Finding Approximate and Constrained Motifs in Graphs

Riccardo Dondi, Guillaume Fertin, Stéphane Vialette

Finding Approximate and Constrained Motifs in 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, Via Donizetti 3, 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. One of the emerging topics in the analysis of biological networks is the inference of motifs inside a network. In the context of metabolic network analysis, a recent approach introduced in [14], represents the network as a vertex-colored graph, while a motif $\mathcal{M}$ is represented 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 $\mathcal{M}$. We investigate three different variants of the initial problem. The first two variants, Min-Add and Min-Substitute, deal with approximate occurrences of a motif in the graph, while the third variant, Constrained Graph Motif (or CGM for short), constrains the motif to contain a given set of vertices. We investigate the classical and parameterized complexity of the three problems. We show that Min-Add and Min-Substitute are NP-hard, even when $\mathcal{M}$ is a set, and the graph is a tree of degree bounded by 4 in which each color appears at most twice. Moreover, we show that Min-Substitute is in FPT when parameterized by the size of $\mathcal{M}$. Finally, we consider the parameterized complexity of the CGM problem, and we give a fixed-parameter algorithm for graphs of bounded treewidth, while we show that the problem is $W[2]$-hard, even if the input graph has diameter 2.

1 Introduction

The problem of analyzing biological networks such as protein-protein interaction networks and metabolic networks has become increasingly relevant in Computational Biology (see for example [5, 12, 13, 17–19]). While the classical approach is based on graph-theoretical topology of the motif, a recent approach introduced in [14] 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 find 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 studied, and some variants have been introduced. The original problem is known to be NP-complete [14], even if the input graph is a tree with maximum degree 3 and the motif is a set [10], and if the input graph is a bipartite graph with maximum degree 4 and the motif is built over two colors only [10]. It is easy to see that Graph Motif admits a polynomial time algorithm when the input graph is a tree and each color occurs at most twice in the input tree. The Graph Motif problem is known to be in FPT, when parameterized by the size of the motif [4, 10, 11], 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 [10]. Recently, the kernelization complexity of the problem has also been considered [1].

Different variants of the Graph Motif problem have been introduced. Such variants either modify the requirement of connectedness [7], or look for approximate occurrences of the motif, where some colors are allowed to be inserted or deleted in an occurrence of the motif [5, 8, 11]. 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 (Min-Add) or the substitution (Min-Substitute) of some colors. 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. In the third variant, Constrained Graph Motif (or CGM, for short), 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 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 $\mathcal{M}$.

The rest of the paper is organized as follows. In Section 2, we give some preliminary definitions and we formally define the problems. In Section 3, we show that Min-Substitute and Min-Add 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. In Section 4, we give an FPT algorithm for Min-Substitute. 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 5.1, we show that CGM is fixed-parameter tractable for graphs of bounded treewidth, while in Section 5.2 we show that CGM is $W[2]$-hard, even if the diameter of the input graph is bounded by 2. Some of the proofs are omitted due to space constraints.

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 $\{u, v\} \in E'$ iff $u, v \in V'$ 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. We recall that a graph is cubic when each vertex has degree 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, by abuse of notation, 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 *colorful* when $\mathcal{M}$ is a set of colors (rather than a multiset).

In this paper, we consider three variants of the Graph Motif problem. For two of them, Min-Add and Min-Substitute, 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 $\mathcal{M}$. Furthermore, we consider a constrained variant of the Graph Motif problem, CGM, where the input consists of a vertex colored graph and a set of mandatory vertices that must belong to any occurrence of motif $\mathcal{M}$.

Let us introduce the first two variants of Graph Motif problem.

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

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

Notice that, in case $p = 0$, both Min-Add and Min-Substitute are equivalent to the Graph Motif problem. As a consequence, Min-Add and Min-Substitute 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 [10]. Furthermore, Min-Add (resp. Min-Substitute) 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). Notice that Min-Add is in FPT, when parameterized by $|\mathcal{M}|$. Indeed, in [11], a variant of Graph Motif, called Multiset Graph Motif With Gaps (MGMG), is considered: given an input graph G and a motif $\mathcal{M}$, we look for an occurrence of $\mathcal{M}$ that is allowed to contain gaps. Note that this is precisely Min-Add, where the gaps represent colors to be added to $\mathcal{M}$. As in [11] it is shown that MGMG is in FPT when parameterized by $|\mathcal{M}|$, we can conclude that

the Min-Add problem is in FPT. Furthermore, in case the motif is colorful, a fixed-parameter algorithm for Min-Add has been given in [5].

