



# Parameterized Dynamic Cluster Editing

Junjie Luo<sup>1,2</sup> · Hendrik Molter<sup>3</sup> · André Nichterlein<sup>3</sup> · Rolf Niedermeier<sup>3</sup>

Received: 12 December 2018 / Accepted: 3 July 2020 / Published online: 25 July 2020  
© The Author(s) 2020

## Abstract

We introduce a dynamic version of the NP-hard graph modification problem CLUSTER EDITING. The essential point here is to take into account dynamically evolving input graphs: having a cluster graph (that is, a disjoint union of cliques) constituting a solution for a first input graph, can we cost-efficiently transform it into a “similar” cluster graph that is a solution for a second (“subsequent”) input graph? This model is motivated by several application scenarios, including incremental clustering, the search for compromise clusterings, or also local search in graph-based data clustering. We thoroughly study six problem variants (three modification scenarios edge editing, edge deletion, edge insertion; each combined with two distance measures between cluster graphs). We obtain both fixed-parameter tractability as well as (parameterized) hardness results, thus (except for three open questions) providing a fairly complete picture of the parameterized computational complexity landscape under the two perhaps most natural parameterizations: the distances of the new “similar” cluster graph to (1) the second input graph and to (2) the input cluster graph.

**Keywords** Graph-based data clustering · Incremental clustering · Compromise clustering · Correlation clustering · Local search · Goal-oriented clustering · NP-hard problems · Fixed-parameter tractability · Parameterized complexity · Kernelization · Multi-choice knapsack

---

An extended abstract of this work appears in the proceedings of the 38th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS '18)[35]. Unfortunately, the conference version contains a claim whose unpublished proof contained an error (see Table 1). This full version contains all proof details.

---

✉ Hendrik Molter  
h.molter@tu-berlin.de

Extended author information available on the last page of the article

## 1 Introduction

The NP-hard CLUSTER EDITING problem [6, 41], also known as CORRELATION CLUSTERING [5], is one of the most popular graph-based data clustering problems in algorithmics. Given an undirected graph, the task is to transform it into a disjoint union of cliques (also known as a cluster graph) by performing a minimum number of edge modifications (deletions or insertions). Being NP-hard, CLUSTER EDITING gained high popularity in studies concerning parameterized algorithmics, e.g. [1, 4, 9, 11, 14, 23, 26, 29, 33]. To the best of our knowledge, to date these parameterized studies mostly focus on a “static scenario”. Chen et al. [14] are an exception by also looking at temporal and multilayer graphs. In their work, the input is a set of graphs (multilayer) or an ordered list of graphs (temporal), in both cases defined over the same vertex set. The goal is to transform each input graph into a cluster graph such that, in the multilayer case, the number of vertices in which any two cluster graphs may differ is upper-bounded, and in the temporal case, the number of vertices in which any two consecutive (with respect to their positions in the list) cluster graphs may differ is upper-bounded.

In contrast to the work of [14], we do *not* assume that all future changes are known. We consider the scenario where, given an input graph, we only know changes that lie immediately ahead, that is, we know the “new” graph that the input graph changes to. Thus we seek to efficiently and effectively adapt an existing solution, namely a cluster graph. Motivated by the assumption that the “new” cluster graph should only change moderately but still be a valid representation of the data, we parameterize both on the number of edits necessary to obtain the “new” cluster graph and on the difference between the “old” and the “new” cluster graph. We finally remark that there have been previous parameterized studies of dynamic (or incremental) graph problems, dealing e.g. with coloring [30], domination [3, 19], and vertex deletion [2, 34] problems.

### 1.1 Mathematical Model

In principle, the input for a dynamic version of a static problem  $X$  are two instances  $I$  and  $I'$  of  $X$ , a solution  $S$  for  $I$ , and an integer  $d$ . The task is to find a solution  $S'$  for  $I'$  such that the distance between  $S$  and  $S'$  is upper-bounded by  $d$ . Often, there is an additional constraint on the size of  $S'$ . Moreover, some distance measure between  $I$  and  $I'$  has often been considered as a parameter for the problem [2, 3, 19, 34]. We arrive at our following “original dynamic version” of CLUSTER EDITING (phrased as decision version).

ORIGINAL DYNAMIC CLUSTER EDITING

**Input:** Two undirected graphs  $G_1$  and  $G_2$  and a cluster graph  $G_c$  over the same vertex set, and two nonnegative integers: a budget  $k$  and a distance upper bound  $d$  such that  $|E(G_1) \oplus E(G_c)| \leq k$ .

**Question:** Is there a cluster graph  $G'$  for  $G_2$  such that

1.  $|E(G_2) \oplus E(G')| \leq k$  and
2.  $\text{dist}(G', G_c) \leq d$ ?

Herein,  $\oplus$  denotes the symmetric difference between two sets and  $\text{dist}(\cdot, \cdot)$  is a generic distance function for cluster graphs, which we discuss later. Moreover,  $G_c$  is supposed to be the “solution” given for the input graph  $G_1$ . However, since the question in this problem formulation is independent from  $G_1$ , we can remove  $G_1$  from the input and arrive at the following simplified version of the problem. For the remainder of this paper we focus on this simplified formulation of DYNAMIC CLUSTER EDITING.

#### DYNAMIC CLUSTER EDITING

**Input:** An undirected graph  $G$  and a cluster graph  $G_c$  over the same vertex set, and two nonnegative integers: a budget  $k$  and a distance upper bound  $d$ .

**Question:** Is there a cluster graph  $G'$  for  $G$  such that

1.  $|E(G) \oplus E(G')| \leq k$  and
2.  $\text{dist}(G', G_c) \leq d$ ?

There are many different distance measures for cluster graphs[37, 38]. Indeed, we will study two standard ways of measuring the distance between two cluster graphs. One is called classification error distance, which measures the number of vertices one needs to move between cliques to make two cluster graphs the same—we subsequently refer to it as *matching-based distance*. The other is called disagreement distance, which is the symmetric distance between two edge sets—we subsequently refer to it as *edge-based distance*. Notably, the edge-based distance upper-bounds the matching-based distance. We give formal definitions in Sect. 2.

## 1.2 Motivation and Related Work

Beyond parameterized algorithmics and static CLUSTER EDITING, dynamic clustering in general has been subject to many studies, mostly in applied computer science[12, 16, 17, 42–44]. We mention in passing that there are also close ties to reoptimization (e.g.,[7, 8, 40]) and parameterized local search (e.g.,[21, 25, 28, 30, 36]).

There are several natural application scenarios that motivate the study of DYNAMIC CLUSTER EDITING. Next, we list four of them.

*Dynamically updating an existing cluster graph* DYNAMIC CLUSTER EDITING can be interpreted to model a smooth transition between cluster graphs, reflecting that “customers” working with clustered data in a dynamic setting may only tolerate a moderate change of the clustering from “one day to another” since “revolutionary” transformations would require too dramatic changes in their work. In this spirit, when employing small parameter values, DYNAMIC CLUSTER EDITING has kind of an evolutionary flavor with respect to the history of the various cluster graphs in a dynamic setting.

*Editing a graph into a target cluster graph* For a given graph  $G$ , there may be many cluster graphs which are at most  $k$  edge modifications away. The goal then is to find one of these which is close to the given target cluster graph  $G_c$  since in a

corresponding application one is already “used to” work with  $G_c$ . Adapting a different point of view, the editing into the target cluster graph  $G_c$  might be too expensive (that is,  $|E(G) \oplus E(G_c)|$  is too big), and one has to find a solution cluster graph with small enough modification costs but being still close to the target  $G_c$ .

*Local search for an improved cluster graph* Here the scenario is that one may have found an initial clustering expressed by  $G_c$ , and one searches for another solution  $G'$  for  $G$  within a certain local region around  $G_c$  (captured by our parameter  $d$ ).

*Editing into a compromise clustering* When focusing on the edge-based distance, one may generalize the definition of DYNAMIC CLUSTER EDITING by allowing  $G_c$  to be any graph (not necessarily a cluster graph). This may be used as a model for “compromise cluster editing” in the sense that the goal cluster graph then is a compromise for a cluster graph suitable for both input graphs since it is close to both of them.

### 1.3 Our Results

We investigate the (parameterized) computational complexity of DYNAMIC CLUSTER EDITING. We study DYNAMIC CLUSTER EDITING as well as two restricted versions where only edge deletions (“Deletion”) or edge insertions (“Completion”) are allowed. We show that all problem variants (notably also the completion variants, whose static counterpart is trivially polynomial-time solvable) are NP-complete even if the input graph  $G$  is already a cluster graph. Table 1 surveys our main complexity results.

The general versions of DYNAMIC CLUSTER EDITING all turn out to be parameterized intractable (W[1]-hard) by the single natural parameter “budget  $k$ ” or

**Table 1** Result overview for DYNAMIC CLUSTER EDITING. We primarily categorize the problem variants by the distance measure (Matching, Edge) they use and secondarily by the allowed modification operation

		Parameter					
Problem variant		$k + d$	$k$	$d$			
Match.	Editing	FPT (PK)	Theorem 3	W[1]-h	Theorem 2	W[1]-h	} Theorem 2
	Deletion	FPT (PK)		<i>open</i>		W[1]-h	
	Completion	FPT (PK)		<i>open</i>		FPT	Theorem 4
Edge	Editing	FPT (PK)	Theorem 3	W[1]-h	Theorem 2	W[1]-h	} Theorem 2
	Deletion	FPT (PK)		FPT	Theorem 4	W[1]-h	
	Completion	FPT (PK)		<i>open</i> <sup>a</sup>		FPT	Theorem 4

NP-completeness for all problem variants (last column) even holds if the input graph  $G$  is a cluster graph. PK stands for polynomial-size problem kernel

<sup>a</sup>In the conference version[35] of this paper we claimed that DCCOMPLETION (EDGE DIST) is in FPT when parameterized by  $k$ . Unfortunately, the unpublished “proof” for this claim contained an error that we could not fix

“distance  $d$ ”; however, when both parameters are combined, we achieve a polynomial-size problem kernel, implying fixed-parameter tractability. We also derive a generic approach, based on a reduction to MULTI-CHOICE KNAPSACK, to derive fixed-parameter tractability for several deletion and completion variants with respect to the parameters budget  $k$  as well as the distance  $d$ .

## 1.4 Organization of the Paper

Our work, after introducing basic notations (Sect. 2), consists of two main parts. In Sect. 3, we provide all our (parameterized) hardness results. In Sect. 4, we develop several positive algorithmic results, namely polynomial-size problem kernels through polynomial-time data reduction, and fixed-parameter solving algorithms. We conclude with a summary and directions for future work (Sect. 5).

## 2 Preliminaries and Problems Variants

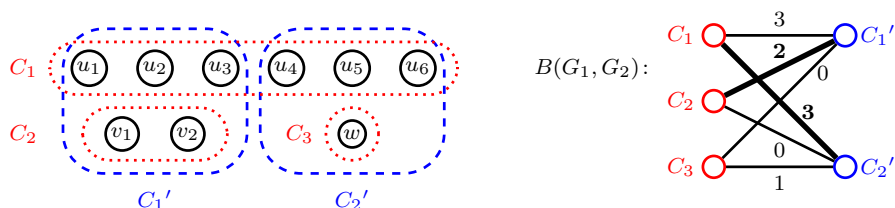
In this section we give a brief overview on concepts and notations of graph theory and parameterized complexity theory that are used in this paper. We also give formal definitions of the distance measures for cluster graphs we use and of our problem variants. We use  $\oplus$  to denote the symmetric difference of two sets, that is, for two sets  $A, B$  we have  $A \oplus B := (A \setminus B) \cup (B \setminus A)$ .

### 2.1 Graph-Theoretic Concepts and Notations

Given an undirected graph  $G = (V, E)$ , we say that a vertex set  $C \subseteq V$  is a *clique* in  $G$  if  $G[C]$  is a complete graph. We say that a vertex set  $C \subseteq V$  is *isolated* in  $G$  if there is no edge  $\{u, v\} \in E$  with  $u \in C$  and  $v \in V \setminus C$ . A  $P_3$  is a path with three vertices. We say that vertices  $u, v, w \in V$  form an induced  $P_3$  in  $G$  if  $G[\{u, v, w\}]$  is a  $P_3$ . We say that an edge  $\{u, v\} \in E$  is part of a  $P_3$  in  $G$  if there is a vertex  $w \in V$  such that  $G[\{u, v, w\}]$  is a  $P_3$ . Analogously, we say that a non-edge  $\{u, v\} \notin E$  is part of a  $P_3$  in  $G$  if there is a vertex  $w \in V$  such that  $G[\{u, v, w\}]$  is a  $P_3$ . A *cluster graph* is simply a disjoint union of cliques. Equivalently, a graph  $G = (V, E)$  is a *cluster graph* if for all  $u, v, w \in V$  we have that  $G[\{u, v, w\}]$  is not a  $P_3$ , or in other words,  $P_3$  is a forbidden induced subgraph for cluster graphs.

### 2.2 Distance Measures for Cluster Graphs

A cluster graph is simply a disjoint union of cliques. We use two basic distance measures for cluster graphs [37, 38]. The first one is called “matching-based distance” and counts how many vertices have to be moved from one clique to another to make two cluster graphs the same (see Fig. 1 for an illustrating example). It is formally defined as follows.



**Fig. 1** An illustration of the matching-based distance measure. On the left side, red dotted boundaries represent cliques in cluster graph  $G_1$ , and blue dashed boundaries represent cliques in cluster graph  $G_2$ . The bipartite graph on the right side is the edge-weighted bipartite graph  $B(G_1, G_2)$ . The maximum-weight matching  $M$  for  $B(G_1, G_2)$  of weight five is formed by the two edges represented by the two bold lines. The edges not in  $M$  indicate how vertices have to be moved from one cluster to another: Moving  $w$  from  $C_3$  to  $C_1$  and  $u_1, u_2, u_3$  from  $C_1$  to  $C_2$  transforms  $G_1$  into  $G_2$  (Color figure online)

**Definition 1** (*Matching-based distance*) Let  $G_1 = (V, E_1)$  and  $G_2 = (V, E_2)$  be two cluster graphs defined over the same vertex set. Let  $B(G_1, G_2) := (V_1 \uplus V_2, E, w)$  be a weighted complete bipartite graph, where each vertex  $u \in V_1$  corresponds to one cluster of  $G_1$ , denoted by  $C_u \subseteq V$ , and each vertex  $v \in V_2$  corresponds to one cluster of  $G_2$ , denoted by  $C_v \subseteq V$ . The weight of the edge between  $u \in V_1$  and  $v \in V_2$  is  $w(\{u, v\}) := |C_u \cap C_v|$ . Let  $W$  be the weight of a maximum-weight matching in  $B(G_1, G_2)$ . The *matching-based distance*  $d_M$  between  $G_1$  and  $G_2$  is  $d_M(G_1, G_2) := |V| - W$ .

The second distance measure is called “edge-based distance” and simply measures the symmetric distance between the edge sets of two cluster graphs.

**Definition 2** (*Edge-based distance*) Let  $G_1 = (V, E_1)$  and  $G_2 = (V, E_2)$  be two cluster graphs defined over the same vertex set. The *edge-based distance* between  $G_1$  and  $G_2$  is  $d_E(G_1, G_2) := |E_1 \oplus E_2|$ .

See Fig. 1 for an example illustration of two cluster graphs  $G_1$  and  $G_2$  defined over the same vertex set  $V = \{u_1, u_2, u_3, u_4, u_5, u_6, v_1, v_2, w\}$ . In  $G_1$  there are three cliques (clusters)  $C_1 = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ ,  $C_2 = \{v_1, v_2\}$  and  $C_3 = \{w\}$ . In  $G_2$  there are two cliques  $C_1' = \{u_1, u_2, u_3, v_1, v_2\}$  and  $C_2' = \{u_4, u_5, u_6, w\}$ . Then in  $B(G_1, G_2)$  we have three vertices on the left side for the cliques in  $G_1$  and two vertices on the right side for the cliques in  $G_2$ . A maximum-weight matching for  $B(G_1, G_2)$  matches  $C_1$  with  $C_1'$  and  $C_2$  with  $C_2'$ , and has weight  $W = 5$ . Thus we have  $d_M(G_1, G_2) = |V| - W = 9 - 5 = 4$ , while  $d_E(G_1, G_2) = 3^2 + 2 \cdot 3 + 1 \cdot 3 = 18$ .

### 2.3 Problem Names and Definitions

In the following we present the six problem variants we are considering. We use DYNAMIC CLUSTER EDITING as a basis for our problem variants. In DYNAMIC CLUSTER

**DELETION** we add the constraint that  $E(G') \subseteq E(G)$  and in **DYNAMIC CLUSTER COMPLETION** we add the constraint that  $E(G) \subseteq E(G')$ . For each of these three variants we distinguish a matching-based version and an edge-based version, where the generic “dist” in the problem definition of **DYNAMIC CLUSTER EDITING** is replaced by  $d_M$  and  $d_E$ , respectively. This gives us a total of six problem variants. We use the following abbreviations for our problem names. The letters “DC” stand for “Dynamic Cluster”, and “Matching Dist” is short for “Matching-Based Distance”. Analogously, “Edge Dist” is short for “Edge-Based Distance”. This yields the following list of studied problems:

- **DYNAMIC CLUSTER EDITING WITH MATCHING-BASED DISTANCE**, abbreviation: **DCEDITING (MATCHING DIST)**.
- **DYNAMIC CLUSTER DELETION WITH MATCHING-BASED DISTANCE**, abbreviation: **DCDELETION (MATCHING DIST)**.
- **DYNAMIC CLUSTER COMPLETION WITH MATCHING-BASED DISTANCE**, abbreviation: **DCCOMPLETION (MATCHING DIST)**.
- **DYNAMIC CLUSTER EDITING WITH EDGE-BASED DISTANCE**, abbreviation: **DCEDITING (EDGE DIST)**.
- **DYNAMIC CLUSTER DELETION WITH EDGE-BASED DISTANCE**, abbreviation: **DCDELETION (EDGE DIST)**.
- **DYNAMIC CLUSTER COMPLETION WITH EDGE-BASED DISTANCE**, abbreviation: **DCCOMPLETION (EDGE DIST)**.

## 2.4 Parameterized Complexity

We use standard notation and terminology from parameterized complexity [15, 18, 22, 39] and give here a brief overview of the most important concepts. A *parameterized problem* is a language  $L \subseteq \Sigma^* \times \mathbb{N}$ , where  $\Sigma$  is a finite alphabet. We call the second component the *parameter* of the problem. A parameterized problem is *fixed-parameter tractable* (in the complexity class **FPT**) if there is an algorithm that solves each instance  $(I, r)$  in  $f(r) \cdot |I|^{O(1)}$  time, for some computable function  $f$ . A parameterized problem  $L$  admits a *polynomial kernel* if there is a polynomial-time algorithm that transforms each instance  $(I, r)$  into an instance  $(I', r')$  such that  $(I, r) \in L$  if and only if  $(I', r') \in L$  and  $|I', r'| \leq r^c$ , for some constant number  $c$ . If a parameterized problem is hard for the parameterized complexity class **W[1]**, then it is (presumably) not in **FPT**. The complexity class **W[1]** is closed under parameterized reductions, which may run in **FPT**-time and additionally set the new parameter to a value that exclusively depends on the old parameter.

