1}{(1-\delta)}} \rceil$. Given an instance I of Maximum Level Motif of size n, let I^k be the instance obtained by applying the product k times. Now, since the problem can be approximated within ratio $2^{\log^\delta n}$ in $\mathcal{O}(n^c)$ time, it follows that there is an algorithm for Maximum Level Motif for instance I^k with performance ratio $2^{\log^\delta n^k}$ that runs in $\mathcal{O}(n^{ck}) = \mathcal{O}(2^{\text{poly}\log n})$ time. But, according to Lemmas 3 and 4, there is an algorithm for instance I with performance ratio $(2^{\log^\delta n^k})^{1/k} \leq (1+\varepsilon)$, and hence we have designed a PTAS algorithm for Maximum Level Motif. The result now follows from Proposition 2. □

Substituting the complexity hypothesis **NP** $\subseteq$ **DTIME**$[2^{\text{poly}\log n}]$ by the classical **P** = **NP** yields the following result (proof - similar to that of Theorem 1 - omitted): no polynomial-time algorithm achieves a constant approximation ratio for Maximum Level Motif (*i.e.*, Maximum Level Motif is not in **APX**), unless **P** = **NP**.

4 Exponential-time algorithms

We give here two exact branch-and-bound algorithms for Maximum Motif in case the target graph is a tree. Let $I = (T, \mathcal{M})$ be an instance of Maximum Motif problem, where the target graph is a tree $T = (V, E)$.

Lemma 5. Maximum Motif *for trees of size n can be solved in $\mathcal{O}(1.62^n \text{ poly}(n))$ time. In case the motif is colorful, the time complexity reduces to $\mathcal{O}(1.33^n \text{ poly}(n))$.*

We briefly present the main ideas of the proof. First, the algorithms choose a vertex $r \in V$ (we assume w.l.o.g. that r is part of the optimal solution), and the tree T is rooted at r. Both algorithms rely on the fact that, once we have computed a set of vertices $V' \subseteq V$ that are part of the optimal solution, we can compute in polynomial time the maximum cardinality subset $L' \subseteq N(V')$, such that $\mathcal{C}(V') \cup \mathcal{C}(L') \subseteq \mathcal{M}$. Hence, we can assume that a branching occurs only at an internal vertex.

The first algorithm considers a candidate internal vertex v_x and branches in two sub-cases associated with v_x: (1) v_x is added to the solution, or (2) v_x is not added to the solution, and the subtree rooted at v_x is removed.

When $\mathcal{M}$ is colorful, we can assume that there exists two vertices that have the same color c. Indeed, if v_x is the only vertex colored $c(v_x)$, then the algorithm never branches on v_x. The algorithm branches in two sub-cases associated with vertex v_x: (1) v_x is added to the solution, and, for each $v_y \in V'$ colored $c(v_x)$, the subtree rooted at vertex v_y is removed, or (2) v_x is not added to the solution T_S, and then the subtree rooted at v_x is removed.

5 Parameterized Complexity

Fixed-parameter tractability plays a central role in parameterized complexity [5, 13]. In this section we present two fixed-parameter tractable algorithms for Maximum Motif. We first describe the perfect family of hash functions used in both algorithms. Then we give an FPT algorithm in case the target graph is a tree. Finally, we present a (slower) algorithm for the general case.

Consider an instance $I = (G, \mathcal{M})$ of Maximum Motif, where $G = (V, E)$ is a graph and $\mathcal{M}$ is a multiset of colors. For a color c_i of $\mathcal{M}$ and a subset $V' \subseteq V$, we denote by $m_{\mathcal{M}}(c_i)$ the number of occurrences of c_i in $\mathcal{M}$ and by $m_{V'}(c_i)$ the number of vertices in V' colored c_i. In the sequel, we assume that $m_{\mathcal{M}}(c_i) \leq m_V(c_i)$ since an occurrence of $\mathcal{M}$ in G has at most $min\{m_{\mathcal{M}}(c_i), m_V(c_i)\}$ occurrences of color c_i. For a subset of vertices $V' \subseteq V$ and a submotif $\mathcal{M}' \subseteq \mathcal{M}$, we say that V' *violates* $\mathcal{M}'$ if $m_{\mathcal{M}'}(c_i) < m_{V'}(c_i)$ for some $c_i \in \mathcal{M}$.

Both algorithms are based on the color-coding technique [1]. We recall the basic definition of perfect hash functions. For a set S, a family F of functions from S to $\{1, 2, \ldots, k\}$ is *perfect* if for any $S' \subseteq S$ of size k, there exists an injective function $f \in F$ from S' to $\{1, 2, \ldots, k\}$. In the sequel, k will stand for the size of a solution for Maximum Motif. Consider a family H of perfect hash functions from $\mathcal{M}$ to the set $\{1_H, 2_H, \ldots, k_H\}$ (we use the subscript H to emphasis that this set is related to the family H). Let $\mathcal{M}'$ be a submotif of size k and let $G' = (V', E')$ be the occurrence of $\mathcal{M}'$ in G. Since H is perfect, there exists an injective function $h \in H$ that assigns to each occurrence of a color in $\mathcal{M}'$ a distinct label in $\{1_H, 2_H, \ldots, k_H\}$.