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

Constrained Graph Motif (CGM) *Input* : A multiset of colors $\mathcal{M}$, a vertex-colored graph $G = (V, E)$, a set of mandatory vertices $V_M \subseteq V$. *Question* : Is there a subset $V' \subseteq V$, such that $G[V']$ is connected, $C(V') = \mathcal{M}$ and $V_M \subseteq V'$?

Given an instance of CGM, define the *optional occurrences* C_o as $C_o = \mathcal{M} \setminus C(V_M)$.

The CGM problem is NP-complete, since the Graph Motif problem is NP-complete [10, 14]. It is easy to see that CGM is fixed-parameter tractable, when the parameter is the size of the motif. Indeed, recall that Graph Motif is fixed-parameter tractable. By recoloring the graph, assigning a unique color to each vertex in V_M, and by modifying accordingly $\mathcal{M}$, we can conclude that each occurrence of $\mathcal{M}$ in G must include all the vertices in V_M.

In Section 5, we investigate the parameterized complexity of the CGM problem, when the parameter is the number of optional occurrences. Notice that the Minimum (Unweighted) Steiner Tree problem is a restriction of the CGM problem, where the non mandatory vertices in the Steiner Tree problem correspond to optional occurrences in CGM. As the Minimum (Unweighted) Steiner Tree problem is W[2]-hard when parameterized by the number of non mandatory vertices [6], it follows that the CGM problem is W[2]-hard when parameterized by the number of optional occurrences.

In Section 5.1, we will consider the case where the input graph has bounded treewidth and we will use a tree decomposition of the graph. Let us recall the definition of tree decomposition of a graph [9, 15]. Given a graph $G = (V, E)$, a tree decomposition of G is a pair $\langle \{X_i : i \in I\}, T \rangle$, such that each X_i is called a *bag*, and T is a tree having as vertices the elements of I and such that:

1. $\cup_{i \in I} X_i = V$;
2. for each edge $\{u, v\} \in E$, there is a bag X_i with $u, v \in X_i$;
3. for each i, j, k in V, if j is on the path from i to k in G, then $X_i \cap X_k \subseteq X_j$.

The width of $\langle \{X_i : i \in I\}, T \rangle$ is equal to $\max\{|X_i| : i \in I\} - 1$ and the treewidth of a graph G is equal to the minimum δ such that G has a tree decomposition of width δ. A tree decomposition $\langle \{X_i, i \in \{1, \ldots, p\}\}, T \rangle$ of a graph G is *nice* (see [15]) when, given a vertex i of the tree decomposition, i has at most two children and the following conditions hold:

1. if i has two children j and k, then $X_i = X_j = X_k$;
2. if i has exactly one child j, then one of the following conditions holds:
   (a) $|X_i| = |X_j| + 1$, and then $X_j \subset X_i$; or

(b) $|X_i| = |X_j| - 1$, and then $X_j \supset X_i$.

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 $\mathcal{M}$ and computes a colorful motif C, recoloring accordingly the vertices of the input graph G. 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 ε is $|\ln(\varepsilon)| \cdot O(e^{|\mathcal{M}|})$.*

3 NP-hardness of Min-Substitute and Min-Add

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

Theorem 1. *The* Min-Substitute *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). Let $G = (V, E)$ be a cubic graph with $V = \{v_1, v_2, \ldots, v_n\}$, the Min-VCC problem asks for a subset $V' \subseteq V$ of size at most p, such that for each $\{v_i, v_j\} \in E$ at least one of v_i, v_j is in V'. Min-VCC is known to be NP-hard [2]. Starting from G, we construct an instance of the Min-Substitute problem which consists of a tree T and a set of colors $\mathcal{M}$. For any vertex $v_i \in V$, let $e_{i,j}$, $1 \leq j \leq 3$, be its 3 incident edges, ordered arbitrarily. The tree $T = (V_T, E_T)$ is defined as follows (see Figure 1):

