Finding and counting vertex-colored subtrees

Sylvain Guillemot, Florian Sikora

Finding and counting vertex-colored subtrees

Sylvain Guillemot[1] and Florian Sikora[2]

[1] Lehrstuhl für Bioinformatik, Friedrich-Schiller Universität Jena, Ernst-Abbe Platz 2, 00743 Jena, Germany, E-mail: sylvain.guillemot@uni-jena.de
[2] Université Paris-Est, LIGM - UMR CNRS 8049, France
E-mail: sikora@univ-mlv.fr

Abstract. The problems studied in this article originate from the Graph Motif problem introduced by Lacroix et al. [17] in the context of biological networks. The problem is to decide if a vertex-colored graph has a connected subgraph whose colors equal a given multiset of colors M. Using an algebraic framework recently introduced by Koutis et al. [15,16], we obtain new FPT algorithms for Graph Motif and variants, with improved running times. We also obtain results on the counting versions of this problem, showing that the counting problem is FPT if M is a set, but becomes #W[1]-hard if M is a multiset with two colors.

1 Introduction

An emerging field in the modern biology is the study of the biological networks, which represent the interactions between biological elements [1]. A network is modeled by a vertex-colored graph, where nodes represent the biological compounds, edges represent their interactions, and colors represent functionalities of the graph nodes. Networks are often analyzed by studying their *network motifs*, which are defined as small recurring subnetworks. Motifs generally correspond to a set of elements realizing a same function, and which may have been evolutionarily preserved. Therefore, the discovery and the querying of motifs is a crucial problem [20], since it can help to decompose the network into functional modules, to identify conserved elements, and to transfer biological knowledge across species.

The initial definition of network motifs involves conservation of the topology and of the node labels; hence, looking for topological motifs is roughly equivalent to subgraph isomorphism, and thus is a computationally difficult problem. However, in some situations, the topology is not known or is irrelevant, which leads to searching for *functional* motifs instead of *topological* ones. In this setting, we still ask for the conservation of the node labels, but we replace topology conservation by the weaker requirement that the subnetwork should form a connected subgraph of the target graph. This approach was advocated by [17] and led to the definition of the Graph Motif problem [10]: given a vertex-colored graph $G = (V, E)$ and a multiset of colors M, find a set $V' \subseteq V$ such that the induced subgraph $G[V']$ is connected, and the multiset of colors of the vertices of V' is equal to M. In the literature, a distinction is made between the *colorful* case

(when M is a set), and the *multiset* case (when M is an arbitrary multiset). Although this problem has been introduced for biological motivations, [3] points out that it may also be used in social or technical networks.

Not surprisingly, Graph Motif is NP-hard, even if G is a bipartite graph with maximum degree 4 and M is built over two colors only [10]. The problem is still NP-hard if G is a tree, but in this case it can be solved in $\mathcal{O}(n^{2c+2})$ time, where c is the number of distinct colors in M, while being W[1]-hard for the parameter c [10]. The difficulty of this problem is counterbalanced by its fixed-parameter tractability when the parameter is k, the size of the solution [17,10,3]. The currently fastest FPT algorithms for the problem run in $\mathcal{O}^*(2^k)$ time for the colorful case, $\mathcal{O}^*(4.32^k)$ time for the multiset case, and use exponential space (the $\mathcal{O}^*$ notation suppresses $poly(n,k)$ factors).

Our contribution is twofold. First, we consider in Section 3 the *decision* versions of the Graph Motif problem, as well as some variants: we obtain improved FPT algorithms for these problems, by using the algebraic framework of *multilinear detection* for arithmetic circuits [15,16], presented in the next section. Second, we investigate in Section 4 the *counting* versions of the Graph Motif problem: instead of deciding if a motif appears in the graph, we now want to count the occurrences of this motif. This allows to assess if a motif is over- or under- represented in the network, by comparing the actual count of the motif to its expected count under a null hypothesis [19]. We show that the counting problem is FPT in the colorful case, but becomes #W[1]-hard for the multiset case with two colors. We refer the reader to [12,11] for definitions related to parameterized counting classes.

2 Definitions

This section contains definitions related to arithmetic circuits, and to the Multilinear Detection (MLD) problem. It concludes by stating Theorem 1, which will be used throughout the paper.

2.1 Arithmetic circuits

In the following, a capital letter X will denote a set of variables, and a lowercase letter x will denote a single variable. If X is a set of variables and $\mathbb{A}$ is a commutative ring, we denote by $\mathbb{A}[X]$ the ring of multivariate polynomials with coefficients in $\mathbb{A}$ and involving variables of X. Given a monomial $m = x_1...x_k$ in $\mathbb{A}[X]$, where the x_is are variables, its *degree* is k, and m is *multilinear* iff its variables are distinct.

An *arithmetic circuit* over X is a pair $\mathcal{C} = (C, r)$, where C is a labeled dag such that (i) the children of each node are totally ordered, (ii) the nodes are labeled either by $op \in \{+, \times\}$ or by an element of X, (iii) no internal node is labeled by an element of X, and where r is a distinguished node of C called the *root*. For a given node u we denote by $N_C(u)$ the set of children (*i.e.* out-neighbors) of u in C. We recall that a node u is called a *leaf* of C iff $N_C(u) = \emptyset$,

an *internal node* otherwise. We denote by $T(\mathcal{C})$ the *size of* $\mathcal{C}$ (defined as the number of arcs), and we denote by $S(\mathcal{C})$ the number of nodes of $\mathcal{C}$ of indegree ≥ 2.

Given a commutative ring $\mathbb{A}$, *evaluating* $\mathcal{C}$ *over* $\mathbb{A}$ *under a mapping* $\phi : X \to \mathbb{A}$ consists in

1. computing, for each node u of C, a value $val(u) \in \mathbb{A}$ as follows:

$$val(u) = \sum_{u' \in N_C(u)} val(u') \text{ if } u \text{ is labeled by } +$$

$$val(u) = \prod_{u' \in N_C(u)} val(u') \text{ if } u \text{ is labeled by } \times$$

$$val(u) = \phi(x) \text{ if } u \text{ is a leaf labeled by } x \in X$$

   where the operations are carried out in $\mathbb{A}$. By convention, empty sums evaluate to $0_{\mathbb{A}}$, and empty products evaluate to $1_{\mathbb{A}}$.
2. returning the value $val(r)$ as the result of the evaluation.