Fix some function $h \in H$. For any $c_i \in \mathcal{M}$, denote by $S_H(c_i) \subseteq \{1_H, 2_H, \ldots, k_H\}$ the set of labels associated with occurrences of color c_i by function h. Furthermore, we associate with each vertex v colored c_i the set of labels $S_H(v) = S_H(c_i)$. Let $V' \subseteq V$, $L_H \subseteq \{1_H, \ldots, k_H\}$, then $\mathcal{C}(S_H, V', L_H)$ is defined as the family of sets $S_H(v) \cap L_H$, with $v \in V'$. Notice that $\mathcal{C}(S_H, V', L_H)$ may contain more occurrences of the same set of labels, *i.e.*, if $v_1, v_2 \in V'$ and $c(v_1) = c(v_2)$, then $(S_H(v_1) \cap L_H) = (S_H(v_2) \cap L_H)$. When, $L_H = \{1_H, \ldots, k_H\}$, we denote $\mathcal{C}(S_H, V', L_H)$ by $\mathcal{C}(S_H, V')$.

Definition 1. *Let $\mathcal{C}(S_H, V', L_H)$ be a family of sets $S_H(v)$ with $v \in V'$ and $L_H \subseteq \{1_H, \ldots, k_H\}$, then $\mathcal{C}(S_H, V', L_H)$ is* feasible *if and only if there exists an injective function p from the sets of $\mathcal{C}(S_H, V', L_H)$ to L_H, so that, for each $S_H(v) \in \mathcal{C}(S_H, V', L_H)$, $p(S_H(v))$ is a label of $S_H(v) \cap L_H$.*

Consider now a family $\mathcal{C}(S_H, V')$ of sets associated with V'. Let c_i be a color of $\mathcal{M}$, then by construction $|S_H(c_i)| \leq m_{\mathcal{M}}(c_i)$. Hence, if $\mathcal{C}(S_H, V')$ is feasible, then V' does not violate $\mathcal{M}$.

We now present an FPT algorithm for the case the target graph is a tree $T = (V, E)$. Let $r \in V$, and we want to compute a solution $T' = (V', E')$ of Maximum Motif, so that $|V'| = k$ and $r \in V'$ (we run the algorithm for each $r \in V$.) Define r as the root of T and, for each internal vertex v of V, define a left-to-right ordering on the children of v. Assume that r is colored $c(r)$. Observe that, since r must belong to T', we can safely remove an occurrence of color $c(r)$ from $\mathcal{M}$. Furthermore, we assume that function h assigns to this occurrence of $c(r)$ label 1_H and that $S_H(r) = \{1_H\}$. Observe that there is no other vertex $u \in V - \{r\}$, so that $S_H(u)$ contains 1_H. We can now give the definition of the rightmost vertex of a subtree T' of T.

Definition 2. *Let $T' = (V', E')$ be a subtree of T. A vertex $v \in V'$ is defined to be the rightmost vertex of T' if and only if (i) v has no children in V' and (ii) for each vertex $u \in V'$ on the path from r to v, V' does not contains the right sibling of u.*

Now, consider a vertex $v \in V$ and a subset L_H of labels in $\{1_H, \ldots, k_H\}$. Define $P_r[v, L_H]$ as follows:

$$P_r[v, L_H] = \begin{cases} 1 & \text{if there exists a subtree } T' = (V', E') \text{ of } T \text{ with } r \in V' \text{ and with} \\ & \text{rightmost vertex } v \text{ and such that } \mathcal{C}(S_H, V', L_H) \text{ is feasible,} \\ 0 & \text{otherwise.} \end{cases}$$

The recurrence to compute $P_r[v, L_H]$ is as follows.

$$P_r[v, L_H] = \bigvee_{u, L'_H} P_r[u, L'_H], \tag{1}$$

where u is either a descendant of a left sibling of v or the parent of v, and $L'_H = L_H - \{i_H\}$, for some $i_H \in S_h(v) \cap L_H$. Notice that $P_r[v, \{1_H\}] = 0$, for each $v \in V - \{r\}$, $P_r[r, \{1_H\}] = 1$, and that $P_r[r, \{i_H\}] = 0$ for each $i_H \in \{2_H, \ldots, k_H\}$.

Lemma 6. *Given a labelling h of the motif $\mathcal{M}$, we can compute in time $\mathcal{O}(n^2 2^k)$ if there is a subtree T' of T of size k that matches a submotif $\mathcal{M}'$ of $\mathcal{M}$.*

Proof. We have to show that $P_r[v, L_H] = 1$ if and only if there exists a subtree $T' = (V', E')$ of T having root r, which is an occurrence of a submotif $\mathcal{M}'$ of $\mathcal{M}$ of size $|L_H|$. Since T' must contain r, we assume that L_H contains 1_H.

First, consider a subtree $T' = (V', E')$ with root r. Let v be the rightmost vertex of T'. From the definition of rightmost vertex, it follows that there is no child of v in V' and that there is no vertex in T' which is a right sibling of a vertex on the path from r to v. Denote by $T'' = (V'', E'')$ the tree obtained from T' by removing v. Let u be the rightmost vertex of T''. By definition of rightmost vertex, u is either the parent of vertex v, denoted by $p(v)$, or a descendant of a child v' of $p(v)$ in T'' (with v' a left sibling of v in T' by definition of rightmost vertex).

Let $\mathcal{M}' = \mathcal{C}(V')$ be the multiset of colors associated with the vertices of T'. Consider $\mathcal{C}(S_H, V', L_H)$, the collection of sets of labels in L_H assigned to V'. Notice that $\mathcal{C}(S_H, V', L_H)$ is feasible, as T' is a solution of Maximum Motif problem. It follows that there is an injective function p that assigns to each set $S_H(u)$, with $u \in V'$, a label i_H in $S_H(u)$. But then, function p assigns label $i_H \in S_H(v)$ to the set $S_H(v)$. It follows that the family of sets $S_H(u)$ with $u \in (V' - \{v\})$ must be feasible when p assigns a label in set $\{1_H, \ldots, k_H\} - \{i_H\}$ to each set $S_H(u)$, with $u \in (V(T') - \{v\})$. Hence $P_r[v, L_H] = 1$.

