Finding Approximate and Constrained Motifs in Graphs *

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

[1] Dipartimento di Scienze dei Linguaggi, della Comunicazione e degli Studi Culturali
Universita degli Studi di Bergamo, Via Donizetti 3, 24129 Bergamo - Italy
riccardo.dondi@unimib.it
[2] Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR CNRS 6241
Universite de Nantes, 2 rue de la Houssiniere, 44322 Nantes Cedex 3 - France
guillaume.fertin@univ-nantes.fr
[3] LIGM, CNRS UMR 8049, Universite Paris-Est,
5 Bd Descartes 77454 Marne-la-Vallee, France
vialette@univ-mlv.fr

Abstract. One of the most relevant topics in the analysis of biological networks is the identi cation of functional motifs inside a network. A recent approach introduced in literature, called Graph Motif, represents the network as a vertex-colored graph, and the motif $\mathcal{M}$ as a multiset of colors. An occurrence of a motif $\mathcal{M}$ in a vertex-colored graph G is a connected induced subgraph of G whose vertex set is colored exactly as $\mathcal{M}$. In this paper we investigate three di erent variants of the Graph Motif problem. The rst two variants, Minimum Adding Motif (Min-Add Graph Motif) and Minimum Substitution Motif (Min-Sub Graph Motif), deal with approximate occurrences of a motif in the graph, while the third variant, Constrained Graph Motif (CGM), constrains the motif to contain a given set of vertices. We investigate the computational and parameterized complexity of the three problems. We show that Min-Add Graph Motif and Min-Sub Graph Motif are both NP-hard, even when $\mathcal{M}$ is a set, and the graph is a tree with maximum degree 4 in which each color appears at most twice. Then, we show that Min-Sub Graph Motif is xed-parameter tractable when parameterized by the size of $\mathcal{M}$. Finally, we consider the parameterized complexity of the CGM problem ; we give a xed-parameter algorithm for graphs of bounded treewidth, and show that the problem is W[2]-hard when parameterized by $|\mathcal{M}|$, even if the input graph has diameter 2.

Keywords: Graph motif; Computational biology; Parameterized complexity; Algorithms; Computational complexity.

* A preliminary version of the paper has appeared in CPM 2011 [9]

1 Introduction

The analysis of biological networks has become increasingly relevant in computational biology. A crucial problem that emerged recently in the analysis of protein-protein interaction networks and metabolic networks is the identi cation of functional motifs inside a network (see for example [6, 13, 14, 20{22]). The classical approach to identify motifs inside a network is based on the graph-theoretical topology of the motif. However the information on the topology is often missing. A recent approach introduced in [15, 6] aims at discovering functional motifs that do not rely on the conservation of the topology, but that are simply connected components of the network. This approach has been formalized as a graph problem (named Graph Motif) in which, given a vertex-colored graph $G = (V; E)$ and a multiset $\mathcal{M}$ of colors, the goal is to nd a subset $V' \subseteq V$ which is connected and whose vertex set is colored exactly as $\mathcal{M}$.

The Graph Motif problem has been widely investigated in the past. The problem is known to be NP-complete [15], even if the input graph is a tree with maximum degree 3, the motif is a set and there exists at most three vertices in the graph G with the same color [11], and if the input graph is a bipartite graph with maximum degree 4 and the motif is built over only two colors [11]. The Graph Motif problem admits a polynomial-time algorithm when the input graph is a tree and each color occurs at most twice in the input tree [11, 23] and when the input graph is a caterpillar [1]. However, if the input graph is a rooted tree of height two, then the problem is NP-complete, even if the motif is a set of colors. The Graph Motif problem is known to be in FPT, when parameterized by the size of the motif [4, 11, 12], while it is W[1]-hard when parameterized by the number of distinct colors in the motif, even in the case the input graph is a tree [11]. Recently, the kernelization complexity of the problem has also been considered [1]. More precisely, in [1] it is proved that the Graph Motif problem does not admit a polynomial kernel, even if restricted to comb graphs[1], unless NP $\subseteq$ co-NP=Poly.

Related Work. Several variants of the Graph Motif problem have been considered in the literature. Such variants either modify the requirement of connectedness [8, 4], or look for approximate occurrences of the motif, where some colors are allowed to be inserted or deleted in an occurrence of the motif [6, 8, 12].

[1] A comb graph is a tree, where all vertices have degree at most 3 and all the vertices of degree 3 lie on a single simple path.

parameterized by the size of the motif) in the general case when the motif is a multiset in [12]. Indeed, in [12], a variant of Graph Motif that allows for insertions and deletions of colors was investigated, and a randomized fixed-parameter algorithm for this variant (hence also for Min-Add Graph Motif) was given.

Other variants of Graph Motif have been proposed. The List-colored Graph Motif problem (LGM) is a variant of Graph Motif where each vertex of the input graph is associated with a list of colors. LGM is NP-complete, since it is a generalization of the Graph Motif problem, but it is fixed-parameter tractable [6, 4, 12] when parameterized by the size of the motif. Two edge-weighted variants of the Graph Motif problem have been considered. The first variant, given an edge-weighted graph, asks for an occurrence of the motif, of minimum weight, in the input graph [12]. This variant admits a fixed-parameter algorithm if parameterized by the size of the motif and the weight of the solution [12]. The second variant asks for an occurrence that minimizes the weight of the edge-cut between the occurrence of the motif and the rest of the graph [5]. This variant is known to be fixed-parameter tractable when parameterized by the weight of the occurrence of $\mathcal{M}$ and by either the maximum degree of the input graph or the treewidth of graph [5].

Following this direction, we consider three variants of the Graph Motif problem. In the first two variants, we relax the constraint that each color of $\mathcal{M}$ must appear in an occurrence of the motif, and we allow for the adding (Minimum Adding Motif problem, Min-Add Graph Motif, introduced in [6]) or the substitution (Minimum Substitution Motif problem, Min-Sub Graph Motif) of some colors in $\mathcal{M}$. These two problems are motivated by the fact that, due to experimental errors, there may not exist an exact occurrence of the motif $\mathcal{M}$ in the graph G.

Then, we consider a third variant of the problem, Constrained Motif (CGM), where we strengthen the requirement of connectedness, constraining some vertices of the input graph to be part of an occurrence of a motif $\mathcal{M}$. This variant is motivated by the fact that, due to a previous knowledge on the structure of the network, we may require some of the vertices to be contained in any occurrence of $\mathcal{M}$. As an example, we can preprocess an instance of Graph Motif and define as mandatory vertex in an occurrence of $\mathcal{M}$ a vertex v that is the only vertex of the input graph colored by some $c \in \mathcal{M}$. Furthermore, we may know that only an element of a metabolic network has some specific function, hence we want to constrain the corresponding vertex to be in an occurrence of the motif.

The rest of the paper is organized as follows. In Section 2, we give some preliminary de nitions and we formally de ne the three problems studied in this paper. In Section 3, we show that Min-Sub Graph Motif and Min-Add Graph Motif are NP-hard, even when $\mathcal{M}$ is a set, the input graph is a tree T of degree bounded by 4 and each color has at most two occurrences in T. Notice that under the same hypotheses, the Graph Motif problem admits a polynomial-time algorithm [11, 23]. In Section 4, we give an FPT algorithm for Min-Sub Graph Motif, when the parameter is the size of the motif. In Section 5, we discuss the parameterized complexity of the CGM problem, when the parameter is the number of colors not belonging to mandatory vertices; in Section 5.2, we show that CGM is xed-parameter tractable for graphs of bounded treewidth, and in Section 5.3 we show that CGM is W[2]-hard, even if the input graph has diameter 2.

2 Preliminaries

In this section, we recall basic notations used in the rest of the paper. Given a graph $G = (V; E)$ and $V' \subseteq V$, 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 $\{u; v\} \in E$. Given a vertex $v \in V$, we denote by $N(v)$ the set of vertices in G adjacent to v. A graph is *cubic* when each vertex has degree exactly 3.

Let G be a connected graph, where every vertex $u \in V(G)$ is assigned a color $c(u)$ from a set $\mathcal{C}$ of colors. For any subset V' of V, let $c(V')$ be the multiset of colors assigned to the vertices in V'. Let $\mathcal{M}$ be a multiset of colors, whose colors are taken from the set $\mathcal{C}$. Given a colored graph G and a subset of vertices $V' \subseteq V(G)$, $c(V')$ is said to *match* a multiset of colors $\mathcal{M}$ if $c(V')$ is equal to $\mathcal{M}$. In this case, we say that V' matches $\mathcal{M}$. Given a subset of vertices $V' \subseteq V(G)$ such that V' matches $\mathcal{M}$ and $G[V']$ is connected, then V' is called an *occurrence* of $\mathcal{M}$ in G. A motif $\mathcal{M}$ is said to be *colorful* when each color appears at most once in $\mathcal{M}$.