Observe that if operations in $\mathbb{A}$ require $\mathcal{O}(t)$ time and $\mathcal{O}(s)$ space, then the above evaluation can be computed in $\mathcal{O}(t.T(\mathcal{C}))$ time and $\mathcal{O}(s.S(\mathcal{C}))$ space. The *symbolic evaluation* of $\mathcal{C}$ is the polynomial $P_{\mathcal{C}} \in \mathbb{Z}[X]$ obtained by evaluating $\mathcal{C}$ over $\mathbb{Z}[X]$ under the identity mapping $\phi : X \to \mathbb{Z}[X]$.

2.2 Multilinear Detection

Informally, the Multilinear Detection problem asks, for a given arithmetic circuit $\mathcal{C}$ and an integer k, if the polynomial $P_{\mathcal{C}}$ has a multilinear monomial of degree k. However, this definition does not give a certificate checkable in polynomial-time, so for technical reasons we define the problem differently.

A *monomial-subtree* of $\mathcal{C}$ is a pair $T = (\mathcal{C}', \phi)$, where $\mathcal{C}' = (C', r')$ is an arithmetic circuit over X whose underlying dag C' is a directed tree, and where $\phi : V_{\mathcal{C}'} \to V_{\mathcal{C}}$ is such that (i) $\phi(r') = r$, (ii) if $u \in V_{\mathcal{C}'}$ is labeled by $x \in X$, then so is $\phi(u)$, (iii) if $u \in V_{\mathcal{C}'}$ is labeled by $+$ then so is $\phi(u)$, and $N_{\mathcal{C}'}(u)$ consists of a single element $v \in N_{\mathcal{C}}(\phi(u))$, (iv) if $u \in V_{\mathcal{C}'}$ is labeled by $\times$, then so is $\phi(u)$, and ϕ maps bijectively $N_{\mathcal{C}'}(u)$ into $N_{\mathcal{C}}(\phi(u))$ by preserving the ordering on siblings. The *variables* of T are the leaves of C' labeled by variables in X. We say that T is *distinctly-labeled* iff its variables are distinct. Intuitively, the (distinctly-labeled) monomial-subtrees of $\mathcal{C}$ with k variables correspond to the (multilinear) monomials of $P_{\mathcal{C}}$ having degree k. Therefore, we formulate the Multilinear Detection problem as follows:

Name: Multilinear Detection (MLD)
Input: An arithmetic circuit $\mathcal{C}$ over a set of variables X, an integer k
Solution: A distinctly-labeled monomial-subtree of $\mathcal{C}$ with k variables.

Solving MLD amounts to decide if $P_{\mathcal{C}}$ has a multilinear monomial of degree k (observe that there are no possible cancellations), and solving #MLD amounts to compute the sum of the coefficients of multilinear monomials of $P_{\mathcal{C}}$ having degree k. The restriction of MLD when $|X| = k$ is called Exact Multilinear Detection (XMLD). In this article, we will rely on the following far-reaching result from [21,16] to obtain new algorithms for Graph Motif:

Theorem 1 ([21,16]). MLD *can be solved by a randomized algorithm which uses* $\tilde{\mathcal{O}}(2^k T(\mathcal{C}))$ *time and* $\tilde{\mathcal{O}}(S(\mathcal{C}))$ *space.*

Here, we use the $\tilde{\mathcal{O}}$ notation to suppress polylogarithmic factors, *i.e.* factors of the form $\mathcal{O}((\log n)^c)$ where n is the instance size and c is a constant. By a "randomized algorithm" with running time $\mathcal{O}(t)$, we mean an algorithm which (i) always answers no on negative instances, (ii) answers yes with probability $\geq \frac{1}{2}$ on positive instances, (iii) always runs in time $\mathcal{O}(t)$ regardless of the random choices made in an execution. We point out that the algorithm of Theorem 1 proceeds by multiple evaluations of the circuit over $\mathbb{Z}$. Therefore, it still applies if the circuit is given as an evaluation oracle over the integers.

3 Finding vertex-colored subtrees

In this section, we consider several variants of the Graph Motif problem, and we obtain improved FPT algorithms for these problems by reduction to MLD. Notably, we obtain $\mathcal{O}^*(2^k)$ time algorithms for problems involving *colorful* motifs, and $\mathcal{O}^*(4^k)$ time algorithms for *multiset* motifs.

3.1 The colorful case

In the colorful formulation of the problem, the graph is vertex-colored, and we seek a subtree with k vertices having distinct colors. This leads to the following formal definition:

Name: Colorful Graph Motif (CGM)
Input: A graph $G = (V, E)$, $k \in \mathbb{N}$, a set C, a function $\chi : V \to C$
Solution: A subtree $T = (V_T, E_T)$ of G s.t. (i) $|V_T| = k$ and (ii) for each $u, v \in V_T$ distinct, $\chi(u) \neq \chi(v)$.

The restriction of Colorful Graph Motif when $|C| = k$ is called Exact Colorful Graph Motif (XCGM). Note that this restriction requires that the vertices of T are bijectively labeled by the colors of C. In [7], the XCGM problem was shown to be solvable in $\mathcal{O}^*(2^k)$ time and space, while it is not difficult to see that the general CGM problem can be solved in $\mathcal{O}^*((2e)^k)$ time and $\mathcal{O}^*(2^k)$ space by color-coding. By using a reduction to Multilinear Detection, we improve upon these complexities. In the following, we let n and m denote the number of vertices and the number of edges of G, respectively.