## 3 Intractability Results

In this section, we first establish **NP**-completeness results.

**Theorem 1** *All considered problem variants of DYNAMIC CLUSTER EDITING are NP-complete, even if the input graph  $G$  is a cluster graph.*

Intuitively, Theorem 1 means that on top of the NP-hard task of transforming a graph into a cluster graph, it is computationally hard to improve an already found cluster graph with respect to being closer to the target cluster graph. Notably, while the dynamic versions of CLUSTER COMPLETION turn out to be NP-complete, it is easy to see that classic CLUSTER COMPLETION is solvable in polynomial time.

In a second part of this section we show W[1]-hardness results both for budget parameter  $k$  and for distance parameter  $d$  for several variants of DYNAMIC CLUSTER EDITING. Formally, we show the following.

**Theorem 2** *The following problems are W[1]-hard when parameterized by the budget  $k$ :*

- DCEEDITING (MATCHING DIST),
- DCEEDITING (EDGE DIST).

The following problems are W[1]-hard when parameterized by the distance  $d$ :

- DCEEDITING (MATCHING DIST),
- DCCDELETION (MATCHING DIST),
- DCEEDITING (EDGE DIST), and
- DCCDELETION (EDGE DIST).

The proof of Theorem 2 is based on several parameterized reductions which are presented in Sect. 3.2. The proof of Theorem 1 is based on nonparameterized polynomial-time many-one reductions (see Sect. 3.1) and some parameterized reductions that also imply NP-hardness (see Sect. 3.2). More precisely, Theorem 1 follows from Lemmas 1, 2, Observation 1, and Lemma 3 presented in Sect. 3.1, as well as Lemmas 4 and 5 presented in Sect. 3.2.

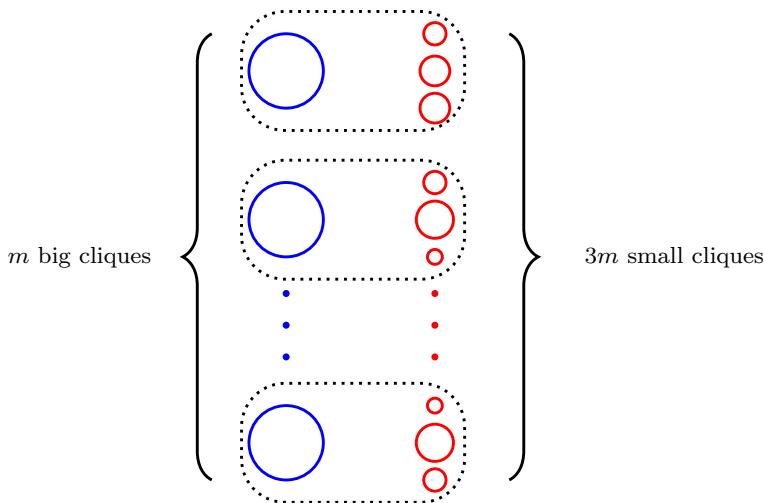
### 3.1 Polynomial-Time Many-One Reductions

We first present two polynomial-time many-one reductions from the strongly NP-hard 3-PARTITION problem[24] for both DCCOMPLETION (MATCHING DIST) and DCCOMPLETION (EDGE DIST) with input graphs that are already cluster graphs. We start with the latter.

**Lemma 1** *DCCOMPLETION (EDGE DIST) is NP-complete, even if the input graph  $G$  is a cluster graph.*

**Proof** We present a polynomial-time reduction from 3-PARTITION, where, given a multi-set of  $3m$  positive integers  $\{a_1, a_2, \dots, a_{3m}\}$  with  $\sum_{1 \leq i \leq 3m} a_i = mB$ ,





**Fig. 2** Illustration of the constructed instance for the proof of Lemma 1. Graph  $G$  has  $m$  big cliques on the left side and  $3m$  small cliques on the right side. Each number  $a_i$  in the instance of 3-PARTITION is represented by a small clique with  $a_i$  vertices on the right side. Every dashed rounded rectangle containing one big clique and three small cliques is a possible group in a solution

for  $1 \leq i \leq 3m$  it holds that  $B/4 < a_i < B/2$ , and the task is to determine whether this multi-set can be partitioned into  $m$  disjoint subsets  $A_1, A_2, \dots, A_m$  such that for each  $1 \leq i \leq m$  it holds that  $\sum_{a_j \in A_i} a_j = B$ . Given an instance  $\{a_1, a_2, \dots, a_{3m}\}$  of 3-PARTITION, we construct an instance  $(G, G_c, k, d)$  of DCCOMPLETION (EDGE DIST) as follows. The construction is illustrated in Fig. 2. For graph  $G$ , we first create  $m$  disjoint *big* cliques, each with  $M = 4(mB)^2$  vertices. Then for every integer  $a_i$ , we create a *small* clique  $C_i$  with  $|C_i| = a_i$  vertices. We set  $G_c$  to be a complete graph. Further, we set  $k := mMB + \frac{m}{2}B^2 - \frac{1}{2} \sum_{1 \leq i \leq 3m} a_i^2$  and  $d := |E(G) \oplus E(G_c)| - k$ .

Next we show that  $\{a_1, a_2, \dots, a_{3m}\}$  is a yes-instance of 3-PARTITION if and only if  $(G, G_c, k, d)$  is a yes-instance of DCCOMPLETION (EDGE DIST).

( $\Rightarrow$ ): Assume that  $\{a_1, a_2, \dots, a_{3m}\}$  is a yes-instance of 3-PARTITION. Then there is a partition  $A_1, A_2, \dots, A_m$  such that for each  $1 \leq i \leq m$  it holds that  $\sum_{a_j \in A_i} a_j = B$ . For each  $A_i$ , we can combine the corresponding three small cliques and one big clique of size  $M$  into one clique. This costs  $MB + \frac{1}{2}(B^2 - \sum_{a_j \in A_i} a_j^2)$  edge insertions. In total, there are

$$mMB + \frac{m}{2}B^2 - \frac{1}{2} \sum_{1 \leq i \leq 3m} a_i^2 = k$$

edge insertions. Hence we get a cluster graph  $G'$  with  $|E(G) \oplus E(G')| = k$  and  $|E(G') \oplus E(G_c)| = |E(G) \oplus E(G_c)| - k = d$ .

( $\Leftarrow$ ): Assume that  $(G, G_c, k, d)$  is a yes-instance of DCCOMPLETION (EDGE DIST) and let  $G'$  be the solution. Since  $k + d = |E(G) \oplus E(G_c)|$ , to get  $G'$  we have to add exactly  $k$  edges to  $G$ . We make the following two observations. First, we can

never combine two big cliques, as otherwise we need at least  $M^2 > k$  edge insertions. Second, every small clique must be combined with a big clique, as otherwise we have at most  $M(mB - 1)$  edge insertions between big cliques and small cliques and at most  $(mB)^2$  edge insertions between small cliques, and in total there are at most  $M(mB - 1) + (mB)^2 = mMB - 3(mB)^2 < k$  edge insertions. Hence, to get solution  $G'$  we must partition all  $3m$  small cliques  $C_1, C_2, \dots, C_{3m}$  in  $G$  into  $m$  groups  $A_1, A_2, \dots, A_m$  and combine all cliques in each group with one big clique.

We can split the edge insertions into two parts  $k = k_1 + k_2$ , where  $k_1 := mMB$  is the number of edge insertions between big cliques and small cliques, and  $k_2 := \sum_{1 \leq i \leq m} \sum_{C_j, C_k \in A_i} |C_j||C_k|$  is the total number of edge insertions between small cliques in each group. We can also write  $k_2$  as

$$k_2 = \frac{1}{2} \sum_{1 \leq i \leq m} \left( \sum_{C_j \in A_i} |C_j| \right)^2 - \frac{1}{2} \sum_{1 \leq i \leq 3m} a_i^2.$$

Recall that  $k = mMB + \frac{m}{2}B^2 - \frac{1}{2} \sum_{1 \leq i \leq 3m} a_i^2$ , so we have that

$$\sum_{1 \leq i \leq m} \left( \sum_{C_j \in A_i} |C_j| \right)^2 = mB^2.$$

Since  $\sum_{1 \leq i \leq 3m} |C_i| = \sum_{1 \leq i \leq 3m} a_i = mB$ , the equality

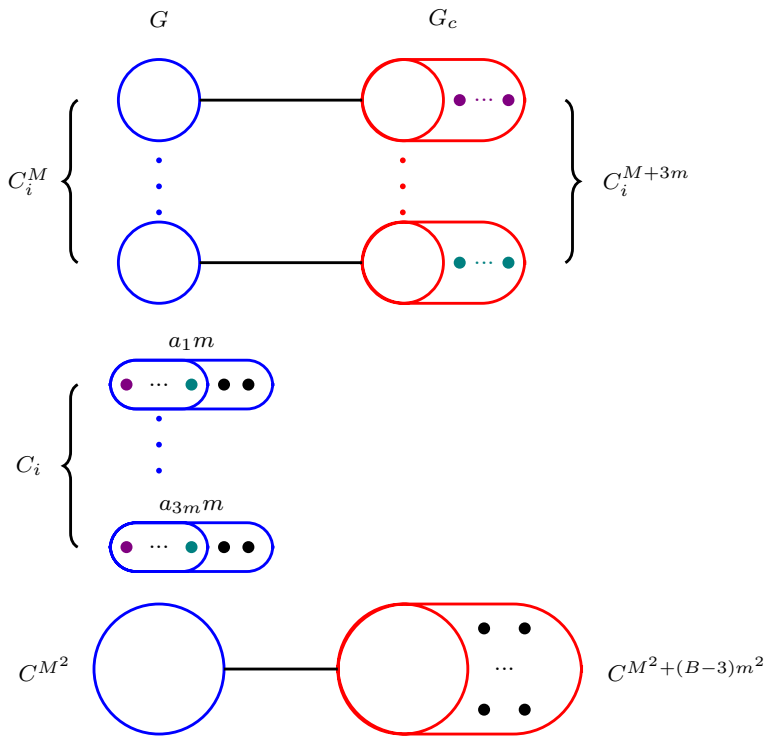
$$\sum_{1 \leq i \leq m} \left( \sum_{C_j \in A_i} |C_j| \right)^2 = mB^2$$

holds only if  $C_1, C_2, \dots, C_{3m}$  can be partitioned into  $m$  disjoint subsets  $A_1, A_2, \dots, A_m$  such that for  $1 \leq i \leq m$  it holds that  $\sum_{C_j \in A_i} |C_j| = B$ . Thus,  $\{a_1, a_2, \dots, a_{3m}\}$  can be partitioned into  $m$  disjoint subsets  $A_1', A_2', \dots, A_m'$  such that for  $1 \leq i \leq m$  it holds that  $\sum_{a_j \in A_i'} a_j = B$ .  $\square$

We continue with DCCOMPLETION (MATCHING DIST). The corresponding NP-hardness reduction uses the same basic ideas as in Lemma 1. The main difference is that in the proof of Lemma 1 we make use of the property that we need to add exactly  $k$  edges. This enforces that every small clique should be combined with a big clique, while in the following proof we need to make use of the matching-based distance to enforce this.

**Lemma 2** DCCOMPLETION (MATCHING DIST) is NP-complete, even if the input graph  $G$  is a cluster graph.

**Proof** We present a polynomial-time reduction from 3-PARTITION, where, given a multi-set of  $3m$  positive integers  $\{a_1, a_2, \dots, a_{3m}\}$  with  $\sum_{1 \leq i \leq 3m} a_i = mB$ , for  $1 \leq i \leq 3m$  it holds that  $B/4 < a_i < B/2$ , and the task is to determine whether



**Fig. 3** Illustration of the constructed instance for the proof of Lemma 2. Blue borders on the left side represent cliques in graph  $G$  and red borders on the right side represent cliques in cluster graph  $G_c$ . Each number  $a_i$  in the instance of 3-PARTITION is represented by a clique  $C_i$  with  $a_i m$  vertices in graph  $G$ , which contains one vertex from every  $C_i^{M+3m}$  and  $a_i(m-1)$  vertices from  $C^{M^2+(B-3)m^2}$ . The maximum-weight matching for  $B(G, G_c)$  is formed by the edges between  $C_i^M$  and  $C_i^{M+3m}$  and the edge between  $C^{M^2+(B-3)m^2}$  and  $C_i^{M+3m}$  (Color figure online)

this multi-set can be partitioned into  $m$  disjoint subsets  $A_1, A_2, \dots, A_m$  such that for each  $1 \leq i \leq m$  it holds that  $\sum_{a_j \in A_i} a_j = B$ . Given an instance  $\{a_1, a_2, \dots, a_{3m}\}$  of 3-PARTITION, we construct an instance  $(G, G_c, k, d)$  of DCCOMPLETION (MATCHING DIST) as follows.

The construction is illustrated in Fig. 3. For graph  $G$ , we first create  $m$  big cliques  $C_1^M, C_2^M, \dots, C_m^M$  each with  $M := 4(mB)^2$  vertices. Then for every integer  $a_i$  in  $\{a_1, a_2, \dots, a_{3m}\}$ , we create a small clique  $C_i$  with  $|C_i| = a_i m$ . Lastly, we create a clique  $C^{M^2}$  with  $M^2$  vertices. For graph  $G_c$ , we create  $m+1$  cliques as follows. For every  $C_i^M$  in  $G$ , we create a clique  $C_i^{M+3m}$  with  $M+3m$  vertices which contains all  $M$  vertices from  $C_i^M$  and one vertex from each  $C_i$  for  $1 \leq i \leq 3m$ . In other words, each  $C_i$  in  $G$  contains exactly one vertex from each  $C_i^{M+3m}$  in  $G_c$  for  $1 \leq i \leq m$ . Lastly, we create a clique  $C^{M^2+(B-3)m^2}$  which contains all remaining vertices, that is,  $M^2$  vertices from  $C^{M^2}$  and vertices from every  $C_i$  for  $1 \leq i \leq 3m$  which are not contained in any  $C_i^{M+3m}$ . Thus  $C^{M^2+(B-3)m^2}$  contains  $M^2 + \sum_{1 \leq i \leq 3m} (a_i - 1)m = M^2 + (B-3)m^2$  vertices. Set  $k := m^2 MB + \frac{m^3}{2} B^2 - \frac{m^2}{2} \sum_{1 \leq i \leq 3m} a_i^2$  and  $d := m^2 B - 3m$ .

It is easy to see that the maximum-weight matching  $M^*$  for  $B(G, G_c)$  is to match  $C_i^M$  with  $C_i^{M+3m}$  for every  $1 \leq i \leq m$  and is to match  $C^{Bm^2}$  with  $C^{(2B-3)m^2}$ . Thus the matching-based distance between  $G$  and  $G_c$  is

$$d_0 = d_M(G, G_c) = \sum_{1 \leq i \leq 3m} a_i m = m^2 B.$$

Now, we show that  $\{a_1, a_2, \dots, a_{3m}\}$  is a yes-instance of 3-PARTITION if and only if  $(G, G_c, k, d)$  is a yes-instance of DCCOMPLETION (MATCHING DIST).

( $\Rightarrow$ ): Assume that  $\{a_1, a_2, \dots, a_{3m}\}$  is a yes-instance of 3-PARTITION. Then there is a partition  $A_1, A_2, \dots, A_m$  such that for  $1 \leq i \leq m$  it holds that  $\sum_{a_j \in A_i} a_j = B$ . We add edges into  $G$  to get a cluster graph  $G'$  as follows. For each  $A_i$ , we combine the corresponding three small cliques for the three integers in  $A_i$  and the big clique  $C_i^M$  into one clique. This costs

$$MmB + m^2 \sum_{a_k, a_j \in A_i} a_j a_k = MmB + \frac{m^2}{2} \left( B^2 - \sum_{a_j \in A_i} a_j^2 \right)$$

edge insertions. In total, there are  $m^2 MB + \frac{m^3}{2} B^2 - \frac{m^2}{2} \sum_{1 \leq i \leq 3m} a_i^2 = k$  edge insertions. Since every small clique  $C_i$ , combined with some big clique  $C_j^M$ , contains one vertex from  $C_j^{M+3m}$ , we obtain

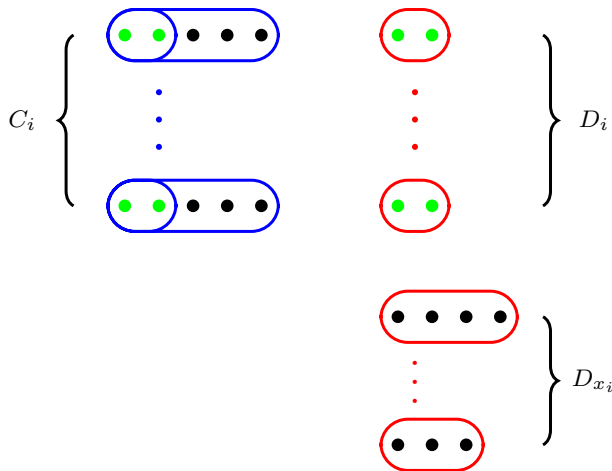
$$d_M(G', G_c) = d_0 - 3m = m^2 B - 3m = d.$$

( $\Leftarrow$ ): Assume that  $(G, G_c, k, d)$  is a yes-instance of DCCOMPLETION (MATCHING DIST) and let  $G'$  be the solution and let  $M'$  be the maximum-weight matching between  $G'$  and  $G_c$ . First note that clique  $C^{M^2}$  has  $M^2$  vertices and  $M^2 > k$ , so we cannot combine  $C^{M^2}$  with any other clique. Since  $M^2 > (B-3)m^2$ , we have also  $|C^{M^2}| > \frac{1}{2} |C^{M^2+(B-3)m^2}|$ . Hence, in the matching  $M'$  clique  $C^{M^2}$  must be matched with  $C^{M^2+(B-3)m^2}$ . Next in the matching  $M'$  every  $C_i^{M+3m}$  in  $G_c$  must be matched with a clique in  $G'$  which contains clique  $C_i^M$ , since otherwise the distance between  $G'$  and  $G_c$  is at least  $M$  and  $M > d$ . This also means that we cannot combine two big cliques  $C_i^M$  and  $C_j^M$ . Since  $d_M(G', G_c) \leq d = d_0 - 3m$ , to get solution  $G'$  every small clique  $C_i$  for  $1 \leq i \leq 3m$  has to be combined with some big clique  $C_j^M$ .

