ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

The Graph Motif problem parameterized by the structure of the input graph☆

Édouard Bonnet [a], Florian Sikora [b,*]

[a] Institute for Computer Science and Control, Hungarian Academy of Sciences, (MTA SZTAKI), Hungary

[b] Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, Paris, France

ARTICLE INFO

Article history:
Received 12 November 2015
Received in revised form 10 November 2016
Accepted 12 November 2016
Available online xxxx

Keywords:
Parameterized complexity
Graph motif problem
Structural parameterization
Computational biology

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 mostly 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. In the many utilizations of Graph Motif, however, the input graph originates from real-life applications 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. For the FPT cases, we also give some kernelization lower bounds as well as some ETH-based lower bounds on the worst case running time. 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.

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 [33]. However, it also 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 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 [40].

☆ An extended abstract of this work appears in IPEC 2015.

* Corresponding author.
E-mail addresses: bonnet.edouard@sztaki.mta.hu (É. Bonnet), florian.sikora@dauphine.fr (F. Sikora).

http://dx.doi.org/10.1016/j.dam.2016.11.016
0166-218X/

Please cite this article in press as: É. Bonnet, F. Sikora, The Graph Motif problem parameterized by the structure of the input graph, Discrete Applied Mathematics (2016), http://dx.doi.org/10.1016/j.dam.2016.11.016

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 is a graph, we denote by $V(G)$ its set of vertices and by $E(G)$ its set of edges. If $G = (V, E)$ is a graph and $S \subseteq V$, $E_G(S)$ denotes the subset of edges of E having both endpoints in S. If $G = (V, E)$ is a graph and $S \subseteq V$ is a subset of vertices, $G[S]$ denotes the subgraph of G induced by S: $(S, E_G(S))$. For a vertex $v \in V$, the set of neighbors of v in G is denoted by $N_G(v)$, and $N_G(S) := (\bigcup_{v \in S} N_G(v)) \setminus S$. We define $N_G[v] := N_G(v) \cup \{v\}$ and $N_G[S] := N_G(S) \cup S$. In all the previous definitions, we will lose the subscript $_G$ whenever the graph G we are referring to is either implicit or irrelevant. 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 set of mutually disjoint 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'.

Multisets. A *multiset* is a generalization of the notion of set where each element may appear more than once. The *multiplicity* of the element x in the multiset M, denoted by $m_M(x)$, is the number of occurrences of x in M. We adopt the natural convention that $m_M(x) = 0$ if x does not belong to M. The cardinality of a multiset M denoted by $|M|$ is its number of elements *with their multiplicity*: $\Sigma_x m_M(x)$. 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 $\forall x$, $m_D(x) = \max(0, m_M(x) - m_N(x))$. We write $M \subseteq N$ if and only if $M \setminus N = \emptyset$ and $M \subset N$ if and only if $M \subseteq N$ and $M \neq N$.

Example 1. Let $M = \{1, 2, 2, 4, 5, 5, 5\}$ and $N = \{1, 1, 1, 2, 2, 3, 3, 4, 5, 5, 5, 5\}$. Then, $|M| = 7$, $|N| = 12$, $M \setminus N = \emptyset$, $N \setminus M = \{1, 1, 3, 3, 5\}$, and $M \subseteq N$.

Graph Motif. The problem is defined as follows:

> Graph Motif
>
> • **Input**: A triple (G, c, M), where $G = (V, E)$ is a graph, $c : V \to \mathcal{C}$ is a coloring of the vertices, and M is a multiset of colors of $\mathcal{C}$.
> • **Output**: A subset $R \subseteq V$ such that
> (1) $G[R]$ is connected and
> (2) $c(R) = M$.

In the above definition, $c(R)$ denotes the multiset of colors of vertices in R. We use that slight abuse of notation for convenience. We will refer to condition (1) as the *connectivity constraint* and to condition (2) as the *multiset constraint*.

Parameterized complexity. 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 [20,38,16] for more details about fixed-parameter tractability). The parameterized complexity hierarchy is composed of the classes $\mathsf{FPT} \subseteq \mathsf{W[1]} \subseteq \mathsf{W[2]} \subseteq \cdots \subseteq \mathsf{W[P]} \subseteq \mathsf{XP}$. The class XP is the set of problems solvable in time $|I|^{f(k)}$, where f is a computable function.

A W[1]-hard problem is not fixed-parameter tractable (unless FPT = W[1]) and one can prove W[1]-hardness by means of a *parameterized reduction* from a W[1]-hard problem. This is a mapping of an instance (I, k) of a problem A_1 in $g(k) \cdot |I|^{O(1)}$ time (for any computable function g) into an instance (I', k') for A_2 such that $(I, k) \in A_1 \Leftrightarrow (I', k') \in A_2$ and $k' \leqslant h(k)$ for some function h.

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.

It is well known that a decidable problem is in FPT if and only if it has a kernel, but this equivalence yields super-polynomial kernels (in general). To design efficient parameterized algorithms, a kernel of polynomial (or even linear) size in k is important. However, some lower bounds on the size of the kernel can be shown under the assumption that the polynomial hierarchy is a proper hierarchy. To show such results, we will use the cross-composition technique developed by Bodlaender et al. [9].

Definition 2 (*Polynomial Equivalence Relation [9]*)**.** An equivalence relation $\mathcal{R}$ on Σ^* is said to be *polynomial* if the following two conditions hold:

(i) There is an algorithm that given two strings $x, y \in \Sigma^*$ decides whether x and y belong to the same equivalence class in time $(|x| + |y|)^{O(1)}$.

(ii) For any finite set $S \subseteq \Sigma^*$ the equivalence relation $\mathcal{R}$ partitions the elements of S into at most $(\max_{x \in S} |x|)^{O(1)}$ classes.

Please cite this article in press as: É. Bonnet, F. Sikora, The Graph Motif problem parameterized by the structure of the input graph, Discrete Applied Mathematics (2016), http://dx.doi.org/10.1016/j.dam.2016.11.016