Proposition 1. CGM *is solvable by a randomized algorithm in* $\tilde{\mathcal{O}}(2^k k^2 m)$ *time and* $\tilde{\mathcal{O}}(kn)$ *space.*

Proof. Let I be an instance of CGM. We construct the following circuit $\mathcal{C}_I$: its set of variables is $\{x_c : c \in C\}$, and we introduce intermediary nodes $P_{i,u}$ for $1 \leq i \leq k, u \in V$, as well as a root node P. Informally, the multilinear monomials of $P_{i,u}$ will correspond to colorful subtrees of G having i vertices, including u. The definitions are as follows:

$$P_{i,u} = \sum_{i'=1}^{i-1} \sum_{v \in N_G(u)} P_{i',u} P_{i-i',v} \text{ if } i > 1, \qquad P_{1,u} = x_{\chi(u)}$$

and $P = \sum_{u \in V} P_{k,u}$. The resulting instance of MLD is $I' = (\mathcal{C}_I, k)$. By applying Theorem 1, and by observing that $T(\mathcal{C}_I) = \mathcal{O}(k^2 m)$ and $S(\mathcal{C}_I) = \mathcal{O}(kn)$, we solve I' in $\tilde{\mathcal{O}}(2^k k^2 m)$ time and $\tilde{\mathcal{O}}(kn)$ space. The correctness of the construction follows by showing by induction on $1 \leq i \leq k$ that: $x_{c_1} ... x_{c_d}$ is a multilinear monomial of $P_{i,u}$ iff (i) $d = i$ and (ii) there exists $T = (V_T, E_T)$ colorful subtree of G such that $u \in V_T$ and $\chi(V_T) = \{c_1, ..., c_d\}$. □

3.2 The multiset case

We now consider several variants of the CGM problem. The first two variants allow for *multiset motifs*: instead of seeking a subtree with distinct colors, we now allow some colors to be repeated but impose a maximum number of occurrences for each color. This problem can be seen as a generalization of the original Graph Motif problem.

Given a multiset M over a set A, and given an element $x \in A$, we denote by $n_M(x)$ the number of occurrences of x in M. Given two multisets M, M', we denote their inclusion by $M \subseteq M'$. We denote by $|M|$ the size of M, where elements are counted with their multiplicities. Given two sets A, B, a function $f : A \to B$ and a multiset X over A, we let $f(X)$ denote the multiset containing the elements $f(x)$ for $x \in X$, counted with multiplicities; precisely, given $y \in B$ we have $n_{f(X)}(y) = \sum_{x \in A : f(x) = y} n_X(x)$.

We now define the two following variants of Colorful Graph Motif, which allow for multiset motifs:

Name: Multiset Graph Motif (MGM)
Input: A graph $G = (V, E)$, an integer k, a set C, a function $\chi : V \to C$, a multiset M over C.
Solution: A subtree $T = (V_T, E_T)$ of G s.t. (i) $|V_T| = k$ and (ii) $\chi(V_T) \subseteq M$.

Name: Multiset Graph Motif With Gaps (MGMG)
Input: A graph $G = (V, E)$, integers k, r, a set C, a function $\chi : V \to C$, a multiset M over C.
Solution: A subtree $T = (V_T, E_T)$ of G s.t. (i) $|V_T| \leq r$ and (ii) there exists

$S \subseteq V_T$ of size k such that $\chi(S) \subseteq M$.

The restriction of Multiset Graph Motif when $|M| = k$ is called Exact Multiset Graph Motif (XMGM). Note that in this case we require that T contains every occurrence of M, *i.e.* $\chi(V_T) = M$. In this way, the XMGM problem coincides with the Graph Motif problem defined in [10,3], while the MGM problem is the parameterized version of the Max Motif problem considered in [9]. The notion of gaps is introduced in [17], and encompasses the notion of insertions and deletions of [7].

Previous algorithms for these problems relied on color-coding [2]; these algorithms usually have an exponential space complexity, and a high time complexity. For the Graph Motif problem, [10] gives a randomized algorithm with an implicit $\mathcal{O}(87^k km)$ running time, while [3] describes a first randomized algorithm running in $\mathcal{O}(8.16^k m)$, and shows a second algorithm with $\mathcal{O}(4.32^k k^2 m)$ running time, using two different speed-up techniques ([4] and [13]). For the Max Motif problem, [9] present a randomized algorithm with an implicit $\mathcal{O}((32e^2)^k km)$ running time. Here again, we can apply Theorem 1 to improve the time and space complexities:

Proposition 2. *1.* MGM *is solvable by a randomized algorithm in* $\tilde{\mathcal{O}}(4^k k^2 m)$ *time and* $\tilde{\mathcal{O}}(kn)$ *space.*

2. MGMG *is solvable by a randomized algorithm in* $\tilde{\mathcal{O}}(4^k r^2 m)$ *time and* $\tilde{\mathcal{O}}(rn)$ *space.*

Proof. Point 1. We modify the circuit of Proposition 1 as follows. For each color $c \in C$ with $n_M(c) = m$, we introduce variables $y_{c,1}, ..., y_{c,m}$, and we introduce a node $Q_c = y_{c,1} + ... + y_{c,m}$. For each vertex $u \in V$, we introduce a variable x_u, and we define $P_{1,u} = x_u Q_{\chi(u)}$. The intuition is that the variables x_u will ensure that we choose different vertices to construct the tree, and that the variables $y_{c,i}$ will ensure that a given color cannot occur more than required. The resulting instance of MLD is $I' = (\mathcal{C}_I, 2k)$, and since $T(\mathcal{C}_I) = \mathcal{O}(k^2 m)$ and $S(\mathcal{C}_I) = \mathcal{O}(kn)$, we solve it in the claimed bounds by Theorem 1. A similar induction as in Proposition 1 shows that: for every $1 \leq i \leq k$, a multilinear monomial of $P_{i,u}$ has the form $x_{v_1} y_{c_1,j_1} ... x_{v_i} y_{c_i,j_i}$, and it is present iff there is a subtree (V_T, E_T) of G such that $u \in V_T$, $V_T = \{v_1, ..., v_i\}$ and $\chi(V_T) = \{\!\{c_1, ..., c_i\}\!\} \subseteq M$.

