

Clustering with Locally Bounded Ignorance

Jaroslav Garvardt 

Institute of Computer Science, University of Jena, Germany

Christian Komusiewicz 

Institute of Computer Science, University of Jena, Germany

Abstract

In CORRELATION CLUSTERING, the input is a graph $G = (V, E)$ with weight function $\omega : \binom{V}{2} \rightarrow \mathbb{Z}_{\geq 0}$ and the task is to partition the vertex set into clusters such that the total weight of edges between clusters and missing edges inside clusters is minimized. Due to close connections between CORRELATION CLUSTERING and EDGE MULTICUT, deciding whether there is a partition with total cost at most k is FPT with respect to k but a polynomial kernel is presumably impossible. We study the influence of the structure of the fuzzy edge graph, that is, the graph induced by the weight-0 edges, on the problem complexity. We show in particular that CORRELATION CLUSTERING admits a polynomial problem kernel when parameterized by $k + d$, where d is the degeneracy of the fuzzy edge graph, and when parameterized by $k + c$, where c is the closure of the fuzzy edge graph. We complement these positive results by showing hardness for several settings where the graph induced by the edges and nonedges has very restricted structure.

2012 ACM Subject Classification Theory of computation $\rightarrow$ Parameterized complexity and exact algorithms; Theory of computation $\rightarrow$ Graph algorithms analysis

Keywords and phrases Graph-based Data Clustering, Cluster Editing, Kernelization

Digital Object Identifier 10.4230/LIPIcs.WG.2026.17

Related Version *Full Version*: <https://doi.org/10.48550/arXiv.2605.13917>

Funding *Jaroslav Garvardt*: Supported by the Carl Zeiss Foundation, Germany, within the project “Interactive Inference”.

1 Introduction

Clustering is the task of grouping similar objects together. A crucial tool for formalizing this task are similarity graphs where the objects are vertices and we draw an edge between two vertices if and only if these objects are similar. Clustering now corresponds to computing a partition of the vertex sets into clusters such that there are few edges between different clusters and many edges inside the clusters. There are different concrete formulations of this problem, with CLUSTER EDITING being one of the most prominent ones [3]. Here, we aim to minimize the total number of misclassified vertex pairs which are defined as edges between different clusters and missing edges inside clusters. This can also be viewed as minimizing the number k of edge insertions and deletions to transform the input graph into a cluster graph, that is, a graph whose connected components are all cliques. These cliques correspond to the output clusters and the number k of necessary modifications is exactly the number of misclassifications. CLUSTER EDITING is NP-hard [26, 29] and cannot be solved in subexponential time under the ETH [25]. This hardness motivates the study from the viewpoint of parameterized algorithms with most studies focusing on FPT-algorithms or kernelizations for the parameter k and related parameters [2, 15, 16, 24, 25, 30].

In many applications, the similarity data is more precise than merely a binary predicate that distinguishes between similar and dissimilar vertex pairs. In particular, the edges and nonedges may be assigned a strictly positive integer weight, with large weights corresponding



© Jaroslav Garvardt and Christian Komusiewicz;
licensed under Creative Commons License CC-BY 4.0

52nd International Workshop on Graph-Theoretic Concepts in Computer Science (WG 2026).

Editors: Jan Goedgebeur and Paweł Rzażewski; Article No. 17; pp. 17:1–17:17

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

to a strong evidence for similarity and dissimilarity, respectively. This problem, known as WEIGHTED CLUSTER EDITING, also has FPT-algorithms and kernelizations for the parameter k [4, 7] but it is harder than CLUSTER EDITING when the input is restricted to special graph classes [17, 14].

While being more realistic, WEIGHTED CLUSTER EDITING still assumes full knowledge in the sense that each vertex pair is either deemed to be similar or dissimilar. To address that for some vertex pairs there may be no information, one may assign weight 0 to them, which means that they can be ignored when computing the number of misclassifications or alternatively, that they can be modified for free. This most general version of the problem is known as CORRELATION CLUSTERING [1, 11].