We can split  $k$  into two parts  $k = k_1 + k_2$ , where  $k_1 := m^2 MB$  is the number of edge insertions between big cliques and small cliques, and  $k_2$  is the total number of edge insertions between small cliques. Similarly to the analysis in Lemma 1, we have that  $k_2 \geq \frac{m^3}{2} B^2 - \frac{m^2}{2} \sum_{1 \leq i \leq 3m} a_i^2$  and the equality holds only if  $\{a_1, a_2, \dots, a_{3m}\}$  can be partitioned into  $m$  disjoint subsets  $A_1, A_2, \dots, A_m$  such that for  $1 \leq i \leq m$  it holds that  $\sum_{a_j \in A_i} a_j = B$ .  $\square$

Observe that when  $G$  is a cluster graph, we can “swap”  $G$  with  $G_c$  and  $k$  with  $d$ :

**Observation 1** When  $G$  is a cluster graph, instance  $(G, G_c, k, d)$  of DCEEDITING (EDGE DIST) is a yes-instance if and only if instance  $(G_c, G, d, k)$  of DCEEDITING (EDGE



**Fig. 4** Illustration of the constructed instance for the proof of Lemma 3. Each clique  $C_i$  in  $G$  contains two green vertices, which form a clique  $D_i$  in  $G_c$ . The instance of EXACT COVER BY 3-SETS is encoded by black vertices. On the left side each  $C_i$  in  $G$  encodes a set  $S_i$ ; on the right side each  $D_{x_i}$  in  $G_c$  encodes the appearance of the element  $x_i$  (Color figure online)

DIST) is a yes-instance. When both  $G$  and  $G_c$  are cluster graphs and  $E(G) \subseteq E(G_c)$ , instance  $(G, G_c, k, d)$  of DCCOMPLETION (EDGE DIST) is a yes-instance if and only if instance  $(G_c, G, d, k)$  of DCDELETION (EDGE DIST) is a yes-instance.

Observe that from Lemma 1 and Observation 1 we can infer NP-hardness for DCDELETION (EDGE DIST) even if  $G$  is a cluster graph. For the matching-based distance, we do not have an analogue of Observation 1. Thus, we provide another reduction showing NP-hardness for DCDELETION (MATCHING DIST) even if  $G$  is a cluster graph.

**Lemma 3** DCDELETION (MATCHING DIST) is NP-complete, even if the input graph  $G$  is a cluster graph.

**Proof** We present a polynomial-time reduction from the NP-hard EXACT COVER BY 3-SETS problem[31], where, given a set  $X$  with  $|X| = 3q$  and a collection  $\mathcal{S}$  of 3-element subsets of  $X$ , the task is to determine whether  $\mathcal{S}$  contains a subcollection  $\mathcal{S}' \subseteq \mathcal{S}$  of size  $q$  that covers every element in  $X$  exactly once. Given an instance  $(X, \mathcal{S})$  of EXACT COVER BY 3-SETS, where  $X = \{x_1, x_2, \dots, x_{3q}\}$  and  $\mathcal{S} = \{S_1, S_2, \dots, S_m\}$ , we construct an instance  $(G, G_c, k, d)$  of DCDELETION (MATCHING DIST) in polynomial time as follows.

The construction is illustrated in Fig. 4. For every set  $S_i = \{x_{i_1}, x_{i_2}, x_{i_3}\}$  in  $\mathcal{S}$ , we create a clique  $C_i = \{v_1^i, v_2^i\} \cup \{x_{i_1}^i, x_{i_2}^i, x_{i_3}^i\}$  in  $G$ . So  $G$  contains  $m$  order-five cliques  $C_1, C_2, \dots, C_m$ . For  $G_c$ , we first create  $m$  cliques  $D_1, D_2, \dots, D_m$  with  $D_i = \{v_1^i, v_2^i\}$ . Then for each element  $x_i$ , we create a clique  $D_{x_i} = \{x_i^j \mid x_i \in S_j\}$ .

For example, if an element  $x_i$  is contained in some set  $S_j$ , then in  $G$  the corresponding clique  $C_j$  for  $S_j$  contains a vertex  $x_i^j$  which is also contained in the clique  $D_{x_i}$  in  $G_c$ . Hence, if there is a subcollection  $\mathcal{S} \simeq$  of size  $q$  that covers every element in  $X$  exactly once, then we can find these  $q$  corresponding cliques in  $G$  and separate them to get  $3q$  new vertices each contained in one different clique  $D_{x_i}$  in  $G_c$ . Finally, we set  $k := 9q$  and  $d := 3m - 3q$ .

Note that the maximum-weight matching  $M^*$  for  $B(G, G_c)$  has to match every  $C_i$  in  $G$  with  $D_i$  in  $G_c$ . Thus  $d_M(G, G_c) = 3m$ . Now we show that  $(X, \mathcal{S})$  is a yes-instance of EXACT COVER BY 3-SETS if and only if  $(G, G_c, k, d)$  is a yes-instance of DCDELETION (MATCHING DIST).

( $\Rightarrow$ ): Assume that  $(X, \mathcal{S})$  is a yes-instance of EXACT COVER BY 3-SETS. Let  $\mathcal{S} \simeq$  be the solution. For every  $S_i \in \mathcal{S} \simeq$ , we find the corresponding clique  $C_i = \{v_1^i, v_2^i\} \cup \{x_{i_1}^i, x_{i_2}^i, x_{i_3}^i\}$  in  $G$  and partition it into four cliques  $\{v_1^i, v_2^i\}$ ,  $\{x_{i_1}^i\}$ ,  $\{x_{i_2}^i\}$ , and  $\{x_{i_3}^i\}$ . Let  $G'$  be the resulting cluster graph. For every such clique  $C_i$ , we delete nine edges to partition it. Thus, overall we need to delete  $9q = k$  edges. Since every element of  $X$  is covered by exactly one set from  $\mathcal{S} \simeq$ , we have that in  $G'$  we get  $3q$  new cliques each with one vertex, and each vertex is contained in a different clique from  $D_{x_1}, D_{x_2}, \dots, D_{x_{3q}}$ . Thus, we have  $d_M(G, G_c) = 3m - 3q = d$ .

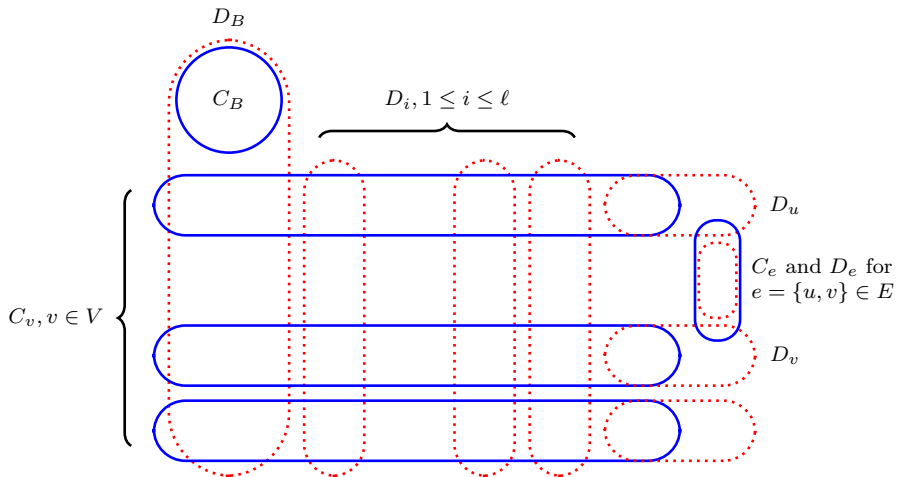
( $\Leftarrow$ ): Assume that  $(G, G_c, k, d)$  is a yes-instance of DCDELETION (MATCHING DIST). Let  $G'$  be the solution and  $M'$  be the maximum-weight matching between  $G'$  and  $G_c$ . Since we can only delete edges to get  $G'$  and every set  $S_i$  can only contain each element from  $X$  once, we get that in  $M'$  any edge incident on  $D_{x_i}$  has weight at most one. Since  $d_M(G, G_c) \leq d = 3m - 3q$ , it has to hold that in  $M'$  every  $D_{x_i}$  is matched with a new clique in  $G'$  and they share exactly one vertex. Thus we need  $3q$  new cliques in  $G'$  to be matched with  $D_{x_1}, D_{x_2}, \dots, D_{x_{3q}}$  in  $G_c$ . To get these  $3q$  new cliques, we need to separate at least  $3q$  vertices from  $C_1, C_2, \dots, C_m$  in  $G$ . Since we can delete at most  $k = 9q$  edges, there have to be  $q$  cliques from  $C_1, C_2, \dots, C_m$  such that we can separate each of them into four parts, where the first part contains  $\{v_1^i, v_2^i\}$  and the remaining three parts each have one vertex. Moreover, these  $3q$  new cliques each share one vertex with one different clique from  $D_{x_1}, D_{x_2}, \dots, D_{x_{3q}}$ . Thus, in the instance  $(X, \mathcal{S})$  of EXACT COVER BY 3-SETS we can find the corresponding  $3q$  sets and they cover each element of  $X$  exactly once.  $\square$

### 3.2 Parameterized Reductions

We first show that DCEEDITING (MATCHING DIST) is W[1]-hard when parameterized by the budget  $k$ .

**Lemma 4** DCEEDITING (MATCHING DIST) is NP-complete and W[1]-hard with respect to the budget  $k$ , even if the input graph  $G$  is a cluster graph.

**Proof** We present a parameterized reduction from CLIQUE, where, given a graph  $G_0$  and an integer  $\ell$ , we are asked to decide whether  $G_0$  contains a complete

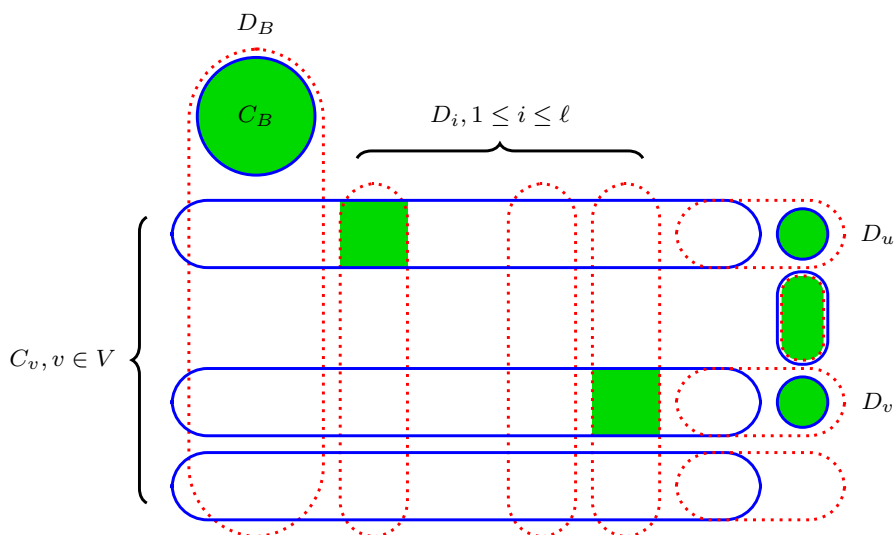


**Fig. 5** Illustration of the constructed instance for the proof of Lemma 4. Blue solid borders represent cliques in  $G$  and red dotted borders represent cliques in  $G_c$ . One horizontal long blue border represents a clique  $C_v$  in  $G$ . It has  $\ell + 2$  parts and each part is contained in one clique of  $G_c$ . The first part contains  $\ell^7$  vertices which are contained in the big clique  $D_B$  of  $G_c$ . The following  $\ell$  parts each contain  $\ell^3$  vertices which are contained in the  $\ell$  cliques  $D_i$  of  $G_c$ , and the last part contains  $\ell^2$  vertices which are contained in  $D_v$  of  $G_c$  (Color figure online)

subgraph of order  $\ell$ . CLIQUE is  $\mathbf{W}[1]$ -hard when parameterized by  $\ell$  [18]. Given an instance  $(G_0, \ell)$  of CLIQUE, we construct an instance  $(G, G_c, k, d)$  of DCEEDITING (MATCHING DIST) as follows.

The construction is illustrated in Fig. 5. Let  $n = |V(G_0)|$ . We first construct  $G$ . For every vertex  $v$  of  $G_0$ , we create a clique  $C_v$  of size  $\ell^7 + \ell^4 + \ell^2$ . For every edge  $e$  of  $G_0$ , we create a clique  $C_e$  of size  $\ell^4 + 2$ . Lastly, we create a big clique  $C_B$  of size  $\ell^8$ . Note that  $G$  is already a cluster graph. Next we construct  $G_c$ . We first create  $\ell$  cliques  $D_i$  of size  $n\ell^3$  for each  $1 \leq i \leq \ell$ . Every  $D_i$  contains  $\ell^3$  vertices in every  $C_v$  in  $G$ . In other words, every  $C_v$  in  $G$  contains  $\ell^3$  vertices in every  $D_i$  in  $G_c$ . Then we create a big clique  $D_B$  which contains all vertices in  $C_B$  and  $\ell^7$  vertices in every  $C_v$ . For every vertex  $v$  of  $G_0$ , we create clique  $D_v$  which contains  $\ell^2$  vertices in  $C_v$ , and one vertex in every  $C_e$  for  $v \in e$ . Lastly, for every edge  $e$  we create  $D_e$  which contains  $\ell^4$  vertices in  $C_e$ . We set  $k := \binom{\ell}{2}(2\ell^4 + 1) + \ell \binom{\ell - 1}{2}$  and we set  $d := d_0 - \ell(\ell - 1)$ , where  $d_0 = d_M(G, G_c)$  is the matching-based distance between  $G$  and  $G_c$ , which is computed as follows.

To compute  $d_M(G, G_c)$ , we need to find an optimal matching in  $B(G, G_c)$ , the weighted bipartite graph between  $G$  and  $G_c$ . First, in an optimal matching  $D_B$  must be matched with  $C_B$  since  $|C_B \cap D_B| = \ell^8 > |C_v \cap D_B| = \ell^7$  for any  $v \in V(G_0)$  and  $C_B \subseteq D_B$ . Similarly,  $D_e$  must be matched with  $C_e$  for every  $e \in E(G_0)$ . Then the remaining  $n$  cliques  $C_v$  in  $G$  need to be matched to  $\ell$  cliques  $D_i$  and  $n$  cliques  $D_v$  in  $G_c$ . Since  $|C_v \cap D_i| = \ell^3 > |C_v \cap D_v| = \ell^2$  for any  $v \in V(G_0)$  and  $1 \leq i \leq \ell$ , it is always better to match  $C_v$  with some  $D_i$ . Since there are only  $\ell$  cliques  $D_i$ , we can choose any  $\ell$  cliques from  $\{C_v \mid v \in V(G_0)\}$  to be matched with  $D_i$  for  $1 \leq i \leq \ell$  and



**Fig. 6** Illustration of a possible solution for the constructed instance (see Fig. 5) in the proof of Lemma 4. Blue solid borders represent cliques in  $G'$  and red dotted borders represent cliques in  $G_c$ . Green shaded areas indicate how cliques of  $G'$  and  $G_c$  are matched. If two horizontal cliques of  $G'$  (blue) are matched with two of the  $\ell$  vertical cliques of  $G_c$ , then the corresponding vertices are part of the clique and hence are adjacent. This means that the cliques corresponding to the edge can be matched in the indicated way (Color figure online)

the remaining  $n - \ell$  cliques to be matched with  $D_v$ . Thus we have many different matchings in  $B(G, G_c)$  which have the same maximum weight, and each of them corresponds to choosing  $\ell$  different cliques from  $\{C_v \mid v \in V(G_0)\}$  to be matched with  $D_i$  for  $1 \leq i \leq \ell$ . For each optimal matching, there are  $\ell$  free cliques  $D_v$  in  $G_c$  which are not matched.

This reduction works in polynomial time. We show that there is a clique of size  $\ell$  in  $G_0$  if and only if there is a cluster graph  $G' = (V, E')$  such that  $|E(G') \oplus E(G)| \leq k$  and  $d_M(G', G_c) \leq d$ .

( $\Rightarrow$ ): Assume that there is a clique  $C^*$  of order  $\ell$  in  $G_0$ . We modify the graph  $G$  as follows. First, for every edge  $e$  in the clique  $C^*$  partition the corresponding clique  $C_e$  in  $G$  into three parts; one part contains all vertices in  $D_e$  and the other two parts each have one vertex. After this we get  $\ell(\ell - 1)$  single vertices. Since  $C^*$  is a clique, all these single vertices can be partitioned into  $\ell$  groups such that each group has  $\ell - 1$  vertices and all these  $\ell - 1$  vertices are contained in the same  $D_v$  for some  $v \in C^*$ . Then for each  $v \in C^*$ , we combine the corresponding  $\ell - 1$  vertices into one clique  $C_v^{\ell-1}$ . Denote the resulting graph by  $G'$ . For an illustration see Fig. 6. Along the way to get  $G'$ , we delete  $\binom{\ell}{2}(2\ell^4 + 1)$  edges and add  $\ell \binom{\ell-1}{2}$  edges, thus  $|E(G) \oplus E(G')| = \binom{\ell}{2}(2\ell^4 + 1) + \ell \binom{\ell-1}{2} = k$ . Next we show that  $d_M(G', G_c) \leq d_0 - \ell(\ell - 1)$ . Recall that an optimal matching in  $B(G, G_c)$  can choose  $\ell$  cliques from  $\{C_v \mid v \in V(G_0)\}$  to be matched with  $D_i$  for  $1 \leq i \leq \ell$ . Now



in  $B(G', G_c)$  we can choose all cliques in  $\{C_v \mid v \in C^*\}$  to be matched with  $D_i$  for  $1 \leq i \leq \ell$ , and then match  $C_v^{\ell-1}$  with  $D_v$  for all  $v \in C^*$ . Then in the new matching we have  $\ell$  additional edges between  $C_v^{\ell-1}$  and  $D_v$  for  $v \in C^*$ , each with weight  $\ell - 1$ . Hence  $d_M(G', G_c) \leq d_0 - \ell(\ell - 1)$ .

( $\Leftarrow$ ): Assume that there is a cluster graph  $G' = (V, E')$  such that  $|E' \oplus E(G)| \leq k$  and  $d_M(G', G_c) \leq d$ . Note that  $k < \ell^7$ , thus  $k < |C_v|$  and  $k < |C_B|$ . Consequently, we can only modify edges between vertices in  $C_e$ . It is easy to see that in any optimal matching in  $B(G', G_c)$ , we still have that clique  $C_B$  must be matched with  $D_B$  and clique  $C_e$  must be matched with  $D_e$  for every  $e \in E(G_0)$ . We should choose  $\ell$  cliques from  $\{C_v \mid v \in V(G_0)\}$  to be matched with  $D_i$  for  $1 \leq i \leq \ell$ , which creates  $\ell$  free cliques  $D_v$ . Hence, to decrease the distance between  $G$  and  $G_c$  or to increase the matching, we have to create new cliques to be matched with these  $\ell$  free cliques  $D_v$ . Note that every  $D_v$  only contains single vertices from  $C_e$  with  $v \in e$  and the vertices contained in  $C_v$ . To create new cliques we need to first separate  $C_e$  to get single vertices and then combine them. To decrease the distance by  $\ell(\ell - 1)$ , we need to separate at least  $\ell(\ell - 1)$  single vertices from  $C_e$ . This will cost at least  $\ell(\ell - 1)(\ell^4 + 1) - \binom{\ell}{2} = \binom{\ell}{2}(2\ell^4 + 1)$  edge deletions if we always separate one  $C_e$  into three parts and get two single vertices. Then we need to combine these single vertices into at most  $\ell$  cliques since there are at most  $\ell$  free cliques  $D_v$ . This will cost at least  $\ell \binom{\ell - 1}{2}$  edge insertions if all these  $\ell(\ell - 1)$  single vertices can be partitioned into  $\ell$  groups and each group has  $\ell - 1$  vertices. Since  $k = \binom{\ell}{2}(2\ell^4 + 1) + \ell \binom{\ell - 1}{2}$ , we have that in the first step we have to choose  $\binom{\ell}{2}$  cliques  $C_e$  and separate them into three parts and all these  $\ell(\ell - 1)$  single vertices are evenly distributed in  $\ell$  free cliques  $D_v$ . This means that in  $G_0$  we can select  $\binom{\ell}{2}$  edges between  $\ell$  vertices and each vertex has  $\ell - 1$  incident edges. Thus there is a clique of size  $\ell$  in  $G_0$ .  $\square$

The next lemma shows that DCEEDITING (EDGE DIST) is  $W[1]$ -hard with respect to  $k$ . The corresponding parameterized reduction is from CLIQUE and shares some similarities with the reduction presented in the proof of Lemma 4 with respect to the edge gadgets. Our proof is based on the following property for instances of DCEEDITING (EDGE DIST) with  $k + d = |E(G) \oplus E(G_c)|$ .