Point 2. We modify the construction of Point 1 by now setting $P_{1,u} = 1 + x_u Q_{\chi(u)}$ for each $u \in V$, and $P = \sum_{u \in V} \sum_{i=1}^{r} P_{i,u}$. Informally, adding the constant 1 to each $P_{1,u}$ permits to ignore some vertices of the subtree, allowing to only select a set S of k vertices such that $\chi(S) \subseteq M$. The correctness of the construction is shown by a similar induction as above. The catch here is that when considering two trees T_1, T_2 obtained from $P_{i',u}, P_{i-i',v}$, their selected vertices will be distinct, but they may have "ignored" vertices in common; we can then find a subset of $E(T_1) \cup E(T_2) \cup \{uv\}$ which forms a tree containing all selected vertices from T_1, T_2. □

3.3 Edge-weighted versions

We consider an edge-weighted variant of the problem, where the subtree is now required to have a given total weight, in addition to respecting the color constraints. This variant has been studied in [6] under the name Edge-Weighted Graph Motif. In our case, we define two problems, depending on whether we consider colorful or multiset motifs.

Name: Weighted Colorful Graph Motif (WCGM)
Input: A complete graph $G = (V, E)$, a function $\chi : V \to C$, a weight function $w : E \to \mathbb{N}$, integers k, r
Solution: A subtree $T = (V_T, E_T)$ of G such that (i) $|V_T| = k$, (ii) χ is injective on V_T, (iii) $\sum_{e \in E_T} w(e) \leq r$.

Name: Weighted Multiset Graph Motif (WMGM)
Input: A complete graph $G = (V, E)$, a function $\chi : V \to C$, a weight function $w : E \to \mathbb{N}$, integers k, r, a multiset M
Solution: A subtree $T = (V_T, E_T)$ of G such that (i) $|V_T| = k$, (ii) $\chi(V_T) \subseteq M$, (iii) $\sum_{e \in E_T} w(e) \leq r$.

We observe that the WMGM problem contains as particular case the Min-CC problem introduced in [8], which seeks a subgraph respecting the multiset motif, and having at most r connected components. Indeed, we can easily reduce Min-CC to WMGM: given the graph G, we construct a complete graph G' with the same vertex set, and we assign a weight 0 to edges of G, and a weight 1 to non-edges of G.

Proposition 3. *1.* WCGM *is solvable by a randomized algorithm in* $\tilde{\mathcal{O}}(2^k k^2 r^2 m)$ *time and* $\tilde{\mathcal{O}}(krn)$ *space.*
2. WMGM *is solvable by a randomized algorithm in* $\tilde{\mathcal{O}}(4^k k^2 r^2 m)$ *time and* $\tilde{\mathcal{O}}(krn)$ *space.*

Proof. We only prove 1, since 2 relies on the same modification as in Proposition 2. The construction of the arithmetic circuit is similar to the construction in Proposition 1. The set of variables is $\{x_c : c \in C\}$, and we introduce nodes $P_{i,j,u}$, for $1 \leq i \leq k$ and $0 \leq j \leq r$, whose multilinear monomials will correspond to colorful subtrees having i vertices including u, and with total weight $\leq j$. The definitions are as follows:

$$P_{i,j,u} = \sum_{i'=1}^{i-1} \sum_{v \in V} \sum_{j'=0}^{j-w(uv)} P_{i',j',u} P_{i-i',j-j'-w(uv),v} \text{ if } i > 1, \qquad P_{1,j,u} = x_{\chi(u)}$$

and $P = \sum_{u \in V} P_{k,r,u}$. The resulting instance of MLD is $I' = (\mathcal{C}_I, k)$, and since $T(\mathcal{C}_I) = \mathcal{O}(k^2 r^2 m)$ and $S(\mathcal{C}_I) = \mathcal{O}(krn)$, we solve it in the claimed bounds by Theorem 1. The correctness of the construction follows by showing that: given $1 \leq i \leq k, 0 \leq j \leq r$, $u \in V$, $x_{c_1} ... x_{c_d}$ is a multilinear monomial of $P_{i,j,u}$ iff (i) $d = i$ and (ii) there exists $T = (V_T, E_T)$ colorful subtree of G with $u \in V_T, \chi(V_T) = \{c_1, ..., c_d\}$ and $\sum_{e \in E_T} w(e) \leq j$. □

4 Counting vertex-colored subtrees

In this section, we consider the counting versions of the problems XCGM and XMGM introduced in Section 3. For the former, we show that its counting version #XCGM is FPT; for the latter, we prove that its counting version #XMGM is #W[1]-hard.

4.1 FPT algorithms for the colorful case

We show that #XCGM is fixed-parameter tractable (Proposition 5). We rely on a general result for #XMLD (Proposition 4), which uses inclusion-exclusion as in [14].

Say that a circuit $\mathcal{C}$ is *k-bounded* iff $P_{\mathcal{C}}$ has only monomials of degree $\leq k$. Observe that given a circuit $\mathcal{C}$, we can efficiently transform it in a k-bounded circuit $\mathcal{C}'$ such that (i) $\mathcal{C}$ and $\mathcal{C}'$ have the same monomials of degree k, (ii) $|\mathcal{C}'| \leq (k+1)^2|\mathcal{C}|$. The following result shows that we can efficiently count solutions for k-bounded circuits with k variables (and thus for general circuits, with an extra $\mathcal{O}(k^2)$ factor in the complexity).

Proposition 4. #XMLD *for k-bounded circuits is solvable in $\mathcal{O}(2^k T(\mathcal{C}))$ time and $\mathcal{O}(S(\mathcal{C}))$ space.*

Proof. Let $\mathcal{C}$ be the input circuit on a set X of k variables. For a monomial m let $Var(m)$ denote its set of variables. Given $S \subseteq X$, let N_S, resp. N'_S, be the number of monomials m of $P_{\mathcal{C}}$ such that $Var(m) = S$, resp. $Var(m) \subseteq S$. Observe that for every $S \subseteq X$, we have $N'_S = \sum_{T \subseteq S} N_T$. Therefore, by Möbius inversion it holds that for every $S \subseteq X$, $N_S = \sum_{T \subseteq S} (-1)^{|S \setminus T|} N'_T$.