In this paper, we consider three variants of the Graph Motif problem. For two of them, Minimum Adding Motif (Min-Add Graph Motif) and Minimum Substitution Motif (Min-Sub Graph Motif), we look for a vertex set V' of $G = (V; E)$, such that $G[V']$ is connected and $c(V')$ is not necessarily equal to $\mathcal{M}$. Let us de ne formally these two variants of Graph Motif.

Min-Add Graph Motif (decision version)
Input : A multiset of colors $\mathcal{M}$ over a set C of colors, a vertex-colored graph $G = (V; E)$ whose vertices are colored by $c : V \to C$, an integer p.
Question : Is there a subset $V' \subseteq V$, such that $G[V']$ is connected, $\mathcal{M} \subseteq c(V')$ and $|c(V') \setminus \mathcal{M}| \leq p$?

Min-Sub Graph Motif (decision version)
Input : A multiset of colors $\mathcal{M}$ over a set C of colors, a vertex-colored graph $G = (V; E)$ whose vertices are colored by $c : V \to C$, an integer p.
Question : Is there a subset $V' \subseteq V$, such that $G[V']$ is connected and $c(V')$ can be obtained from $\mathcal{M}$ with at most p substitutions?

Notice that for Min-Sub Graph Motif, $|c(V')| = |\mathcal{M}|$. Notice also that, in case $p = 0$, both Min-Add Graph Motif and Min-Sub Graph Motif are equivalent to the Graph Motif problem. As a consequence, Min-Add Graph Motif and Min-Sub Graph Motif are both NP-hard when the motif is colorful, the input graph consists of a tree T and each color has at most 3 occurrences in T [11]. Furthermore, by the NP-completeness of Graph Motif, it follows that Min-Add Graph Motif (resp. Min-Sub Graph Motif) cannot be approximated within any approximation factor, and does not admit any fixed-parameter tractable algorithm, when the parameter is the number of added colors (resp. the number of substitutions).

As discussed in Section 1, the Graph Motif problem admits a polynomial time algorithm when the input graph is a tree T and each color has at most two occurrences in T. Hence, we will focus on the computational complexity of Min-Add Graph Motif and Min-Sub Graph Motif for this restriction. Furthermore, since Min-Add Graph Motif is fixed-parameterized tractable when parameterized by $|\mathcal{M}|$ (see Section 1), we will focus on the parameterized complexity of Min-Sub Graph Motif, when parameterized by $|\mathcal{M}|$.

Let us now consider a different variant of the Graph Motif problem, called Constrained Graph Motif (CGM).

of Minimum (Unweighted) Steiner Tree we simply ignore the color of the vertices). A solution of the Minimum (Unweighted) Steiner Tree consists of a set V' of at most k non mandatory vertices that connects the mandatory vertices. The corresponding solution of CGM contains the vertices $V' \cup V_M$, and possibly other vertices to obtain an occurrence of $\mathcal{M}$. On the other side, consider a solution V^* of CGM . Then, the set $V' = V^* \setminus V_M$ consists of vertices colored by residue colors, and $|V'| = k$. Notice that V' allows for the connection of mandatory vertices of G. As the Minimum (Unweighted) Steiner Tree problem is W[2]-hard when parameterized by the number of non mandatory vertices [7], it follows that the CGM problem is W[2]-hard when parameterized by the number of residue colors.

In the rest of the paper, in order to extend some results from the case when $\mathcal{M}$ is colorful to the general case, we use the recoloring technique introduced in [4], based on the color-coding technique [3]. The recoloring technique starts from a general motif $\mathcal{M}$ on a set $\mathcal{C}$ of colors and computes a colorful motif $\mathcal{M}^*$ by recoloring accordingly the vertices of the input graph G, as follows: for each color $c \in \mathcal{C}$ that occurs h times in $\mathcal{M}$, de ne the colors $c_1^*;::::;c_h^*$ in $\mathcal{M}^*$. For each vertex v such that $c(v) = c$, recolor v with one of the color $c_1^*;::::;c_h^*$, randomly with probability $\frac{1}{h}$. Let V' be an occurrence of $\mathcal{M}$ in the graph G, then V' achieves a *colorful recoloring* if $c(V')$ is colorful after the recoloring of $\mathcal{M}$ and G. In [4], the following result was shown:

Lemma 1 (Betzler et al. [4]). *Given a motif $\mathcal{M}$, the number of trials to achieve a colorful recoloring of $\mathcal{M}$ with an error probability of " is* $|\ln(")| \cdot O(e^{|\mathcal{M}|})$.

Notice that, at the cost of an increase in the time complexity of the algorithm, the result of Lemma 1 can be derandomized using families of perfect hash functions [3].

3 NP-hardness of **Min-Sub Graph Motif** and **Min-Add Graph Motif**

In this section, we prove that Min-Sub Graph Motif and Min-Add Graph Motif are both NP-hard, even if the input graph is a tree, the motif is colorful and each color occurs at most twice in the input tree. Recall that, under the same hypotheses, the Graph Motif problem admits a polynomial-time algorithm, while Graph Motif is NP-hard even if the input graph is a tree, the motif is colorful and each color has at most three occurrences in the input tree.

Theorem 1. *The Min-Sub Graph Motif problem is* NP*-hard, even when the input graph is a tree of maximum degree* 4*, each color occurs at most twice in the input graph and the motif is colorful.*

Proof. We give a reduction from the Minimum Vertex-Cover on Cubic Graphs problem (Min-VCC) to Min-Sub Graph Motif. Let $G = (V; E)$ be a cubic graph with $V = \{v_1; v_2; :::; v_n\}$ and p be an integer ; the Min-VCC problem asks for a subset $V' \subseteq V$ of cardinality at most p, such that for each $\{u; v\} \in E$ at least one of $u; v$ is in V'. Min-VCC is known to be NP-hard [2]. Starting from G, we construct an instance of the Min-Sub Graph Motif problem which consists of a tree T and a set of colors $\mathcal{M}$. For any vertex $v_i \in V$, let $e_{i,x}$, $1 \leq x \leq 3$, be its 3 incident edges, ordered arbitrarily. The tree $T = (V_T; E_T)$ is de ned as follows (see Figure 1):

$$\begin{aligned} V_T = &\{l_i; a_{i,1}; a_{i,2} : 1 \leq i \leq n\} \cup \\ &\{s_i : 1 \leq i \leq p\} \cup \\ &\{t_i : 1 \leq i \leq n + 1\} \cup \\ &\{e_{i,x} : 1 \leq i \leq n \wedge 1 \leq x \leq 3\} \end{aligned}$$

and

$$\begin{aligned} E_T = &\{\{l_i; l_{i+1}\} : 1 \leq i < n\} \cup \\ &\{\{s_i; s_{i+1}\} : 1 \leq i < p\} \cup \\ &\{\{t_i; t_{i+1}\} : 1 \leq i < n + 1\} \cup \\ &\{\{l_n; t_1\}\} \cup \\ &\{\{t_{n+1}; s_1\}\} \cup \\ &\{\{l_i; a_{i,1}\}; \{a_{i,1}; a_{i,2}\} : 1 \leq i \leq n\} \cup \\ &\{\{a_{i,2}; e_{i,x}\} : 1 \leq i \leq n \wedge 1 \leq x \leq 3\}: \end{aligned}$$

Clearly, this construction gives us a tree of maximum degree 4. Let us describe the colors assigned to each vertex of $V(G)$. Each vertex l_i, $1 \leq i \leq n$, is assigned a unique color $c(l_i)$, each vertex s_i, $1 \leq i \leq p$, is assigned a unique color $c(s_i)$, and each vertex t_i, $1 \leq i \leq n + 1$, is assigned a unique color $c(t_i)$. The two vertices $a_{i,1}$, $a_{i,2}$, $1 \leq i \leq n$, are assigned the same color $c(v_i)$. Finally, each vertex $e_{i,x}$ in V_T, $1 \leq i \leq n$ and $1 \leq x \leq 3$, associated to an edge $\{v_i; v_j\}$ in E, is assigned color $c_{i,j}$. Each color occurs at most twice in T, as each color $c_{i,j}$ is associated to two vertices of T, while each color $c(v_i)$ is associated to two vertices $a_{i,1}$ and $a_{i,2}$. $\mathcal{M}$ is a set of colors de ned as follows: $\mathcal{M} = \{c(l_i) : 1 \leq i \leq$

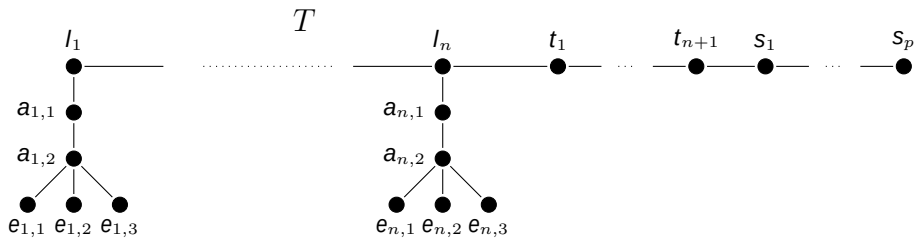

Fig. 1. Illustration of the reduction from Min-VCC to Min-Sub Graph Motif.

