The Graph Motif Problem Parameterized by the Structure of the Input Graph

Édouard Bonnet[*1] **and Florian Sikora**[2]

1 **Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI), Budapest, Hungary**
bonnet.edouard@sztaki.mta.hu

2 **PSL, Université Paris-Dauphine, LAMSADE UMR CNRS 7243, France**
florian.sikora@dauphine.fr

Abstract

The Graph Motif problem was introduced in 2006 in the context of biological networks. It consists of deciding whether or not a multiset of colors occurs in a connected subgraph of a vertex-colored graph. Graph Motif has been analyzed from the standpoint of parameterized complexity. The main parameters which came into consideration were the size of the multiset and the number of colors. Though, in the many applications of Graph Motif, the input graph originates from real-life and has structure. Motivated by this prosaic observation, we systematically study its complexity relatively to graph structural parameters. For a wide range of parameters, we give new or improved FPT algorithms, or show that the problem remains intractable. Interestingly, we establish that Graph Motif is W[1]-hard (while in W[P]) for parameter max leaf number, which is, to the best of our knowledge, the first problem to behave this way.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems

Keywords and phrases Parameterized Complexity, Structural Parameters, Graph Motif, Computational Biology

Digital Object Identifier 10.4230/LIPIcs.IPEC.2015.319

1 Introduction

The Graph Motif problem has received a lot of attention during the last decade. Informally, Graph Motif is defined as follows: given a graph with arbitrary colors on the nodes and a multiset of colors called the motif, the goal is to decide if there exists a subset of vertices of the graph such that (1) the subgraph induced by this subset is connected and (2) the colors on the subset of vertices match the motif, i.e. each color appears the same number of times as in the motif. Originally, this problem is motivated by applications in biological network analysis [24]. However, it proves useful in social or technical networks [4] or in the context of mass spectrometry [8].

Studying biological networks allows a better characterization of species, by determining small recurring subnetworks, often called *motifs*. Such motifs can correspond to a set of nodes realizing some function, which may have been evolutionary preserved. Thus, it is crucial to determine these motifs to identify common elements between species and transfer the biological knowledge. Graph Motif corresponds to topology-free queries and can be

* This work is partially supported by ERC Starting Grant PARAMTIGHT (n. 280152).

seen as a variant of a graph pattern matching problem with the sole topological requirement of connectedness. Such queries were also studied extensively for sequences during the last thirty years, and with the increase of knowledge about biological networks, it is relevant to extend these queries to networks [30].

2 Preliminaries and previous work

For any two integers $x < y$, we set $[x, y] := \{x, x+1, \ldots, y-1, y\}$, and for any positive integer x, $[x] := [1, x]$. If $G = (V, E)$ is a graph and $S \subseteq V$ a subset of vertices, $G[S]$ denotes the subgraph of G induced by S. For a vertex $v \in V$, the set of neighbors of v in G is denoted by $N_G(v)$, or simply $N(v)$, and $N_G(S) := (\bigcup_{v \in S} N(v)) \setminus S$ and will often be written just $N(S)$. We define $N[v] := N(v) \cup \{v\}$ and $N[S] := N(S) \cup S$. We say that a vertex v *dominates* a set of vertices S if $S \subseteq N[v]$. A set of vertices R *dominates* another set of vertices S if $S \subseteq N[R]$. If $G = (V, E)$ is a graph and $V' \subseteq V$, $G - V'$ denotes the graph $G[V \setminus V']$. A *universal vertex* v, in a graph $G = (V, E)$, is such that $N_G[v] = V$. A *matching* of a graph is a mutually disjoint set of edges. In an explicitly bipartite graph $G = (V_1 \cup V_2, E)$, we call a matching of size $\min(|V_1|, |V_2|)$ a *perfect matching*. A *cluster graph* (or simply, *cluster*) is a disjoint union of cliques. A *co-cluster graph* (or, *co-cluster*) is the complement graph of a cluster graph. If $\mathcal{C}$ is a class of graphs, the *distance to* $\mathcal{C}$ of a graph G is the minimum number of vertices to remove from G to get a graph in $\mathcal{C}$.

If $f : A \to B$ is a function and $A' \subseteq A$, $f_{|A'}$ denotes the restriction of f to A', that is $f_{|A'} : A' \to B$ such that $\forall x \in A'$, $f_{|A'}(x) := f(x)$. Similarly, if E is a set of edges on vertices of V and $V' \subseteq V$, $E_{|V'}$ is the subset of edges of E having both endpoints in V'.

Graph Motif and multisets. Graph Motif is defined as follows:

Graph Motif

- **Input**: A triple (G, c, M), where $G = (V, E)$ is a graph, $c : V \to \mathcal{C}$ gives some color of $|\mathcal{C}|$ to the vertices, and M is a multiset of colors of $\mathcal{C}$.
- **Output**: A subset $P \subseteq V$ such that (1) $G[P]$ is connected and (2) $c(P) = M$.

We will refer to condition (1) as the *connectivity constraint* and to condition (2) as the *multiset constraint*. For convenience, if $S \subseteq V$, $c(S)$ will denote the multiset of colors of vertices in S.

The *multiplicity* of element x in multiset M, denoted by $m_M(x)$ is the number of occurences of x in M. The cardinality of a multiset M denoted by $|M|$ is its number of elements *with their multiplicity*: $\Sigma_{x \in M} m_M(x)$. If M and N are two multisets, $M \cup N$ is the multiset A such that $\forall x$, $m_A(x) = m_M(x) + m_N(x)$, and $M \setminus N$ is the multiset D such that $\forall x \in M$, $m_D(x) = \max(0, m_M(x) - m_N(x))$ (and $\forall x \notin M$, $m_D(x) = 0$). We write $M \subseteq N$ iff $M \setminus N = \emptyset$ and $M \subset N$ iff $M \subseteq N$ and $M \neq N$. For example, let $M = \{1, 2, 2, 4, 5, 5, 5\}$ and $N = \{1, 1, 1, 2, 2, 3, 3, 4, 5, 5, 5, 5\}$. Here, $|M| = 7$, $|N| = 12$, $M \setminus N = \emptyset$, $N \setminus M = \{1, 1, 3, 3, 5\}$, and $M \subseteq N$.