- $V_T = \{l_i, a_{i,1}, a_{i,2} : 1 \leq i \leq n\} \cup \{s_i : 1 \leq i \leq p\} \cup \{t_i : 1 \leq i \leq n+1\} \cup \{e_{i,j} : 1 \leq i \leq n \ \wedge \ 1 \leq j \leq 3\}$;
- $E_T = \{\{l_i, l_{i+1}\} : 1 \leq i < n\} \cup \{\{s_i, s_{i+1}\} : 1 \leq i < p\} \cup \{\{t_i, t_{i+1}\} : 1 \leq i < n+1\} \cup \{\{l_n, t_1\}\} \cup \{\{t_{n+1}, s_1\}\} \cup \{\{l_i, a_{i,1}\}, \{a_{i,1}, a_{i,2}\} : 1 \leq i \leq n\} \cup \{\{a_{i,2}, e_{i,j}\} : 1 \leq i \leq n \ \wedge \ 1 \leq j \leq 3\}$.

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 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 $e_{i,j} = \{v_i, v_j\}$

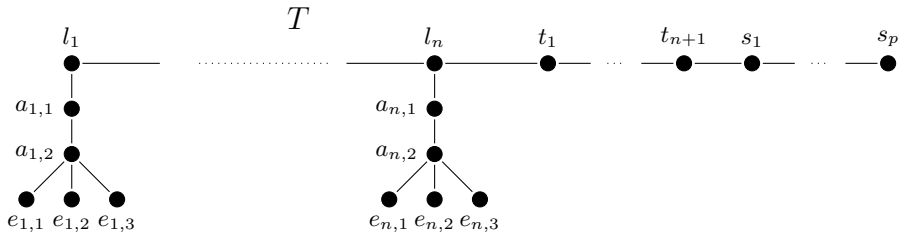

Fig. 1. Illustration of the reduction from MIN-VCC to MIN-SUBSTITUTE.

in E, is assigned color $c(e_{i,j})$. Each color occurs at most twice in T, as each color $c(e_{i,j})$ is associated to two vertices of T, while each color $c(v_i)$ is associated to vertices $a_{i,1}$, $a_{i,2}$. $\mathcal{M}$ is a set of colors defined as follows: $\mathcal{M} = \{c(l_i) : 1 \leq i \leq n\} \cup \{c(s_i) : 1 \leq i \leq p\} \cup \{c(v_i) : 1 \leq i \leq n\} \cup \{c(e_{i,j}) : e_{i,j} \in E\}$. Notice that no occurrence of a color $c(t_i)$, $1 \leq i \leq n+1$, belongs to $\mathcal{M}$.

Starting from a vertex cover $V' \subseteq V$ of G of size at most p, a solution $V_{T'}$ of MIN-SUBSTITUTE, that substitutes p colors from $\mathcal{M}$, is obtained as follows. Given an edge $e_{i,j} = \{v_i, v_j\}$, define $e^{min}_{i,j} = \min\{i, j\}$. The vertex set $V_{T'}$ defined as follows:
$V_{T'} = \{l_i, a_{i,1} : 1 \leq i \leq n\} \cup \{t_i : 1 \leq i \leq p - |V'|\} \cup \{a_{i,2} : v_i \in V'\} \cup \{e_{i,x} : c(e_{i,x}) = c(e_{i,j}) \wedge\ i = e^{min}_{i,j}\}$.

By construction and since V' is a vertex cover, $V_{T'}$ induces a subtree of T. It is easy to see that, given $C(V_{T'}) = \mathcal{M}'$, $\mathcal{M}'$ can be obtained from $\mathcal{M}$ by p substitutions.