**Observation 2** *If an instance  $(G, G_c, k, d)$  of DCEEDITING (EDGE DIST) (DCDELETION (EDGE DIST) or DCCOMPLETION (EDGE DIST)) has the property that  $k + d = |E(G) \oplus E(G_c)|$ , then any solution  $G'$  satisfies that  $|E(G) \oplus E(G')| = k$ ,  $|E(G') \oplus E(G_c)| = d$ , and  $E(G) \oplus E(G') \subseteq E(G) \oplus E(G_c)$ .*

**Proof** On the one hand, for any graph  $G'$ , we have that

$$|E(G') \oplus E(G)| + |E(G') \oplus E(G_c)| \geq |E(G) \oplus E(G_c)| = k + d.$$

On the other hand, a solution  $G'$  satisfies that  $|E(G') \oplus E(G)| \leq k$  and  $|E(G') \oplus E(G_c)| \leq d$ . Thus we have that  $|E(G) \oplus E(G')| = k$  and  $|E(G') \oplus E(G_c)| = d$ . Let  $S_1 := (E(G) \oplus E(G')) \setminus (E(G) \oplus E(G_c))$  and  $S_2 := (E(G) \oplus E(G')) \setminus S_1$ . Then

$$|E(G') \oplus E(G)| = |S_1| + |S_2|. \quad (1)$$

Next we show that

$$|E(G') \oplus E(G_c)| = |E(G) \oplus E(G_c)| + |S_1| - |S_2| = k + d + |S_1| - |S_2|. \quad (2)$$

Let us consider

$$\begin{aligned} E(G') \oplus E(G_c) &= (E(G) \oplus (E(G) \oplus E(G'))) \oplus E(G_c) \\ &= (E(G) \oplus (S_1 \cup S_2)) \oplus E(G_c) \\ &= (E(G) \oplus E(G_c)) \oplus (S_1 \cup S_2) \\ &= ((E(G) \oplus E(G_c)) \setminus (S_1 \cup S_2)) \cup ((S_1 \cup S_2) \setminus (E(G) \oplus E(G_c))) \\ &= (E(G) \oplus E(G_c)) \setminus S_2 \cup S_1, \end{aligned}$$

where the last equation holds since

$$S_1 = (E(G) \oplus E(G')) \setminus (E(G) \oplus E(G_c)) \Rightarrow S_1 \cap (E(G) \oplus E(G_c)) = \emptyset$$

and

$$S_2 = (E(G) \oplus E(G')) \setminus S_1 \Rightarrow S_2 \subseteq E(G) \oplus E(G_c).$$

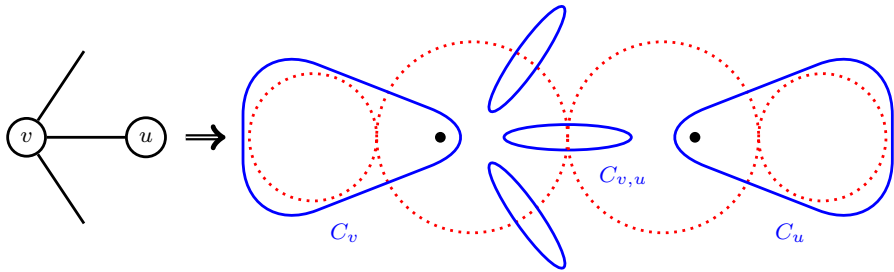
Hence,  $|E(G') \oplus E(G_c)| = |E(G) \oplus E(G_c)| + |S_1| - |S_2|$ . If  $S_1 \neq \emptyset$ , then combine Eqs. (1) and (2) and we get

$$|E(G') \oplus E(G)| + |E(G') \oplus E(G_c)| = k + d + 2|S_1| > k + d,$$

which is a contradiction. Thus we conclude that  $S_1 = \emptyset$  and hence  $E(G) \oplus E(G') \subseteq E(G) \oplus E(G_c)$ .  $\square$

From this result we can conclude that, when  $k + d = |E(G) \oplus E(G_c)|$ , the only way to get a solution  $G'$  is to find a subset of  $E(G) \oplus E(G_c)$  with size exactly  $k$  such that modifying the edges of this subset in  $G$  yields a cluster graph.

**Lemma 5** DCEEDITING (EDGE DIST) is NP-complete and  $W[1]$ -hard with respect to the budget  $k$ , even if the input graph  $G$  is a cluster graph and  $k + d = |E(G) \oplus E(G_c)|$ .



**Fig. 7** Illustration of the constructed instance for the proof of Lemma 5. On the left side there are two connected vertices  $v$  and  $u$  in  $G_0$  and three edges incident on  $v$ . On the right side we have the corresponding parts in  $G$  and  $G_c$ . Red dotted circles represent cliques in  $G_c$  and blue solid borders represent cliques in  $G$  (Color figure online)

**Proof** We present a parameterized reduction from CLIQUE, where given a graph  $G_0$  and an integer  $\ell$ , we are asked to decide whether  $G_0$  contains a complete subgraph of order  $\ell$ . CLIQUE is  $W[1]$ -hard when parameterized by  $\ell$  [18]. Given an instance  $(G_0, \ell)$  of CLIQUE, we construct an instance  $(G, G_c, k, d)$  of DCEEDITING (EDGE DIST) as follows. The construction is illustrated in Fig. 7. We set  $L_1 := \ell^7 + 1$  and  $L_2 := \ell^2$ . We first construct  $G$ . For every vertex  $v$  of  $G_0$ , we create a clique  $C_v$  of size  $L_1 + 1 = \ell^7 + 2$ , and for every edge  $e$  of  $G_0$ , we create a clique  $C_e$  of size  $2L_2$ . Note that  $G$  is already a cluster graph. Next, we construct  $G_c$ . For every vertex  $v$  of  $G_0$ , let  $C_e^1, C_e^2, \dots, C_e^p$  be all cliques of size  $2L_2$  in  $G$  which represent all edges incident on  $v$ . For each vertex  $v$  of  $G_0$ , we create two cliques in  $G_c$ . One of them contains  $L_1$  vertices of  $C_v$ , the other contains the one remaining vertex of  $C_v$ , called the *single vertex* of  $C_v$ , and  $L_2$  vertices from every  $C_e^i$  for  $1 \leq i \leq p$  (see also Fig. 7). Set

$$k := \ell L_1 + \ell(\ell - 1)L_2 + \ell \binom{\ell - 1}{2} L_2^2 + \binom{\ell}{2} L_2^2 \quad (3)$$

and set  $d := |E(G) \oplus E(G_c)| - k$ . This reduction works in polynomial time.

Now we show that there is a clique of size  $\ell$  in  $G_0$  if and only if there is a cluster graph  $G' = (V, E')$  such that  $|E' \oplus E(G)| \leq k$  and  $|E' \oplus E(G_c)| \leq d$ . To simplify the proof, we assume  $\ell \geq 3$  in the following.

( $\Rightarrow$ ): Assume that there is a clique  $C^*$  of size  $\ell$  in  $G_0$ . We modify graph  $G$  in the following two steps. We first partition cliques in  $G$  according to  $C^*$  by deleting edges as follows. For every vertex  $v$  in  $C^*$ , find the clique  $C_v$  and delete edges between the single vertex in  $C_v$  and the remaining  $L_1$  vertices. For every edge  $e$  in  $C^*$ , find the clique  $C_e$  in  $G$  and delete edges to partition the clique into two parts, each with  $L_2$  vertices. In the first step we delete  $\ell L_1 + \binom{\ell}{2} L_2^2$  edges. The next step is to combine some cliques by adding edges. For every vertex  $v$  in  $C^*$ , we combine the single vertex from  $C_v$  and  $\ell - 1$  cliques of size  $L_2$  into one clique. In this step we add  $\ell(\ell - 1)L_2 + \ell \binom{\ell - 1}{2} L_2^2$  edges. Thus in total we modify  $k$  edges.

( $\Leftarrow$ ): Assume that there is a cluster graph  $G' = (V, E')$  such that  $|E(G') \oplus E(G)| \leq k$  and  $|E(G') \oplus E(G_c)| \leq d$ . Since  $k + d = |E(G) \oplus E(G_c)|$ , we have  $|E(G) \oplus E(G')| = k$  and  $|E(G') \oplus E(G_c)| = d$ . Thus, to get the solution  $G'$  we have to modify exactly  $k$  edges from  $E(G) \oplus E(G_c)$ . As a result, we only have the following four kinds of operations:

1. partition a clique  $C_v$  in  $G$  into two parts, one with the single vertex and the other with  $L_1$  vertices, which costs  $L_1 = \ell^7 + 1$  edge deletions;
2. partition a clique  $C_e$  in  $G$  into two parts, each with  $L_2$  vertices contained in one clique in  $G_c$ , which costs  $L_2^2 = \ell^4$  edge deletions;
3. combine the single vertex of  $C_v$  with some cliques of size  $L_2$  which come from partitioning clique  $C_e$  into two parts, which costs  $aL_2 = a\ell^2$  edge insertions for some integer  $a$ ;
4. combine some cliques of size  $L_2$  which come from partitioning clique  $C_e$  into two parts, which costs  $\binom{b}{2}L_2^2 = \binom{b}{2}\ell^4$  edge insertions for some integer  $b$ .

First, we claim that there must be  $\ell$  cliques of size  $L_1 + 1$  in  $G$  that have been partitioned. Note that  $k = \ell^8 + \frac{1}{2}\ell^7 - \ell^6 + \frac{1}{2}\ell^5 + \ell^4 - \ell^3 + \ell$ , where the last additive term  $\ell$  can only come from partitioning  $\ell$  cliques of size  $L_1 + 1$  in  $G$ . In addition, there cannot be more than  $\ell$  cliques of size  $L_1 + 1$  in  $G$  that have been partitioned, since  $(\ell + 1)L_1 > k$  (assuming  $\ell \geq 3$ ). Thus exactly  $\ell$  cliques of size  $L_1 + 1$  in  $G$  have to be partitioned and we get  $\ell$  single vertices. This costs  $\ell L_1$  edge deletions, which is the first additive item of Eq. (3).

Next, we claim that at least  $\ell(\ell - 1)$  cliques of size  $L_2$  are combined with these  $\ell$  single vertices we got in the last step. This is because the second term of  $k$ ,  $\ell(\ell - 1)L_2$ , is strictly less than  $\ell^4$ , and hence can only come from the third kind of operation, combining the single vertex with cliques of size  $L_2$ . Suppose that  $\ell(\ell - 1) + \delta$  cliques of size  $L_2$  are combined with these single vertices for some  $\delta \geq 0$ . Then we need  $(\ell(\ell - 1) + \delta)L_2$  edge insertions. Note that the second additive term of Eq. (3) is  $\ell(\ell - 1)L_2$ .

Then, we need to partition at least  $\binom{\ell}{2} + \frac{[\delta]}{2}$  cliques of size  $2L_2$  so that we can combine them with single vertices. Denote by  $f_1(\ell, \delta)$  the number of edge deletions this separation cost. Clearly,  $f_1(\ell, \delta) \geq \left(\binom{\ell}{2} + \frac{[\delta]}{2}\right)L_2^2$ . Notice that the last additive term of Eq. (3) is  $\binom{\ell}{2}L_2^2$ .

Finally, when we combine a single vertex with more than one clique of order  $L_2$ , then we also need to add edges between these cliques. Denote by  $f_2(\ell, \delta)$  the number of edge insertions between these cliques. Since we have  $\ell(\ell - 1) + \delta$  cliques of size  $L_2$  and  $\ell$  single vertices, and every clique is combined with one single vertex, it follows that  $f_2(\ell, \delta) \geq \ell \binom{\ell - 1}{2}L_2^2$ . Notice that the third additive term of Eq. (3) is  $\ell \binom{\ell - 1}{2}L_2^2$ .

Overall, we need

$$\ell L_1 + (\ell(\ell - 1) + \delta)L_2 + f_2(\ell, \delta) + f_1(\ell, \delta) \geq k$$

edge modifications. Equality only holds if  $\delta = 0$ ,  $f_1(\ell, \delta) = \binom{\ell}{2} L_2^2$ , and  $f_2(\ell, \delta) = \ell \binom{\ell - 1}{2} L_2^2$ . Here  $f_2(\ell, \delta) = \ell \binom{\ell - 1}{2} L_2^2$  means that we can partition all  $\ell(\ell - 1)$  cliques of size  $L$  into  $\ell$  parts, each with  $\ell - 1$  cliques, and then combine all  $\ell - 1$  cliques in each part with one single vertex. Moreover,  $f_1(\ell, \delta) = \binom{\ell}{2} L_2^2$  means that all these  $\ell(\ell - 1)$  cliques of order  $L$  come from partitioning  $\binom{\ell}{2}$  cliques of order  $2L_2$ . Then, in  $G_0$  we have  $\ell$  vertices (corresponding to these  $\ell$  single vertices) and  $\binom{\ell}{2}$  edges (corresponding to these  $\binom{\ell}{2}$  cliques of order  $2L_2$ ) such that each vertex has  $\ell - 1$  incident edges from these  $\binom{\ell}{2}$  edges. Hence, these  $\ell$  vertices form a clique in  $G_0$ .  $\square$

Note that in the reduction of Lemma 5 the constructed graph  $G$  is a cluster graph. According to Observations 1 and 2, this reduction can also be used to prove  $W[1]$ -hardness with respect to the distance  $d$ .

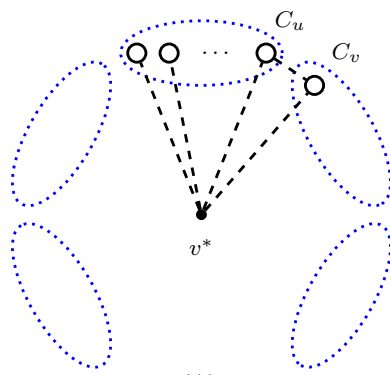
**Corollary 1**  $DC_{\text{EDITING}}(\text{EDGE DIST})$  is  $NP$ -complete and  $W[1]$ -hard with respect to the distance  $d$ , even if the input graph  $G$  is a cluster graph and  $k + d = |E(G) \oplus E(G_c)|$ .

The following result also exploits the property that we need exactly  $k$  edge modifications when  $k + d = |E(G) \oplus E(G_c)|$ .

**Lemma 6**  $DC_{\text{DELETION}}(\text{EDGE DIST})$  is  $W[1]$ -hard with respect to the distance  $d$ , even when  $k + d = |E(G) \oplus E(G_c)|$ .

**Proof** We present a parameterized reduction from  $MULTICOLORED\ CLIQUE$ . In  $MULTICOLORED\ CLIQUE$ , we are given an integer  $\ell$  and a graph where every vertex is colored with one of  $\ell$  colors. The task is to find a clique of order  $\ell$  containing one vertex of each color.  $MULTICOLORED\ CLIQUE$  is  $W[1]$ -hard with respect to  $\ell$  [20]. Let  $(G_0 = (V, E), \ell)$  be an instance of  $MULTICOLORED\ CLIQUE$ . We construct an instance  $(G, G_c, k, d)$  of  $DC_{\text{DELETION}}(\text{EDGE DIST})$  as follows. For every vertex  $v$  in  $G_0$ , create a clique  $C_v$  with  $2\ell$  vertices in  $G_c$ . Add a special clique with one vertex  $v^*$  in  $G_0$ . For graph  $G$ , first copy  $G_c$  and then add more edges as follows: add edges between  $v^*$  and all other vertices in  $G$ , and for every edge  $\{u, v\}$  in  $G_0$ , add all edges between vertices in  $C_u$  and vertices in  $C_v$ . Set  $d = 2\ell^2 + 4\ell^2 \binom{\ell}{2}$  and  $k = |E(G) \oplus E(G_c)| - d$ . This reduction works in polynomial time and the construction is illustrated in Fig. 8.

**Fig. 8** Illustration of the constructed instance for the proof of Lemma 6. A circle represents a clique in  $G_c$ . Circles in the same blue dotted ellipse mean that the corresponding vertices in  $G_0$  have the same color. Dotted edges represent the additional edges in  $G$ . Vertex  $v^*$  is connected to all other vertices in  $G$ . For an edge  $\{u, v\}$  in  $G_0$ , the corresponding two cliques  $C_u$  and  $C_v$  are connected in  $G$  (Color figure online)



Note that  $k + d = |E(G) \oplus E(G_c)|$  and according to Observation 2 a solution  $G'$  for instance  $(G, G_c, k, d)$  has to delete exactly  $k$  edges from  $E(G) \oplus E(G_c)$  from  $G$ , which is equivalent to adding exactly  $d$  edges from  $E(G) \oplus E(G_c)$  to  $G_c$ . Next we show that there is a multicolored clique of order  $\ell$  in  $G_0$  if and only if there is a cluster graph  $G' = (V, E(G'))$  such that  $|E(G') \oplus E(G)| \leq k$  and  $|E(G') \oplus E(G_c)| \leq d$ .

( $\Rightarrow$  : ) Suppose that there is a multicolored clique  $C_0$  of size  $\ell$  in  $G_0$ , then for all vertices in  $C_0$  find the corresponding cliques in  $G_c$ , and combine these  $\ell$  cliques and vertex  $v^*$  into one big clique. Denote the resulting graph by  $G'$ . To get graph  $G'$  from  $G$ , we need to delete  $|E(G) \oplus E(G_c)| - (2\ell^2 + 4\ell^2 \binom{\ell}{2}) = k$  edges, and all these edges are in  $E(G) \oplus E(G_c)$ . In this way we get a new cluster graph  $G'$  such that  $|E(G') \oplus E(G)| = k$  and  $|E(G') \oplus E(G_c)| = d$ .