Assume now that $P_r[v, L_H] = 1$. We will prove the results by induction. Since $P_r[v, L_H] = 1$, by Recurrence (1) it follows that there must exists a vertex $u \in V'$ and a label $i_H \in S_H(v)$, so that $P_r[u, L_H - \{i_H\}] = 1$. By induction hypothesis, it follows that there is a subtree of $T'' = (V'', E'')$ of T having root r, so that T'' has size $|L_H| - 1$, u is the rightmost vertex of T'' and $\mathcal{C}(S_H, V'', L_H - \{i_H\})$ is feasible. Hence, by construction, also $\mathcal{C}(S_H, V', L_H)$ is feasible. We will show that v is adjacent to a vertex of T''. By definition of rightmost vertex, u is either the parent of vertex v, denoted by $p(v)$, or a descendant of a child v' of $p(v)$ in T'' (with v' a left sibling of

v in T'). In the former case clearly u and v are adjacent. In the latter case, that is u is not $p(v)$, since T'' must be rooted at r, $p(v)$ belongs to T'', hence v is adjacent to a vertex of T''.

Observe that, if $P[v, \{1_H, \ldots, k_H\}] = 1$, it follows that there is a subtree $T' = (V', E')$ containing the root of T, so that each $\mathcal{C}(V')$ is assigned a distinct label in $\{1_H, \ldots, k_H\}$. By construction V' does not violate $\mathcal{M}$, hence $\mathcal{C}(V')$ is a submotif of $\mathcal{M}$ of size k.

Now, we consider the time complexity of the algorithm. Observe that there exists $\mathcal{O}(n2^k)$ values of the form $P[v, K']$, with $v \in V(T)$ and $L'_H \subseteq \{1_H, \ldots, k_H\}$. Now, in order to compute value $P[v, K']$, we have to check at most $\mathcal{O}(nk)$ other values $P[u, K'']$. Hence the time complexity is $\mathcal{O}(n^2 k 2^k)$. □

Observe that we have to choose $\mathcal{O}(n)$ possible roots. Furthermore, since the family of perfect hash functions has size $\mathcal{O}(\log n)\, 2^{\mathcal{O}(k)}$, it follows that the algorithm time complexity is $\mathcal{O}(k2^k n^3 \log n)\, 2^{\mathcal{O}(k)}$.

Next we describe a parameterized algorithm when the instance of Maximum Motif consists in a graph $G = (V, E)$ and a motif $\mathcal{M}$. The algorithm for this case consists in combining two perfect families of hash functions, and then applying a strategy similar to that presented in [6, 4].

Consider two different perfect families of hash functions: a family H from $\mathcal{M}$ to $\{1_H, \ldots, k_H\}$, as we have previously introduced in this section, and a family F from the set V to $\{1_F, \ldots, k_F\}$. By the property of the family of perfect hash functions, we know that there is a function $f \in F$ such that the vertices of G that belong to a solution of size k are associated with distinct labels of $\{1_F, \ldots, k_F\}$. Similarly, we know that there is a function $h \in H$ such that the occurrences of colors of $\mathcal{M}$ that belong to an optimal solution, are associated with different labels of $\{1_H, \ldots, k_H\}$. Observe that each family of perfect hash functions consists of $\mathcal{O}(\log n)\, 2^{\mathcal{O}(k)}$ functions. Hence, we can combine all the possible pairs (f, h) of functions, with $f \in F$ and $h \in H$, in time $\mathcal{O}(\log^2 n)\, 4^{\mathcal{O}(k)}$.

Recall that, for each color $c_i \in \mathcal{M}$, $S_H(c_i)$ denotes the set of labels associated with occurrences of color c_i by function h, and that, given v is colored c_i, $S_H(v) = S(c_i)$. Now, for each $v \in V$ and for each subset $L \subseteq \{1_F, \ldots, k_F\}$, define $M_L(v)$ as the family of all sets of labels $H' \subseteq \{1_H, \ldots, k_H\}$ so that there exists an occurrence V', with $v \in V'$, where the set of labels in $\{1_F, \ldots, k_F\}$ that f assigns to V' is exactly L and such that $\mathcal{C}(S_H, V', H')$ is feasible. Now, we present a method called *Batch procedure* for computing $M_L(v)$, similar to that introduced in [6, 4]. Assume that we have computed the family of sets $M_{L'}(v)$, with $L' \subseteq L \setminus f(v)$, we apply the following procedure.

Batch Procedure(L, v):

- Define C_H to be the family of all pairs (H', L') such that $H' \subseteq \{1_H, \ldots, k_H\} - \{i_H\}$ for some $i_H \in S_H(v_i)$, $L' \subseteq L \setminus \{f(v)\}$, and $H' \in M_{L'}(u)$ for some $u \in N(v)$.
- Run through all pairs of (H', L'), (H'', L'') in C_H and determine whether $H' \cap H'' = \emptyset$ and $H' \cup H'' \subseteq \{1_H, \ldots, k_H\} - \{i_H\}$, for some $i_H \in S_H(v_i)$, and whether $L' \cap L'' = \emptyset$. If there is such a pair, add $(H' \cup H'', L' \cup L'')$ to C_H and repeat this step. Otherwise, continue to the next step.
- Set $M_L(v)$ to be all the sets of labels $H' \cup \{i_H\}$, where $i_H \in S_H(v_i) - H'$, $(H', L') \in C_H$ and $L' = L \subseteq -\{f(v)\}$.