$n\} \cup \{c(s_i) : 1 \leq i \leq p\} \cup \{c(v_i) : 1 \leq i \leq n\} \cup \{c_{i,j} : \{v_i; v_j\} \in E\}$. Notice that no occurrence of a color $c(t_i)$, $1 \leq i \leq n+1$, belongs to $\mathcal{M}$.

Consider a vertex cover $V' \subseteq V$ of G of cardinality at most p, then we show that there exists a solution $V_{T'}$ of Min-Sub Graph Motif, that substitutes p colors from $\mathcal{M}$, as follows. The vertex set $V_{T'}$ de ned as follows:

$$\begin{aligned} V_{T'} = {} & \{l_i; a_{i,1} : 1 \leq i \leq n\} \cup \\ & \{t_i : 1 \leq i \leq p - |V'|\} \cup \\ & \{a_{i,2} : v_i \in V'\} \cup \\ & \{e_{i,x} : c(e_{i,x}) = c_{i,j}; 1 \leq i \leq n \wedge 1 \leq x \leq 3 \wedge i = \min(i; j)\}: \end{aligned}$$

By construction and since V' is a vertex cover, $V_{T'}$ induces a subtree of T. If we let $\mathcal{M}'$ stand for $c(V_{T'})$, we have $|\mathcal{M}'| = |\mathcal{M}|$, and $\mathcal{M}'$ can be obtained from $\mathcal{M}$ with p substitutions. Indeed there are exactly $|V'| \leq p$ colors appearing twice in $\mathcal{M}'$ and exactly once in $\mathcal{M}$, and these are exactly the colors $c(v_i)$, $v_i \in V'$. Each of these colors can be obtained from $\mathcal{M}$ by substituting a color $c(s_j)$ with a color $c(v_i)$. Furthermore, there are $p - |V'|$ colors $c(t_i)$, $1 \leq i \leq p - |V'|$, in $\mathcal{M}' \setminus \mathcal{M}$, each obtained by substituting a color $c(s_j)$ with a color $c(t_i)$.

Let us now consider a solution $V_{T'}$ of Min-Sub Graph Motif, where $c(V_{T'}) = \mathcal{M}'$, $|\mathcal{M}'| = |\mathcal{M}|$, and $\mathcal{M}'$ can be obtained from $\mathcal{M}$ with at most p substitutions. First, we show that $V_{T'}$ does not contain any vertex of the set $\{s_i : 1 \leq i \leq p\}$. Indeed, assume that a vertex s_i is part of $V_{T'}$; then, by construction, the set of vertices $\{t_j : 1 \leq j \leq n+1\}$ must belong to $V_{T'}$, and since $\mathcal{M}$ does not contain occurrences of any color $c(t_j)$, $1 \leq j \leq n+1$, it follows that $\mathcal{M}'$ requires at least $n+1$ substitutions. Notice that $n+1 > p$, as each vertex cover V' of G has cardinality at most n. Hence, we can assume that $V_{T'}$ does not contain any vertex in the set $\{s_i : 1 \leq i \leq p\}$. It follows that all the colors $c(s_i)$,

$1 \leq i \leq p$, in $\mathcal{M}$ must be substituted, and, since by hypothesis $\mathcal{M}'$ can be obtained from $\mathcal{M}$ with at most p substitutions, it follows that only the colors $c(s_i)$, $1 \leq i \leq p$, are substituted. Hence $\{l_i; a_{i,1} : 1 \leq i \leq n\} \subseteq V_{T'}$ and $\mathcal{M}' \supseteq \{c_{i,j} : \{v_i; v_j\} \in E\}$. Since $T[V_{T'}]$ must be connected, it follows that each vertex colored $c_{i,j}$ must be connected to some vertex $a_{i,2} \in V_{T'}$ colored by $c(v_i)$. De ne $V' = \{v_i : a_{i,2} \in V_{T'}\}$; it follows that V' is a vertex cover of G of cardinality at most p, which completes the proof. □

Next, in Theorem 2 we focus on the NP-hardness of Min-Add Graph Motif .

Theorem 2. *The Min-Add Graph Motif problem is* NP*-hard, even when the input graph is a tree of maximum degree* 4*, each color occurs at most twice in the input graph and the motif is colorful.*

Proof. We show the result by presenting a reduction from Minimum Vertex-Cover on Cubic Graphs (Min-VCC) to Min-Add Graph Motif (for a de nition of Min-VCC we refer the reader to the proof of Theorem 1). Notice that the reduction is similar to that given in Theorem 1. As for the reduction of Theorem 1, starting from $G = (V; E)$ we construct in polynomial-time an instance of the Min-Add Graph Motif problem which consists of a tree T and a set of colors $\mathcal{M}$. For any vertex $v_i \in V$, let $e_{i,x}$, $1 \leq x \leq 3$, be its 3 incident edges, ordered arbitrarily. First, we describe the construction of $T = (V_T; E_T)$. The tree T is de ned as follows (see Figure 2):

$$\begin{aligned} V_T = &\{l_i; a_{i,1}; a_{i,2} : 1 \leq i \leq n\} \cup \\ &\{e_{i,x} : 1 \leq i \leq n \wedge 1 \leq x \leq 3\} \end{aligned}$$

and

$$\begin{aligned} E_T = &\{\{l_i; l_{i+1}\} : 1 \leq i < n\} \cup \\ &\{\{l_i; a_{i,1}\}; \{a_{i,1}; a_{i,2}\} : 1 \leq i \leq n\} \cup \\ &\{\{a_{i,2}; e_{i,x}\} : 1 \leq i \leq n \wedge 1 \leq x \leq 3\}. \end{aligned}$$

Clearly, this construction gives us a tree of maximum degree 4. Now, let us describe the colors assigned to each vertex of V_T. Each vertex l_i, $1 \leq i \leq n$, is assigned a unique color $c(l_i)$, $1 \leq i \leq n$. The two vertices $a_{i,1}$ and $a_{i,2}$, $1 \leq i \leq n$, are assigned the same color $c(v_i)$. Finally, each vertex $e_{i,x}$, $1 \leq i \leq n$ and $1 \leq x \leq 3$, is associated to an edge $\{v_i; v_j\}$ in E and it is assigned color $c_{i,j}$. Each color occurs at most twice in T,

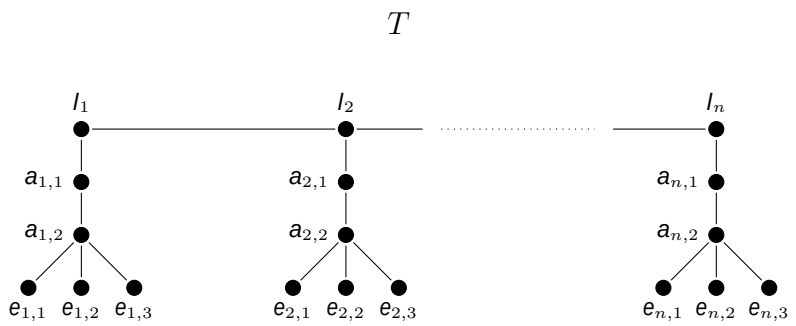

Fig. 2. Illustration of the reduction from Min-VCC to Min-Add Graph Motif.

as each color $c_{i,j}$ is associated to two vertices of V_T, while color $c(v_i)$ is associated to vertices $a_{i,1}$ and $a_{i,2}$. $\mathcal{M}$ is a set of colors de ned as follows: $\mathcal{M} = \{c(l_i) : 1 \leq i \leq n\} \cup \{c(v_i) : 1 \leq i \leq n\} \cup \{c_{i,j} : \{v_i; v_j\} \in E\}$.