Parameterized Complexity and ETH. A parameterized problem (I, k) is said *fixed-parameter tractable* (or in the class FPT) w.r.t. (with respect to) parameter k if it can be solved in $f(k) \cdot |I|^c$ time (in *fpt-time*), where f is any computable function and c is a constant (see [28, 14] for more details about fixed-parameter tractability). The parameterized complexity hierarchy is composed of the classes $\mathsf{FPT} \subseteq \mathsf{W}[1] \subseteq \mathsf{W}[2] \subseteq \cdots \subseteq \mathsf{W}[\mathsf{P}] \subseteq \mathsf{XP}$. The class XP contains problems solvable in time $|I|^{f(k)}$, where f is an unrestricted function.

A powerful technique to design parameterized algorithms is *kernelization*. In short, kernelization is a polynomial-time self-reduction algorithm that takes an instance (I, k) of a parameterized problem P as input and computes an equivalent instance (I', k') of P such that $|I'| \leqslant h(k)$ for some computable function h and $k' \leqslant k$. The instance (I', k') is called a *kernel* in this case. If the function h is polynomial, we say that (I', k') is a polynomial kernel.

The *Exponential Time Hypothesis* (ETH) is a conjecture by Impagliazzo et al. [21] asserting that there is no $2^{o(n)}$-time algorithm for 3-SAT on instances with n variables. The so-called sparsification lemma, also proved in [21], shows that if ETH turns out to be true, then there is no $2^{o(n+m)}$-time algorithm solving 3-SAT where m is the number of clauses.

Previous work. Many results about the complexity of Graph Motif are known. The problem is NP-hard even with strong restrictions. For instance, it remains NP-hard for bipartite graphs of maximum degree 4 and motifs containing two colors only [15], or for trees of maximum degree 3 and when the motif is colorful (that is, no color occurs more than once) [15], or for rooted trees of depth 2 [2]. However, the problem is solvable in polynomial time when the graph is a caterpillar [2], or when both the number of colors in the motif and the treewidth of the graph are bounded by a constant [15].

As Graph Motif is intractable even for very restricted classes of graphs, and considering that, in practice, the motif is supposed to be small compared to the graph, the parameterized complexity of Graph Motif relatively to the size of the motif has been tackled. It is indeed in FPT when parameterized by the size of the motif. At least seven different papers gave an FPT algorithm [15, 4, 20, 23, 5, 30, 29]. The best (randomized) algorithm runs in time $O^*(2^k)$ where the O^* notation suppresses polynomial factors [5, 30] and works well in practice for small values of k, even with hundreds of millions of edges [6]. The current best deterministic algorithm takes time $O^*(5.22^k)$ [29]. However, an algorithm running in time $O^*((2-\epsilon)^k)$ would break the 2^n barrier in solving Set Cover instances with n elements [5]. Besides, it is unlikely that Graph Motif admits a polynomial kernel, even on a restricted class of trees [2]. Ganian also proved that the problem is in FPT when the parameter is the size of a minimum vertex cover of the graph [17]. Actually, his algorithm is given for a smaller parameter called twin-cover. Ganian also show that Graph Motif can be solved in $O^*(2^k)$ for graphs with neighborhood diversity k [18]. On the negative side, the problem is W[1]-hard relatively to the number of colors, even for trees [15]. To deal with the huge rate of noise in the biological data, many variants of the problem has been introduced. For example, the approach of Dondi *et al.* requires a solution with a minimum number of connected components [13], while the one of Betzler *et al.* asks for a 2-connected solution [4]. In other variants stemming purely from bio-informatics, some colors can be added to, substituted or subtracted from the solution [10, 13].

In light of the previous paragraphs, it is clear that the complexity of Graph Motif is well known for different versions and constraints on the problem itself. However, only few works take into account the structure of the input graph. We believe that this an interesting direction since Graph Motif has applications in real-life problems, where the input is not random. For example, some biological networks have been shown scale-free or with small diameter [1]. We will therefore introduce a systematic study with respect to structural graph parameters [22, 16]. We believe that this is also of theoretical interest, to understand how a given parameter influences the complexity of the problem.

Organization. In Section 3, we improve the known FPT algorithms with parameter distance to clique, vertex cover number, and edge clique cover number. We also give a parameterized

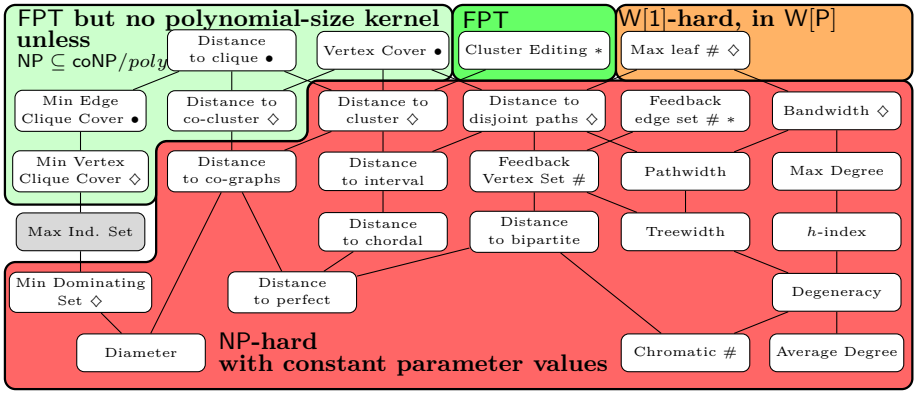

Figure 1 Hasse diagram of the relationship between different parameters ([22]). Two parameters are connected by a line if the parameter below can be polynomially upper-bounded in the parameter above. For example, *vertex cover* is above *distance to disjoint paths* since deleting a vertex cover produces an independent set, hence a set of disjoint paths. Therefore, positive results propagate upwards, while negative results propagate downwards. Results marked by ♢ are obtained in this paper, those marked with • are improvement of existing results, and those marked with ∗ are corollaries of existing results.