( $\Leftarrow$  : ) Suppose that there is a cluster graph  $G'$  such that  $|E(G') \oplus E(G)| \leq k$  and  $|E(G') \oplus E(G_c)| \leq d$ . Since  $k + d = |E(G) \oplus E(G_c)|$ , it has to hold that  $|E(G') \oplus E(G)| = k$  and  $|E(G') \oplus E(G_c)| = d$ . Since  $d = 2\ell^2 + 4\ell^2 \binom{\ell}{2}$ , and except for  $v^*$ , every clique in  $G_c$  has  $2\ell$  vertices,  $2\ell^2$  in  $d$  must come from adding edges between  $v^*$  and  $\ell$  cliques in  $G_c$ . Since  $G'$  is a cluster graph, there must be edges between every pair of these  $\ell$  cliques in  $G$ , which means that there is a multicolored clique of order  $\ell$  in  $G_0$ .  $\square$

The final two results show  $\mathbf{W}[1]$ -hardness with respect to the distance  $d$  for DCEEDITING (MATCHING DIST) and DCDELETION (MATCHING DIST).

**Lemma 7** DCEEDITING (MATCHING DIST) is  $\mathbf{W}[1]$ -hard with respect to the distance  $d$ .

**Proof** We present a parameterized reduction from CLIQUE ON REGULAR GRAPHS, where given a regular graph  $G^* = (V, E)$  with vertex degree  $r$  with  $r < \frac{n}{2}$ , and a number  $k^*$  with  $k^* \leq r$ , we are asked to decide whether  $G^*$  contains a clique of size  $k^*$ . CLIQUE ON REGULAR GRAPHS is  $\mathbf{W}[1]$ -hard with respect to  $k^*$  [10]. Given an instance  $(G_0, k_0, r)$

of CLIQUE ON REGULAR GRAPHS, we construct an instance  $(G, G_c, d, k)$  of DCEEDITING (MATCHING DIST) as follows.

Graph  $G$  is the same as  $G_0$  and graph  $G_c = (V, \binom{V}{2})$  is a complete graph. Set  $d := k_0$  and  $k := \frac{n(n-1-r)}{2} - k_0(n + k_0 - 2r - 2)$ . The construction can trivially be done in polynomial time. In the following we show that there is a clique of size  $k_0$  in  $G_0$  if and only if  $(G, G_c, d, k)$  is a yes-instance of DCEEDITING (MATCHING DIST).

( $\Rightarrow$ ): Assume that there is a clique of order  $k_0$  in  $G_0$ ; we construct a graph  $G'$  which consists of two cliques, where one of them contains the vertices from the clique of order  $k_0$  in  $G_0$ ; the other, denoted by  $C_{\max}$ , contains the remaining vertices and has order  $n - k_0$ . Next we compute  $|E(G) \oplus E(G')|$ , which consists of two parts:

- $D(k_0)$ : the set of edges between vertices in  $C_{\max}$  and the remaining vertices, and
- $A(k_0)$ : the set of added edges between vertices in  $C_{\max}$ .

Since the vertices outside  $C_{\max}$  form a clique, every such vertex has  $r - k_0 + 1$  edges connected to vertices in  $C_{\max}$ . Thus  $|D(k_0)| = k_0(r - k_0 + 1)$ . To determine  $|A(k_0)|$ , we count the sum of the degrees of vertices in  $C_{\max}$ . Before adding edges to  $C_{\max}$ , the sum is  $(n - k_0)r$ . After adding edges the sum should be  $(n - k_0)(n - k_0 - 1) + |D(k_0)|$ . So the number of edges which need to be added to  $C_{\max}$  is

$$|A(k_0)| = \frac{(n - k_0)(n - k_0 - 1) + |D(k_0)| - (n - k_0)r}{2}.$$

Then we get the size of the modification set for  $G'$ :

$$|E(G) \oplus E(G')| = |D(k_0)| + |A(k_0)| = \frac{n(n - 1 - r)}{2} - k_0(n + k_0 - 2r - 2) = k.$$

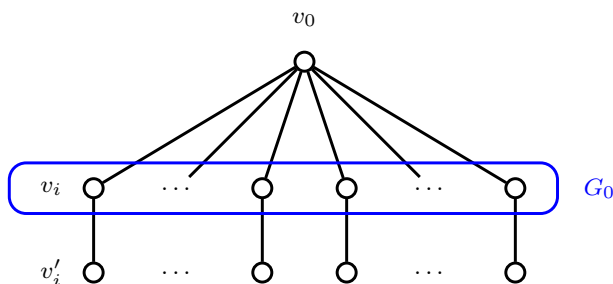
( $\Leftarrow$ ): To simplify the following proof, we define three functions:

- $g_1(x) := x(r - x + 1)$ ,
- $g_2(x) := \frac{(n-x)(n-x-1) + g_1(x) - (n-x)r}{2}$ , and
- $f(x) := g_1(x) + g_2(x) = \frac{n(n-1-r)}{2} - x(n + x - 2r - 2)$ .

Since  $r < \frac{n}{2}$ , we have that  $f(x)$  is monotonically decreasing and  $f(k_0) = k$ .

Suppose that there is no clique of size  $k_0$  in  $G_0$ . We need to show that there is no cluster graph  $G'$  satisfying both  $|E(G) \oplus E(G')| \leq k$  and  $d(G_c, G') \leq d$ . Suppose towards a contradiction that there is such a cluster graph  $G'$ . Denote the largest cluster in  $G'$  as  $C_{\max}$ . Since  $d_M(G_c, G') \leq d$ , we have that  $|V(C_{\max})| \geq n - k_0$ . Define

- $D$ : the set of edges between vertices in  $C_{\max}$  and the remaining vertices, and
- $A$ : the set of added edges between vertices in  $C_{\max}$ .



**Fig. 9** Illustration of constructed graph  $G$  in the proof of Lemma 8. Edges in  $G_0$  are not shown. Every vertex  $v_i$  is connected to its copy  $v'_i$  and the universal vertex  $v_0$

To get the clique  $C_{\max}$  from  $G$ , we have to delete all edges in  $D$  and add all edges in  $A$ , thus  $|E(G) \oplus E(G')| \geq |D| + |A|$ . We distinguish the following two cases:

*Case 1*  $|C_{\max}| = n - k_0$ . Every vertex outside  $C_{\max}$  has at least  $r - k_0 + 1$  edges connected to vertices in  $C_{\max}$ , and since there is no clique of order  $k_0$  in  $G_0$ , among all vertices outside  $C_{\max}$ , there is at least one vertex which has more than  $r - k_0 + 1$  edges connected to vertices in  $C_{\max}$ . This means that  $|D| > g_1(k_0)$  and  $|A| > g_2(k_0)$ . Thus, we have:

$$|E(G) \oplus E(G')| \geq |D| + |A| > g_1(k_0) + g_2(k_0) = f(k_0) = k.$$

*Case 2*  $|C_{\max}| > n - k_0$ .

Suppose that  $|C_{\max}| = n - k'$ , where  $k' < k_0$  is the number of all vertices outside  $C_{\max}$ . Now we have  $|D| \geq g_1(k')$  and  $|A| \geq g_2(k')$ , and

$$|E(G) \oplus E(G')| \geq |D| + |A| \geq g_1(k') + g_2(k') = f(k') > f(k_0) = k.$$

The last inequality holds since  $f(k)$  is monotonically decreasing.

In both cases we have that there is no solution for instance  $(G, G_c, d, k)$ . □

The above reduction cannot be used to show  $\mathbf{W}[1]$ -hardness with respect to  $d$  for DCDELETION (MATCHING DIST) since both edge insertions and edge deletions are needed. Next we show that DCDELETION (MATCHING DIST) remains  $\mathbf{W}[1]$ -hard with respect to  $d$ .

**Lemma 8** DCDELETION (MATCHING DIST) is  $\mathbf{W}[1]$ -hard with respect to the distance  $d$ .

**Proof** We present a parameterized reduction from CLIQUE ON REGULAR GRAPHS, where, given a regular graph  $G^* = (V, E)$  with vertex degree  $r$  with  $r < \frac{n}{2}$ , and number  $k^*$  with  $k^* \leq r$ , we are asked to decide whether  $G^*$  contains a clique of size  $k^*$ . CLIQUE ON REGULAR GRAPHS is known to be  $\mathbf{W}[1]$ -hard with respect to  $k^*$  [10]. Given an instance  $(G_0, \ell, r)$  of CLIQUE ON REGULAR GRAPHS, where  $G_0$  is a regular graph



with vertex degree  $r$ , we construct an instance  $(G, G_c, d, k)$  of DCDELETION (MATCHING DIST) as follows. The construction is illustrated in Fig. 9.

Let  $\{v_1, v_2, \dots, v_n\}$  be the vertex set of  $G_0$ . For graph  $G$ , we first copy the whole graph  $G_0$ . Then we add a universal vertex and a private neighbor for each original vertex: we add a universal vertex  $v_0$  and add an edge between  $v_0$  and every vertex  $v_i$  in  $G_0$ , and for every vertex  $v_i$  in  $G_0$ , we add vertex  $v'_i$  and add an edge between  $v_i$  and  $v'_i$ . The graph  $G_c$  has the same vertex set as  $G$ . Moreover, graph  $G_c$  contains edges between  $v_i$  and  $v'_i$  for all  $1 \leq i \leq n$ . That is,  $G_c$  consists of  $n + 1$  cliques:  $C_0$  with  $V(C_0) = \{v_0\}$  and  $C_i$  with  $V(C_i) = \{v_i, v'_i\}$  for  $1 \leq i \leq n$ . Set  $k = n + \frac{m}{2} - \binom{\ell}{2}$

and  $d = \ell$ . Next we show that there is clique of size  $\ell$  in  $G_0$  if and only if the constructed instance  $(G, G_c, d, k)$  is a yes-instance of DCDELETION (MATCHING DIST).

( $\Rightarrow$ ): Assume that there is clique  $C^*$  of order  $\ell$  in  $G_0$ . Then in  $G$  we first delete edges  $\{v_i, v'_i\}$  for all  $v_i \in V(C^*)$  and delete edges  $\{v_0, v_i\}$  for all  $v_i \in \{v_1, v_2, \dots, v_n\} \setminus V(C^*)$ . Second, delete all edges between vertices in  $\{v_1, v_2, \dots, v_n\}$  except for edges between vertices in  $C^*$ . We delete  $n$  edges in the first step and  $\frac{m}{2} - \binom{\ell}{2}$  edges in the second step, since  $G_0$  is a regular graph and  $C^*$

is a clique. By deleting these  $n + \frac{m}{2} - \binom{\ell}{2} = k$  edges, we get a cluster graph  $G'$  which contains  $n + 1$  cliques:  $C'_i$  with  $V(C'_i) = \{v'_i\}$  for  $v_i \in V(C^*)$ ,  $C'_j$  with  $V(C'_j) = \{v_j, v'_j\}$  for  $v_i \in \{v_1, v_2, \dots, v_n\} \setminus V(C^*)$ , and  $C'_0$  with  $V(C'_0) = \{v_0\} \cup V(C^*)$ . Thus  $d_M(G', G_c) = \ell$ .

( $\Leftarrow$ ): Assume that  $(G, G_c, d, k)$  is a yes-instance of DCDELETION (MATCHING DIST) and let  $G'$  be a solution. Since we can only delete edges, for every pair of edges  $\{v_i, v'_i\}$  and  $\{v_i, v_0\}$  for  $v_i \in \{v_1, v_2, \dots, v_n\}$ , we have to delete one of them because  $\{v_0, v'_i\} \notin E(G)$ . This means that for every  $v_i \in \{v_1, v_2, \dots, v_n\}$  vertex  $v_i$  is either in the same clique with  $v'_i$  or with  $v_0$ . Suppose that in  $G'$  there are  $p \leq \ell$  vertices from  $\{v_1, v_2, \dots, v_n\}$  which are in the same clique with  $v_0$ . Then these  $p$  vertices must form a clique  $C'$  in  $G_0$ . To get  $G'$ , we have to delete edges  $\{v_i, v'_i\}$  for all  $v_i \in V(C')$  and edge  $\{v_i, v_0\}$  for every vertex  $v_i \in \{v_1, v_2, \dots, v_n\} \setminus V(C')$ . This costs  $n$  edge deletions. Moreover, we have to delete all edges between vertices in  $\{v_1, v_2, \dots, v_n\}$  except for edges between vertices in  $C'$ . This costs  $\frac{m}{2} - \binom{p}{2}$  edge deletions. Overall we have

$$|E(G) \oplus E(G')| = n + \frac{m}{2} - \binom{p}{2}.$$

Since  $G'$  is a solution, we have that  $|E(G) \oplus E(G')| \leq k = n + \frac{m}{2} - \binom{\ell}{2}$ . Hence,  $p \geq \ell$  and  $G_0$  contains a clique of order  $\ell$ .  $\square$

We now have shown all intractability results stated in Theorem 2.

## 4 Fixed-Parameter Tractability Results

In this section, we complement the hardness results of Sect. 3 by identifying tractable cases for the considered variants of DYNAMIC CLUSTER EDITING. We first show that all problem variants admit a polynomial kernel for the combination of the two parameters budget  $k$  and distance  $d$ . Then we present further FPT-results with respect to single parameters.

### 4.1 Polynomial Kernels for the Combined Parameter $(k + d)$

In this subsection, we present polynomial kernels with respect to the parameter combination  $(k + d)$  for all considered variants of DYNAMIC CLUSTER EDITING : Formally, we prove the following theorem.

**Theorem 3** *The following problems admit an  $O(k^2 + d^2)$ -vertex kernel:*

- DCEEDITING (MATCHING DIST),
- DCDELETION (MATCHING DIST), and
- DCCOMPLETION (MATCHING DIST).

*The following problems admit an  $O(k^2 + k \cdot d)$ -vertex kernel:*

- DCEEDITING (EDGE DIST),
- DCDELETION (EDGE DIST), and
- DCCOMPLETION (EDGE DIST).

*All kernels can be computed in  $O(|V|^3)$  time.*

We describe polynomial-time data reduction rules that each take as input an instance  $(G = (V, E), G_c = (V, E_c), k, d)$  and output a reduced instance. We say that a data reduction rule is *correct* if the reduced instance is a yes-instance if and only if the original instance is a yes-instance (of the corresponding problem variant). A data reduction rule works for all problem variants that fit a given restriction. For example, the restriction Editing/Deletion (given for Reduction Rule 2a) addresses the problems: DCEEDITING (EDGE DIST), DCEEDITING (MATCHING DIST), DCDELETION (EDGE DIST), and DCDELETION (MATCHING DIST). If no restriction is given, then a data reduction rule works for all problem variants. In the correctness proof of each data reduction rule, we assume that all previous rules are not applicable.

The first rule formalizes an obvious constraint on the solvability of the instance (for all problem variants). The correctness of this rule is obvious.

**Reduction Rule 1** *If  $k < 0$  or  $d < 0$ , then output NO.*<sup>1</sup>

<sup>1</sup> Formally, this does not fit the definition of a data reduction rule, but we can assume that instead of NO the rule outputs a trivial no-instance of constant size.

We next use some well-known data reduction rules for classic CLUSTER EDITING[26] to get a graph which consists of isolated cliques plus one vertex set of size  $k^2 + 2k$  that does not contain any isolated cliques. These rules remove edges that are part of  $k + 1$  induced  $P_3$ s and add edges between non-adjacent vertex pairs that are part of  $k + 1$  induced  $P_3$ s. The correctness proofs are straightforward adaptations of the correctness proofs of these rules for classic CLUSTER EDITING. The reason we use these data reduction rules instead of rules used for linear-vertex kernels for classic CLUSTER EDITING[11, 13, 27] is that these rules do not eliminate any possible solution. Thus, the presented rules perform edge edits that are provably part of every optimal edge modification set.

**Reduction Rule 2a** (Editing/Deletion) *If there are  $k + 1$  induced  $P_3$ s in  $G$  that contain a common edge  $\{u, v\} \in E$ , then remove that edge from  $E$  and decrease  $k$  by one.*

**Reduction Rule 2b** (Completion) *If there are  $k + 1$  induced  $P_3$ s in  $G$  that contain a common edge  $\{u, v\} \in E$ , then output NO.*

**Reduction Rule 3a** (Editing/Completion) *If there are  $k + 1$  induced  $P_3$ s in  $G$  that contain a common non-edge  $\{u, v\} \notin E$ , then add that edge to  $E$  and decrease  $k$  by one.*

**Reduction Rule 3b** (Deletion) *If there are  $k + 1$  induced  $P_3$ s in  $G$  that contain a common non-edge  $\{u, v\} \notin E$ , then output NO.*

**Lemma 9** *Reduction Rules 2a, 2b, 3a, and 3b are correct.*

**Proof** Let  $I = (G = (V, E), G_c = (V, E_c), k, d)$  be an instance of a problem variant of DYNAMIC CLUSTER EDITING and let  $I^*$  be the instance after applying any of the four reduction rules. It is obvious that if  $I$  is a no-instance, then  $I^*$  is also a no-instance. In the following we show that if  $I$  is a yes-instance, then so is  $I^*$  for Reduction Rule 2a and Reduction Rule 2b. The correctness of Reduction Rule 3a and Reduction Rule 3b follows by symmetric arguments.

We now show for Reduction Rule 2a that the removed edge has to be in any solution for instance  $I$ . This implies that if  $I$  is a yes-instance of a completion variant of the problem, then Reduction Rule 2b is not applicable. Assume for the sake of contradiction that  $I$  is a yes-instance and that  $I^*$  is a no-instance. Then there is a cluster graph  $G'$  with  $|E(G') \oplus E| \leq k$  that is a solution for  $I$  and contains edge  $\{u, v\}$ . However, we know that there are  $k + 1$  vertices  $w_1, w_2, \dots, w_{k+1}$  such that  $G[\{u, v, w_i\}]$  is a  $P_3$  for all  $1 \leq i \leq k + 1$  (otherwise the rule would not be applicable). To destroy these  $P_3$ s without removing edge  $\{u, v\}$  we need at least  $k + 1$  edge additions or deletions. This is a contradiction to the assumption that  $|E(G') \oplus E| \leq k$ .  $\square$

As for classic CLUSTER EDITING we can upper-bound the number of vertices that are part of  $P_3$ s, leading to the following reduction rule.

**Reduction Rule 4** *If there are more than  $k^2 + 2k$  vertices in  $V$  that are all contained in some induced  $P_3$  in  $G$ , then output NO.*

**Lemma 10** *Reduction Rule 4 is correct.*

**Proof** Let  $I = (G = (V, E), G_c = (V, E_c), k, d)$  be an instance of a problem variant of DYNAMIC CLUSTER EDITING where Reduction Rule 2a / Reduction Rule 2b and Reduction Rule 3a / Reduction Rule 3b are not applicable. We show that if Reduction Rule 4 is applicable, then  $I$  is a no-instance.

Let  $R \subseteq V$  denote the set of vertices in  $V$  that are each contained in an induced  $P_3$  in  $G$ . Assume for the sake of contradiction that  $I$  is a yes-instance and that Reduction Rule 4 is applicable, that is,  $|R| > k^2 + 2k$ . Then there is a cluster graph  $G'$  with  $|E(G') \oplus E| \leq k$  that is a solution for  $I$ . For each  $\{u, v\} \subseteq R$ , let  $R_{uv}$  denote the set of vertices  $w$  such that  $G[\{u, v, w\}]$  is a  $P_3$ . Since the aforementioned rules are not applicable, we know that  $|R_{uv}| \leq k$ . We further know that  $R \subseteq \bigcup_{\{u,v\} \in E(G') \oplus E} (\{u, v\} \cup R_{uv})$ . It follows that  $|R| \leq k(k+2) = k^2 + 2k$ . This is a contradiction to the assumption that Reduction Rule 4 is applicable.  $\square$

In classic CLUSTER EDITING we can just remove all isolated cliques from the graph. This is not always possible in our setting because of the distance constraints to  $G_c$ . However, if there is a vertex set that forms an isolated clique both in  $G$  and  $G_c$ , then we can remove it since it has no influence on  $k$  or  $d$  in any problem variant. This is formalized in the next rule. We omit a formal correctness proof.