Now, consider a vertex cover $V' \subseteq V$ of G of size p. Then there exists a solution $T' = (V_{T'}; E_{T'})$ of Min-Add Graph Motif that adds at most p colors, de ned as follows. The subtree T' of T is induced by the vertex set $V_{T'}$ de ned as follows:

$$\begin{aligned} V_{T'} = &\{l_i; a_{i,1} : 1 \leq i \leq n\} \cup \\ &\{a_{i,2} : v_i \in V'\} \cup \\ &\{e_{i,x} : c(e_{i,x}) = c_{i,j}; 1 \leq i \leq n \wedge 1 \leq x \leq 3 \wedge i = \min(i; j)\}. \end{aligned}$$

By construction and since V' is a vertex cover, $V_{T'}$ induces a subtree of T. Furthermore, $c(V'_T) = \mathcal{M}' \supseteq \mathcal{M}$. Indeed each color $c(l_i)$, $1 \leq i \leq n$, each color $c(v_i)$, $1 \leq i \leq n$, and each color $c_{i,j}$, where $\{v_i; v_j\} \in E$, appears in $\mathcal{M}'$. Notice that by construction $\mathcal{M}' \setminus \mathcal{M}$ is the set colors $c(v_i)$, with $v_i \in V'$. As a consequence $|\mathcal{M}' \setminus \mathcal{M}| \leq p$.

Let us consider now a solution $V_{T'}$ of Min-Add Graph Motif, where $c(V_{T'}) = \mathcal{M}'$ and let $|\mathcal{M}' \setminus \mathcal{M}| \leq p$. First, notice that each color in the set $\{c(l_i) : 1 \leq i \leq n\} \cup \{c_{i,j} : \{v_i; v_j\} \in E\}$ occurs once in $\mathcal{M}$. Since each color in the set $\{c_{i,j} : \{v_i; v_j\} \in E\}$ is associated to a leaf of T, it follows that none of the colors in $\{c(l_i) : 1 \leq i \leq n\} \cup \{c_{i,j} : \{v_i; v_j\} \in E\}$ will be in $\mathcal{M}' \setminus \mathcal{M}$. As a consequence, it follows that $\mathcal{M}' \setminus \mathcal{M}$ contains only those colors $c(v_i)$, where v_i is a vertex of the graph G, and in this case $\mathcal{M}'$ will contain two occurrences of $c(v_i)$ and the vertices $a_{i,1}$, $a_{i,2}$ will be in $V_{T'}$. Since each $c_{i,j} \in \mathcal{M}'$, it follows that each vertex having color $c_{i,j}$ is adjacent in $V_{T'}$ to a vertex $a_{i,2}$, colored by $c(v_i)$. De ning $V' = \{v_i \in V : a_{i,2} \in V_{T'}; 1 \leq i \leq n\}$, it follows that V' is a vertex cover of G of cardinality at most p, which completes the proof. □

4 Parameterized Complexity of **Min-Sub Graph Motif**

In this section, we discuss the parameterized complexity of Min-Sub Graph Motif, when parameterized by $|\mathcal{M}|$. We recall that, as discussed in Section 2, Min-Sub Graph Motif and Min-Add Graph Motif are not in FPT when parameterized by the size of the solution, and that Min-Add Graph Motif is in FPT when parameterized by $|\mathcal{M}|$. Hence we consider the parameterized complexity of the Min-Sub Graph Motif problem, when the parameter is $|\mathcal{M}|$, and we exhibit a fixed-parameter algorithm for this case.

Let us first consider the case where the motif $\mathcal{M}$ is colorful (i.e., $\mathcal{M}$ is a set). The algorithm is based on dynamic programming. Let $(G = (V; E); \mathcal{M})$ be an instance of Min-Sub Graph Motif. First, we assume that each connected component P of G has size at least $|\mathcal{M}|$, otherwise we can remove P from G as there is no solution of Min-Sub Graph Motif that includes a vertex of P.

Define a *k-partial occurrence* of a (multi)set of colors C in G as a set of k vertices $V_T \subseteq V$ such that $G[V_T]$ is connected and there exists a subset $V_{T,C} \subseteq V_T$, that matches a subset $C' \subseteq C$.

Given a vertex v of the input graph G, a subset of colors $C \subseteq \mathcal{M}$, and a value $k > 0$, define $S[v; C; k]$ as the minimum value z required by a k-partial occurrence V_T of C in G such that:

1. $v \in V_T$,
2. $|V_{T,C}| = q$, and
3. $z + q = k$.

Now, let us define the dynamic programming recurrence to compute $S[v; C; k]$:

$$S[v; C; k] = \min_{C' \subseteq C, u \in N(v), k_1 + k_2 = k} \{ S[v; C'; k_1] + S[u; C \setminus C'; k_2] \} : \quad (1)$$

For the base cases: $S[v; C; 1] = 0$, when $c(v) \in C$, for each $C \subseteq \mathcal{M}$, and $v \in V$, and $S[v; C; 1] = 1$ when $c(u) \notin C$, for each $C \subseteq \mathcal{M}$. We now prove the correctness of Recurrence (1).

Lemma 2. *Let $(G; \mathcal{M})$ be an instance of Min-Sub Graph Motif, v be a vertex of G, and C be a subset of $\mathcal{M}$. There is a k-partial occurrence V_T of C in G, with $v \in V_T$ and $|V_{T,C}| = q$, if and only if there exists an entry $S[v; C; k] = z$, where $z + q = k$.*

Proof. Let us consider the case $S[v; C; k] = z$. In the basic case, $S[v; C; 1] \in \{0; 1\}$ depending whether $c(v) \in C$ or $c(v) \notin C$. Let

$S[v; C'; k_1] = z_1$ and $S[u; C \setminus C'; k_2] = z_2$, where $z_1 + z_2 = z$, $k = k_1 + k_2$ and u, v are adjacent in G. By induction, there exists a k_1-partial occurrence V_{T_1} of C' such that $v \in V_{T_1}$, with $|V_{T_1}| = k_1$ and $|V_{T_1,C'}| = q_1$, where $q_1 + z_1 = k_1$, and there exists a k_2-partial occurrence V_{T_2} of $C \setminus C'$ such that $u \in V_{T_2}$, with $|V_{T_2}| = k_2$ and $|V_{T_2,C\setminus C'}| = q_2$, where $q_2 + z_2 = k_2$.

Consider a color c and assume that there exists a vertex $v_1 \in V_{T_1}$ and a vertex $v_2 \in V_{T_2}$, such that $c(v_1) = c(v_2) = c$. Observe that at most one of v_1, v_2 can match a color of C' and $C \setminus C'$, as these two sets are disjoint. Furthermore, notice that we can assume that if a color c is matched by a vertex $V_{T,C}$, then either $c \in c(V_{T_1,C'})$ (hence it contributes to the value q_1) or either $c \in c(V_{T_2,C'})$ (hence it contributes to the value q_2). It follows that $q = q_1 + q_2$. Notice that the two sets V_{T_1} and V_{T_2} may share some vertices. In this case, the k-partial occurrence V_T of C is obtained by replacing the vertices in $V_{T_1} \cap V_{T_2}$ with $|V_{T_1} \cap V_{T_2}|$ vertices adjacent to the connected component of G induced by $V_{T_1} \cup V_{T_2}$.

Assume that there exists a k-partial occurrence V_T of C in G such that $|V_{T,C}| = q$. Let us first consider the basic cases. If we are considering a single vertex v, then $S[v; C'; 1] = 0$ if $c(v) \in C'$, that is the color $c(v)$ is matched, while $S[v; C'; 1] = 1$ if $c(v) \notin C'$.

Assume that $|V_T| = k > 1$; we can then find a vertex $u \in V_T$ where $\{u; v\} \in E$, and a set $C' \subseteq C$, such that V_T can be split into a k_1-partial occurrence V_{T_1} of C' in G, with $|V_{T_1,C'}| = q_1$ and $v \in V_{T_1}$, and a k_2-partial occurrence V_{T_2} of $C \setminus C'$ in G, with $|V_{T_2,C\setminus C'}| = q_2$ and $u \in V_{T_2}$, where it holds $k_1 + k_2 = k$ and $V_{T_1} \cap V_{T_2} = \emptyset$. The bipartition of C into C' and $C \setminus C'$ is such that if a color c is in $c(V_{T_1})$, then $c \in C'$, else $c \in C \setminus C'$. By construction, $q = q_1 + q_2$. By induction, $S[v; C; k_1] = z_1$ and $S[u; C \setminus C'; k_2] = z_2$, where $z_1 = k_1 - q_1$, and $z_2 = k_2 - q_2$. Applying Recurrence (1), it follows that $S[v; C; k] = z$. □

An optimal solution for Min-Sub Graph Motif can be computed from the minimal value z in the entries $S[v; \mathcal{M}; |\mathcal{M}|]$, with $v \in V$. Indeed, an $|\mathcal{M}|$-partial occurrence V_T of $\mathcal{M}$ is a solution of Min-Sub Graph Motif requiring $|\mathcal{M}| - |V_{T,|\mathcal{M}|}|$ substitutions.