algorithm for the parameter distance to co-cluster which nicely reuses the FPT algorithms for both vertex cover number and distance to clique and another algorithm for parameter vertex clique cover number. These last two algorithms are noteworthy since a bounded distance to co-cluster or a bounded vertex clique cover number do not imply a bounded neighborhood diversity, a parameter for which Graph Motif was already known to be in FPT. We also show that a polynomial kernel for the aforementioned parameters is unlikely. In Section 4, we show that Graph Motif remains hard on graphs of constant distance to disjoint paths, or constant bandwidth, or constant distance to cluster, or constant dominating set number. More surprisingly, we establish that Graph Motif is W[1]-hard (but in W[P]) for the parameter max leaf number. To the best of our knowledge, there is no previously known problem behaving similarly when parameterized by max leaf number. Indeed, graphs with bounded max leaf number are really simple and, for instance, all the problems studied in [16] are FPT for this parameter. These positive and negative results draw a tight line between tractability and intractability (see Figure 1). Due to space constraints, some proofs (marked with ★) are deferred to the full version of the paper.

3 FTP algorithms and lower bound in the size of kernels

In this section, we improve or establish new FPT algorithms for several parameters. We also give a lower bound on the size of the kernel for all those parameters except *cluster editing number*. Figure 1 summarizes those results.

3.1 Cluster editing and linear neighborhood diversity

The cluster editing number of a graph is the number of edge deletions or additions required to get a cluster graph. It can be computed in time $O^*(1.62^k)$ [7]. We will use a known result involving another parameter called neighborhood diversity introduced by Lampis [25]. A graph has neighborhood diversity k if there is a partition of its vertices into at most k sets

such that all the vertices in each set *have the same type*. And, two vertices u and v *have the same type* iff $N(v) \setminus \{u\} = N(u) \setminus \{v\}$. We say that a graph parameter κ has *linear* (resp. *exponential*) *neighborhood diversity* if, for every positive integer k, all the graphs G such that $\kappa(G) \leqslant k$ have neighborhood diversity ck (resp. c^k) for some constant c. We say that a parameter κ has *unbounded neighborhood diversity*, if there is *no* function f such that all graphs G with $\kappa(G) \leqslant k$ have neighborhood diversity $f(k)$.

▶ **Theorem 1** ([18]). Graph Motif *can be solved in* $O^*(2^k)$ *on graphs with neighborhood diversity* k.

The following result is a direct consequence of the fact that, restricted to connected graphs, cluster editing has linear neighborhood diversity.

▶ **Corollary 2.** Graph Motif *can be solved in* $O^*(8^k)$*, where* k *is the cluster editing number.*

Proof. Let $(G = (V, E), c, M)$ be any instance of Graph Motif. We can assume that G is connected, otherwise we run the algorithm in each connected component of G. Let X be the set of vertices which are an endpoint of an edited edge (deleted or added) and let G' be the cluster graph obtained by the k edge editions. We may observe that $|X| \leqslant 2k$ and that the number of maximal cliques $C_1, \ldots, C_l$ in G' is bounded by k (otherwise, G could not be connected). For each $i \in [l]$, and for each vertex $v \in C_i \setminus X$, $N[x] = C_i$. Thus the neighborhood diversity of G is bounded by $|X| + l \leqslant 2k + k = 3k$. So, we can run the algorithm for bounded neighborhood diversity [18] and it takes time $O^*(2^{3k})$. ◀

3.2 Parameters with exponential neighborhood diversity

The next three parameters that we consider are *distance to clique*, *size of a minimum vertex cover*, and *size of a minimum edge clique cover*. For the first two, a value of k entails that the neighborhood diversity is at most $k + 2^k$; and neighborhood diversity 2^k for the third one. Therefore, Ganian has already given an algorithm running in double exponential time for these parameters ($O^*(2^{k+2^k})$ or $O^*(2^{2^k})$, see Theorem 1, [17, 18]). We improve this bound to single exponential time $2^{O(k)}$ (more precisely $O^*(8^k)$) for distance to clique and to $2^{O(k \log k)}$ for the vertex cover and edge clique cover numbers. The latter running time is sometimes called *slightly superexponential* FPT time [26]. Then, we prove that for each of those three parameters, a polynomial kernel is unlikely.

As a preparatory lemma for the algorithm parameterized by distance to clique, we show that a variant of Set Cover with thresholds is solvable in time $O^*(2^n)$, where n is the size of the universe. In the problem that we call here Colored Set Cover with Thresholds, one is given a triple $(\mathcal{U}, \mathcal{S} = \mathcal{C}_1 \uplus \ldots \uplus \mathcal{C}_l, (a_1, \ldots, a_l))$ where $\mathcal{U}$ is a ground set of n elements, $\mathcal{S}$ is a set of subsets of $\mathcal{U}$ partitioned into l classes called *colors* and $(a_1, \ldots, a_l)$ is a tuple of l positive integers called *threshold vector*. The goal is to find a set cover $\mathcal{T} \subseteq \mathcal{S}$ (not necessarily minimum) such that for each $i \in [l]$, the number of sets with color i (that is, in $\mathcal{C}_i$) in $\mathcal{T}$ is at most a_i.

▶ **Lemma 3.** Colored Set Cover with Thresholds *with* n *elements and* m *sets can be solved in time* $O(nm2^n + nm)$.