Let us consider now a solution $V_{T'}$ of MIN-SUBSTITUTE, 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 a vertex of the set $\{s_i : 1 \leq i \leq p\}$. Indeed, assume that a vertex s_i is part of $V_{T'}$; 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 size 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 $\mathcal{M}' \supseteq \{c(e_{i,j}) : e_{i,j} \in E\}$. Since $T[V_{T'}]$ must be connected, it follows that each vertex colored $c(e_{i,j})$ must be connected to some vertex $a_{i,2} \in V_{T'}$ colored by $c(v_i)$. Define $V' = \{v_i : a_{i,2} \in V_{T'}\}$; then V' is a cover of G of size at most p, which completes the proof. □

Theorem 2. *The* MIN-ADD *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. (*Sketch*) The result follows from a reduction from Min-VCC similar to that of Theorem 1. Given an instance of Min-VCC, an instance $(T, \mathcal{M})$ of Min-Add is constructed as follows. $T = (V_T, E_T)$ is defined as follows (see Fig. 2):

- $V_T = \{l_i, a_{i,1}, a_{i,2} : 1 \leq i \leq n\} \cup \{e_{i,j} : 1 \leq i \leq n \ \wedge \ 1 \leq j \leq 3\}$;
- $E_T = \{\{l_i, l_{i+1}\} : 1 \leq i < n\} \cup \{\{l_i, a_{i,1}\}, \{a_{i,1}, a_{i,2}\} : 1 \leq i \leq n\} \cup \{\{a_{i,2}, e_{i,j}\} : 1 \leq i \leq n \ \wedge \ 1 \leq j \leq 3\}$.

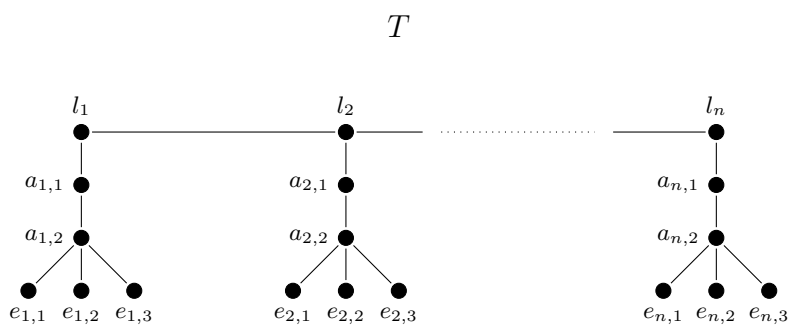

Fig. 2. Illustration of the reduction from Min-VCC to Min-Add.

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 $a_{i,1}$, $a_{i,2}$, $1 \leq i \leq n$, 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 $e_{i,j} = \{v_i, v_j\}$ in E, is assigned color $c(e_{i,j})$. $\mathcal{M}$ is a set of colors defined as follows: $\mathcal{M} = \{c(l_i) : 1 \leq i \leq n\} \cup \{c(v_i) : 1 \leq i \leq n\} \cup \{c(e_{i,j}) : e_{i,j} \in E\}$.

It can be proved that starting from a vertex cover $V' \subseteq V$ of G, we can compute in polynomial time a solution $V_{T'}$ of Min-Add such that $C(V_{T'}) \supseteq \mathcal{M}$ and $|C(V_{T'})| \leq |\mathcal{M}| + |V'|$. Conversely, starting from a solution $V_{T'}$ of Min-Add such that $C(V_{T'}) \supseteq \mathcal{M}$ and $|C(V_{T'})| \leq |\mathcal{M}| + p$, we can compute a vertex cover V' of G such that $|V'| \leq p$. □

4 Parameterized Complexity of Min-Substitute

In this section, we discuss the parameterized complexity of Min-Substitute, when parameterized by $|\mathcal{M}|$. We recall that Min-Substitute is not in FPT, as discussed in Section 2, when parameterized by the the size of the solution (i.e., the number of substituted colors).

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-Substitute. Instead of computing directly a solution for Min-Substitute, we compute a solution for a slightly different problem, where we visit the vertices of a connected component of G, allowing to visit some vertices

more than once. Let v be a vertex of the input graph G, let $C \subseteq \mathcal{M}$ be a subset of colors, let k be the number of vertices of the solution of Min-Substitute we are looking for, and define $S[v, C, k]$ as the minimum value z required by a visit of a connected set V_T of vertices of G such that:

1. $v \in V_T$;
2. exactly k visits of vertices in V_T are done;
3. $C(V_T)$ matches q colors of C, where $z = k - q$.