The time complexity of the algorithm is $O(3^{|\mathcal{M}|}n^4)$. Indeed, there are at most $3^{|\mathcal{M}|}$ bipartitions of a subset C of $\mathcal{M}$, for each possible subset C of $\mathcal{M}$, while, for each vertex $v \in V$, there are at most n choices for vertex u adjacent to v, and there are at most n choices for the value of k_1 (and hence of k_2). In order to extend the results to a multiset, we apply the recoloring technique described in [4] and briefly recalled in Section 2. Combining Lemma 2 with Lemma 1, we obtain a randomized

xed-parameter algorithm for Min-Sub Graph Motif of time complexity $O^*((3e)^{O(|\mathcal{M}|)})$.

5 Parameterized Complexity of CGM

In this section, we consider the parameterized complexity of CGM, where the parameter is the number k of residue colors, that is $k = |R$ there

th]TJ/F43 7.9701 Tf 8.288.31.6361 T

5.2 A Fixed-Parameter Algorithm for Graphs of Bounded Treewidth

First, we show that CGM is in FPT when the input graph is a tree.

Theorem 3. *The CGM problem, parameterized by the number k of residue colors, is in FPT when the input graph is a tree.*

Proof. Let $(T = (V; E); \mathcal{M}; V_M)$ be an instance of CGM, where T is a tree. Let k be the number of residue colors. First x a vertex $v' \in V_M$ as the root of T. Now, denote by $V_A \subseteq V$ a set of vertices that must be part of any feasible solution of CGM . Let V_A be the uniquely de ned set of vertices such that $V_A \cup V_M$ is connected in T.

Obviously, each vertex in V_A must be part of any feasible solution of CGM over instance $(T; \mathcal{M}; V_M)$. Notice that if $c(V_A) \not\subseteq \mathcal{M}$, then there is no solution to CGM.

Let $\mathcal{M}' = \mathcal{M} \setminus c(V_A)$. Let T^* be a tree obtained from T by contracting to a single vertex r^* the subtree $T[V_A]$. Since T is rooted at v' and $v' \in V_A$, it follows that r^* is the root of T^*. Assign a new color $c(r^*)$ to the root of r^* and de ne $\mathcal{M}^* = \mathcal{M}' \cup \{c(r^*)\}$. It follows that the CGM problem over input $(T; \mathcal{M}; V_M)$ is reduced to the Graph Motif problem over input $(T^*; \mathcal{M})$. As $|\mathcal{M}^*| \leq k + 1$, and as the Graph Motif problem is xed-parameter tractable when the parameter is the size of the motif [12], the result follows. □

Next, we describe a xed-parameter algorithm for CGM for graphs of bounded treewidth. Let $(G = (V; E); \mathcal{M}; V_M)$ be an instance of CGM, and let us rst consider the case where the motif $\mathcal{M}$ is colorful.

Recall that we denote by k the number of residue colors, and by the treewidth of graph G. The algorithm is based on a nice tree decomposition of G (see Section 2 for a de nition). We also consider a slightly more general problem, where we look for an occurrence of a motif, such that the occurrence consists of at most + 1 connected components. The di erent connected components are induced by a partition of a bag X_i of the nice tree decomposition. Given a vertex i of the nice tree decomposition of G, we denote by $T[i]$ the subtree of the nice tree decomposition rooted at i and we let $V(T[i]) = \{u \in X_j : j \in T[i]\}$.

Now, consider a set X_i, $1 \leq i \leq p$, of the nice tree decomposition $\langle\{X_i; i \in \{1; : : : ; p\}\}; T\rangle$. From the de nition of treewidth, it follows that $|X_i| \leq$ + 1. Now, let us de ne a mapping function f_i associated to the vertices of X_i, as follows.

De nition 2. *Let X_i, $1 \leq i \leq p$, be a bag of the nice tree decomposition of G. A mapping function f_i from X_i to $\{0;1;:::; \ +1\}$ is* feasible *when*

1. *$f_i(v) \neq 0$ for each mandatory vertex v in X_i,*
2. *for each pair of vertices $u;v \in X_i$ such that $c(u) = c(v)$, then $f_i(u) = 0$ or $f_i(v) = 0$,*
3. *for any xed l, with $1 \leq l \leq \ +1$, de ne $X_i^l = \{v \in X_i : f(v_i) = l\}$ and $X_i' = \cup_{h=1}^{l} X_i^h$; then X_i^l is a maximal connected component of $G[X_i']$.*

Informally, a feasible mapping f_i represents a partition of a subset $X_i' \subseteq X_i$ into at most $\ +1$ connected components, where $f_i(v) = p \neq 0$ implies that v belongs to the p-th connected component of X_i', while $f_i(v) = 0$ implies that v does not belong to X_i'.

De nition 3. *Let W be a set of vertices of $V(T[i])$, consisting of the connected components $W_1;W_2;:::;W_z$. Let f_i be a feasible mapping from X_i to $\{0;1;::: \ +1\}$, then W is mapped (or partitioned)* according *to f_i if:*

1. *for each p, $1 \leq p \leq z$, $W_p \cap X_i \neq \emptyset$, and there exists exactly one l, $1 \leq l \leq \ +1$, such that $W_p \cap X_i = X_i^l$;*
2. *for each l, $1 \leq l \leq \ +1$, such that $X_i^l \neq \emptyset$, there exists exactly one p, $1 \leq p \leq z$, such that $X_i^l = W_p \cap X_i$.*

Notice that by De nition 3, if a vertex u of W is not in X_i, then there exists a vertex v in $W \cap X_i$ such that v and u are in the same connected component W_x of W, v is assigned some label $l \neq 0$, and all the vertices of $W_x \cap X_i$ are assigned the same label l.

Given two sets X_i and X_j of a nice tree decomposition, and two feasible mappings $f_i : X_i \to \{0;:::; \ +1\}$ and $f_j : X_j \to \{0;:::; \ +1\}$, then f_i and f_j are *consistent* if, for each $v \in X_i \cap X_j$, $f_i(v) = f_j(v)$.

Let i be a vertex of the nice tree decomposition, with exactly one child j, such that $|X_i| = |X_j| + 1$ and $X_j \subset X_i$, with $v \in X_i \setminus X_j$, then a feasible mapping f_i is an *extension* of a feasible mapping f_j, when either:

1. $f_i(v) = 0$, or

2. $f_i(v) = l$, $l \in \{1;:::; \ +1\}$, $f_i(u) \neq l$ for each $u \in X_i \cap X_j$, and f_i, f_j are consistent, or

3. there exists a value $l \in \{1;:::; \ +1\}$ such that
   (a) $f_i(v) = l$,

(b) for any $z \in X_i \cap X_j$, if $f_j(z) = 0$ then $f_i(z) = 0$,
(c) for any $z \in X_i \cap X_j$ such that $f_j(z) \neq 0$, if $f_i(z) \neq f_j(z)$ then $f_i(z) = l$.

Given a feasible mapping f_i from X_i to $\{0; 1; :::\ + 1\}$, de ne $c(X_i; f_i) = \{c \in R_c : \exists v \in X_i;\ c(v) = c \wedge f_i(v) \neq 0\}$.

De ne the value $S[i; f_i; C']$, where i is a vertex of the nice tree decomposition of G, f_i is a feasible mapping function from X_i to $\{0; 1; :::\ + 1\}$ and $C' \subseteq R_c$ is a subset of the residue colors. $S[i; f_i; C'] = 1$ when there exists a set W of vertices in the nice tree decomposition rooted at i, such that the vertices of W can be partitioned according to f_i, where each mandatory vertex of $T[i]$ is in W, and the set of residue colors in $c(W)$ is C' ; else $S[i; f_i; C'] = 0$. Next, we describe how to compute $S[i; f_i; C']$ by dynamic programming, depending on three di erent cases of a nice tree decomposition.