Proof. We order the sets of $\mathcal{S}$ such that sets of the same color appear consecutively, say, first the sets of $\mathcal{C}_1$, then the sets of $\mathcal{C}_2$, and so on. The order within the sets of a same color is not important and is chosen arbitrarily. We denote the sets resultantly ordered by $S_1, \ldots, S_m$ and function c maps the index of a set to its color. Therefore, $c(j) = i$ means that set S_j has

color i ($S_j \in C_i$). We fill by dynamic programming the table T, where $T[U, j]$ is meant to contain the minimum number of sets in $\mathcal{C}_{c(j)}$ among any subset of $\{S_1, \ldots, S_j\}$ that covers $U \subseteq \mathcal{U}$ and respects the threshold vector.

As an initialization step, for each $U \subseteq \mathcal{U}$, we set $T[U, 1] = 1$ if $U \subseteq S_1$, and $T[U, 1] = \infty$ otherwise. For each $j \in [2, m]$, assuming that $T[U', j-1]$ was already filled for every $U' \subseteq \mathcal{U}$, we distinguish two cases to fill $T[U, j]$. If S_j is the first set of the color class $\mathcal{C}_{c(j)}$ then:

$$T[U,j] = \begin{cases} 0 \text{ if } T[U,j-1] < \infty & \text{(* discard } S_j \text{ *)} \\ 1 \text{ if } T[U,j-1] = \infty \text{ and } T[U \setminus S_j, j-1] < \infty & \text{(* add } S_j \text{ *)} \\ \infty \text{ otherwise} \end{cases}$$

Otherwise S_j is not the first set in $\mathcal{C}_{c(j)}$ and:

$$T[U,j] = \min \begin{cases} T[U,j-1] & \text{(* discard } S_j \text{ *)} \\ v+1 \text{ if } v < a_{c(j)} \text{ and } \infty \text{ otherwise} & \text{(* add } S_j \text{ *)} \end{cases}$$

with $v = T[U \setminus S_j, j-1]$.

A standard induction shows that the instance is positive iff $T[\mathcal{U}, m] \neq \infty$. The only costly operation in filling one entry of table T is the set difference which can be done in $O(n)$. If we want to produce an actual solution (and not solely decide the problem), we can add one bit in each entry $T[U, j]$ signaling whether or not S_j should be taken. Should the instance be positive, it then takes time $O(nm)$ to reconstruct a solution from a filled table T. Therefore, the running time is $O(n|T| + nm) = O(nm2^n + nm)$. ◀

▶ **Theorem 4.** Graph Motif *can be solved in* $O^*(8^k)$*, where* k *is the distance to clique.*

Proof. Let $(G = (V, E), c : V \to \mathcal{C}, M)$ be any instance of Graph Motif and assume R is a solution, that is $G[R]$ is connected and $c(R) = M$. If there is no solution, our algorithm will detect it eventually. We first compute a set $S \subseteq V$ of size k such that $C := V \setminus S$ is a clique. This can be done in time $O^*(2^k)$ by branching over the two endpoints of a *non-edge*, or even in $O^*(1.2738^k)$ by applying the state-of-the-art algorithm for Vertex Cover on the complementary graph [11]. Running through all the 2^k subsets of S, one can guess the subset $S' = R \cap S$ of S which is in the solution R, and $S_1, S_2, \ldots, S_{k'}$ be the $k' \leqslant k$ connected components of $G[S']$. It must hold that $c(S') \subseteq M$, otherwise R would not be a solution. Now, the problem boils down to finding a non-empty (an empty subset would mean that $S' = R$ which can be easily checked) subset $C' \subseteq C$ such that $G[S' \cup C']$ is connected and $c(C') \subseteq M \setminus c(S')$. Then, the set $S' \cup C'$ can be extended into a solution by adding vertices of $C \setminus C'$ with the right colors. The graph $G[S' \cup C']$ is connected iff each connected component S_j of $G[S']$ has at least one neighbor in $N(C')$. We build an equivalent instance of Colored Set Cover with Thresholds in the following way. The ground set $\mathcal{U}$ is of size k' with one element x_j per connected component S_j of $G[S']$. For each vertex v in C colored by i, there is a set S_v colored by i such that $x_j \in S_v$ iff $N(v) \cap S_j \neq \emptyset$. For each color i, the threshold a_i is set to the multiplicity of i in $M \setminus c(S')$. If there are more than one set with the same color and the same elements, we keep only one copy of this colored set. The number of sets is therefore at most $2^{k'}|\mathcal{C}|$. So, it takes time $O(k'2^{k'}|\mathcal{C}|(2^{k'}+1)) = O^*(4^{k'})$ to solve this instance, hence an overall worst case running time of $O^*(2^k + 2^k 4^k) = O^*(8^k)$. ◀

▶ **Theorem 5.** Graph Motif *can be solved in* $O^*(2^{2k \log k})$ *on graphs with a vertex cover of size* k*.*

Proof. We start similarly to the previous algorithm. We compute a minimum vertex cover S of G in time $O^*(2^k)$ (or $O^*(1.2738^k)$ [11]), and then guess in time $O^*(2^k)$ the subset $S' = S \cap R$, where R is a fixed solution. Again, we denote by $S_1, S_2, \ldots, S_{k'}$ the connected components of $G[S']$. We remove $c(S')$ from the motif and we remove from V the set I'

of the vertices of the independent set $I := V \setminus C$ which have no neighbor in S'. Now, by the transformation presented in the algorithm parameterized by distance to clique, the problem could be made equivalent to a constrained version of Colored Set Cover with Thresholds where the intersection graph (with an edge between two sets if they have a non-empty intersection) of the solution has to be connected. Unfortunately, it is not clear whether or not this variant can be solved in time $2^{O(n)}$. Thus, at this point, we have to do something different.