Notice that a vertex of V_T may be visited more than once, while the overall number of visits must be k. Now, let us define the dynamic programming recurrence to compute $S[v, C, k]$.

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

For the base cases: $S[u, C', 1] = 0$, when $c(u) \in C'$, for each $C' \subseteq C$ and $u \in V$, and $S[u, C', 1] = 1$ when $c(u) \notin C'$. Now, let us prove the correctness of Recurrence (1).

Lemma 2. *Let $(G, \mathcal{M})$ be an instance of* Min-Substitute*, let v be a vertex of G, and let C be a subset of $\mathcal{M}$. There is a visit of a connected vertex set V_T of G, such that $v \in V_T$, the vertices of V_T are visited k times, and $C(V_T)$ matches q colors of C, iff there exists an entry $S[v, C, k] = z$, where $z = k - q$.*

An optimal solution for the Min-Substitute problem can be found as follows. We look for the minimal value z in the entries $S[v, \mathcal{M}, |\mathcal{M}|]$, with $v \in V$. Notice that this value may be associated to a visit of a connected vertex set V_T, where some of the vertices may be visited repeatedly. Each repeated visit of a vertex represents a color to be substituted, since $\mathcal{M}$ is colorful. It follows that we can compute a feasible solution for Min-Substitute by replacing these repeated visits by some connected components adjacent to V_T without increasing the number of substitutions.

The time complexity of the algorithm is $O^*(3^{|\mathcal{M}|})$, as we have to consider all possible subsets $C \subseteq \mathcal{M}$ and for each subset C we have to consider all possibile bipartitions of C. Indeed, there are $O(3^{|\mathcal{M}|})$ possible bipartitions of all possible subsets C of $\mathcal{M}$. In order to extend the results to a multiset, we apply the recoloring technique described in [4]. Combining Lemma 2 with Lemma 1, and we get that Min-Substitute, parameterized by $|\mathcal{M}|$, can be solved in time $O^*((3e)^{O(|\mathcal{M}|)})$.

5 Parameterized Complexity of CGM

In this section, we consider the parameterized complexity of CGM, where the parameter is the number k of optional occurrences C_o, that is $k = |C_o|$, where $C_o = \mathcal{M} \setminus C(V_M)$. First, in Section 5.1, we show that CGM is fixed-parameter tractable, when the input graph is of bounded treewidth ; then, in Section 5.2, we prove that CGM is W[2]-hard, even when the input graph is of diameter 2.

5.1 An FPT Algorithm for Graphs of Bounded Treewidth

Here, we describe a fixed-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 first consider the case where the motif $\mathcal{M}$ is colorful.

Denote by k the number of optional occurrences and by δ the treewidth of graph G. The algorithm is based on a nice tree decomposition of G (see Section 2 for the definition of nice tree decomposition of a graph). We also consider a slightly more general problem, where instead of requiring that an occurrence of a motif consists of a single connected component, we may have an occurrence consisting of at most $\delta+1$ connected components, where the different connected components are induced by a partition of a bag X_i of the nice tree decomposition. Given a vertex i of the tree decomposition of G, denote by $T[i]$ the subtree of the nice tree decomposition rooted at i and 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 $\langle\{X_i, i \in \{1, \ldots, p\}\}, T\rangle$. From the definition of treewidth, it follows that $|X_i| \leq \delta + 1$. Now, let us define a mapping function f_i associated to the vertices of X_i, as follows.