Since $\mathcal{C}$ is k-bounded, N_X is the number of multilinear monomials of $P_{\mathcal{C}}$ having degree k. Now, each value N'_S can be computed by evaluating $\mathcal{C}$ under the mapping $\phi : X \to \mathbb{Z}$ defined by $\phi(v) = 1$ if $v \in S$, $\phi(v) = 0$ if $v \notin S$. By the Möbius inversion formula, we can thus compute the desired value N_X in $\mathcal{O}(2^k T(\mathcal{C}))$ time and $\mathcal{O}(S(\mathcal{C}))$ space. □

It is worth mentioning that Proposition 4 generalizes several counting algorithms based on inclusion-exclusion, such as the well-known algorithm for #Hamiltonian Path of [14], as well as results of [18]. Indeed, the problems considered in these articles can be reduced to counting multilinear monomials of degree n for circuits with n variables (where n is usually the number of vertices of the graph), which leads to algorithms running in $\mathcal{O}^*(2^n)$ time and polynomial space.

Let us now turn to applying Proposition 4 to the #XCGM problem. Recall that we defined in Proposition 1 a circuit $\mathcal{C}_I$ for the general CGM problem; we will have to modify it slightly for the purpose of counting solutions.

Proposition 5. #XCGM *is solvable in $\mathcal{O}(2^k k^3 m)$ time and $\mathcal{O}(k^2 n)$ space.*

Proof. Let I be an instance of XCGM. A *rooted solution* for I is a pair (u, T) where T is a solution of XCGM on I and u is a vertex of T (which must be seen as the root of the tree). The solutions of XCGM on I are also called *unrooted solutions*. Let $N_r(I)$ and $N_u(I)$ be the number of rooted, resp. unrooted, solutions for I. We will show how to compute $N_r(I)$ in the claimed time and space bounds; since $N_u(I) = \frac{N_r(I)}{k}$, the result will follow.

To compute N_r, observe first that we cannot apply Proposition 4 to the circuit $\mathcal{C}_I$ of Proposition 1. Indeed, the circuit $\mathcal{C}_I$ counts the ordered subtrees, and not the unordered ones. Therefore, we need to modify the circuit in the following way: at each vertex v of V_T, we examine its children by increasing color. This leads us to define the following circuit $\mathcal{C}'_I$: suppose w.l.o.g. that $C = \{1, ..., k\}$, introduce nodes $P_{i,j,u}$ for each $1 \leq i \leq k, 1 \leq j \leq k+1, u \in V$, variables x_i for each $1 \leq i \leq k$, and define:

$$P_{1,j,u} = x_{\chi(u)}, \quad P_{i,j,u} = 0 \text{ if } i \geq 2, j = k+1$$

$$P_{i,j,u} = P_{i,j+1,u} + \sum_{i'=1}^{i-1} \sum_{v \in N_G(u):\chi(v)=j} P_{i',j+1,u} P_{i-i',1,v} \text{ if } i \geq 2, 1 \leq j \leq k$$

Let us also introduce a root node $P = \sum_{u \in V} P_{k,1,u}$. Given $1 \leq i, j \leq k$ and $u \in V$, let $\mathcal{S}_{i,j,u}$ denote the set of pairs (u, T) where (i) T is a properly colored subtree of I containing u and having i vertices, (ii) the neighbors of u in T have colors $\geq j$. It can be shown by induction on i that: there is a bijection between $\mathcal{S}_{i,j,u}$ and the multilinear monomials of $P_{i,j,u}$. Therefore, the number of multilinear monomials of P is equal to N_r; since $T(\mathcal{C}'_I) = \mathcal{O}(k^3 m), S(\mathcal{C}'_I) = \mathcal{O}(k^2 n)$ and since $\mathcal{C}'_I$ is k-bounded, it follows by Proposition 4 that N_r can be computed in $\mathcal{O}(2^k k^3 m)$ time and $\mathcal{O}(k^2 n)$ space. □

4.2 Hardness of the multiset case

In this subsection, we show that #XMGM is #W[1]-hard. For convenience, we first restate the problem in terms of *vertex-distinct embedded subtrees*.

Let $G = (V, E)$ and $H = (V', E')$ be two multigraphs. An *homomorphism* of G into H is a pair $\phi = (\phi_V, \phi_E)$ where $\phi_V : V \to V'$ and $\phi_E : E \to E'$, such that if $e \in E$ has endpoints x, y then $\phi_E(e)$ has endpoints $\phi_V(x), \phi_V(y)$. An *embedded subtree* of G is a pair $\mathcal{T} = (T, \phi_V, \phi_E)$ where $T = (V_T, E_T)$ is a tree, and (ϕ_V, ϕ_E) is an homomorphism from T into G. We say that $\mathcal{T}$ is a *vertex-distinct* embedded subtree of G (a "vdst" of G) if ϕ_V is injective. We say $\mathcal{T}$ is an *edge-distinct* embedded subtree of G (an "edst" of G) iff ϕ_E is injective. We restate XMGM as follows:

Name: Exact Multiset Graph Motif (XMGM)
Input: A graph $G = (V, E)$, an integer k, a set C, a function $\chi : V \to C$, a multiset M over C s.t. $|M| = k$.
Solution: A vdst (T, ϕ_V, ϕ_E) of G s.t. $\chi \circ \phi_V(V_T) = M$.

We first show the hardness of two intermediate problems (Lemma 1). Before defining these problems, we need the following notions. Consider a multigraph $G = (V, E)$. Consider a partition $\mathcal{P}$ of V into $V_1, ..., V_k$, and a tuple $t \in [r]^k$. A $(\mathcal{P}, t)$-*mapping* from a set A is an injection $\psi : A \to V \times [r]$ such that for every $x \in A$, if $\psi(x) = (v, i)$ with $v \in V_j$, then $1 \leq i \leq t_j$. From ψ, we define its *reduction* as the function $\psi^r : A \to V$ defined by $\psi^r(x) = v$ whenever $\psi(x) = (v, i)$. We also define a tuple $T(\psi) = (n_1, ..., n_k) \in [r]^k$ such that for each $i \in [k]$, $n_i = \max_{v \in V_i} |\{x \in A : \psi^r(x) = v\}|$.