Let $R_d = \{r_1, r_2, \ldots, r_l\} \subseteq R \setminus S'$ be a minimal (inclusion-wise) set of vertices such that $G[S' \cup R_d]$ is connected. We can observe that $l \leqslant k' \leqslant k$. We guess in time $O^*(l!B_l)$ (where B_l is the l-th Bell number, i.e., the number of partitions of a set of size l) an ordered partition $P := \langle A_1, A_2, \ldots, A_l \rangle$ of the connected components $\{S_1, \ldots, S_{k'}\}$ such that, for each $i \in [l]$, (1) r_i has at least one neighbor in each connected component of A_i and (2) if $i \geqslant 2$, r_i has at least one neighbor in a connected component of $\bigcup_{1 \leqslant j < i} A_j$. Note that such an ordered partition always exists since $G[S' \cup R_d]$ is connected. Now, we build the bipartite graph $B = (P \cup M', F)$, where $M' = M \setminus c(S')$ and there is an edge between $A_i \in P$ and each copy of color $c \in M'$ iff there is a vertex $v \in I$ colored by c in the original graph G and such that (1) v has at least one neighbor in each connected component of A_i and (2) if $i \geqslant 2$, v has at least one neighbor in a connected component of $\bigcup_{1 \leqslant j < i} A_j$. By construction, $\{\{A_i, c(r_i)\} \mid i \in [l]\}$ is a maximum matching of size $|P| = l$ in graph B. Thus, we compute in polynomial time a maximum matching $\{\{A_i, c_i\} \mid i \in [l]\}$ in B. Then, we obtain a solution to the Graph Motif instance by taking, for each $i \in |l]$ any vertex v_i colored by c_i and having (1) at least one neighbor in each connected component of A_i and (2) if $i \geqslant 2$, at least one neighbor in a connected component of $\bigcup_{1 \leqslant j < i} A_j$. This can also be done in polynomial time and the existence of such a v_i is guaranteed by the construction of graph B. Then, we complete set $S' \cup \bigcup_{i \in [l]} \{v_i\}$ into a solution by taking any vertices in $I \setminus I'$ with the right colors. As $l! \leqslant l^l$, $B_l \leqslant (\frac{l}{2})^l$ (even $B_l < (\frac{0.792l}{\ln(l+1)})^l$ [3]), and $l \leqslant k$ the overall running time is $O^*(2^k + 2^k k! B_k) = O^*(k^k k^k) = O^*(2^{2k \log k})$. ◀

In the Edge Clique Cover problem, one asks, given a graph $G = (V, E)$ and an integer k, for k subsets $C_1, \ldots, C_k \subseteq V$, such that $\forall i \in [k]$, $G[C_i]$ is a clique, and $\forall e \in E$, e lies in a clique C_i for some $i \in [k]$. The set $\{C_1, \ldots, C_k\}$ is called an *edge clique cover* of G. The *edge clique cover number* of a graph G is the smallest k such that G has an edge clique cover of size k. Edge Clique Cover admits a kernel of size 2^k [19] and, as observed in [12], it can be solved by dynamic programming in time $2^{O(n+m)}$. Therefore, it can be solved in time $2^{O(2^k + 2^{2k})}$, that is $2^{2^{O(k)}}$. On the negative side, Edge Clique Cover cannot be solved in time $2^{2^{o(k)}}$ under ETH [12]. Thus, the algorithm of Ganian [18] is essentially optimal if the edge clique cover is not given. But, we may imagine that the instance comes with an optimal or close to optimal edge clique cover, or that we have a good heuristic to compute it (a polynomial time approximation with sufficiently good ratio is unlikely [27]).

▶ **Theorem 6.** Graph Motif *can be solved in time* $2^{2^{O(k)}}$*, where* k *is the edge clique cover number, and in time* $O^*(2^{2k \log k + k})$ *if an edge clique cover of size* k *is given as part of the input.*

Proof. Let $(G = (V, E), c, M)$ be any instance of Graph Motif. If not given, we first compute an edge clique cover $\{C_1, \ldots, C_k\}$ of size k in G, in time $2^{2^{O(k)}}$ [19]. We guess in time $O^*(2^k)$ the exact subset $\{C'_1, \ldots, C'_{k'}\} \subseteq \{C_1, \ldots, C_k\}$ of cliques C_i such that $C_i \cap R$ is non-empty, for a fixed solution R. Now, we turn the instance into an equivalent instance where the motif has size $|M| + k'$ and the graph has at most $|V| + k'$ vertices and a vertex cover of size k'. The new graph is a bipartite graph $B = (A \cup W, F)$ such that A contains one vertex

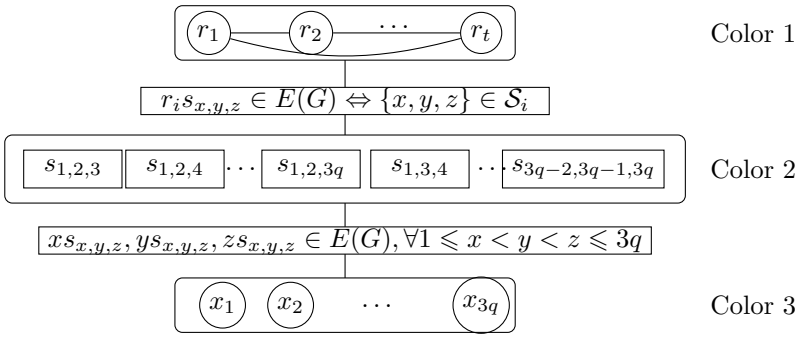

Figure 2 Illustration of the construction of G. The motif consists of 1 occurrence of color 1, q of color 2 and $3q$ of color 3.