Definition 1. *Let X_i be a bag of the nice tree decomposition of G. A mapping function f_i from X_i to $\{0, 1, \ldots, \delta + 1\}$ 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. *define $X_i^l = \{v \in X_i : f(v_i) = l\}$, $l \in \{1, \ldots, \delta + 1\}$, and $X_i' = \cup_l X_i^l$, then X_i^l is a maximal connected component of $G[X_i']$.*

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

Definition 2. *Let W be a set of vertices of $V(T[i])$, consisting of the connected components $W_1, W_2, \ldots, W_z$. Let f_i be a feasible mapping of X_i in $\{0, 1, \ldots \delta+1\}$, 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, $1 \leq l \leq \delta + 1$, such that $W_p \cap X_i = X_i^l$*
2. *for each l, $1 \leq l \leq \delta + 1$, 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 Definition 2, 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 $f_i : X_i \to \{0, \ldots, \delta + 1\}$ and $f_j : X_j \to \{0, \ldots, \delta + 1\}$, 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. $f_i(v) = l$, $l \in \{1, \ldots, \delta + 1\}$, $f_i(u) \neq l$ for each $u \in X_i \cap X_j$, and f_i, f_j are consistent; or
3. there exists a value $l \in \{1, \ldots, \delta + 1\}$ such that
   (a) $f_i(v) = l$;
   (b) if $f_j(z) = 0$, then $f_i(z) = 0$, for $z \in X_i \cap X_j$;
   (c) if $f_i(z) \neq f_j(z)$, for $z \in X_i$ and $f_j(z) \neq 0$, then $f_i(z) = l$.

Given a feasible mapping f_i of X_i in $\{0, 1, \ldots \delta + 1\}$, define $c(X_i, f_i) = \{c \in C_o : \exists v \in X_i,\ c(v) = c \wedge f_i(v) \neq 0\}$.

Let us define 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 of the set X_i in $\{0, 1, \ldots \delta + 1\}$ and $C' \subseteq C_o$ be a subset of the set of optional occurrences. $S[i, f_i, C'] = 1$ 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, each mandatory vertex of $T[i]$ is in W, and the set of optional occurrences of $c(W)$ 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 the three different cases of a nice tree decomposition.

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

$$S[i, f_i, C'] = \bigvee_{f_j, f_k, C_j, C_k} S[j, f_j, C_j] \wedge S[k, f_k, C_k],$$

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

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, $C' = C_j \cup \{c(v)\}$ and $c(v) \notin C_j$, when $f_i(v) \neq 0$ and $v \notin V_M$, and $C' = C_j$ when $f_i(v) = 0$ or $v \in V_M$.

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 there is a vertex $z \in X_i \cap X_j$, such that $f_j(z) = f_j(v)$, with $v \in X_j \setminus X_i$, when $f_j(v) \neq 0$.

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

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 of X_i, 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 optional occurrences in $c(W)$. Then $S[i, f_i, C'] = 1$.*

Lemma 4. *Let $S[i, f_i, C'] = 1$ for a feasible mapping function f_i of X_i in $\{0, 1, \ldots, \delta+1\}$, then there exists a set W of vertices in $V(T[i])$ such that the set of optional occurrences 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.*

Theorem 3 shows how the values $S[i, f_i, C']$ are used to compute the existence of a feasible solution for CGM.

Theorem 3. *Let $(G = (V, E), \mathcal{M}, V_M)$ be an instance of the* CGM *problem. Then there is a solution W of* CGM *over instance of $(G, \mathcal{M}, V_M)$ iff there is a vertex i of the nice tree decomposition and a feasible function f_i that maps X_i in $\{0, x\}$, with $x \in \{1, \ldots, \delta+1\}$, such that $S[i, f_i, C_o] = 1$ and such $V_M \subseteq V(T[i])$.*

Proof. Assume that there is a vertex i of the nice tree decomposition and a feasible function f_i that maps X_i in $\{0, x\}$, with $x \in \{1, \ldots, \delta + 1\}$, such that $S[i, f_i, C_o] = 1$ 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 C_o 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 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 $S[i, f_i, C_o] = 1$ for some feasible function f_i that maps X_i in $\{0, x\}$, with $x \in \{1, \ldots, \delta + 1\}$. □

Now, we discuss the time complexity of the above algorithm. Denote by n the size of V. Given a vertex i and the associated set X_i of the nice tree decomposition, the number of possible mapping functions of X_i into $\{0, \ldots, \delta+1\}$ is $O(\delta^\delta)$. 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 $O(\delta^\delta n 2^k)$ entries $S[i, f_i, C]$. Given a mapping function f_i of X_i into $\{0, \ldots, \delta + 1\}$, computing an entry $S[i, f_i, C]$, given the entries of the children (or the child) of i, requires time at most $O(\delta^{2\delta} 2^{2k})$ (notice that the worst case occurs when i has two children). Hence the total time complexity is $O(\delta^{3\delta} n 2^{3k})$.