**Reduction Rule 5** *If there is a vertex set  $C \subseteq V$  that is an isolated clique in  $G$  and  $G_c$ , then remove all vertices in  $C$  from  $G$  and  $G_c$ .*

Now we introduce four new problem-specific reduction rules that will allow us to upper-bound the sizes of all remaining isolated cliques and their number in a function depending on  $k + d$ . The next rules deal with large isolated cliques and allow us to either remove them or to conclude that we face a no-instance. We first state the reduction rule for the matching-based distance problem variants and then turn to the edge-based distance variants.

**Reduction Rule 6a** (Matching-based distance) *If there is a vertex set  $C \subseteq V$  with  $|C| > k + 2d + 2$  that is an isolated clique in  $G$ , then*

- *if for each vertex set  $C' \subseteq V$  that is an isolated clique in  $G_c$  we have that  $|C \cap C'| \leq d$ , then answer NO,*
- *otherwise, if there is a vertex set  $C' \subseteq V$  that is an isolated clique in  $G_c$  and  $|C \cap C'| > d$ , then remove vertices in  $C$  from  $G$  and  $G_c$  and decrease  $d$  by  $|C \setminus C'|$ . Furthermore, if  $d \geq 0$ , then add a set  $C_d$  of  $k + d + 1$  fresh vertices to  $V$ . Add all edges between vertices in  $C_d$  to  $G$  and add all edges between vertices in  $C_d \cup (C' \setminus C)$  to  $G_c$  (if not already present).*

**Lemma 11** *Reduction Rule 6a is correct.*

**Proof** Let  $I = (G = (V, E), G_c = (V, E_c), k, d)$  be an instance of a problem variant of DYNAMIC CLUSTER EDITING that uses the matching-based distance and let  $I^*$  be the instance after applying the reduction rule. Note that if there is a vertex set  $C \subseteq V$  with  $|C| > k + 2d + 2 > k + 1$  that is an isolated clique in  $G$ , then we can neither divide this clique into smaller cliques nor can we add any vertex to this clique, since then more than  $k$  edge modifications would be necessary (or it is not allowed in the case of deletion or completion). This means that if  $I$  is a yes-instance and  $G'$  is the solution for  $I$ , then for all  $\{u, v\} \in E(G') \oplus E$  we have that  $\{u, v\} \cap C = \emptyset$  or, in other words,  $C$  is also an isolated clique in  $G'$ .

We first argue that if for each isolated clique  $C'$  in  $G_c$  we have that  $|C \cap C'| \leq d$ , then we face a no-instance. Assume for contradiction that  $I$  is a yes-instance and  $G'$  is the solution for  $I$ . Then we know that  $C$  is also an isolated clique in  $G'$  and no matter to which clique  $C'$  in  $G_c$  the clique  $C$  in  $G'$  is matched, we always have that  $|C \setminus C'| > d$  and hence the matching-based distance between  $G'$  and  $G_c$  is too large. This is a contradiction to the assumption that  $I$  is a yes-instance.

Now assume that there is an isolated clique  $C'$  in  $G_c$  with  $|C \cap C'| > d$ . We show that if  $I$  is a yes-instance, then  $I^*$  is a yes-instance. Let  $I$  be a yes-instance and let  $G'$  be a solution for  $I$ . Then we know that  $C$  is also an isolated clique in  $G'$ . If  $C$  in  $G'$  is not matched to  $C'$  in  $G_c$ , then the matching-based distance between  $G'$  and  $G_c$  is larger than  $d$ . Hence, we can assume that  $C$  in  $G'$  is matched to  $C'$  in  $G_c$ .

For the next argument we introduce the following terminology. If in an optimal solution an isolated clique  $C$  in  $G'$  is matched to an isolated clique  $C'$  in  $G_c$ , then we say that this match contributes  $|C \setminus C'| + |C' \setminus C|$  to the matching-based distance between  $G'$  and  $G_c$ .

Now we look at the instance  $I^* = (G^* = (V^*, E^*), G_c^* = (V^*, E_c^*), k^*, d^*)$  where the vertices in  $C$  are removed from  $G$  and  $G_c$ . The reduction rule further reduces  $d$  by  $|C \setminus C'|$  and introduces a new isolated clique  $C_d$  of size  $k + d + 1$  to  $G$ . In  $G_c$  we have that  $C_d \cup (C' \setminus C)$  is an isolated clique. We claim that  $G^{\star'} = (V^*, E^* \oplus (E(G') \oplus E))$  is a valid solution for  $I^*$ . First, note that  $G^{\star'}$  is a cluster graph. Since the removed clique and the added clique both have order larger than  $k$ , we can conclude that  $G^{\star'}$  is the graph that results from removing the clique  $C$  from  $G'$  and then adding the clique  $C_d$ . Concerning the matching-based distance, we can replace the match between  $C$  and  $C'$  in  $G'$  and  $G_c$  by the match between  $C_d$  and  $C_d \cup (C' \setminus C)$  in  $G^{\star'}$  and  $G_c^*$ , respectively. Note that the contribution of the match between  $C$  and  $C'$  in  $G'$  and  $G_c$ , respectively, minus  $|C \setminus C'|$  (the value by which  $d$  is decreased by the reduction rule) is the same as the contribution of the match between  $C_d$  and  $C_d \cup (C' \setminus C)$  in  $G^{\star'}$  and  $G_c^*$ , respectively. Hence, we can conclude that the matching-based distance between  $G'$  and  $G_c$  is the same as the matching-based distance between  $G^{\star'}$  and  $G_c^*$ . It follows that  $I^*$  is a yes-instance.

By a symmetric argument it follows that if  $I^*$  is a yes-instance, then  $I$  is a yes-instance.  $\square$

**Reduction Rule 6b** (Edge-based distance) *If there is a vertex set  $C \subseteq V$  with  $|C| > k + 1$  that is an isolated clique in  $G$ , then decrease  $d$  by  $|E_c| + \binom{|C|}{2} - 2|E(G_c[C])| - |E(G_c[V \setminus C])|$  and remove the vertices in  $C$  from  $G$  and  $G_c$ .*

**Lemma 12** *Reduction Rule 6b is correct.*

**Proof** Let  $I = (G = (V, E), G_c = (V, E_c), k, d)$  be an instance of a problem variant of DYNAMIC CLUSTER EDITING that uses the edge-based distance and let  $I^*$  be the instance after applying the reduction rule. Note that if there is a vertex set  $C \subseteq V$  with  $|C| > k + 1$  that is an isolated clique in  $G$ , then we can neither divide this clique into smaller cliques nor can we add any vertex to this clique, since then more than  $k$  edge modifications would be necessary (or it is not allowed in the case of deletion or completion). This means that if  $I$  is a yes-instance and  $G'$  is the solution for  $I$ , then for all  $\{u, v\} \in E(G') \oplus E$  we have that  $\{u, v\} \cap C = \emptyset$  or, in other words,  $C$  is also an isolated clique in  $G'$ . This implies that removing  $C$  from  $G'$  and  $G_c$  decreases  $d$  by the number of edges between vertices in  $C$  that are present in  $G'$  but not present in  $G_c$  plus the number of edges in  $G_c$  that have one endpoint in  $C$  and one endpoint in  $V \setminus C$  (note that no such edges are present in  $G'$ ). The number of edges between vertices in  $C$  that are present in  $G'$  is clearly  $\binom{|C|}{2}$  and the number of edges between vertices in  $C$  that are present in  $G_c$  is  $|E(G_c[C])|$ . The number of edges in  $G_c$  that have one endpoint in  $C$  and one endpoint in  $V \setminus C$  is the total number of edges in  $G_c$  minus the number of edges in  $G_c$  between vertices in  $C$  and the edges in  $G_c$  between vertices in  $V \setminus C$ . Hence, we get  $d = \binom{|C|}{2} - |E(G_c[C])| + |E_c| - (|E(G_c[C])| + |E(G_c[V \setminus C])|)$ , which yields the decrease conducted by the reduction rule. Note that this number is independent of  $G'$ . It follows that  $G'[V \setminus C]$  is a solution for  $I^*$ . Thus,  $I^*$  is a yes-instance.

By an analogous argument we get that if  $I^*$  is a yes-instance, then  $I$  is also a yes-instance.  $\square$

If none of the previous rules are applicable, then we know that there are no large cliques left in the graph. The next rules allow us to conclude that we face a no-instance if there are too many small cliques left.

**Reduction Rule 7a** (Matching-based distance) *If there are more than  $2k + d$  isolated cliques in  $G$ , then output NO.*

**Lemma 13** *Reduction Rule 7a is correct.*

**Proof** Let  $I = (G = (V, E), G_c = (V, E_c), k, d)$  be an instance of a problem variant of DYNAMIC CLUSTER EDITING that uses the matching-based distance. We show that if Reduction Rule 5 is not applicable and there are  $2k + d + 1$  isolated cliques  $C_1, C_2, \dots, C_{2k+d+1} \subseteq V$  in  $G$ , then  $I$  is a no-instance.

Assume for the sake of contradiction that  $I$  is a yes-instance. Then there is a cluster graph  $G'$  with  $|E(G') \oplus E| \leq k$  that is a solution for  $I$ . Since Reduction Rule 5 is not applicable we have that for isolated cliques  $C_i$  in  $G$  with  $1 \leq i \leq 2k + d + 1$  the vertex set  $C_i$  is not an isolated clique in  $G_c$ . Each edge modification in  $E(G') \oplus E$  when applied to  $G$  can reduce the number of isolated cliques in  $G$  that are not isolated cliques in  $G_c$  by at most 2. This happens when two isolated cliques of  $G$  are joined in  $G'$  and the union of their vertices is an isolated clique in  $G_c$ . It is easy to check that this is the best case. It follows that after at most  $k$  edge modifications,  $G'$  still has at least  $d + 1$  isolated cliques that are not isolated cliques in  $G_c$ . Thus the matching-based distance cannot be decreased to  $d$  which is a contradiction to the assumption that we face a yes-instance.  $\square$

**Reduction Rule 7b** (Edge-based distance) *If there are more than  $2(k + d)$  isolated cliques in  $G$ , then output NO.*

**Lemma 14** *Reduction Rule 7b is correct.*

**Proof** Let  $I = (G = (V, E), G_c = (V, E_c), k, d)$  be an instance of a problem variant of DYNAMIC CLUSTER EDITING that uses the edge-based distance. We show that if Reduction Rule 7b is not applicable and there are  $2(k + d) + 1$  isolated cliques  $C_1, C_2, \dots, C_{2k+d+1} \subseteq V$  in  $G$ , then  $I$  is a no-instance.

Let  $M = E \oplus E_c$ . Since Reduction Rule 7b is not applicable, we have that for isolated cliques  $C_i$  in  $G$  with  $1 \leq i \leq 2(k + d) + 1$  the vertex set  $C_i$  is not an isolated clique in  $G_c$ . It follows that for all  $C_i$  there is a vertex  $u \in C_i$  and a vertex  $v \in V$  such that  $\{u, v\} \in M$ . This implies that  $|M| > k + d$  and, hence, we face a no-instance.  $\square$

In the following we show that the rules we presented decrease the number of vertices of the instance to a number polynomial in  $k + d$ .

**Lemma 15** *Let  $(G = (V, E), G_c = (V, E_c), k, d)$  be an instance of any one of the considered problem variants of DYNAMIC CLUSTER EDITING that uses the matching-based distance. If none of the suitable data reduction rules applies, then  $|V| \in O(k^2 + d^2)$  and  $|E| \in O(k^3 + d^3)$ .*

**Proof** Let  $I = (G = (V, E), G_c = (V, E_c), k, d)$  be an instance of a problem variant of DYNAMIC CLUSTER EDITING that uses the matching-based distance.

Since Reduction Rule 4 is not applicable, we know that there are at most  $k^2 + 2k$  vertices in  $G$  that are not part of an isolated clique. It is also known that there are  $O(k^3)$  edges between those vertices[26]. Further, since Reduction Rule 7a is not applicable, we know that there are at most  $2k + d$  isolated cliques in  $G$ . Since Reduction Rule 6a is not applicable, we know that each isolated clique has size at most  $k + 2d + 2$ . This yields a maximum number of  $3k^2 + 2d^2 + 5dk + 2d + 6k \in O(k^2 + d^2)$  vertices and  $O(k^3 + d^3)$  edges.  $\square$

**Lemma 16** *Let  $(G = (V, E), G_c = (V, E_c), k, d)$  be an instance of any one of the considered problem variants of DYNAMIC CLUSTER EDITING that uses the edge-based distance. If none of the suitable data reduction rules applies, then  $|V| \in O(k^2 + k \cdot d)$  and  $|E| = O(k^3 + k^2 \cdot d)$ .*

**Proof** Let  $I = (G = (V, E), G_c = (V, E_c), k, d)$  be an instance of a problem variant of DYNAMIC CLUSTER EDITING that uses the edge-based distance.

Since Reduction Rule 4 is not applicable, we know that there are at most  $k^2 + 2k$  vertices in  $G$  that are not part of an isolated clique. It is also known that there are  $O(k^3)$  edges between those vertices[26]. Further, since Reduction Rule 7b is not applicable, we know that there are at most  $2(k + d)$  isolated cliques in  $G$ . Since Reduction Rule 6b is not applicable, we know that each isolated clique has size at most  $k + 1$ . This yields a maximum number of  $2dk + 2d + 3k^2 + 4k \in O(k^2 + k \cdot d)$  vertices and  $O(k^3 + k^2 \cdot d)$  edges.  $\square$

Finally, we can apply all data reduction rules exhaustively in  $O(|V|^3)$  time.

**Lemma 17** *Let  $(G = (V, E), G_c = (V, E_c), k, d)$  be an instance of any one of the considered problem variants of DYNAMIC CLUSTER EDITING. Then the respective reduction rules can be exhaustively applied in  $O(|V|^3)$  time.*

**Proof** We first exhaustively apply Reduction Rule 1, Reduction Rule 2a, Reduction Rule 2b, Reduction Rule 3a, Reduction Rule 3b, and Reduction Rule 4. These rules are well-known data reduction rules for classic CLUSTER EDITING and it is known that these rules can exhaustively be applied in  $O(|V|^3)$  time[26] if the graph is represented by an adjacency matrix.

It is easy to check that none of the remaining rules introduces new induced  $P_3$ s to  $G$ . Hence, we know that once these rules (except Reduction Rule 1) are exhaustively applied, then they will not be applicable after any of the other rules is applied.

From now on we assume that the graph  $G$  is represented in the following way. Adjacencies between vertices that are part of an induced  $P_3$  are represented in an adjacency matrix. We know that all other vertices are contained in isolated cliques. We store a list of cliques and also a map from vertices to the isolated clique they are contained in. We assume that  $G_c$  is represented in the same way. It is easy to check that this new representation can be computed in  $O(|V|^3)$  time from the adjacency matrix.

Using the new representation of  $G$  and  $G_c$ , we can apply Reduction Rule 5 exhaustively in  $O(|V|^2)$  time. We can check in  $O(|V|)$  time whether Reduction Rule 6a or Reduction Rule 6b is applicable and, if so, apply the rule in  $O(|V|^2)$  time. After each application of Reduction Rule 6a we exhaustively apply Reduction Rule 5. Since each application of Reduction Rule 6a or Reduction Rule 6b decreases the number of vertices in  $V$  by at least one, we can apply these rules exhaustively in  $O(|V|^3)$  time. Using our representation of  $G$ , we can apply Reduction Rule 7a and Reduction Rule 7b in  $O(|V|)$  time.

Altogether, we obtain an overall running time in  $O(|V|^3)$ .  $\square$



It is easy to see that Theorem 3 directly follows from Lemmas 15, 16, and 17. We remark that the number of edges that are not part of an isolated clique can be bounded by  $O(k^3)$ [26].

## 4.2 Fixed-Parameter Tractable Cases for Single Parameters

In this subsection, we show that several variants of DYNAMIC CLUSTER EDITING are fixed-parameter tractable with respect to either the budget  $k$  or the distance  $d$ .

**Theorem 4** *DCDELETION (EDGE DIST) is in FPT when parameterized by the budget  $k$ .*

All our FPT results in Sect. 4.2 are using the same approach: We reduce (in FPT time) the input to an instance of MULTI-CHOICE KNAPSACK (MCK), formally defined as follows.

MULTI-CHOICE KNAPSACK (MCK)

**Input:** A family of  $\ell$  mutually disjoint sets  $S_1, \dots, S_\ell$  of items, a weight  $w_{i,j}$  and a profit  $p_{i,j}$  for each item  $j \in S_i$ , and two integers  $W$  and  $P$ .

**Question:** Is it possible to select one item from each set  $S_i$  such that the total profit is at least  $P$  and the total weight is at most  $W$ ?

MCK is known to be solvable in pseudo-polynomial time by dynamic programming:

**Lemma 18** [32, Section 11.5] *MCK can be solved in  $O(W \cdot \sum_{i=1}^{\ell} |S_i|)$  time.*

As our approach is easier to explain with the edge-based distance, we start with this case and afterwards show how to extend it to the matching-based distance. As already exploited in our reductions showing NP-hardness (see Theorem 1), all variants of DYNAMIC CLUSTER EDITING carry some number-problem flavor. Our generic approach will underline this flavor: We will focus on cases where we can partition the vertex set of the input graph into parts such that we will neither add nor delete an edge between two parts. Moreover, we require that the parts are “easy” enough to list all Pareto-optimal (with respect to  $k$  and  $d$ ) solutions in FPT-time (this is usually achieved by some kernelization arguments). However, even with these strict requirements we cannot solve the parts independently from each other: The challenge is that we have to select for each part an appropriate Pareto-optimal solution. Finding a feasible combination of these part-individual solutions leads to a knapsack-type problem (in this case MCK). Indeed, this is common to all studied variants of DYNAMIC CLUSTER EDITING.

The details for our generic four-step approach (for edge-based distance) are given subsequently. In order to apply this approach on a concrete problem variant,

we have to show (using problem-specific arguments) how the requirements in the first two steps can be met.