$v(C'_i)$ per clique C'_i (so, A is a vertex cover of graph B of size $k' \leqslant k$), $W = C'_1 \cup \ldots \cup C'_{k'} \subseteq V$, and there is an edge in F between $v(C'_i) \in A$ and $w \in W$ iff $w \in C'_i$. Each vertex in W keeps the color it had in G. A fresh color c is given to the k' vertices of A, and color c is added to the motif M with multiplicity k'. Then, we run the algorithm parameterized by the vertex cover number of Theorem 5. This algorithm has an overall running time of $O^*(2^k 2^{2k \log k})$, if the edge clique cover is given, and $2^{2^{O(k)}}$ otherwise. ◀

Ganian [17], Theorem 5 and Theorem 4 prove that Graph Motif is in FPT if the parameter is the vertex cover number or the distance to clique. Therefore, the problem has a kernel [28]. Though, the size of this kernel is *a priori* not known. We show that the corresponding kernels cannot be polynomial unless NP $\subseteq$ coNP/*poly*.

▶ **Theorem 7.** *Unless* NP $\subseteq$ coNP/*poly*, Graph Motif *has no polynomial kernel when parameterized by the vertex cover number or the distance to clique, even for* (i) *motifs with only 3 colors and* (ii) *when the motif is colorful.*

Proof. We only give the proof for (i). The second item (ii) can be proven similarly following the ideas of [5].

We will define an OR-cross-composition [9] from the NP-complete X3C problem, stated as follows: given and integer q, a set $X = \{x_1, x_2, \ldots, x_{3q}\}$ and a collection $\mathcal{S} = \{S_1, \ldots, S_{|\mathcal{S}|}\}$ of 3-elements subsets of X, the goal is to decide if $\mathcal{S}$ contains a subcollection $\mathcal{T} \subseteq \mathcal{S}$ such that $|\mathcal{T}| = q$ and each element of X occurs in exactly one element of $\mathcal{T}$. Given t instances, $(X_1, \mathcal{S}_1), (X_2, \mathcal{S}_2), \ldots, (X_t, \mathcal{S}_t)$, of X3C, we define our equivalence relation $\mathcal{R}$ such that any strings that are not encoding valid instances are equivalent, and $(X_i, \mathcal{S}_i), (X_j, \mathcal{S}_j)$ are equivalent iff $|X_i| = |X_j|$ and $|\mathcal{S}_i| = |\mathcal{S}_j|$. Hereafter, we assume that $X_i = [3q]$ and $\mathcal{S}_i = \{S_1, \ldots, S_{|\mathcal{S}_i|}\}$, for any $i \in [t]$. We will build an instance (G, c, M) of Graph Motif parameterized by the vertex cover or the distance to clique, where G is the input graph, c the coloring function and M the motif, such that there is a solution for Graph Motif iff there is an $i \in [t]$ such that there is a solution for $(X_i, \mathcal{S}_i)$. We will now describe how to build such instance of Graph Motif. The graph G consists of t nodes $r_1, r_2, \cdots, r_t$ forming a clique. There are also $O((3q)^3)$ nodes $s_{x,y,z}, 1 \leqslant x < y < z \leqslant 3q$, with an edge between r_i and $s_{x,y,z}$ iff the 3-element subset $\{x, y, z\}$ exists in $\mathcal{S}_i$. Finally, there are $3q$ nodes $x_i, 1 \leqslant i \leqslant 3q$, and there is an edge between x_i and every subset $s_{x,y,z}$ where x_i occurs (see also Figure 2). The coloration is $c(r_i) = 1$, for all $1 \leqslant i \leqslant t$, $c(s_{x,y,z}) = 2$ for all $1 \leqslant x < y < z \leqslant 3q$, and $c(x_i) = 3, 1 \leqslant i \leqslant 3q$. The multiset M consists of 1 occurrence of the color 1, q occurrences of color 2 and $3q$ occurrences of color 3.

It is easy to see that $\{s_{x,y,z} | 1 \leqslant x < y < z \leqslant 3q\} \cup \{x_i | 1 \leqslant i \leqslant 3q\}$ is a vertex cover for G and that its removal leaves only a clique, and that its size is polynomial in $3q$ and hence in the size of the largest instance.

Let us show that there is a solution for our instance of Graph Motif iff at least one of the $(X_i, \mathcal{S}_i)$'s has a solution of size q.

($\Leftarrow$) Suppose that $(X_i, \mathcal{S}_i)$ has a solution $\mathcal{T}_i$ of size q. We set $P = \{r_i\} \cup \{s_{x,y,z} \mid \{x, y, z\} \in \mathcal{T}_i\} \cup \{x_i | 1 \leqslant i \leqslant 3q\}$. One can easily check that $G[P]$ is connected and that $c(P) = M$.

($\Rightarrow$) Suppose that there is a solution $P \subseteq V$ such that $G[P]$ is connected and $c(P) = M$. Due to the motif, only one of the nodes r_i is in P and all nodes x_i are in P. We claim that there is then a solution $\mathcal{T}_i$ in $(X_i, \mathcal{S}_i)$, where i is the index of the only node r_i in P. We add in $\mathcal{T}_i$ the q sets $\{x, y, z\}$ such that $s_{x,y,z} \in P$. By the connectivity constraint, these sets all occurs in the instance i s.t. $r_i \in P$. Let us now prove that $\mathcal{T}_i$ covers exactly all the elements of X_i. Since P is a solution, the nodes $s_{x,y,z}$ in P correspond to a partition of X. Otherwise, one of the node x_i will not be connected. ◀

3.3 Parameters with unbounded neighborhood diversity

This section disproves the idea that Graph Motif is only tractable for classes with bounded neighborhood diversity. Indeed, we show that Graph Motif is in FPT parameterized by the size of a *vertex clique cover* or by the distance to co-cluster. The former algorithm creates a win/win based on König's theorem applied to a bounded number of auxiliary bipartite graphs. The latter is simpler and use as subroutines the algorithms parameterized by vertex cover number and distance to clique.