Lemma 7. *Given a vertex $v \in V$ and $L \subseteq \{1_F, \ldots, k_F\}$, the batch procedure computes correctly $M_L(v)$, assuming $M_{L'}(u)$ is given for each u adjacent to v and for each $L' \subseteq L \setminus \{f(v)\}$.*

Notice that function h assigns a distinct label in $\{1_H, \ldots, k_H\}$ to each occurrence of a color in a submotif $\mathcal{M}'$, with $|\mathcal{M}'| = k$. Consider $M_L(v) = \{1_H, 2_H, \ldots, k_H\}$ with $L = \{1_F, 2_F, \ldots, k_F\}$. The

set of vertices in V' associated with labels $\{1_F, 2_F \ldots, k_F\}$ are then associated with colors having labels in $\{1_H, 2_H \ldots, k_H\}$. Hence, $C(V')$ does not violate $\mathcal{M}$.

Lemma 8. *Given labeling functions $h : \mathcal{M} \to \{1, \ldots, k\}$ and $f : V \to \{1, \ldots, k\}$, the batch procedure determines in $\mathcal{O}(2^{5k}kn^2)$ time whether there exists a solution of* Maximum Motif *of size k.*

Since each perfect family of hash functions consists of $\mathcal{O}(\log n)\, 2^{\mathcal{O}(k)}$, the overall time complexity of the algorithm is $\mathcal{O}(2^{5k}kn^2 \log^2 n)\, 4^{\mathcal{O}(k)}$.

References

1. N. Alon, R. Yuster, and U. Zwick, *Color coding*, Journal of the ACM **42** (1995), no. 4, 844–856.
2. N. Betzler, M.R. Fellows, C. Komusiewicz, and R. Niedermeier, *Parameterized algorithms and hardness results for some graph motif problems*, Proc. 19th Annual Symposium on Combinatorial Pattern Matching (CPM), Pisa, Italy (P. Ferragina and G.M. Landau, eds.), Lecture Notes in Computer Science, vol. 2089, Springer, 2008, To appear.
3. R. Diestel, *Graph theory*, second ed., Graduate texts in Mathematics, no. 173, Springer-Verlag, 2000.
4. R. Dondi, G. Fertin, and S. Vialette, *Weak pattern matching in colored graphs: Minimizing the number of connected components*, Proc.10th Italian Conference on Theoretical Computer Science (ICTCS, Roma, Italy, World-Scientific, 2007, pp. 27–38.
5. R. Downey and M. Fellows, *Parameterized complexity*, Springer-Verlag, 1999.
6. M. Fellows, G. Fertin, D. Hermelin, and S. Vialette, *Sharp tractability borderlines for finding connected motifs in vertex-colored graphs*, Proc. 34th International Colloquium on Automata, Languages and Programming (ICALP), Wroclaw, Poland (L. Arge, C. Cachin, T. Jurdzinski, and A. Tarlecki, eds.), Lecture Notes in Computer Science, vol. 4596, Springer, 2007, pp. 340–351.
7. J. Hein, T. Jiang, L. Wang, and K. Zhang, *On the complexity of comparing evolutionary trees*, Discrete Applied Mathematics **71** (1996), 153–169.
8. T. Jiang and M. Li, *On the approximation of shortest common supersequences and longest common subsequences*, SIAM Journal on Computing **24** (1995), 1122–1139.
9. D. Karger, R. Motwani, and G.D.S. Ramkumar, *On approximating the longest path in a graph*, SIAM Journal on Computing **24** (1995), 1122–1139.
10. B.P. Kelley, R. Sharan, R.M. Karp, T. Sittler, D. E. Root, B.R. Stockwell, and T. Ideker, *Conserved pathways within bacteria and yeast as revealed by global protein network alignment*, Proceedings of the National Academy of Sciences **100** (2003), no. 20, 11394–11399.
11. M. Koyutürk, A. Grama, and W. Szpankowski, *Pairwise local alignment of protein interaction networks guided by models of evolution*, Proc. 9th Annual International Conference on Research in Computational Molecular Biology (RECOMB), Cambridge, MA, USA (S. Miyano, J. P. Mesirov, S. Kasif, S. Istrail, P. A. Pevzner, and M. S. Waterman, eds.), Lecture Notes in Bioinformatics, vol. 3500, Springer, 2005, pp. 48–65.
12. V. Lacroix, C.G. Fernandes, and M.-F. Sagot, *Motif search in graphs: application to metabolic networks*, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **3** (2006), no. 4, 360–368.
13. R. Niedermeier, *Invitation to fixed parameter algorithms*, Lecture Series in Mathematics and Its Applications, Oxford University, Press, 2006.
14. C.H. Papadimitriou and M. Yannakakis, *Optimization, approximation and complexity classes*, Journal of Computer and System Sciences **43** (1991), 425–440.
15. J. Scott, T. Ideker, R.M. Karp, and R. Sharan, *Efficient algorithms for detecting signaling pathways in protein interaction networks*, Journal of Computational Biology **13** (2006), 133–144.
16. R. Sharan, T. Ideker, B. Kelley, R. Shamir, and R.M. Karp, *Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data*, Proc. 8th annual international conference on Computational molecular biology (RECOMB), San Diego, California, USA (P.E. Bourne and D. Gusfield, eds.), ACM Press, 2004, pp. 282–289.
17. R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker, *Conserved patterns of protein interaction in multiple species*, Proc. Natl Acad. Sci. USA **102** (2005), no. 6, 1974–1979.