Case 1) Assume that vertex i has two children j and h (recall that $X_i = X_j = X_h$), then

$$S[i; f_i; C'] = \bigvee_{f_j, f_h, C_j, C_h} S[j; f_j; C_j] \wedge S[h; f_h; C_h];$$

where f_i, f_j, f_h are all feasible and consistent, $C' = (C_j \cup C_h)$ and $c(X_i; f_i) = C_j \cap C_h$.

Case 2) Assume that i has exactly one child j, such that $X_i = X_j \cup \{v\}$, then

$$S[i; f_i; C'] = \bigvee_{f_j, C_j} S[j; f_j; C_j];$$

where f_i and f_j are feasible, f_i is an extension of f_j, and (i) $C' = C_j \cup \{c(v)\}$ and $c(v) \notin C_j$, if $f_i(v) \neq 0$ and $v \in V_M$; (ii) $C' = C_j$ otherwise.

Case 3) Assume that X_i has exactly one child X_j, such that $X_i = X_j \setminus \{v\}$, then

$$S[i; f_i; C'] = \bigvee_{f_j} S[j; f_j; C'];$$

where f_i and f_j are feasible and consistent, and, when $f_j(v) \neq 0$, there is a vertex $z \in X_i \cap X_j$, such that $f_j(z) = f_j(v)$.

For the base cases (when X_i is a leaf of the nice tree decomposition), de ne $S[i; f_i; C'] = 1$ if there is a partition of the vertices of X_i according to the feasible function f_i, and if $c(X_i; f_i) = C'$; otherwise, de ne $S[i; f_i; C'] = 0$.

First, we prove the correctness of the above recurrences, then we discuss the time complexity of the algorithm.

Lemma 3. *Let f_i be a feasible mapping function from X_i to $\{0; 1; : : : ; \ + 1\}$, and let W be a set of vertices in $V(T[i])$, such that W contains all the mandatory vertices in $V(T[i])$, W can be mapped according to f_i and C' is the set of residue colors in $c(W)$. Then $S[i; f_i; C'] = 1$.*

Proof. Let C' be a set of residue colors and let $T[i]$ be the subtree of T rooted at vertex i. Consider a set of vertices W in $V(T[i])$ such that the set of residue colors of $c(W)$ is C'. Let us consider a function f_i, such that the vertices of W are mapped according to function f_i. In what follows, we will prove by induction that $S[i; f_i; C'] = 1$. The lemma holds in the base case, by de nition, when i is a leaf of the nice tree decomposition. By the induction hypothesis, assume that the lemma holds for any vertices j of $T[i]$ di erent from the root of $T[i]$. Now, we will consider three possible cases for the tree decomposition.

Assume that we are in Case 1) of the tree decomposition. Recall that $X_i = X_j = X_h$. By the property of tree decomposition, as the sets of the tree decomposition containing a vertex of G must be connected, any vertices in $V(T[j]) \cap V(T[h])$ must belong also to X_i. It follows that each color c in C' is covered either by a vertex in X_i, or by a vertex v of exactly one of $V(T[j])$, $V(T[h])$ (notice that $v \in X_i$). Let C_j, C_h be the two sets of residue colors covered by the vertices of $V(T[j])$ and $V(T[h])$ respectively. Now, consider the subset W_y of vertices of W contained in $V(T[y])$, $y \in \{j; h\}$, and let f_y be a function consistent with f_i. Since the vertices of W can be mapped according to f_i, the vertices of W_y can be partitioned according to f_y. By induction hypothesis, it follows that $S[j; f_j; C_j] = 1$ and $S[h; f_h; C_h] = 1$, hence it follows that $S[i; f_i; C'] = 1$.

Assume that we are in Case 2) of the tree decomposition. Recall that $X_i = X_j \cup \{v\}$. Let us consider two cases, depending whether v belongs to W. First, if v does not belong to W, then $f_i(v) = 0$ and each vertex of W belongs to $V(T[j])$, hence considering the function f_j such that f_i is an extension of f_j, it holds by induction hypothesis $S[j; f_j; C'] = 1$. Applying Case 2) of the recurrence, there is a mapping f_i such that f_i and f_j are consistent, and $S[i; f_i; C'] = 1$. Second, if v belongs to W, it follows that the set of residue colors covered by W with vertices in $V(T[j])$ is $C_j = C' \setminus$

$\{c(v)\}$ if $v \in V_M$, else $C_j = C'$. Now consider the connected components of W induced by f_i. After removing v, the connected components will (possibly) split into at most $<$ + 1 connected components $S_1;:::;S_\gamma$. Each connected component S_h, $1 \leq h \leq$, contains a vertex y_h adjacent to v and, by the property of tree decomposition, $y_h \in X_i \cap X_j$. Consider a function f_j of X_j such that f_i is an extension of f_j. Then, the set of vertices in $S \setminus \{v\}$ is mapped according to f_j, and by induction hypothesis $S[j; f_j; C_j] = 1$. Applying Case 2) of the recurrence, $S[i; f_i; C'] = 1$.

Assume nally that we are in Case 3) of the tree decomposition. Since $X_j = X_i \cup \{v\}$, then the set of residue colors covered by vertices of W in $V(T[i])$ is equal to the set of residue colors covered by vertices of W in $V(T[j])$. Observe that, by the property of the tree decomposition, all the vertices adjacent to v in G belong to $V(T[j])$. If v does not belong to W, then de ning f_j as the function consistent with f_i and such that $f_j(v) = 0$, it holds $S[j; f_j; C'] = 1$ and hence $S[i; f_i; C'] = 1$. If, on the contrary, v belongs to W, either $C' = \emptyset$ and $f_i(u) = 0$, for each $u \in X_i$, or there must be a vertex u of $W \cap X_i$ adjacent to v in W such that $f_i(u) \neq 0$, otherwise v is isolated and cannot be mapped according to a mapping f_i. Hence we can de ne a mapping function f_j which is consistent with f_i, such that $f_i(u) = f_j(u) = f_j(v)$ and W can be partitioned according to f_j. By induction hypothesis $S[j; f_j; C'] = 1$, and applying Case 3) of the recurrence, it follows that $S[i; f_i; C'] = 1$. □

Lemma 4. *Let $S[i; f_i; C'] = 1$ for a feasible mapping function f_i from X_i to $\{0;1;:::;\ + 1\}$, then there exists a set W of vertices in $V(T[i])$ such that the set of residue colors in $c(W)$ is C', W contains all the mandatory vertices in $V(T[i])$ and the vertices of W can be mapped according to f_i.*

Proof. First notice that, since each mapping function f_h considered, with h a vertex of $T[i]$, is feasible, then each mandatory vertex that belongs to X_h is assigned by f_h a label $l \neq 0$. By construction, it follows that $S[i; f_i; C'] = 1$ only if each mandatory vertex is assigned a label di erent from 0, hence it must belong to any set of vertices that can be mapped according to f_i.

The lemma holds in the base case, by de nition, when i is a leaf of the nice tree decomposition. By the induction hypothesis, we assume that the property holds for each vertex j in $T[i]$ di erent from i, that is if $S[j; f_j; C_j] = 1$, then there exists a set of vertices W' in $V(T[i])$ such that the set of residue colors of $c(W)$ is C_j and such that the vertices of W' can be mapped according to f_j, for some feasible function f_j. Now, let us consider the three possible cases for the tree decomposition.

Assume that we are in Case 1), that is vertex i has exactly two children j and h in the nice tree decomposition of G. Notice that, by the property of a tree decomposition, all the vertices that are contained in both $V(T[j])$ and $V(T[h])$ must belong to X_i. Since $S[i; f_i; C']$ = 1, it follows that there must exist two sets C_j, C_h such that $C' = C_j \cup C_h$ and $C_j \cap C_h = c(X_i; f_i)$, and two feasible functions f_j, f_h, both consistent with f_i, such that $S[j; f_j; C_j]$ = 1 and $S[h; f_h; C_h]$ = 1. This implies that there exists two sets of vertices W_j, W_h in $V(T[j])$ and $V(T[h])$ such that the set of residue colors of $c(W_j)$ and $c(W_h)$ are C_j and C_h respectively, and such that W_j and W_h can be mapped according to functions f_j, f_h respectively. Since f_i must be consistent with f_j and f_h, and since $X_i = X_j = X_h$, it follows that there exists a set of vertices W in $V(T[i])$ such that the set of residue colors of $c(W)$ is C' and such that W can be mapped according to the mapping function f_i.