Given two tuples $t, t' \in [r]^k$, denote $t \leq t'$ iff $t_i \leq t'_i$ for each $i \in [k]$. Note that for a $(\mathcal{P}, t)$-mapping ψ, we always have $T(\psi) \leq t$ since ψ is injective. We say that a $(\mathcal{P}, t)$-*labeled edst for* G is a tuple (T, ψ_V, ψ_E) where (i) $T = (V_T, E_T)$ is a tree, (ii) ψ_V is a $(\mathcal{P}, t)$-mapping from V_T, (iii) (T, ψ^r_V, ψ_E) is an edst of G. Our intermediate problems are defined as follows:

Name: Multicolored Embedded Subtree-1 (MEST $-$ 1)
Input: Integers k, r, a k-partite multigraph G with partition $\mathcal{P}$, a tuple $t \in [r]^k$
Solution: A $(\mathcal{P}, t)$-labeled edst (T, ψ_V, ψ_E) for G s.t. $|V_T| = r$ and $T(\psi_V) = t$.

The MEST $-$ 2 problem is defined similarly, except that we do not require that $T(\psi_V) = t$ (and thus we only have $T(\psi_V) \leq t$). While we will only need #MEST $-$ 2 in our reduction for #XMGM, we first show the hardness of #MEST $-$ 1, then reduce it to #MEST $-$ 2.

Lemma 1. #MEST$-$1 *and* #MEST$-$2 *are* #W[1]*-hard for parameter* (k, r).

The proof is omitted due to space constraints.

Proposition 6. #XMGM *is* #W[1]*-hard for parameter* k.

Proof. We reduce from #MEST $-$ 2, and conclude using Lemma 1. Let $I = (k, r, G, t)$ be an instance of #MEST $-$ 2, where $G = (V, E)$ is a multigraph, and let $\mathcal{S}_I$ be its set of solutions. From G, we construct a graph H as follows: (i) we subdivide each edge $e \in E$, creating a new vertex $a[e]$, (ii) we substitute each vertex $v \in V_i$ by an independent set formed by t_i vertices $b[v, 1], ..., b[v, t_i]$. We let A be the set of vertices $a[e]$ and B the set of vertices $b[v, i]$, we therefore have a bipartite graph $H = (A \cup B, F)$. We let $I' = (H, 2r - 1, C, \chi, M)$, where $C = \{1, 2\}$, χ maps A to 1 and B to 2, and M consists of $r - 1$ occurrences of 1 and r occurrences of 2.

Then I' is our resulting instance of #XMGM, and we let $\mathcal{S}_{I'}$ be its set of solutions. Notice that by definition of χ and M, $\mathcal{S}_{I'}$ is the set of vdst (T, ϕ_V, ϕ_E) of H containing $r - 1$ vertices mapped to A and r vertices mapped to B. We now show that we have a parsimonious reduction, by describing a bijection $\Phi : \mathcal{S}_I \to \mathcal{S}_{I'}$. Consider $\mathcal{T} = (T, \psi_V, \psi_E)$ in $\mathcal{S}_I$; we define $\Phi(\mathcal{T}) = (T', \phi_V, \phi_E)$ as follows:

- For each edge $e = uv \in E(T)$, we have $f_e := \psi_E(e) \in E(G)$: we then subdivide e, creating a new vertex x_e. Let T' be the resulting tree;

– For each vertex x_e, we define $\phi_V(x_e) = a[f_e]$. For each other vertex u of T', we have $u \in V(T)$, let $(v, i) = \psi_V(u)$; we then set $\phi_V(u) = b[v, i]$ (this is possible since if $v \in V_j$ then $1 \leq i \leq t_j$, by definition of ψ_V).

From ϕ_V, we then define ϕ_E in a natural way. Then $\mathcal{T}' = \Phi(\mathcal{T})$ is indeed in $\mathcal{S}_{I'}$: (i) $\mathcal{T}'$ is a vertex distinct subtree of H (by definition of ϕ_V and since $\mathcal{T}$ was edge-distinct, the values $\phi_V(x_e)$ are distinct; by injectivity of ψ_V, the other values $\phi_V(u)$ are distinct); (ii) it has $r-1$ vertices mapped to A and r vertices mapped to B. To prove that Φ is a bijection, we describe the inverse correspondence $\Psi : \mathcal{S}_{I'} \to \mathcal{S}_I$. Consider $\mathcal{T}' = (T', \phi_V, \phi_E)$ in $\mathcal{S}_{I'}$; we define $\Psi(\mathcal{T}') = (T, \psi_V, \psi_E)$ as follows. Let A', B' be the vertices of T' mapped to A, B respectively. Let i be the number of nodes of A' which are leaves: since the nodes of A' have degree 1 or 2 in T' depending on whether they are leaves or internal nodes, we then have $|E(T')| \leq i + 2(r - 1 - i) = 2r - i - 2$; since $|E(T')| = 2r - 2$, we must have $i = 0$. It follows that all leaves of T' belong to B'; from T', by contracting each vertex of A' in T' we obtain a tree T with r vertices. We then define ψ_V, ψ_E as follows: (i) given $u \in B'$, if $\phi_V(u) = b[v, j]$, then $\psi_V(u) = (v, j)$; (ii) given $e = uv \in E(T)$, there corresponds two edges $ux, vx \in E(T')$ with $x \in A'$, and we thus have $\phi_V(x) = a[f]$, from which we define $\psi_E(e) = f$. It is easily seen that the resulting $\mathcal{T} = \Psi(\mathcal{T}')$ is in $\mathcal{S}_I$, and that the operations Φ and Ψ are inverse of each other. □

5 Conclusion

In this paper, we have obtained improved FPT algorithms for several variants of the Graph Motif problem. Reducing to the Multilinear Detection problem yielded a significant reduction of the base of the exponent in the time complexity, as well as a polynomial space complexity. We have also considered the counting versions of these problems, for the first time in the literature.

We would like to mention two open questions of theoretical interest. First, we would like to know whether the $\mathcal{O}^*(4^k)$ running times obtained for multiset motifs can be further reduced. Second, while we have shown that #XMGM was #W[1]-hard for a motif with two colors, we leave open its complexity for one color. Note that this problem amounts to count the k-vertex subtrees of an (uncolored) graph.