Appendix

Proof (of Lemma 1). Consider a solution $T' = (V', E')$ of Maximum Level Motif for instance I, and assume that T' does not contain the root r of T. Notice that T' must be a rooted subtree of T, and let $y \in V'$ be the root of T'. Now consider the unique path $P = (r, x'_1, \ldots, x'_p = y)$, from the root r to y. Two vertices x'_i and x'_j of P, $1 \leq i \neq j \leq p$, have distinct colors, since they belong to different levels of T. Moreover, each vertex x'_i, with $1 \leq i \leq p-1$, has a distinct color from each vertex $v \in V'$, since vertices x'_i and v belong to different levels of T. Define T'' as the subtree of T induced by the set of vertices $V'' = V' \cup \bigcup_{i=1}^{p-1} x_i$. Notice that T'' contains the root r of T, and by construction $|V''| \geq |V'|$. □

Proof (Of Lemma 2). Write $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ and assume that T_1 and T_2 are rooted at r_1 and r_2, respectively. Let $I_1 \times I_2 = (T_{1,2}, \mathcal{M}_{1,2})$ and write $T_{1,2} = (T_{1,2}, E_{1,2})$. First, we show that $T_{1,2}$ is a rooted tree. Indeed, $T_{1,2}[V_{1,2,r}]$ is isomorphic to T_1 and each vertex in $V_{1,2} - V_{1,2,r}$ belongs to a subtree rooted at some $v_i(r_2) \in V_{1,2,r}$. Furthermore, $T_{1,2}$ is rooted by definition.

Now, we show that two vertices of $T_{1,2}$ have the same color only if they are at the same level in $T_{1,2}$. Let $u_1(u_2)$ and $v_1(v_2)$ be two vertices of $T_{1,2}$ such tat $c(u_1(u_2)) = c(v_1(v_2)) = c_a(c_b)$. If $u_1 = v_1$ we are done so that we may now assume $u_1 \neq v_1$. Therefore, we must have $c(u_1) = c(v_1) = c_a$. Furthermore, observe that, by construction, all vertices in $u_1[T_2]$ and $v_1[T_2]$ are colored $c_a(c_x)$ for some color $c_x \in \mathcal{M}$. Consider the subtree $T_{1,2}[V_{1,2,r}]$ induced $V_{1,2,r}$. Since $T_{1,2}[V_{1,2,r}]$ is isomorphic to T_1, each vertex $x_i(r_2) \in V_{1,2,r}$ has color $c(x_i)(c(r_2))$. Now, since all vertices of $u_1[T_2]$ and $v_1[T_2]$ are colored $c_a(c_x)$, it follows that the root $x_i(r_2)$ of $u_1[T_2]$ and the root $x_j(r_2)$ of $v_1[T_2]$ have the same color $c_a(c(r_2))$. Then it follows that $x_i(r_2)$ and $x_j(r_2)$ must be at the same level l_1 of $T_{1,2}$ since they both belong to $V_{1,2,r}$ and $T_{1,2}[V_{1,2,r}]$ is isomorphic to T_1, where x_i and x_j must be both at level l_1.

Now, consider the subtrees $u_1[T_2]$ and $v_1[T_2]$ isomorphic to T_2. Recall, that vertices $u_1(u_2)$ and $v_1(v_2)$ of $T_{1,2}$ are both colored $c_a(c_b)$. As previously observed, all vertices $u_1(u_j)$ in $u_1[T_2]$ and $v_1[T_2]$ are associated with colors $c_a(c(u_j))$ for some $u_j \in V_2$. Since $I_2 = (T_2, \mathcal{M}_2)$ is an instance of Maximum Level Motif, vertices u_2 and v_2 must be at the same level l_2 in T_2 since $c(u_2) = c(v_2) = c_b$. Then, since $u_1[T_2]$ and $v_1[T_2]$ are both isomorphic to T_2, $u_1(u_2)$ and $v_1(v_2)$ are both at level l_2 in $u_1[T_2]$ and $v_1[T_2]$, respectively. It follows that both $u_1(u_2)$ and $v_1(v_2)$ are at level $l_1 + l_2$ in $T_{1,2}$.

Finally, let us consider the motif $\mathcal{M}_{1,2}$. By construction, $\mathcal{M}_{1,2}$ is a set, hence it is colorful. □

Proof (of Lemma 3). We prove the lemma by induction on k. The result is certainly valid for $k = 1$. Let $k \geq 2$ and assume that the lemma holds for all $1 \leq k' \leq k-1$. Let $T_S = (V_{T_S}, E_{T_S})$ be a solution of Maximum Level Motif for instance I, with $V_{T_S} = \{v_1, v_2, \ldots, v_z\}$. Observe that T_S is a subtree of T and that all vertices in V_{T_S} have distinct colors since $\mathcal{M}$ is colorful. By Lemma 1, we can assume that the root r of T is part of V_{T_S}. We now construct a solution T_{S^k} for instance I^k as follows.