In the Vertex Clique Cover problem (also known as Clique Partition), one asks, given a graph $G = (V, E)$ and an integer k, for a *partition* of the vertices into k subsets $C_1, \ldots, C_k \subseteq V$, such that $\forall i \in [k]$, $G[C_i]$ is a clique. The set $\{C_1, \ldots, C_k\}$ is called an *vertex clique cover* of G. The *vertex clique cover number* of a graph G is the smallest k such that G has an vertex clique cover of size k. This problem is equivalent to the Graph Coloring problem since a graph as a vertex clique cover of size k iff its complement is k-colorable. Therefore, Vertex Clique Cover is unlikely to be in XP. However, if a vertex clique cover comes with the input, we show that Graph Motif is in FPT for parameter vertex clique cover number. One can notice that Graph Motif is NP-hard in 2-colorable graphs. This is a striking example of how easier can Graph Motif be on the denser counterpart of two complementary classes.

To realize that vertex clique cover number has unbounded neighborhood diversity, think of the complement of a bipartite graph. The vertex clique cover is of size 2 but the neighborhood diversity could be arbitrary; for parameter distance to co-cluster, think of the complementary of a cluster graph with an unbounded number of cliques.

▶ **Theorem 8 (★).** Graph Motif *can be solved in time* $O^*(2^{4k \log(2k)})$ *where* k *is the vertex clique cover number, provided that the vertex clique cover is given as part of the input.*

▶ **Theorem 9 (★).** Graph Motif *can be solved in* $O^*(2^{2k \log k})$*, where* k *is the distance to co-cluster.*

4 Parameters for which Graph Motif is hard

In this section, we provide several parameters for which Graph Motif is not in XP, unless P = NP. In other words, the problem is NP-hard even for fixed values of the parameter. We

also prove that the problem remains W[1]-hard for parameter max leaf number. Figure 1 summarizes these results.

4.1 Deletion set numbers

We study parameters which correspond to the minimum number of vertices to remove to make the graph belong to a restricted class. We will show that Graph Motif remains NP-hard for constant values of those parameters. More precisely, the colorful restriction of Graph Motif is hard even if we can obtain a set of disjoint paths by removing 1 vertex, a cluster graph by removing 1 vertex, and an acyclic graph by removing 0 edge.

▶ **Theorem 10** ([15]). Graph Motif *is* NP*-hard even when G is a tree of maximum degree 3 and the motif is colorful.*

▶ **Corollary 11.** Graph Motif *is* NP*-hard even for graphs with feedback edge set 0 and when the motif is colorful.*

▶ **Theorem 12 (★).** Graph Motif *is* NP*-hard even (i) for graphs with distance* 1 *to disjoint paths and when the motif is colorful and (ii) for graphs with bandwidth* 4 *and when the motif is colorful.*

▶ **Theorem 13 (★).** Graph Motif *is* NP*-hard even for graphs with distance* 1 *to cluster and when the motif is colorful.*

4.2 Dominating set number

Being given a small dominating set of the graph cannot help in solving Graph Motif. For any instance $(G = (V, E), c, M)$, one may add a universal new vertex v to G, and color it with a color which does not appear in motif M. The minimum dominating set $\{v\}$ is of size 1. Vertex v cannot be part of the solution due to its color, so answering the new problem is as hard as solving the original instance. Though, this could be considered as cheating since a vertex whose color is not in M can immediately be discarded from the graph. We show that even when $\forall v \in V$, $c(v) \in M$, graphs with dominating set of size 2 can be hard to solve.

▶ **Theorem 14 (★).** Graph Motif *is* NP*-hard even for graphs with a minimum dominating set of size* 2 *and when the motif is colorful.*

4.3 Max leaf number

The *max leaf number* of a graph G, denoted $\mathrm{ml}(G)$ is the maximum number of leaves (i.e., vertices of degree 1) in a spanning tree of G. Therefore, if G is itself a tree, then $\mathrm{ml}(G)$ is simply the number of leaves of G. We will show that Graph Motif is in XP (even in W[P]) and is W[1]-hard with parameter max leaf number. In fact, we will even prove that it is W[1]-hard on trees with parameter *number of leaves in the tree plus number of distinct colors in the motif*. This strenghtens the previously known result that the problem is W[1]-hard on trees with parameter number of distinct colors in the motif [15].

▶ **Theorem 15 (★).** Graph Motif *can be solved in time* $O^*(16^k n^{10k}) = n^{O(k)}$*, where* $k = ml(G)$ *and is even in* W[P] *with respect to that parameter.*

▶ **Theorem 16 (★).** Graph Motif *is* W[1]*-hard with respect to the max leaf number plus the number of colors, even on trees.*

5 Conclusion and open problems

Figure 1 sums up the parameterized complexity landscape of Graph Motif with respect to structural parameters. For parameter maximum independent set the complexity status of Graph Motif remains unknown. Even when the problem is in FPT, polynomial kernels tend to be unlikely; be it for the natural parameter even on comb graphs or for the vertex cover number or the distance to clique. Is it the case for parameter cluster editing number?

The sparsification lemma [21] together with a straightforward reduction from 3-SAT shows that, under ETH, Graph Motif cannot be solved in time $2^{o(n)}$ on graphs with n vertices. Thus, for every parameter k bounded by n, an algorithm solving Graph Motif in $2^{o(k)}$ would disprove ETH. This is the case of four out of six parameters for which we have given an FPT algorithm; cluster editing and edge clique cover numbers are only bounded by n^2. On the one hand, it says that our algorithm running in $2^{O(k)}$ for parameter distance to clique is probably close to optimal. On the other hand, for parameter vertex cover number, for instance, we have still some room for improvement between the $2^{O(k \log k)}$-upper bound and the $2^{o(k)}$-lower bound under ETH. Can we improve the algorithm to time $2^{O(k)}$, or, on the contrary, show a stronger lower bound of $2^{o(k \log k)}$ (potentially using [26])?