1. When necessary, apply polynomial-time data reduction rules from Sect. 4.1.  
Partition the input graph  $G = (V, E)$  into different parts  $G_1, G_2, \dots, G_{\ell+1}$  for some  $\ell \leq |V|$  such that
  - in  $G$  there is no edge between the parts and
  - if there is a solution, then there exists a solution where no edge between two parts will be inserted or deleted.
2. Compute for each part  $G_i = (V_i, E_i)$ ,  $1 \leq i \leq \ell$ , a set  $S_i \subseteq \mathbb{N}^2$  encoding “cost” and “gain” of all “representative” solutions for  $G_i$ . The size of the set  $S_i$  has to be upper-bounded in a function  $f$  of the parameter  $p$ . (Here,  $p$  will be either  $k$  or  $d$ .)  
More precisely, select a family  $\mathcal{E}_i$  of  $f(p)$  edge sets such that for each edge set  $E'_i \subseteq \binom{V_i}{2}$  in  $\mathcal{E}_i$  the graph  $G'_i = (V_i, E'_i \oplus E_i)$  is a cluster graph achievable with the allowed number of modification operations ( $G'_i = (V_i, E_i \setminus E'_i)$  for edge deletions and  $G'_i = (V_i, E_i \cup E'_i)$  for edge insertions). For each such edge set  $E'_i$ , add to  $S_i$  a tuple containing the cost ( $= |E'_i|$ ) and “decrease” the distance from  $G_i$  to the target cluster graph  $G_c$ . More formally, for edge insertions add  $(|E'_i|, |E'_i \cap E_c| - |E'_i \setminus E_c|)$  to  $S_i$  or for edge deletions add  $(|E'_i|, |E'_i \setminus E_c| - |E'_i \cap E_c|)$  to  $S_i$ , where  $E_c$  is the edge set of  $G_c$ . Note that we allow  $E'_i = \emptyset$ , that is, if  $G_i$  is a cluster graph, then  $S_i$  contains the tuple  $(0, 0)$ .  
The set  $S_i$  has to fulfill the following property: If there is a solution, then there is a solution  $G'$  such that restricting  $G'$  to  $V_i$  yields a tuple in  $S_i$ . More precisely, we require that  $(|E(G'[V_i]) \oplus E_i|, |(E(G'[V_i]) \oplus E_i) \cap E_c| - |(E(G'[V_i]) \oplus E_i) \setminus E_c|) \in S_i$ .
3. Create an MCK instance  $I$  with  $W := k$ ,  $P := |E \oplus E_c| - d$ , and the sets  $S_1, S_2, \dots, S_\ell$  where the tuples in the sets correspond to the items with the first number in the tuple being its weight and the second number being its profit.
4. Return true if and only if  $I$  is a yes-instance.

Note that the requirement in Step 1 implies that a part is a collection of connected components in  $G$ . Furthermore, note that the part  $G_{\ell+1}$  will be ignored in the subsequent steps. Thus  $G_{\ell+1}$  contains all vertices which are not contained in an edge of the edge modification set. Observe that  $\ell \leq n$ . Hence, we have  $\sum_{i=1}^{\ell} |S_i| \in O(f(p) \cdot n)$ . (The parameter  $p$  will be either  $k$  or  $d$ .) Moreover, as  $k$  and  $d$  are smaller than  $n^2$ , it follows that  $W < n^2$  and thus, by Lemma 18, the MCK instance  $I$  created in Step 3 can be solved in  $f(p) \cdot n^3$  time in Step 4. This yields the following.

**Observation 3** *If the partition in Step 1 and the sets  $S_i$  in Step 2 can be computed in  $g(p) \cdot n^c$  time with  $|S_i| \leq f(p)$  for some functions  $f, g$  and constant  $c$ , then the above four-step approach runs in  $g(p) \cdot n^c + f(p) \cdot n^3$  time.*

Note that Step 1 and Step 2 are different for every problem variant we consider. There are, however, some similarities between the variants where only edge insertions are allowed.

### 4.2.1 Edge-Based Distance

We use the above mentioned approach to show that  $\text{DCDELETION (EDGE DIST)}$  and  $\text{DCCOMPLETION (EDGE DIST)}$  are both fixed-parameter tractable with respect to  $k$  and that  $\text{DCCOMPLETION (EDGE DIST)}$  is also fixed-parameter tractable with respect to  $d$ . Since Step 1 and Step 2 are easiest to explain for the edge-deletion variant, we start with  $\text{DCDELETION (EDGE DIST)}$ . Note that the requirements of Step 1 and Step 2 seem impossible to achieve in FPT-time when allowing edge insertions and deletions. Indeed, as shown in Theorem 2, the corresponding edge-edit variants are  $\text{W[1]}$ -hard with respect to the studied (single) parameters  $k$  and  $d$ , respectively.

**Lemma 19**  $\text{DCDELETION (EDGE DIST)}$  can be solved in  $O(4^{k^4} \cdot n^3)$  time and thus is in FPT when parameterized by the budget  $k$ .

**Proof** We first apply the known data reduction rules for  $\text{CLUSTER EDITING}$  (see discussion after Theorem 3). As a result, we end up with a graph where at most  $k^2 + 2k$  vertices are contained in an induced  $P_3$ ; all other vertices form a cluster graph with each clique containing at most  $k$  vertices. Denote by  $G$  the resulting graph.

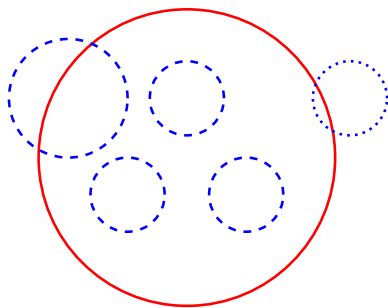
Now we apply our generic four-step approach. Thus we need to provide the details how to implement Step 1 and Step 2. We define the parts  $G_1, G_2, \dots, G_\ell, G_{\ell+1}$  of Step 1 as follows: The first part  $G_1 = (V_1, E_1)$  contains the graph induced by all vertices contained in a  $P_3$ . Each of the cliques in the cluster graph  $G[V \setminus V_1]$  forms another part  $G_i, 2 \leq i \leq \ell$ . Finally, set  $G_{\ell+1} = (\emptyset, \emptyset)$ , that is, we include all vertices in the subsequent steps of our generic approach. Clearly, each part contains less than  $2k^2$  vertices. Moreover, observe that there are no edges between the parts.

As to Step 2, we add, for every edge set  $E'_i \subseteq E_i$  with  $G'_i = (V_i, E_i \setminus E'_i)$  being a cluster graph, a tuple  $(|E'_i|, |E'_i \setminus E_c| - |E'_i \cap E_c|)$  to  $S_i$ . As this enumerates all possible solutions for  $G_i$ , the requirement in Step 2 is fulfilled.

As for the running time, since each part  $G_i$  contains less than  $2k^2$  vertices, we have for each edge set that  $|E_i| \leq 2k^4$  and hence  $|S_i| \leq |2^{E_i}| \leq 4^{k^4}$ . Together with Observation 3 we get the statement of the lemma.  $\square$

Next we show that  $\text{DCCOMPLETION (EDGE DIST)}$  is in FPT with respect to  $d$ . Before applying our generic approach, we make some observations. Since we can only insert edges in  $\text{DCCOMPLETION (EDGE DIST)}$ , we can make all vertices in a connected component pairwise adjacent.

**Observation 4** Let  $(G, G_c, k, d)$  be an instance of  $\text{DCCOMPLETION (EDGE DIST)}$  or of  $\text{DCCOMPLETION (MATCHING DIST)}$ . If  $G$  is not a cluster graph, then there is an equivalent instance  $(G^*, G_c, k^*, d)$  with  $G^*$  being a cluster graph and  $k^* < k$ .



**Fig. 10** Illustration of a possible combination of cliques in  $G$  (blue dashed or dotted circles). The red solid circle denotes a clique in  $G_c$ . Less than half of the vertices in the clique represented by the dotted (rightmost) circle are contained in the clique represented by the red solid circle. Lemma 21 states that the clique represented by the dotted circle will not be combined with a clique represented by a dashed circle (Color figure online)

**Proof** Since we are only allowed to insert edges and the solution graph is required to be a cluster graph, it follows that every connected component of  $G$  has to be made into a clique. Call the resulting cluster graph  $G^*$  and denote by  $k'$  the number of inserted edges. Clearly,  $(G, G_c, k, d)$  is a yes-instance if and only if  $(G^*, G_c, k - k', d)$  is a yes-instance.  $\square$

Observation 4 allows us in the following to assume that all connected components in  $G$  are cliques. The main difference between DCDELETION (EDGE DIST) and DCCOMPLETION (EDGE DIST) is that in the former we can partition cliques independently and thus the partition for Step 1 is straightforward. In the latter problem, Step 1 requires some preliminary observations. We subsequently show that for DCCOMPLETION (EDGE DIST) we can partition the graph  $G$  according to cliques in  $G_c$  such that no edges between the parts are added.

We assign every clique  $C$  in  $G$  a number that indicates the “best match” in  $G_c$ , that is, the clique  $D$  in  $G_c$ —if existing—that contains more than half of the vertices of  $C$ . More formally, let  $\mathcal{C}$  be the set of all cliques in  $G$  and  $\mathcal{D} = \{D_1, D_2, \dots, D_q\}$  be the cliques in  $G_c$ . We define a function  $T : \mathcal{C} \rightarrow \{0, 1, \dots, q\}$  mapping a clique  $C \in \mathcal{C}$  to a number between 0 and  $q$  as follows:

$$T(C) := \begin{cases} i & \text{if } \exists i : |C \cap D_i| > \frac{1}{2}|C|; \\ 0 & \text{otherwise.} \end{cases}$$

We say that we *merge* two cliques  $C_i$  and  $C_j$  when we add all edges between  $C_i$  and  $C_j$ . We show next that we can assume that we only merge cliques  $C_i$  and  $C_j$  with  $T(C_i) = T(C_j)$  (see Fig. 10 for an illustration).

**Definition 3** Let  $(G, G_c, k, d)$  be a yes-instance of DCCOMPLETION (EDGE DIST) such that  $G$  is a cluster graph. A solution  $G' = (V, E')$ ,  $E' \supseteq E$ , is called *standard* if for each clique  $C'$  in  $G'$  the following holds: If  $C_1 \neq C_2$  are two cliques in  $G$  and  $C_1, C_2 \subseteq C'$ , then  $T(C_1) = T(C_2) > 0$ .

We start with a technical lemma that essentially states that merging two cliques  $C_i$  and  $C_j$  with  $T(C_i) \neq T(C_j)$  or  $T(C_i) = 0$  will not decrease the edge-based distance.

**Lemma 20** *Let  $(G, G_c, k, d)$  be an instance of DCCOMPLETION (EDGE DIST), let  $C_0, C_1, \dots, C_r$  be isolated cliques in  $G$  with  $r \geq 1$ , and let  $E^* := \{\{u, v\} \mid u \in C_0, v \in C_1 \cup \dots \cup C_r\}$ . If  $T(C_0) = 0$  or if  $T(C_0) \neq T(C_i)$  for all  $1 \leq i \leq r$ , then  $|E^* \cap E(G_c)| \leq |E^* \setminus E(G_c)|$ .*

**Proof** We first show the statement for  $r = 1$ . Let  $\mathcal{D} = \{D_1, D_2, \dots, D_q\}$  be the cliques in  $G_c$ . Observe that  $|E^* \cap E(G_c)| = \sum_{i=1}^q |C_0 \cap D_i| \cdot |C_1 \cap D_i|$ . Furthermore,

$$|E^* \setminus E(G_c)| = \sum_{i=1}^q |C_0 \setminus D_i| \cdot |C_1 \cap D_i| = \sum_{i=1}^q |C_0 \cap D_i| \cdot |C_1 \setminus D_i|.$$

If  $T(C_0) = 0$ , then, by definition of  $T$ , we have  $|C_0 \setminus D_i| \geq |C_0 \cap D_i|$  for all  $1 \leq i \leq q$ . Thus,  $|E^* \cap E(G_c)| \leq |E^* \setminus E(G_c)|$ . By symmetry, this follows also from  $T(C_1) = 0$ .

It remains to consider the case that  $T(C_0) = j_0 \neq j_1 = T(C_1)$  for some  $j_0, j_1 \in \{1, 2, \dots, q\}$ . By the definition of  $T$ , we have  $|C_0 \setminus D_i| \geq |C_0 \cap D_i|$  for all  $i \in \{1, \dots, q\} \setminus \{j_0, j_1\}$ . Thus,

$$\sum_{i \in \{1, \dots, q\} \setminus \{j_0, j_1\}} |C_0 \cap D_i| \cdot |C_1 \cap D_i| \leq \sum_{i \in \{1, \dots, q\} \setminus \{j_0, j_1\}} |C_0 \setminus D_i| \cdot |C_1 \cap D_i|.$$

It remains to show that

$$\sum_{i \in \{j_0, j_1\}} |C_0 \cap D_i| \cdot |C_1 \cap D_i| \leq \sum_{i \in \{j_0, j_1\}} |C_0 \setminus D_i| \cdot |C_1 \cap D_i|. \quad (4)$$

To this end, observe that

$$C_0 \cap D_{j_0} \subseteq C_0 \setminus D_{j_1} \quad \text{and} \quad C_0 \cap D_{j_1} \subseteq C_0 \setminus D_{j_0}. \quad (5)$$

Furthermore,

$$|C_0 \setminus D_{j_1}| > |C_0 \setminus D_{j_0}| \quad \text{and} \quad |C_1 \setminus D_{j_1}| < |C_1 \setminus D_{j_0}|. \quad (6)$$

From the above, we can deduce that

$$\begin{aligned} & |C_0 \cap D_{j_0}| \cdot |C_1 \cap D_{j_0}| + |C_0 \cap D_{j_1}| \cdot |C_1 \cap D_{j_1}| \\ & \stackrel{(5)}{\leq} |C_0 \setminus D_{j_1}| \cdot |C_1 \cap D_{j_0}| + |C_0 \setminus D_{j_0}| \cdot |C_1 \cap D_{j_1}| \\ & \stackrel{(6)}{\leq} |C_0 \setminus D_{j_0}| \cdot |C_1 \cap D_{j_0}| + |C_0 \setminus D_{j_1}| \cdot |C_1 \cap D_{j_1}|. \end{aligned}$$

This completes the proof for the case  $T(C_0) \neq T(C_1)$ .

The proof for the case  $r > 1$  is now straightforward. We have

$$|E^* \cap E(G_c)| = \sum_{j=1}^r \sum_{i=1}^q |C_0 \cap D_i| \cdot |C_j \cap D_i|,$$

$$|E^* \setminus E(G_c)| = \sum_{j=1}^r \sum_{i=1}^q |C_0 \setminus D_i| \cdot |C_j \cap D_i|.$$

Using the above argument for  $r = 1$  for every  $C_j$ ,  $1 \leq j \leq r$ , we have that

$$\sum_{i=1}^q |C_0 \cap D_i| \cdot |C_j \cap D_i| \leq \sum_{i=1}^q |C_0 \setminus D_i| \cdot |C_j \cap D_i|.$$

Thus,  $|E^* \cap E(G_c)| \leq |E^* \setminus E(G_c)|$ .  $\square$

**Lemma 21** *Let  $(G, G_c, k, d)$  be a yes-instance of DCCOMPLETION (EDGE DIST). Then there exists a standard solution  $G'$  for  $(G, G_c, k, d)$ .*

**Proof** Let  $G'$  be a solution for  $(G, G_c, k, d)$ . Assume that  $G'$  is not a *standard* solution. Let  $C'$  be a clique in  $G'$  which is not standard. (A clique  $C'$  is called a *standard* clique if either it only contains one clique from  $\mathcal{C}$  or there is a number  $i$  with  $1 \leq i \leq q$  such that all cliques from  $\mathcal{C}$  contained in  $C'$  are assigned the value  $i$  under the function  $T$ .) Next we show that we can modify  $G'$  to get a *standard* solution by showing that each non-standard clique can be divided into some standard cliques and the resulting cluster graph is still a solution. We consider two cases distinguishing whether or not  $C'$  contains a clique  $C$  from  $G$  with  $T(C) = 0$ .

If  $C'$  contains a clique  $C$  from  $G$  with  $T(C) = 0$ , then splitting  $C$  from  $C'$  gives a cheaper solution that, by Lemma 20, also has distance at most  $d$  from  $G_c$ .

Now consider the case that there is no clique  $C$  from  $G$  with  $T(C) = 0$  contained in  $C'$ . Let  $S = \{C_1, C_2, \dots, C_{r'}, C_{r'+1}, \dots, C_r\}$  be the set of cliques from  $G$  contained in  $C'$  with  $C = C_1$ . Let  $S_1 = \{C_1, C_2, \dots, C_{r'}\}$  be the set that contains all cliques  $C_i \in S$  with  $T(C_i) = T(C_1)$  and  $S_2 = S \setminus S_1$ . We divide  $C'$  into two cliques  $C_{S_1}$  and  $C_{S_2}$  and get a new cluster graph  $G^*$ , where  $C_{S_1}$  contains all vertices in  $C_1, C_2, \dots, C_{r'}$  and  $C_{S_2}$  contains all vertices in  $C_{r'+1}, C_{r'+2}, \dots, C_r$ . We show that  $G^*$  is also a solution.

Let  $E_{\Delta}^i$  be the set of edges between vertices in  $C_{S_1}$  and vertices in  $C_i$  for each  $r' + 1 \leq i \leq r$ . Let  $E_{\Delta} = \bigcup_{r'+1 \leq i \leq r} E_{\Delta}^i$ . Then to get  $G^*$  from  $G'$ , we need to delete all edges in  $E_{\Delta}$ . Now we split edges in  $E_{\Delta}^i$  into two parts. Let  $E_{\Delta}^i = E_+^i \uplus E_-^i$ , where  $E_+^i = E_{\Delta}^i \cap (E(G) \oplus E(G_c))$  and  $E_-^i = E_{\Delta}^i \cap \left( \binom{V}{2} \setminus (E(G) \oplus E(G_c)) \right)$ . Since  $T(C_i) \neq T(C_1)$  for any  $r' + 1 \leq i \leq r$ , it follows from Lemma 20 that  $|E_+^i| \leq |E_-^i|$ . Then for the distance upper bound, we have that

$$\begin{aligned}
 |E(G^*) \oplus E(G_c)| &= |E(G') \oplus E(G_c)| - \sum_{r'+1 \leq i \leq r} |E_+^i| + \sum_{r'+1 \leq i \leq r} |E_-^i| \\
 &= |E(G') \oplus E(G_c)| - \sum_{r'+1 \leq i \leq r} (|E_+^i| - |E_-^i|) \\
 &\leq |E(G') \oplus E(G_c)| \\
 &\leq d.
 \end{aligned}$$

For the modification budget, we have that

$$|E(G) \oplus E(G^*)| = |E(G) \oplus E(G')| - |E_\Delta| \leq k.$$

Hence,  $G^*$  is also a solution. We can continue to divide clique  $C_{S_2}$  in  $G^*$  in the same way such that every sub-clique is standard. Thus, for every non-standard clique in  $G'$ , we can divide it into several standard cliques and the new cluster graph is still a solution.  $\square$

Note that for a clique  $C_i$  in  $G$  which has exactly half of its vertices from one clique in  $G_c$  and the remaining half of its vertices from another clique in  $G_c$ , no clique in  $G$  has the same type as it. Lemma 21 allows us to focus on standard solutions. This allows us to show our next fixed-parameter tractability result.

**Lemma 22** *DCCOMPLETION (EDGE DIST) can be solved in  $O(d^{d+1} \cdot n^3)$  time and thus is in FPT when parameterized by the distance  $d$ .*

**Proof** We apply our generic four-step approach and thus need to provide the details how to implement Step 1 and Step 2.