First, consider the subtree of T^k which consists of the set $V_{T_S,r'}$ of vertices $v_1(r'), v_2(r'), \ldots, v_z(r')$, where each $v_i(r')$ is the root of a subtree of T^k isomorphic to T^{k-1}. Observe that, by construction, the set of vertices $V_{T_S,r'}$ induces a subtree $T^K[V_{T_S,r'}]$ of T^k. Since vertices $v_1, v_2, \ldots, v_z$ have all distinct colors in T, then it follows that $v_1(r'), v_2(r'), \ldots, v_z(r')$ have distinct colors a well. Let $v_i[T^{k-1}]$ and $v_j[T^{k-1}]$, $1 \leq i < j \leq z$, be two subtrees of T isomorphic to T^{k-1} rooted at $v_i(r')$ and $v_j(r')$. Observe that any two vertices $x \in v_i[T^{k-1}]$ and $y \in v_j[T^{k-1}]$ cannot have the same color, since $c(v_i) \neq c(v_j)$. Now, consider a subtree rooted at $v_i(r')$, with $1 \leq i \leq z$. By induction hypothesis, there is a solution $T_{S^{k-1}}$ of Maximum Level Motif over instance $I^{k-1} = (T^{k-1}, \mathcal{M}^{k-1})$, such that $|T_{S^{k-1}}| \geq |T_S|^{k-1}$. Notice that, by Lemma 1, we can assume that $T_{S^{k-1}}$ contains the root of T^{k-1}. Now we build solution T_{S^k}, by adding, for each $v_i(r')$, with $1 \leq i \leq z$, a subtree of $v_i[T^{k-1}]$ isomorphic to $T_{S^{k-1}}$. Since T_S^k consists of $|T_S|$ such subtrees, it follows immediately that the inequality holds.

Finally, notice that the solution we have built is a feasible solution for Maximum Level Motif for instance I^k. First, T_S^k is connected by construction. Furthermore, each vertex of T_S^k has a distinct color. Indeed, we have shown that this holds for any two vertices that are not in the same subtree $v_i[T^{k-1}]$. By induction hypothesis, since $T_{S^{k-1}}$ is a feasible solution of Maximum Level Motif over instance I^{k-1}, it follows that two vertices that belong to the same subtree $t_i[T^{k-1}]$ must have distinct colors. □

Proof (of Proposition 2). We prove that the problem is **APX**-hard by modifying the L-reduction for Maximum Motif on bounded tree presented in Prop. 1. Let $G = (V, E)$ be an instance of Independent Set on cubic graph. Write $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$. For each $v_i \in V$, let us denote by $E(v_i)$ the three edges of E that are incident to v_i. Furthermore, denote by $e(v_i, j)$ the j-th edges of $E(v_i)$, $1 \leq j \leq 3$, the reference order is arbitrary. We now show how to construct the corresponding instance of Maximum Level Motif. This instance consists in a rooted vertex-colored tree $T = (V_T, E_T)$ and a colorful motif $\mathcal{M}$.

The tree T is defined as follows:

$$\begin{aligned} V_T =& \{r\} \cup \{b_i, x_{i,I}, x_{i,C}, l_i : 1 \leq i \leq n\} \ \cup \\ & \{e_{i,j} : 1 \leq i \leq n \ \wedge \ 1 \leq j \leq 3\} \\ E_T =& \{\{r, b_i\}, \{b_i, x_{i,I}\}, \{b_i, x_{i,C}\}, : 1 \leq i \leq n\} \ \cup \\ & \{\{x_{i,C}, e_{i,j}\} : 1 \leq i \leq n \ \wedge \ 1 \leq j \leq 3\} \\ & \{\{x_{i,I}, l_i\} : 1 \leq i \leq n\} \end{aligned}$$

Vertex r is colored $c(r)$, vertex b_i, $1 \leq i \leq n$, is colored $c(b_i)$, the two vertices $x_{i,C}$ and $x_{i,I}$, $1 \leq i \leq n$, are colored $c(x_i)$, vertex l_i, $1 \leq i \leq n$, is colored $c(l_i)$, vertex $e_{i,j}$, $1 \leq i \leq n$ and $1 \leq j \leq 3$, is colored $c(e_k)$, where $e_k = e(v_i, j)$. Write $\mathcal{C}$ for the set of all colors that occur in T (notice that each color in $\mathcal{C}$ occurs at most three times in T). The motif $\mathcal{M}$ is defined by $\mathcal{M} = \mathcal{C}$, and is hence colorful.

First, observe that this in an instance of Maximum Level Motif. Indeed, the tree T is rooted and has 4 level and all the leaves are at level 4. Two vertices have the same color either if they both are at level 3, i.e. a pair $(x_{i,C}, x_{i,I})$, or if they are both at level 4, i.e. they are leaves associated with the same edge e_k.

In what follows, we show that there exists a solution S of Independent Set on cubic graph of size t if and only if there exists a solution Maximum Level Motif of size $1 + 2n + m + t$.

Let S be a solution of Independent Set, we define a solution T' of Maximum Level Motif as follows. Consider the subset $V' \subseteq V$ defined as by

$$\begin{aligned} V_T = &\{r\} \cup \{b_i : 1 \leq i \leq n\} \ \cup \\ &\{x_{i,I}, l_i : v_i \in V'\} \ \cup \\ &\{x_{i,C} : v_i \notin V'\} \ \cup \\ &\{e_{i,j} : e \in E \ \wedge \ \min(e) = e(v_i, j)\}. \end{aligned}$$

V' induces a subtree T' in T. Furthermore, $\mathcal{C}(V'_T) = \mathcal{M}' \subseteq \mathcal{M}$ is defined by deleting in $\mathcal{M}$ every colors $c(l_i)$ such that $v_i \notin V'$.