References

1 Eric Alm and Adam P. Arkin. Biological Networks. *Current Opinion in Structural Biology*, 13(2):193–202, 2003.

2 Abhimanyu M. Ambalath, Radheshyam Balasundaram, Chintan Rao H., Venkata Koppula, Neeldhara Misra, Geevarghese Philip, and M. S. Ramanujan. On the Kernelization Complexity of Colorful Motifs. In *Proc. of the 5th IPEC*, volume 6478 of *LNCS*, pages 14–25. Springer, 2010.

3 Daniel Berend and Tamir Tassa. Improved bounds on bell numbers and on moments of sums of random variables. *Probab. and Math. Statist.*, 30(2):185–205, 2010.

4 Nadja Betzler, René van Bevern, Michael R. Fellows, Christian Komusiewicz, and Rolf Niedermeier. Parameterized algorithmics for finding connected motifs in biological networks. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 8(5):1296–1308, 2011.

5 Andreas Björklund, Petteri Kaski, and Lukasz Kowalik. Probably optimal graph motifs. In *Proc. of the 30th International Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 20 of *LIPIcs*, 2012.

6 Andreas Björklund, Petteri Kaski, Lukasz Kowalik, and Juho Lauri. Engineering motif search for large graphs. In Ulrik Brandes and David Eppstein, editors, *Proc. of the 17th ALENEX*, pages 104–118. SIAM, 2015.

7 Sebastian Böcker. A golden ratio parameterized algorithm for cluster editing. *J. Discrete Algorithms*, 16:79–89, 2012.

8 Sebastian Böcker, Florian Rasche, and Tamara Steijger. Annotating Fragmentation Patterns. In *Proc. of the 9th International Workshop Algorithms in Bioinformatics (WABI)*, volume 5724 of *LNCS*, pages 13–24. Springer, 2009.

9 Hans L. Bodlaender, Bart M. P. Jansen, and Stefan Kratsch. Kernelization lower bounds by cross-composition. *SIAM J. Discrete Math.*, 28(1):277–305, 2014.

10 Sharon Bruckner, Falk Hüffner, Richard M. Karp, Ron Shamir, and Roded Sharan. Topology-Free Querying of Protein Interaction Networks. *Journal of Computational Biology*, 17(3):237–252, 2010.

11 Jianer Chen, Iyad A. Kanj, and Ge Xia. Improved upper bounds for vertex cover. *Theoretical Computer Science*, 411(40–42):3736 – 3756, 2010.

12 Marek Cygan, Marcin Pilipczuk, and Michal Pilipczuk. Known algorithms for EDGE CLIQUE COVER are probably optimal. In *Proc. of Symposium on Discrete Algorithms, SODA 2013*, pages 1044–1053. SIAM, 2013.

13 Riccardo Dondi, Guillaume Fertin, and Stéphane Vialette. Complexity issues in vertex-colored graph pattern matching. *J Discr Algo*, 9(1):82–99, 2011.

14 Rodney G. Downey and Michael R. Fellows. *Fundamentals of Parameterized Complexity*. Springer, 2013.

15 Michael R. Fellows, Guillaume Fertin, Danny Hermelin, and Stéphane Vialette. Upper and lower bounds for finding connected motifs in vertex-colored graphs. *J. Comput. Syst. Sci.*, 77(4):799–811, 2011.

16 Michael R. Fellows, Daniel Lokshtanov, Neeldhara Misra, Matthias Mnich, Frances A. Rosamond, and Saket Saurabh. The complexity ecology of parameters: An illustration using bounded max leaf number. *Theory Comput. Syst.*, 45(4):822–848, 2009.

17 Robert Ganian. Twin-cover: Beyond vertex cover in parameterized algorithmics. In *Proc. of the 6th International Symposium Parameterized and Exact Computation IPEC 2011*, volume 7112 of *LNCS*, pages 259–271. Springer, 2011.

18 Robert Ganian. Using neighborhood diversity to solve hard problems. *CoRR*, abs/1201.3091, 2012.

19 Jens Gramm, Jiong Guo, Falk Hüffner, and Rolf Niedermeier. Data reduction and exact algorithms for clique cover. *ACM Journal of Experimental Algorithmics*, 13, 2008.

20 Sylvain Guillemot and Florian Sikora. Finding and counting vertex-colored subtrees. *Algorithmica*, 65(4):828–844, 2013.

21 Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? *J. Comput. Syst. Sci.*, 63(4):512–530, 2001.

22 Christian Komusiewicz and Rolf Niedermeier. New races in parameterized algorithmics. In *Proc. of Mathematical Foundations of Computer Science MFCS 2012*, volume 7464 of *LNCS*, pages 19–30. Springer, 2012.

23 Ioannis Koutis. Constrained multilinear detection for faster functional motif discovery. *Inf. Process. Lett.*, 112(22):889–892, 2012.

24 Vincent Lacroix, Cristina G. Fernandes, and Marie-France Sagot. Motif search in graphs: application to metabolic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(4):360–368, 2006.

25 Michael Lampis. Algorithmic meta-theorems for restrictions of treewidth. *Algorithmica*, 64(1):19–37, 2012.

26 Daniel Lokshtanov, Dániel Marx, and Saket Saurabh. Slightly superexponential parameterized problems. In *Proc. of SODA 2011*, pages 760–776, 2011.

27 Carsten Lund and Mihalis Yannakakis. On the hardness of approximating minimization problems. *J. ACM*, 41(5):960–981, 1994.

28 Rolf Niedermeier. *Invitation to Fixed Parameter Algorithms*. Lecture Series in Mathematics and Its Applications. Oxford University Press, 2006.

29 Ron Y. Pinter, Hadas Shachnai, and Meirav Zehavi. Deterministic parameterized algorithms for the graph motif problem. In *Proc. of Mathematical Foundations of Computer Science MFCS 2014*, volume 8635 of *LNCS*, pages 589–600. Springer, 2014.

30 Ron Y. Pinter and Meirav Zehavi. Algorithms for topology-free and alignment network queries. *J. Discrete Algorithms*, 27:29–53, 2014.