Assume that we are in Case 2), that is, vertex i has exactly one child j in the nice tree decomposition of G, and $X_i = X_j \cup \{v\}$. Let f_j be the mapping function of X_j such that f_i is an extension of f_j. Now assume that $f_i(v) = l$ and $f_i(u) \neq l$ for each $u \neq v$. By induction, there exists a set W_j of vertices in $V(T[j])$ such that the set of residue colors of $c(W_j)$ is C_j (notice that C_j is equal to $C' \setminus c(v)$ if $f_i(v) \neq 0$ and $v \in V_M$, and $C_j = C'$ otherwise), and such that the vertices of W_j can be mapped according to f_j. Then there exists a set W of vertices such that the set of residue colors of $c(W)$ is C' and such that W is mapped according to f_i. Assume that $f_i(v) = l$ and that there exists a vertex $u \in X_i \cap X_j$ such that $f_i(u) = l$. By construction, for each vertex $u \in X_j$ such that $f_j(u) \neq f_i(u)$, we must have $f_i(u) = f_i(v) = l$. Furthermore, for each $u; w \in X_i \cap X_j$, such that $f_j(u) = f_j(w)$, it holds $f_i(u) = f_i(w)$. Now, by the induction hypothesis, since $S[j; f_j; C_j]$ = 1, it follows that there exists a set of vertices W_j in $V(T[j])$ such that the set of residue colors of $c(W_j)$ is C_j, and such that the vertices of W_j are mapped according to function f_j. Since f_i is an extension of f_j, by construction, the set of vertices mapped in l by f_i are in a connected component, hence there exists an occurrence W in $V(T[i])$ such that the set of residue colors of $c(W)$ is C' (notice that C_j is equal to $C' \setminus c(v)$ if $v \in V_M$, and $C_j = C'$ otherwise), and such that the vertices are mapped according to f_i.

Assume that we are in Case 3), that is, vertex i has exactly one child j in the nice tree decomposition of G, and $X_i = X_j \setminus \{v\}$. Let f_j be the function over X_j such that f_i and f_j are consistent. If $f_j(v) = 0$, since f_i and f_j must be consistent, it follows that the lemma holds. If vertex v is such that $f_j(v) = l$, with $l \neq 0$, by construction there is a vertex

u of $X_i \cap X_j$ with $f_i(u) = f_j(u) = l$ connected to v. By the induction hypothesis, it follows that there is a set of vertices W in $V(T[j])$ such that the set of optional occurrences of $c(W)$ is C_j, and such that the vertices of W can be mapped according to function f_j. Hence the vertices of W are mapped also according to function f_i, since v is in the same connected component as u, and f_i and f_j are consistent. □

Lemma 5 shows how the values $S[i; f_i; C']$ are used to compute the existence of a feasible solution of CGM.

Lemma 5. *Let $(G = (V; E); \mathcal{M}; V_M)$ be an instance of the CGM problem. Then there is a solution W for CGM over instance of $(G; \mathcal{M}; V_M)$ if and only if there is a vertex i of the nice tree decomposition of G and a feasible function f_i that maps X_i to $\{0; x\}$, $1 \leq x \leq \quad +1$, such that $S[i; f_i; R_c] = 1$ and $V_M \subseteq V(T[i])$.*

Proof. Assume that there is a vertex i of the nice tree decomposition of G and a feasible function f_i that maps X_i to $\{0; x\}$, $1 \leq x \leq \quad + 1$, such that $S[i; f_i; R_c] = 1$ and all the mandatory vertices of G are in $T[i]$. By Lemma 4, it follows that there is a set of vertices W in $V(T[i])$ that contains all the mandatory vertices of G, such that the set of optional occurrences in $c(W)$ is R_c and such that the vertices of W can be mapped according to f_i. Furthermore, notice that W consists of a single connected component. Hence W is a solution of CGM.

Consider the case where there is a solution W of CGM over an instance $(G = (V; E); \mathcal{M}; V_M)$. Consider a vertex i of the tree decomposition of G such that all the vertices of W are contained in $V(T[i])$. By Lemma 3, it follows that $S[i; f_i; R_c] = 1$ for some feasible function f_i that maps X_i to $\{0; x\}$, with $1 \leq x \leq \quad + 1$. □

Now, we discuss the time complexity of the above algorithm. Denote by n the cardinality of V. Given a vertex i and the associated set X_i of the nice tree decomposition, the number of possible mapping functions from X_i to $\{0; : : : ; \quad + 1\}$ is $O(\ ^{\delta})$. The number of possible subsets C' is $O(2^k)$. Since the number of vertices of a nice tree decomposition is $O(n)$, it follows that we have $O(\ ^{\delta} n 2^k)$ entries $S[i; f_i; C]$. Given a mapping function f_i from X_i to $\{0; : : : ; \quad + 1\}$, computing an entry $S[i; f_i; C]$, given the entries of the child(ren) of i, requires time at most $O(\ ^{2\delta} 2^{2k})$ (notice that the worst case occurs when i has two children). Hence the total time complexity is $O(\ ^{3\delta} n 2^{3k})$.

The algorithm can be extended to the case when a motif is a multiset of colors, applying the recoloring technique presented in Sec-

tion 2 (see Lemma 1). As a consequence, we obtain a randomized xed-parameter algorithm for CGM over graphs of treewidth of time complexity $O(|\ln(")|\ ^{\delta} n 2^{4.4427k})$.

5.3 Hardness of Parameterization

The CGM problem parameterized by the number of optional occurrences is W[2]-hard, as stated in Section 2. Here we strengthen the result, showing that the problem is W[2]-hard even when the input graph has diameter 2.

Notice that the diameter of the input graph is a parameter of particular interest in biological networks. Indeed, many biological networks exhibit the so called \small world network phenomenon" [24]. In particular, it has been observed that metabolic networks have usually a small diameter, in many cases ranging from 3 to 5 [16].

Theorem 4. *The CGM problem, parameterized by the number of optional occurrences, is W[2]-hard, even when the input graph has diameter 2.*

Proof. We give a parameterized reduction from Minimum Set Cover (Min-SC) to CGM. Let us rst introduce the Min-SC problem. Given a universe $U = \{u_1; :::; u_n\}$, a collection of sets $\mathcal{S} = \{S_1; :::; S_m\}$ over U and an integer k, the goal of Min-SC is to compute a collection $\mathcal{S}'$ of at most k sets of $\mathcal{S}$, such that $\bigcup_{S'_i \in \mathcal{S}'} S'_i = U$. Min-SC is known to be W[2]-hard when the parameter is k [18].

Let $(U; \mathcal{S})$ be an instance of Min-SC, de ne a corresponding instance $(G = (V; E); \mathcal{M}; V_M)$ of the CGM problem as follows. The graph G is de ned as follows (see Figure 3):

$$
\begin{aligned}
V &= \{r\} \cup \{r'\} \cup \{v_{u,j} : 1 \leq j \leq n\} \cup \{v_{S,i} : 1 \leq i \leq m\} \\
E &= \{r; r'\} \cup \{\{r; v_{S,i}\}; \{\{r'; v_{S,i}\} : 1 \leq i \leq m\} \cup \\
&\quad \{v_{S,i}; v_{u,j}\} : 1 \leq i \leq m \wedge u_j \in S_i\} \cup \{r'; v_{u,j}\} : 1 \leq j \leq n\}:
\end{aligned}
$$

It is easy to see that G has diameter 2 since by construction, vertex r' is connected to every other vertex in V. Let us now de ne the coloring of V. Vertices r and r' are both colored by $c(r)$, vertex $v_{S,i}$, $1 \leq i \leq m$, is colored by $c(S)$, and vertex $v_{u,j}$, $1 \leq j \leq n$, is colored by $c(u_j)$. The motif $\mathcal{M}$ is a multiset containing one occurrence of color $c(r)$, one occurrence of each color $c(u_j)$, $1 \leq j \leq n$, and k occurrences of color $c(S)$. Let us now de ne the set of mandatory vertices: $V_M = V \setminus (\{v_{S,i} : 1 \leq i \leq m\} \cup \{r'\})$.