Conversely, let $T' = (V', E')$ be a solution of Maximum Level Motif of size $1 + 2n + m + t$. Notice that $c(V') = \mathcal{M}' \subseteq \mathcal{M}$. Without loss of generality, suppose that T' is maximal for inclusion (adding any adjacent vertex to T' result in a submotif that does not occur in T). By Lemma 1, we can assume that $r \in V'$, hence $\mathcal{M}'$ contains color $c(r)$. Furthermore, notice that $b_i \in V'$, $1 \leq i \leq n$, since adding any of these missing vertices would result in a larger connected component T'' of T, such that $c(T'') \subseteq \mathcal{M}$, thereby contradicting the maximality of T'. Then it follows that $c(b_i) \in \mathcal{M}'$, $1 \leq i \leq n$. Therefore, still by maximality, V' contains exactly one of $x_{i,C}$ and $x_{i,I}$, $1 \leq i \leq n$, and hence $\mathcal{M}'$ contains color $c(x_i)$, $1 \leq i \leq n$. Pursuing our maximality argument, if $x_{i,C} \in V'_T$ then in case $x_{i,I} \in V'$, $1 \leq i \leq n$, $l_i \in V'$, and hence $\mathcal{M}'$ contains in addition color $c(l_i)$, $1 \leq i \leq n$. Furthermore, we may assume that $c(e) \in \mathcal{M}'$ for all $e \in E$, *i.e.*, submotif $\mathcal{M}'$ contains the color associated with each edge of G. Indeed, suppose that there is not a vertex associated with color $c(e)$, say $e = \{v_i, v_j\}$, in T', that is $c(e)$ is not part of $\mathcal{M}'$. Then, by maximality of T' and $\mathcal{M}'$, it follows that $\{x_{i,I}, l_i\} \subseteq V'$ and $\{x_{j,I}, l_j\} \subseteq V'$, and hence that $x_{i,C}, x_{j,C}, \notin V'$. Therefore, $V'' = V' - \{x_{i,I}, l_i\} \cup \{x_{i,C}\} \cup e_{i,p}$, with $c(e_{i,p}) = c(e)$, induces a subtree in T and $\mathcal{C}(V'') = \mathcal{M}'' = (\mathcal{M}' - \{c(l_i)\}) \cup \{c(e)\}$. Applying the above procedure will eventually result in a submotif that contains the color associated with each edge of G.

Then it follows that $\{v_i : x_{i,C} \in V'\}$ is a vertex cover of G, and hence $\{v_i : x_{i,I} \in V'\}$ is an independent set in G.

Since G is a cubic graph, it follows that $k \geq \frac{1}{4}$ and $m = \frac{3}{2}n$. Then it follows that the described reduction is indeed a L-reduction [14] from Independent Set for cubic graphs to Maximum Level Motif for trees, which proves the proposition. □

Proof (of Lemma 5). Let $I = (T, \mathcal{M})$ be an instance of Maximum Motif, where $T = (V, E)$. Given a tree T', denote by $L(T')$ the set of leaves of T'. First, the algorithm chooses a vertex $r \in V$, and the tree T is rooted at r. Notice that the algorithm is iterated for each possible choice of r.

Lemma 9. *Let $T' = (V', E')$ be a subtree of T, let $V_S \subseteq (V' - L(T'))$, with $\mathcal{C}(V_S) \subseteq \mathcal{M}$, and let T_S be the subtree of T' induced by V_S. Then, we can compute in polynomial time the maximum cardinality submotif $\mathcal{M}' \in \mathcal{M}$, so that there is a set of leaves $L' \subseteq L(T')$, with $\mathcal{C}(L') = \mathcal{M}'$, and so that $c(V_S \cup L') \subseteq \mathcal{M}$.*

Proof. Denote by $m_{T_S}(c_i)$ (resp. $m_{L(T')}(c_i)$) the number of occurrences in $\mathcal{C}(T_S)$ (resp. $\mathcal{C}(L(T'))$) of color c_i, with $c_i \in \mathcal{C}$. For each $c_i \in \mathcal{C}$, denote by $m_{\mathcal{M}}(c_i)$ the occurrences of color c_i in $\mathcal{M}$. Observe that $m_{T_S}(c_i) \leq m_{\mathcal{M}}(c_i)$. Let $l_{T_S}(c_i) = m_{\mathcal{M}}(c_i) - m_{T_S}(c_i)$. Then $\mathcal{M}'$ can be computed by taking independently for each color $c_i \in \mathcal{C}$, $\min(l_{T_S}(c_i), m_{L(T')}(c_i))$ occurrences of color c_i. □

Consider $T' = (V', E')$ and $v_x \in V' - L(T')$. Let $T_S = (V_S, E_S)$ with $V_S \subseteq V'$ and $c(V_S) \subseteq \mathcal{M}$. Let $T'(v_x)$ be the subtree of T' rooted at v_x. Then, if $v_x \in (V' - V_S)$, each vertex of the subtree $T'(v_x)$ is not in of V_S.

Let $T_S = (V_S, E_S)$ be a feasible solution of Maximum Motif over instance $I' = (T', \mathcal{M})$, where $T' = (V', E')$ is a subtree of T. An internal vertex $v_x \in V'$ not included in V_S, and adjacent to a vertex in V_S, is called a *candidate* vertex for T_S. A feasible solution T'_S for Maximum Motif is said to *extend* T_S, if it can be computed starting from T_S. The algorithm considers a candidate vertex v_x of V'. The algorithm branches in two sub-cases associated with vertex v_x:

1. v_x is added to the solution T_S;
2. v_x is not added to the solution S, and the subtree $T'(x)$ is removed from T'.