When a motif is a multiset of colors, we apply the recoloring technique presented in Lemma 1. As a consequence, CGM can be solved with error probability ε in time $O(|\ln(\varepsilon)|\delta^{3\delta} n 2^{4.4427k})$, for graphs of treewidth δ.

5.2 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 is of diameter 2.

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

Proof. (*Sketch*) We give a parameterized preserving reduction from Minimum Set Cover (Min-SC). Given a universe $U = \{u_1, \ldots, u_n\}$ and a collection of sets $\mathcal{S} = \{S_1, \ldots, S_m\}$ over U, CGM asks for 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 [16]. Let $(U, \mathcal{S})$ be

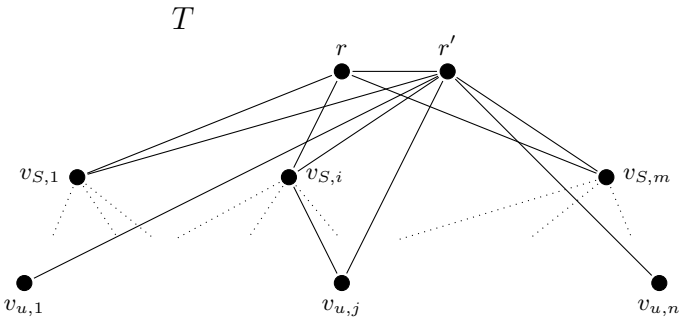

Fig. 3. Illustration of the reduction from Min-SC to CGM ; notice that element $u_j \in S_i$.

an instance of Min-SC, we define a corresponding instance $(G = (V, E), \mathcal{M}, V_M)$ of the CGM problem (see Fig. 3). The graph G of diameter 2 is defined as follows:

- $V = \{r\} \cup \{r'\} \cup \{v_{S,i} : 1 \leq i \leq m\} \cup \{v_{u,j} : 1 \leq j \leq n\}$;
- $E = \{\{r, r'\}\} \cup \{\{r, v_{S,i}\} : 1 \leq i \leq m\} \cup \{\{r', v_{S,i}\} : 1 \leq i \leq m\} \cup \{\{v_{S,i}, v_{u,j}\} : 1 \leq i \leq m \wedge u_j \in S_i\} \cup \{\{r', v_{u,j}\} : 1 \leq j \leq n\}$.

Vertex r and vertex r' are both colored by $c(r)$, vertex $v_{S,i}$ is colored by $c(S)$, $1 \leq i \leq m$, and vertex $v_{u,j}$ is colored by $c(u_j)$, $1 \leq j \leq n$. 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 $c(S)$. Finally, $V_M = V \setminus (\{v_{S,i} : 1 \leq i \leq m\} \cup \{r'\})$.

Then, it is possible to show that, given a solution of Min-SC of size at most k, we can compute in polynomial time a solution of CGM over instance $(G = (V, E), \mathcal{M}, V_M)$. Similarly, it is possible to show that given an occurrence

V_T of motif $\mathcal{M}$ in G, we can compute in polynomial time a solution of MIN-SC of size at most k. By construction, a solution of CGM over instance $(G = (V, E), \mathcal{M}, V_M)$ contains exactly k optional occurrences. Hence the reduction is parameter preserving, thus implying that CGM is $W[2]$-hard. □