By Observation 4, we can assume that our input graph is a cluster graph. Furthermore, exhaustively apply Reduction Rule 6b to delete too big cliques and denote by  $G$  the resulting cluster graph. Let  $\mathcal{C}$  be the set of all cliques in  $G$  and  $\mathcal{D} := \{D_1, D_2, \dots, D_q\}$  be the set of all cliques in  $G_c$ . We partition  $G$  into  $q + 1$  groups  $G_1, G_2, \dots, G_q, G_{q+1}$  with  $G_i = G[V_i]$ , where  $V_i := \{C \in \mathcal{C} \mid T(C) = i\}$  for  $1 \leq i \leq q$  and  $V_{q+1} := \{C \in \mathcal{C} \mid T(C) = 0\}$ . So  $G_{q+1}$  contains all cliques with value 0 under the function  $T$ . According to Lemma 21, if there is a solution, then there is a solution only combining cliques within every group  $G_i$  for  $1 \leq i \leq q$ . This shows that with  $\ell = q$  the requirements of Step 1 of our generic approach are met.

Next we describe Step 2, that is, for every part  $G_i$ , we show how to compute a set  $S_i$  corresponding to all “representative” solutions. To this end, we distinguish two cases:  $G_i$  contains at most  $d + 1$  cliques or at least  $d + 2$  cliques. If  $G_i$  contains at most  $d + 1$  cliques, then we can brute-force all possibilities to partition the cliques and merge the cliques in each partition. There are less than  $(d + 1)^{d+1}$  possibilities to do so and for each possibility we add to  $S_i$  a tuple representing the cost and gain of making all cliques in  $S$  into a clique.

If  $G_i$  contains at least  $d + 2$  cliques, then we show that we need to merge all cliques in  $G_i$ : If not all cliques are merged into one clique, then we have a solution with two parts without any edge between the two parts (each part can be a single clique or a cluster graph). Let  $p$  and  $q$  be the number of vertices in the two

parts that are also in  $D_i$ . Since there are at least  $d + 2$  cliques, each containing at least one vertex from  $D_i$ , it follows that  $p + q \geq d + 2$ ,  $p \geq 1$ , and  $q \geq 1$ . Thus, at least  $p \cdot q \geq d + 1$  edges in  $D_i$  (and thus in  $G_c$ ) are not in our solution, a contradiction to the fact that the solution needs to have a distance of at most  $d$  to  $G_c$ . Hence, we only need to add one tuple to  $S_i$  encoding the cost and gain of making  $G_i$  into one clique. Together with Observation 3 we get the statement of the lemma.  $\square$

## 4.2.2 Matching-Based Distance

We next discuss how to adjust our generic four-step approach for DCCOMPLETION (MATCHING DIST). The main difference to the edge-based distance variants is an additional search tree of size  $O(d^{d+2})$  in the beginning. Each leaf of the search tree then corresponds to a simplified instance where we have additional knowledge on the matching defining the distance of a solution to  $G_c$ . With this additional knowledge, we can apply our generic four-step approach in each leaf, yielding the following.

**Lemma 23** DCCOMPLETION (MATCHING DIST) can be solved in  $O(d^{d+2} \cdot n^3)$  time and thus is in FPT when parameterized by the distance  $d$ .

**Proof** We apply our generic four-step approach and thus need to provide the details how to implement Step 1 and Step 2.

We can assume that our input graph is a cluster graph. Let  $\mathcal{C}$  be the set of all cliques in  $G$  and let  $\mathcal{D} := \{D_1, D_2, \dots, D_q\}$  be the set of all cliques in  $G_c$ . Then we partition all cliques in  $\mathcal{C}$  into two classes  $\mathcal{C}_1$  and  $\mathcal{C}_2$ , where every clique in  $\mathcal{C}_1$  has the property that all its vertices are contained in one clique in  $\mathcal{D}$  and every clique in  $\mathcal{C}_2$  contains vertices from at least two different cliques in  $\mathcal{D}$ . Observe that  $|\mathcal{C}_2| \leq d$  as otherwise the input is a no-instance. Similarly, every clique in  $\mathcal{C}_2$  contains vertices from at most  $d + 1$  different cliques in  $\mathcal{D}$  as otherwise the input is a no-instance.

This allows us to do the following branching step. For each clique in  $\mathcal{C}_2$  we try out all “meaningful” possibilities to match it to a clique in  $\mathcal{D}$ , where “meaningful” means that the cliques in  $\mathcal{C}_2$  and  $\mathcal{D}$  should share some vertices or we decide to not match the cliques of  $\mathcal{C}_2$  to any clique in  $\mathcal{D}$ . For each clique this gives us  $d + 2$  possibilities and hence we have at most  $d^{d+2}$  different cases each of which defines a mapping  $M : \mathcal{C}_2 \rightarrow \mathcal{D} \cup \{\emptyset\}$  that maps a clique in  $\mathcal{C}_2$  to the clique in  $\mathcal{D}$  it is matched to.

Given the mapping  $M$  from cliques in  $\mathcal{C}_2$  to cliques in  $\mathcal{D}$  or  $\emptyset$ , we partition  $G$  into  $q + 1$  groups  $G_1, G_2, \dots, G_q, G_{q+1}$  with  $G_i = G[V_i]$ , where  $V_i := \{C \in \mathcal{C}_1 \mid M(C) \subseteq D_i\} \cup \{C \in \mathcal{C}_2 \mid M(C) = D_i\}$  and  $V_{q+1} := \{C \in \mathcal{C}_2 \mid M(C) = \emptyset\}$ .

If there is a solution with a matching that uses the matches given by  $M$ , then there is a solution only combining cliques within every group  $G_i$ ,  $1 \leq i \leq q$ , since all cliques in  $G_i$  that are not matched by  $M$  are completely contained in  $D_i$  and hence would not be merged with cliques in  $G_j$  for some  $i \neq j$ . This shows that with  $\ell = q$  the requirements of Step 1 of our generic approach are met.



Next we describe Step 2, that is, for every part  $G_i$ , we show how to compute a set  $S_i$  corresponding to all “representative” solutions. Note that all except for at most  $d$  cliques from  $G_i$  need to be merged into one clique that is then matched with  $D_i$ , otherwise the matching distance would be too large. For each clique in  $G_i$  that is not completely contained in  $D_i$  we already know that it is matched to  $D_i$ , hence we need to merge all cliques of this kind to one clique  $C_i^*$ . Each clique in  $G_i$  that is completely contained in  $D_i$  and has size at least  $d + 1$  also needs to be merged to  $C_i^*$ , otherwise the matching distance would be too large. For all cliques of  $G_i$  that are completely contained in  $D_i$  with size  $x$  for some  $1 \leq x \leq d$ , we merge all but  $d$  cliques to  $C_i^*$ . This leaves us with one big clique  $C_i^*$  and  $d^2$  cliques of size at most  $d$  each. Now we can brute-force all possibilities to merge some of the remaining cliques to  $C_i^*$ . There are less than  $d^d$  possibilities to do so and for each possibility we add to  $S_i$  a tuple representing the cost and gain of merging the cliques according to the partition. Together with Observation 3 we get the statement of the lemma.  $\square$

We now have shown all fixed-parameter tractability results stated in Theorem 4.

## 5 Conclusion

Our work provides a first thorough (parameterized) analysis of DYNAMIC CLUSTER EDITING, addressing a natural dynamic setting for graph-based data clustering. We deliver both (parameterized) tractability and intractability results. Our positive algorithmic results (fixed-parameter tractability and polynomial kernelization) are mainly of classification nature. Hence, to get practically useful algorithms, one needs to further improve our running times, a challenge for future research.

A key difference between DYNAMIC CLUSTER EDITING and static CLUSTER EDITING is that all six variants of DYNAMIC CLUSTER EDITING remain NP-hard even when the input graph is already a cluster graph (see Theorem 1). Moreover, DYNAMIC CLUSTER EDITING (both matching- and edge-based distance) is W[1]-hard with respect to the budget  $k$  (see Theorem 2) whereas CLUSTER EDITING is fixed-parameter tractable with respect to  $k$ . An obvious approach to solve DYNAMIC CLUSTER EDITING is to compute (almost) all cluster graphs achievable with at most  $k$  edge modifications, then pick from this set of cluster graphs one at distance at most  $d$  to the target cluster graph. However, listing these cluster graphs is computationally expensive. Indeed, our W[1]-hardness results indicate that we might not do much better than using this simple approach.

Note that we refrain from stating results on potential containment in W[1] for all our W[1]-hardness results. There are two reasons for that. First, for all our W[1]-hard cases, it is easy to show containment in XP, that is, polynomial-time solvability for constant parameter values. For an algorithmic point of view, there is not much further gain by knowing containment in W[1]. Second, we indeed studied containment in W[1] with quite some technical but not particularly inspiring expenditure (using machine models for W[1][22]). We showed W[1]-completeness for the problems DCE (EDGE DIST) and DCDELETION (EDGE DIST) parameterized by  $d$ . For all other problems, however, we failed to show this with reasonable effort.

We mention in passing that our results partly transfer to the “compromise clustering” problem, where, given two input graphs, one wants to find a “compromise” cluster graph that is close enough (in terms of edge-based distance) to both input graphs. It is easy to see that our fixed-parameter tractability results carry over if one of these two input graphs is already a cluster graph. A direction for future research is to examine whether our results can also be adapted to the case where both input graphs are arbitrary. Furthermore, we left open the parameterized complexity of DYNAMIC CLUSTER EDITING (deletion variant and completion variant) with matching-based distance as well as DYNAMIC CLUSTER EDITING (completion variant) with edge-based distance when parameterized by the budget  $k$ , see Table 1 in Sect. 1. Moreover, the existence of polynomial-size problem kernels for our fixed-parameter tractable cases for single parameters (budget  $k$  or distance  $d$ ) is open.

**Acknowledgements** Open Access funding provided by Projekt DEAL. We are grateful to two anonymous Algorithmica reviewers whose constructive feedback helped us to significantly improve our presentation. JL was partially supported by CAS-DAAD Joint Fellowship Program for Doctoral Students of UCAS and partially supported by the DFG, Project AFFA (BR 5207/1). Work done while with TU Berlin. HM was supported by the DFG, Project MATE (NI 369/17).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Abu-Khzam, F.N.: On the complexity of multi-parameterized cluster editing. *J. Discrete Algorithms* **45**, 26–34 (2017)
2. Abu-Khzam, F.N., Egan, J., Fellows, M.R., Rosamond, F.A., Shaw, P.: On the parameterized complexity of dynamic problems. *Theor. Comput. Sci.* **607**, 426–434 (2015)
3. Abu-Khzam, F.N., Cai, S., Egan, J., Shaw, P., Wang, K.: Turbo-charging dominating set with an FPT subroutine: further improvements and experimental analysis. In: *Proceedings of the 14th Annual Conference on Theory and Applications of Models of Computation, TAMC 2017*, volume 10185 of LNCS, pp. 59–70. Springer (2017)
4. Abu-Khzam, F.N., Egan, J., Gaspers, S., Shaw, A., Shaw, P.: Cluster editing with vertex splitting. In: *Proceedings of the 5th International Symposium on Combinatorial Optimization, ISCO 2018*, volume 10856 of LNCS, pp. 1–13. Springer (2018)
5. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. *Mach. Learn.* **56**, 89–113 (2004)
6. Ben-Dor, A., Shamir, R., Yakhini, Z.: Clustering gene expression patterns. *J. Comput. Biol.* **6**(3–4), 281–297 (1999)
7. Böckenhauer, H., Hromkovič, J., Mömke, T., Widmayer, P.: On the hardness of reoptimization. In: *Proceedings of the 34th Conference on Theory and Practice of Computer Science, SOFSEM 2008*, volume 4910 of LNCS, pp. 50–65. Springer (2008)

8. Böckenhauer, H., Burjons, E., Raszyk, M., Rossmanith, P.: Reoptimization of parameterized problems. CoRR, abs/1809.10578 (2018)
9. Böcker, S., Baumbach, J.: Cluster Editing. In: Proceedings of the 9th Conference on Computability in Europe, CiE 2013, volume 7921 of LNCS, pp. 33–44. Springer (2013)
10. Cai, L.: Parameterized complexity of cardinality constrained optimization problems. *Comput. J.* **51**(1), 102–121 (2008)
11. Cao, Y., Chen, J.: Cluster editing: kernelization based on edge cuts. *Algorithmica* **64**(1), 152–169 (2012)
12. Charikar, M., Chekuri, C., Feder, T., Motwani, R.: Incremental clustering and dynamic information retrieval. *SIAM J. Comput.* **33**(6), 1417–1440 (2004)
13. Chen, J., Meng, J.: A  $2k$  kernel for the cluster editing problem. *J. Comput. Syst. Sci.* **78**(1), 211–220 (2012)
14. Chen, J., Molter, H., Sorge, M., Suchý, O.: Cluster editing in multi-layer and temporal graphs. In: Proceedings of the 29th International Symposium on Algorithms and Computation (ISAAC '18), volume 123 of LIPIcs, pp. 24:1–24:13. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2018)
15. Cygan, M., Fomin, F.V., Kowalik, Ł., Lokshantov, D., Marx, D., Pilipczuk, M., Pilipczuk, M., Saurabh, S.: *Parameterized Algorithms*. Springer, Berlin (2015)
16. Dey, T.K., Rossi, A., Sidiropoulos, A.: Temporal clustering. In: Proceedings of the 25th Annual European Symposium on Algorithms, ESA 2017, Volume 87 of LIPIcs, pp. 34:1–34:14. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2017)
17. Dong, X., Frossard, P., Vandergheynst, P., Nefedov, N.: Clustering with multi-layer graphs: a spectral perspective. *IEEE Trans. Signal Process.* **60**(11), 5820–5831 (2012)
18. Downey, R.G., Fellows, M.R.: *Fundamentals of Parameterized Complexity*. Springer, Berlin (2013)
19. Downey, R.G., Egan, J., Fellows, M.R., Rosamond, F.A., Shaw, P.: Dynamic dominating set and turbo-charging greedy heuristics. *Tsinghua Sci. Technol.* **19**(4), 329–337 (2014)
20. Fellows, M.R., Hermelin, D., Rosamond, F., Vialette, S.: On the parameterized complexity of multiple-interval graph problems. *Theor. Comput. Sci.* **410**(1), 53–61 (2009)
21. Fellows, M.R., Fomin, F.V., Lokshantov, D., Rosamond, F.A., Saurabh, S., Villanger, Y.: Local search: Is brute-force avoidable? *J. Comput. Syst. Sci.* **78**(3), 707–719 (2012)
22. Flum, J., Grohe, M.: *Parameterized Complexity Theory*, Volume XIV of Texts in Theoretical Computer Science, An EATCS Series. Springer, Berlin (2006)
23. Fomin, F.V., Kratsch, S., Pilipczuk, M., Pilipczuk, M., Villanger, Y.: Tight bounds for parameterized complexity of cluster editing with a small number of clusters. *J. Comput. Syst. Sci.* **80**(7), 1430–1447 (2014)
24. Garey, M.R., Johnson, D.S.: *Computers and Intractability—A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, New York (1979)
25. Gaspers, S., Kim, E.J., Ordyniak, S., Saurabh, S., Szeider, S.: Don't be strict in local search! In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 486–492. AAAI Press (2012)
26. Gramm, J., Guo, J., Hüffner, F., Niedermeier, R.: Graph-modeled data clustering: exact algorithms for clique generation. *Theory Comput. Syst.* **38**(4), 373–392 (2005)
27. Guo, J.: A more effective linear kernelization for cluster editing. *Theor. Comput. Sci.* **410**(8–10), 718–726 (2009)
28. Guo, J., Hartung, S., Niedermeier, R., Suchý, O.: The parameterized complexity of local search for TSP, more refined. *Algorithmica* **67**(1), 89–110 (2013)
29. Hartung, S., Hoos, H.H.: Programming by optimisation meets parameterised algorithmics: a case study for cluster editing. In: Proceedings of the 9th International Conference on Learning and Intelligent Optimization, LION 2015, Volume 8994 of LNCS, pp. 43–58. Springer (2015)
30. Hartung, S., Niedermeier, R.: Incremental list coloring of graphs, parameterized by conservation. *Theor. Comput. Sci.* **494**, 86–98 (2013)
31. Karp, R.M.: Reducibility among combinatorial problems. In: Miller, R.E., Thatcher, J.W., Bohlinger, J.D. (eds.) *Complexity of Computer Computations*, pp. 85–103. Springer, Boston (1972)
32. Kellerer, H., Pferschy, U., Pisinger, D.: *Knapsack Problems*. Springer, Boston (2004)
33. Komusiewicz, C., Uhlmann, J.: Cluster editing with locally bounded modifications. *Discrete Appl. Math.* **160**(15), 2259–2270 (2012)
34. Krithika, R., Sahu, A., Tale, P.: Dynamic parameterized problems. *Algorithmica* **80**(9), 2637–2655 (2018)

35. Luo, J., Molter, H., Nichterlein, A., Niedermeier, R.: Parameterized dynamic cluster editing. In: Proceedings of the 38th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS '18, Volume 122 of LIPIcs, pp. 46:1–46:15 (2018)
36. Marx, D., Schlotter, I.: Parameterized complexity and local search approaches for the stable marriage problem with ties. *Algorithmica* **58**(1), 170–187 (2010)
37. Meilă, M.: Comparing clusterings: an axiomatic view. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 577–584. ACM (2005)
38. Meilă, M.: Local equivalences of distances between clusterings—a geometric perspective. *Mach. Learn.* **86**(3), 369–389 (2012)
39. Niedermeier, R.: *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, Oxford (2006)
40. Schieber, B., Shachnai, H., Tamir, G., Tamir, T.: A theory and algorithms for combinatorial reoptimization. *Algorithmica* **80**(2), 576–607 (2018)
41. Shamir, R., Sharan, R., Tsur, D.: Cluster graph modification problems. *Discrete Appl. Math.* **144**(1–2), 173–182 (2004)
42. Tang, W., Lu, Z., Dhillon, I.S.: Clustering with multiple graphs. In: Proceedings of the Ninth IEEE International Conference on Data Mining, ICDM 2009, pp. 1016–1021 (2009)
43. Tantipathananandh, C., Berger-Wolf, T.Y.: Finding communities in dynamic social networks. In: Proceedings of the IEEE 11th International Conference on Data Mining, ICDM 2011, pp. 1236–1241 (2011)
44. Tantipathananandh, C., Berger-Wolf, T.Y., Kempe, D.: A framework for community identification in dynamic social networks. In: Proceedings of the 13th SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2007, pp. 717–726. ACM (2007)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Junjie Luo<sup>1,2</sup> · Hendrik Molter<sup>3</sup> · André Nichterlein<sup>3</sup> · Rolf Niedermeier<sup>3</sup>

Junjie Luo  
junjie.luo@tu-berlin.de; luojunjie@amss.ac.cn

André Nichterlein  
andre.nichterlein@tu-berlin.de

Rolf Niedermeier  
rolf.niedermeier@tu-berlin.de

<sup>1</sup> Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

<sup>2</sup> School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

<sup>3</sup> Algorithmics and Computational Complexity, Faculty IV, TU Berlin, Berlin, Germany