First, let us prove that given a solution of Min-SC of cardinality at most k, then we can compute in polynomial-time an occurrence of motif

$\mathcal{M}$ in G. Let $\mathcal{S}'$ be a solution of Min-SC of cardinality at most k. We can assume that $\mathcal{S}'$ contains exactly k sets, otherwise we can add some sets of $\mathcal{S}$ to $\mathcal{S}'$. Let $\mathcal{S}' = \{S_{q_1}; S_{q_2}; \ldots S_{q_k} : 1 \leq q_i \leq m\}$. Let us de ne T as the subgraph of G induced by the following vertices: $V_T = \{r\} \cup \{v_{S,q_i} : 1 \leq i \leq k\} \cup \{u_j : 1 \leq j \leq n\}$. V_T is an occurrence of $\mathcal{M}$, as $c(V_T)$ contains color $c(r)$, k occurrences of color $c(S)$ and one occurrence $c(u_j)$, for each $u_j \in U$. Moreover, T is connected, since $\mathcal{S}'$ covers the elements in U, hence each vertex $v_{u,j}$ is adjacent to at least one vertex of v_{S,q_i}, $1 \leq i \leq k$.

Now, we prove that given an occurrence V_T of motif $\mathcal{M}$ in G, then we can compute in polynomial-time a solution of Min-SC of cardinality at most k. Let V_T be an occurrence of motif $\mathcal{M}$ in G. Observe that V_T must contain each mandatory vertex. Since r is a mandatory vertex and $\mathcal{M}$ contains exactly one occurrence of color $c(r)$, it follows that $r' \notin V_T$, as r' is not mandatory and is colored by $c(r)$.

Now, consider the subset $V_{S'} \subset V_T$ of vertices $v_{S,i}$, $1 \leq i \leq n$. Notice that V_T must contain exactly k such vertices, as they are all colored by $c(S)$ and as $\mathcal{M}$ contains exactly k occurrences of color $c(S)$. Furthermore, since $G[V_T]$ is connected and since each vertex $v_{u,j}$, $1 \leq j \leq n$, is mandatory, hence it must be in $G[V_T]$, it follows that each vertex $v_{u,j}$, $1 \leq j \leq n$, is connected to some vertex in $V_{S'}$. De ning $\mathcal{S}' = \{S_{q_i} : v_{S,q_i} \in V_T\}$, it follows that $\mathcal{S}'$ covers the universe set U and by construction has cardinality k.

Notice that, by construction, the number of optional occurrences in a solution of CGM over instance $(G = (V; E); \mathcal{M}; V_M)$ is exactly k. Hence the reduction is parameter preserving, thus implying that CGM is W[2]-hard.

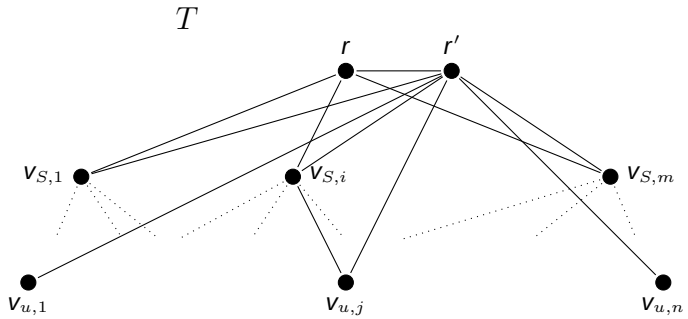

Fig. 3. Illustration of the reduction from Min-SC to CGM; notice that element $u_j \in S_i$.

□

6 Conclusion

In this paper, we have considered three variants of the Graph Motif problem, namely Min-Add Graph Motif, Min-Sub Graph Motif and CGM. We have investigated the computational and parameterized complexity of the three problems. More precisely, we have shown that Min-Add Graph Motif and Min-Sub Graph Motif are NP-hard even if the input graph is a tree T, the motif $\mathcal{M}$ is colorful and each color occurs at most twice in T. Furthermore, we have given an FPT algorithm for Min-Sub Graph Motif, where the parameter is the size of the motif. Finally, we have investigated the parameterized complexity of the CGM problem, when parameterized by the number of optional occurrences. We have shown that CGM is fixed-parameter tractable when the input graph has bounded treewidth, while it is W[2]-hard even if the input graph has diameter 2.

Future directions include the design of faster FPT algorithms for the problems as well as heuristics that can be applied effectively on real instances. Kernelization has been investigated only for Graph Motif [1]. Following the same approach, it would be interesting to investigate the kernelization complexity for some variants of Graph Motif discussed in this paper. Furthermore, we have shown that Min-Add Graph Motif and Min-Sub Graph Motif are NP-hard even if the input graph is a tree T, the motif $\mathcal{M}$ is colorful and each color occurs at most twice in T. Consider the case that each color occurs at most once in T. In this case it is easy to see that Min-Add Graph Motif is in P, while an open problem is to investigate the computational complexity of Min-Sub Graph Motif. Finally, it would be interesting to consider other variants of the Graph Motif problem, in particular a generalization of the considered variants that looks for an occurrence of a motif allowing for insertions, deletions, and substitutions.

Acknowledgments

We would like to thank the anonymous reviewers for very helpful comments that led to significant improvements in the presentation of the paper.

References

1. M. A. Abhimanyu, B. Radheshyam, C. Rao H, V. Koppula, N. Misra, G. Philip, and Ramanujan M.S., On the kernelization complexity of colorful motifs, In *IPEC 2010*, pp. 14{25, 2010.

2. P. Alimonti, V. Kann, Some APX-completeness results for cubic graphs, *Theor. Comput. Sci.* 237(1-2), 123 { 134, 2000.
3. N. Alon, R. Yuster, and U. Zwick, Color coding, *Journal of the ACM* 42(4), 844{856, 1995.
4. N. Betzler, M.R. Fellows, C. Komusiewicz, and R. Niedermeier, Parameterized algorithmics for nding connected motifs in biological networks, *IEEE/ACM Trans. Comput. Biology Bioinform.* 8(5), 1296 {1308, 2011.
5. S Bocker, F. Rasche, T. Steijger, Annotating Fragmentation Patterns. In *WABI 2009*, pp. 13-24, 2009.
6. S. Bruckner, F. Hu ner, R.M. Karp, R. Sharan, R. Shamir, Torque: Topology-free querying of protein interaction networks, *Journal of Computational Biology* 17, pp. 237{252 2010.
7. M. Cesati, *Compendium of parameterized problems*, http://bravo.ce.uniroma2.it/home/cesati/research/compendium.pdf.
8. R. Dondi, G. Fertin, and S. Vialette, Complexity issues in vertex-colored graph pattern matching. *J. Discrete Algorithms* 9(1), 82{99, 2011
9. R. Dondi, G. Fertin, and S. Vialette, Finding approximate and constrained motifs in graphs, In *CPM 2011*, pp. 388-401, 2011.
10. R. Downey and M. Fellows, *Parameterized complexity*, Springer-Verlag, 1999.
11. M. Fellows, G. Fertin, D. Hermelin, and S. Vialette, Upper and lower bounds for nding connected motifs in vertex-colored graphs, *JCSS* 77(4), pp. 799{811, 2011.
12. S. Guillemot, and F. Sikora, Finding and counting vertex-colored subtrees, *Algorithmica*, to appear.
13. 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, *Proc. Nat. Acad. Sci.* 100(20), 11394{11399, 2003.
14. M. Koyutu rk, A. Grama, and W. Szpankowski, Pairwise alignment of protein interaction networks guided by models of evolution, *Journal of Computational Biology* 13(2), 182{199, 2006.
15. 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(4), 360{368, 2006.
16. O. Mason, and M. Verwoerd, Graph Theory and Networks in Biology, *IET Systems Biology* 1(2), 89{119, 2007.
17. R. Niedermeier, *Invitation to xed-parameter algorithms*, Lecture Series in Mathematics and Its Applications, Oxford University, Press, 2006.
18. A. Paz and S. Moran, Non deterministic polynomial optimization problems and their approximations, *Theor. Comput. Sci.* 15, 251{277, 1981.
19. R. Rizzi, F. Sikora, Some results on more exible versions of Graph Motif , In *CSR 2012*, to appear.
20. J. Scott, T. Ideker, R.M. Karp, and R. Sharan, E cient algorithms for detecting signaling pathways in protein interaction networks, *Journal of Computational Biology* 13, 133{144, 2006.
21. R. Sharan, T. Ideker, B. Kelley, R. Shamir, and R.M. Karp, Identi cation of protein complexes by comparative analysis of yeast and bacterial protein interaction data, *Journal of Computational Biology* 12, pp. 835{846, 2005.
22. 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. Nat. Acad. Sci.* 102(6), 1974{1979, 2005.

23. F. Sikora, *Aspects algorithmiques de la comparaison d'elements biologiques*, PhD. Thesis, Universite Paris-Est, 2011.
24. D. J. Watts, and S. H. Strogatz, Collective dynamics of \small-world" networks, *Nature* 393, 440{442, 1998.