References

1. Ambalath, A. M. , Balasundaram, R., Chintan, R., Koppula, V., Misra, N., Philip, G., Ramanujan M.S.: On the Kernelization Complexity of Colorful Motifs. In: Raman, V., Saurabh, S. (eds.) IPEC 2010. LNCS, vol. 6478, pp. 14–25. Springer, Heidelberg (2010)
2. Alimonti, P., Kann, V.: Some APX-Completeness Results for Cubic Graphs. Theor. Comput. Sci. 237(1-2), 123 – 134 (2000)
3. Alon, N., Yuster, R., Zwick, U.: Color Coding. Journal of the ACM 42(4), 844–856 (1995)
4. Betzler, N., Fellows, M.R., Komusiewicz, C., Niedermeier, R.: Parameterized Algorithms and Hardness Results for Some Graph Motif Problems. In: Ferragina, P., Landau, G. (eds.) CPM 2008. LNCS, vol. 5029, pp. 31–43. Springer, Heidelberg (2008)
5. Bruckner, S., Hüffner, F., Karp, R.M., Sharan, R., Shamir, R.: Topology-Free Querying of Protein Interaction Networks. In: Batzoglou, S. (ed.) RECOMB 2009. LNCS, vol. 5541, pp. 74–89. Springer, Heidelberg (2009)
6. Cesati, M.: Compendium of parameterized problems, `http://bravo.ce.uniroma2.it/home/cesati/research/compendium.pdf`
7. Dondi, R., Fertin, G., Vialette, S.: Weak Pattern Matching in Colored Graphs: Minimizing the Number of Connected Components. In: Italiano, G. F., Moggi, E., Laura, L. (eds.) ICTCS 2007, pp. 27–38. World Scientific, Singapore (2007)
8. Dondi, R., Fertin, G., Vialette, S.: Maximum Motif in Vertex-Colored Graphs. In: Kucherov, G., Ukkonen, E. (eds.) CPM 2009. LNCS, vol. 5577, pp. 221-235. Springer, Heidelberg (2009)
9. Downey, R., Fellows, M.: Parameterized Complexity. Springer, Heidelberg (1999)
10. Fellows, M., Fertin, G., Hermelin, D., Vialette S.: Sharp Tractability Borderlines for Finding Connected Motifs in Vertex-Colored Graphs. In: Arge, L., Cachin, C., Jurdzinski, T., Tarlecki, A. (eds.) ICALP 2007. LNCS, vol. 4596, pp. 340–351. Springer, Heidelberg (2007)
11. Guillemot, S., Sikora, F.: Finding and Counting Vertex-Colored Subtrees. In: Hliněný, P., Kucera, A. (eds.) MFCS 2010. LNCS, vol. 6281, pp. 405–416. Springer, Heidelberg (2010)
12. Kelley, B. P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., Ideker, T., Conserved Pathways within Bacteria and Yeast as Revealed by Global Protein Network Alignment. Proc. Nat. Acad. Sci. 100(20), 11394–11399 (2003)
13. Koyutürk, M., Grama, A., Szpankowski, W.: Pairwise Local Alignment of Protein Interaction Networks Guided by Models of Evolution. In: Miyano, S., Mesirov, J.P., Kasif, S., Istrail, S., Pevzner, P.A., Waterman. M.S. (eds.) RECOMB 2005. LNCS, vol. 3500, pp. 48–65. Springer, Heidelberg (2005)
14. Lacroix, V., Fernandes, C.G., Sagot, M. F.: Motif Search in Graphs: Application to Metabolic Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 3(4), 360–368 (2006)

15. Niedermeier, R.: Invitation to Fixed-Parameter Algorithms. Oxford University Press, Oxford (2006)
16. Paz, A., Moran, S.: Non Deterministic Polynomial Optimization Problems and Their Approximations. Theor. Comput. Sci. 15, 251–277 (1981)
17. Scott, J., Ideker, T., Karp, R.M., Sharan, R.: Efficient Algorithms for Detecting Signaling Pathways in Protein Interaction Networks. Journal of Computational Biology 13, 133–144 (2006)
18. Sharan, R., Ideker, T., Kelley, B., Shamir, R., Karp, R. M.: Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data. In: Bourne, P. E., Gusfield, D. (eds.) RECOMB 2004, pp. 282–289. ACM Press, New York (2004)
19. Sharan, R., Suthram, S., Kelley, R., Kuhn, T., McCuine, S., Uetz, P., Sittler, Karp, R.M., Ideker, T.: Conserved Patterns of Protein Interaction in Multiple Species. Proc. Nat. Acad. Sci. 102(6), 1974–1979 (2005)