From a practical point of view, it would be interesting to evaluate the performance of the algorithms described in this article. In particular, how do they compare to implementations based on color-coding or ILPs [7,5]? In this respect, two important questions are whether the algorithm of Theorem 1 can be efficiently derandomized, and whether it can be adapted to efficiently recover a solution without backtracking.

References

1. E. Alm and A.P. Arkin. Biological networks. *Curr. Opin. Struct. Biol.*, 13(2):193–202, 2003.

2. N. Alon, R. Yuster, and U. Zwick. Color-coding. *Journal of the ACM*, 42(4):844–856, 1995.
3. N. Betzler, M.R. Fellows, C. Komusiewicz, and R. Niedermeier. Parameterized Algorithms and Hardness Results for Some Graph Motif Problems. In *CPM 2008*, volume 5029 of *LNCS*, pages 31–43, 2008.
4. A. Björklund, T. Husfeldt, P. Kaski, and M. Koivisto. Fourier meets möbius: fast subset convolution. In *STOC 2007*, pages 67–74, 2007.
5. G. Blin, F. Sikora, and S. Vialette. GraMoFoNe: a Cytoscape plugin for querying motifs without topology in Protein-Protein Interactions networks. In *BICoB 2010*, pages 38–43, 2010.
6. S. Böcker, F. Rasche, and T. Steijger. Annotating Fragmentation Patterns. In *WABI 2009*, volume 5724 of *LNBI*, pages 13–24, 2009.
7. S. Bruckner, F. Hüffner, R.M. Karp, R. Shamir, and R. Sharan. Topology-Free Querying of Protein Interaction Networks. In *RECOMB 2009*, volume 5541 of *LNCS*, pages 74–89, 2009.
8. R. Dondi, G. Fertin, and S. Vialette. Weak pattern matching in colored graphs: Minimizing the number of connected components. In *ICTCS 2007*, pages 27–38, 2007.
9. R. Dondi, G. Fertin, and S. Vialette. Maximum Motif Problem in Vertex-Colored Graphs. In *CPM 2009*, volume 5577 of *LNCS*, pages 221–235, 2009.
10. M.R. Fellows, G. Fertin, D. Hermelin, and S. Vialette. Sharp Tractability Borderlines for Finding Connected Motifs in Vertex-Colored Graphs. In *ICALP 2007*, volume 4596 of *LNCS*, pages 340–351, 2007.
11. J. Flum and M. Grohe. The Parameterized Complexity of Counting Problems. *SIAM Journal on Computing*, 33(4):892–922, 2004.
12. J. Flum and M. Grohe. *Parameterized Complexity Theory*. Springer-Verlag, 2006.
13. F. Hüffner, S. Wernicke, and T. Zichner. Algorithm Engineering For Color-Coding To Facilitate Signaling Pathway Detection. In *APBC 2007*, pages 277–286, 2007.
14. R.M. Karp. Dynamic-programming meets the principle of inclusion and exclusion. *Oper. Res. Lett.*, 1:49–51, 1982.
15. I. Koutis. Faster Algebraic Algorithms for Path and Packing Problems. In *ICALP 2008*, volume 5125 of *LNCS*, pages 575–586, 2008.
16. I. Koutis and R. Williams. Limits and Applications of Group Algebras for Parameterized Problems. In *ICALP 2009*, volume 5555 of *LNCS*, pages 653–664, 2009.
17. V. Lacroix, C.G. Fernandes, and M.-F. Sagot. Motif Search in Graphs: Application to Metabolic Networks. *Trans. Comput. Biol. Bioinform.*, 3(4):360–368, 2006.
18. J. Nederlof. Fast Polynomial-Space Algorithms Using Möbius Inversion: Improving on Steiner Tree and Related Problems. In *ICALP 2009*, volume 5555 of *LNCS*, pages 713–725, 2009.
19. S. Schbath, V. Lacroix, and M.-F. Sagot. Assessing the exceptionality of coloured motifs in networks. *EURASIP JBSB*, pages 1–9, 2009.
20. R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24:427–433, 2006.
21. R. Williams. Finding paths of length k in $O^*(2^k)$ time. *IPL*, 109(6):315–318, 2009.

6 Appendix

6.1 End of proof of Proposition 1

Given a set $S \subseteq C$, define the multilinear monomial $\pi_S := \prod_{c \in S} x_c$. Given $u \in V(T)$ and $S \subseteq C$, an (u, S)-*solution* is a subtree $T = (V_T, E_T)$ of G, such that $u \in V_T$, T is distinctly colored by χ, and $\chi(V_T) = S$. We show by induction on $1 \leq i \leq k$ that: π_S is a multilinear monomial of $P_{i,u}$ iff (i) $|S| = i$ and (ii) there exists an (u, S)-solution. This is clear when $i = 1$; now, suppose that $i \geq 2$, and assume that the property holds for every $1 \leq j < i$.

Suppose that $|S| = i$ and that $T = (V_T, E_T)$ is an (u, S)-solution, let us show that π_S is a multilinear monomial of $P_{i,u}$. Let v be a neighbor of u in T, then removing the edge uv from T produces two trees T_1, T_2 with T_1 containing u and T_2 containing v. These two trees are distinctly colored, let S_1, S_2 be their respective color sets, and let i_1, i_2 be their respective sizes. Since T_1 is an (u, S_1)-solution, π_{S_1} is a multilinear monomial of $P_{i_1,u}$ by induction hypothesis. Since T_2 is a (v, S_2)-solution, π_{S_2} is a multilinear monomial of $P_{i_2,v}$ by induction hypothesis. It follows that $\pi_S = \pi_{S_1}\pi_{S_2}$ is a multilinear monomial of $P_{i_1,u}P_{i_2,v}$, and thus of $P_{i,u}$.