Notice that, since v_x is an internal vertex of T', the subtree $T'(x)$ ha size at lest 2. Hence the number of vertices of V' that the algorithm has to consider is decreased by 1 in Case 1) and by at least 2 in Case 2). Let I be an instance of Maximum Motif, consisting of a motif $\mathcal{M}$ of size m and a tree T with n vertices. Observe that $m \leq n$. Denote by $Z(n)$ the worst case time complexity of the algorithm. Then $Z(n) = Z(n-1) + Z(n-2) + \mathcal{O}(n)$. It follows that $Z(n) \leq \mathcal{O}(1.62^n \text{ poly}(n))$.

Consider now the case when the motif $\mathcal{M}$ is a set of colors. Consider a candidate vertex $v_x \in V'$ for T_S, colored $c(v_x)$. Assume that v_x is the only vertex of $V' - L(T')$ colored $c(v_x)$. Then, v_x is added to V_S. Indeed, since v_x is candidate and $\mathcal{M}$ is a set, there is no vertex in T_S that has color $c(x)$ and a vertex y of T' colored $c(x)$ must be a leaf.

The algorithm considers the following cases associated with a candidate vertex v_x for T_S:

1. v_x is added to the solution T_S; then, for each $v_y \in V'$ colored $c(v_x)$, the subtree rooted at vertex v_y is removed.
2. v_x is not added to the solution T_S; then the subtree of T' rooted at v_x is removed from T'.

Observe that only Case 1) is modified, as Case 2) is identical to the case when $\mathcal{M}$ is a multiset. In Case 1), v_x is added to T_S and, since there exists at least one internal vertex of V' colored $c(v_x)$, the number of vertices that the algorithm has still to consider is decreased by 3. Then, $Z(n) = Z(n-2) + Z(n-3) + \mathcal{O}(n)$. It follows that $Z(n) \leq \mathcal{O}(1.33^n \text{ poly}(n))$.

Let $T_S = (V_S, E_S)$ a solution constructed by the algorithm. Then, Lemma 9 is applied in order to add the maximum number of vertices x of $V - V_S$, so that x is adjacent to some vertex of V_S. □

Proof (of Lemma 7). Consider the family C_H computed by the batch procedure. Let $(H', L') \in CH$, where $H' \subseteq \{1_H, \ldots, k_H\} - \{i_H\}$ for some $i_H \in S_H(v)$ and $L' \subseteq L \setminus f(v)$. By construction, $H' = H'_1 \cup \cdots \cup H'_t$, where each $H'_i \subseteq \{1_H, \ldots, k_H\} - \{i_H\}$, $1 \leq i \leq t$, is associated by function h with a submotif $\mathcal{M}'_i$ that has an occurrences in a set V'_i, so that V'_i includes a vertex adjacent to v. Notice that each V'_i is associated with a set of labels $L_i \subseteq \{1_F, \ldots, k_F\}$, so that $L_i \cap L_j = \emptyset$ for each V_j with $j \neq j$. Hence, all the connected components $G[V'_1]$, …, $G[V'_t]$ are pairwise disjoint,

and $\{v\} \cup V_1 \cdots \cup V_n$ is connected. It follows that $H' \cup \{i_h\}$ is then feasible and it is associated with vertices having labels L', so L' belongs to $M_L(v)$.

Consider now L', a set of labels in $M_L(v)$, so that $L' \cup f(v)$ is part of $M_L(v)$. Observe that, by definition, there exists a set of vertices V', associated with set of labels L', so that the function f assigns to V' the set of labels in L'. Consider now the connected components induced by sets V_1, ..., V_t where $V_1 \cup V_2 \cdots \cup V_t = V'$. Since $V' \cup \{v\}$ must be a connected component, each V_i must a vertex adjacent to v. Each connected component V_i is associated with the a set of labels L_i, so that $L_i \cap L_j \neq \emptyset$, for each $j \neq i$. Now, the batch procedure will compute the pair (H', L') in its second step, and $H' \cup \{i_H\}$ will be added to CH. □

Proof (of Lemma 8). First, we will show that a set $M_L(v)$ is computed by batch procedure in time $\mathcal{O}(2^{4k}kn)$. The first step of batch procedure searches at most $2^k n$ families of subsets H' of labels in $\{1_H, \ldots, k_H\}$, for each $i_H \in S_H(v)$. Notice that $|S_H(v)| \leq k$. Each family consists of at most 2^k sets. Hence, the first step requires $\mathcal{O}(2^{2k}kn)$.

For the second step of the batch procedure, observe that there are at most 2^{2k} set of label-subset pairs H' and L', so the second step is repeated 2^{2k} times. Each iteration of this step can be computed in time $\mathcal{O}(2^k n)$, hence the second step require time $\mathcal{O}(2^{4k}kn)$. Accounting also for the third step, the overall time complexity for of one invocation of the batch procedure is $\mathcal{O}(2^{4k}k + 2^{2k}kn) = \mathcal{O}(2^{4k}kn)$.

According to Lemma 7, the batch procedure must be invoked at most $2^k n$ times in order to obtain $\mathcal{M}_L(v)$ for every $v \in V$ and every label subset $L' \subseteq \{1_F, \ldots, k_F\}$, hence an overall time complexity of $\mathcal{O}(2^{5k}kn^2)$. □