Conversely, suppose that π_S is a multilinear monomial of $P_{i,u}$. By definition of $P_{i,u}$, there exists $1 \leq i' \leq i-1$ and $v \in N_G(u)$ such that π_S is a multilinear monomial of $P_{i',u}P_{i-i',v}$. We can then partition S into S_1, S_2, with π_{S_1} multilinear monomial of $P_{i',u}$ and π_{S_2} multilinear monomial of $P_{i-i',v}$. Induction hypothesis therefore implies that (i) $|S_1| = i'$ and $|S_2| = i - i'$, (ii) there exists an (u, S_1)-solution $T_1 = (V_1, E_1)$ and a (v, S_2)-solution $T_2 = (V_2, E_2)$. Since S_1, S_2 are disjoint, it follows that $|S| = i$, which proves (i); besides, V_1, V_2 are disjoint, and thus $T = (V_1 \cup V_2, E_1 \cup E_2 \cup \{uv\})$ is an (u, S)-solution, which proves (ii).

6.2 Proof of Lemma 1

We first reduce #Multicolored Clique to #MEST $-$ 1. Our source problem #Multicolored Clique is the counting version of Multicolored Clique, which is easily seen to be #W[1]-hard. Let $I = (G, k)$ be an instance of the problem, where $G = (V, E)$ has a partition $\mathcal{P}$ into classes $V_1, ..., V_k$. Our target instance is $I' = (k, r, H, t)$ with $r = k^2 - k + 1$ and $t = (k, k-1, ..., k-1)$. The graph H is obtained by splitting every edge e in two parallel edges; then H is a k-partite multigraph with partition $\mathcal{P}$. Let $\mathcal{S}_I$, $\mathcal{S}_{I'}$ be the solution sets of I and I' respectively. Let $\mathcal{K}_k$ be the multigraph with k vertices $1, ..., k$, and with two parallel edges between distinct vertices; its partition is $\mathcal{P}_k$ consisting of the sets $\{1\}, ..., \{k\}$. Let $\mathcal{U}_k$ denote the set of $(\mathcal{P}_k, t)$-labeled edsts $(\mathcal{T}, \psi_V, \psi_E)$ for $\mathcal{K}_k$ such that $T(\psi_V) = t$. Observe that $\mathcal{U}_k \neq \emptyset$: since every vertex of $\mathcal{K}_k$ has degree $2(k-1)$, it follows that $\mathcal{K}_k$ has an Eulerian path starting at 1, which visits k times the vertex 1, and each other vertex $k-1$ times. We claim that $|\mathcal{S}_{I'}| = |\mathcal{U}_k||\mathcal{S}_I|$, which will prove the correctness of the reduction. To this aim, we will describe a bijection $\Phi : \mathcal{S}_{I'} \times \mathcal{U}_k \to \mathcal{S}_I$.

Consider a pair $P = (C, \mathcal{T}) \in \mathcal{S}_{I'} \times \mathcal{U}_k$ with $\mathcal{T} = (T, \psi_V, \psi_E)$ and $C = \{v_1, ..., v_k\}$ multicolored clique of G (with $v_i \in V_i$). Let $\phi = (\phi_V, \phi_E)$ be the homomorphism of $\mathcal{K}_k$ into H which maps i to v_i, and the parallel edges accordingly. We then define $\mathcal{T}' = \Phi(P)$ by $\mathcal{T}' = (T, \psi'_V, \psi'_E)$, where (i) ψ'_V is defined so that if $\psi_V(u) = (v, i)$ and if $\phi_V(v) = w$ then $\psi'_V(u) = (w, i)$, (ii) $\psi'_E = \psi_E \circ \phi_E$. We verify that $\mathcal{T}' \in \mathcal{S}_I$: indeed, it is a $(\mathcal{P}, t)$-labeled edst of G and $T(\psi'_V) = t$ (since we have composed with injective functions ϕ_V, ϕ_E). To prove that Φ is a bijection, we define the inverse function $\Psi : \mathcal{S}_I \to \mathcal{S}_{I'} \times \mathcal{U}_k$ as follows. Consider $\mathcal{T}' = (T, \psi'_V, \psi'_E)$ $(\mathcal{P}, t)$-labeled edst of G, with $T(\psi'_V) = t$. This equality yields vertices $v_1 \in V_1, ..., v_k \in V_k$ such that $|(\psi^r_V)^{-1}(v_i)| = t_i$. Let $C = \{v_1, ..., v_k\}$, then C is a multicolored clique of G: indeed, $H[C]$ has at most $k^2 - k$ edges, and since ψ'_E is injective it must have exactly $k^2 - k$ edges, implying that $G[C]$ is a complete graph. We can then define (ψ_V, ψ_E) from (ψ'_V, ψ'_E) by "projecting" v_i on i, and the parallel edges accordingly (for instance, if $\psi'_V(u) = (v_i, j)$ then $\psi_V(u) = (i, j)$). We finally define $P = \Psi(\mathcal{T}')$ by $P = (C, \mathcal{T})$ where $\mathcal{T} = (T, \psi_V, \psi_E)$. It is easy to see that $P \in \mathcal{S}_{I'} \times \mathcal{U}_k$, and that Φ and Ψ are inverse of each other.

We now give a Turing-reduction of #MEST-1 to #MEST-2. Given a tuple $t \in [r]^k$, we define the instance $I_t = (k, r, G, t)$, and we let $\mathcal{S}_t, \mathcal{S}'_t$ be its solution sets for #MEST-1, #MEST-2 respectively. Let $N_t = |\mathcal{S}_t|$ and $N'_t = |\mathcal{S}'_t|$. We have for every $t \in [r]^k$: $N'_t = \sum_{t' \leq t} N_{t'}$, which yields by Möbius inversion that for every $t \in [r]^k$: $N_t = \sum_{t' \leq t} \mu(t, t') N'_{t'}$ [3]. Therefore, we can compute a value N_t using $\mathcal{O}(2^k)$ oracle calls for #MEST $- 2$, thereby solving #MEST $- 1$. □

[3] where $\mu(t, t')$ is 0 if there exists $i \in [k]$ s.t. $t_i - t'_i > 1$, and is otherwise equal to $(-1)^r$ where r is the number of $i \in [k]$ s.t. $t_i - t'_i = 1$.