The additional freedom of ignoring some vertex pairs comes at an algorithmic cost, however: CORRELATION CLUSTERING is equivalent to EDGE MULTICUT in the sense that there are easy reductions from CORRELATION CLUSTERING to EDGE MULTICUT and vice versa that preserve the cost k of the optimal solution [11]. With the parameterized complexity status of EDGE MULTICUT being a major open problem at the time, this prompted Bodlaender et al. [5] to investigate a combined parameter that takes the structure of the weight-0 edges, called *fuzzy* edges, into account. The idea was to consider the graph with vertex set V containing only the fuzzy edges and to use the size of a vertex cover r of this graph as an additional parameter. Intuitively, the size of the vertex cover is a measure of the amount of missing information. Bodlaender et al. showed that CORRELATION CLUSTERING admits a kernel with $\mathcal{O}(k^2 + r)$ vertices [5]. Subsequently, it was shown that EDGE MULTICUT and with it CORRELATION CLUSTERING are fixed-parameter tractable with respect to k [6, 27], which implied that the additional parameter r was not necessary to obtain an FPT-algorithm or a problem kernel. Maybe as a consequence, the parameterized complexity of CORRELATION CLUSTERING with respect to k plus structural parameterizations of the fuzzy edge graph was not further considered, despite the importance of the problem. We revisit this line of research motivated by the following observation: While the relation to EDGE MULTICUT implies FPT-algorithms with respect to k it does not give *polynomial* kernels for CORRELATION CLUSTERING, and in fact hardness results for the related MULTIWAY CUT problem [10] imply that such kernels presumably do not exist.¹ In other words, to obtain polynomial kernels for k it is necessary to take the structure of the fuzzy edge graph into account.

Our Results

We consider two parameterizations of the fuzzy edge graph that are both upper-bounded by the maximum degree of the fuzzy edge graph. Thus, in contrast to the global parameter vertex cover number r , these parameters measure the amount of missing information only locally. The first parameter is the degeneracy d of the fuzzy edge graph, which is the smallest number d such that every subgraph of the fuzzy edge graph contains at least one vertex with degree at most d . Notably, d is also upper-bounded by the vertex cover number (and even the treewidth) of the fuzzy edge graph, making the parameterization by $k + d$ smaller than the one by $k + r$. Our first main result is that we can design a polynomial problem kernel for this parameter.

► **Theorem 1.1.** *CORRELATION CLUSTERING parameterized by $k + d$, where d is the fuzzy degeneracy of the input fuzzy graph, admits a kernel with $30k^3d$ vertices.*

¹ We will show nonexistence of a kernel in a restricted setting, so we refrain from providing a more detailed argument for this fact.

The second parameter is the closure c of the fuzzy edge graph. This graph parameter, introduced by Fox et al. [13] is based on the observation that in many graphs, vertices with many common neighbors tend to be adjacent. Formally, a graph is c -closed whenever any pair of vertices with at least c common neighbors is adjacent and the closure of a graph is the smallest number c such that the graph is c -closed. This parameter proved useful as secondary parameter for many different problems [13, 18, 19, 20, 21, 22, 23]. The closure c is not related to the degeneracy of a graph; for example the closure of a complete graph is 0 while the degeneracy is $n - 1$. Hence, it provides an independent way of quantifying the local amount of uncertainty. For example if the uncertainty stems from the fact that there is one set of vertices $S \subseteq V$ for which we could not empirically determine pairwise similarities inside the set, while all other relations are known, then the closure of the fuzzy edge graph would be 0 even if S is large. Our second main result is a polynomial problem kernel for $k + c$.

► **Theorem 1.2.** *CORRELATION CLUSTERING parameterized by $k + c$, where c is the fuzzy closure of the input fuzzy graph, admits a kernel with $30k^2c^2$ vertices.*

Of course, it is perfectly reasonable to consider similar parameterizations for the nonfuzzy vertex pairs, for example the degeneracy of the graph consisting of the edges or the structure of the nonedges. Here, we obtain the following negative result which says that we cannot hope for a polynomial kernel even if the degeneracy and the closure of the graph consisting of the real edges and the degree of the graph consisting of the nonedges are both constant. This result holds even if all real edges and nonedges have unit weight.

► **Theorem 1.3.** *CORRELATION CLUSTERING is NP-hard and does not admit a polynomial kernel for the solution size k , unless $\text{NP} \subseteq \text{coNP}/\text{poly}$, even if the nonedges form a matching and the graph consisting of the real edges has degeneracy 2 and closure 2.*

We also show that CORRELATION CLUSTERING remains NP-hard even if the graph consisting of the real edges is a binary tree or a star.

► **Theorem 1.4.** *CORRELATION CLUSTERING is NP-hard even if the graph consisting of the real edges is a binary tree or a star.*

Due to space constraints, the proof of Theorem 1.4 is deferred to the full version of this work.

2 Preliminaries

Graph Notation. We consider undirected graphs. For a graph $G = (V, E)$ with vertex set V and edge set E , we denote $n := |V|$ and $m := |E|$. For a vertex v , we let $N_G(v) := \{u \mid \{u, v\} \in E\}$ and $N_G[v] := N_G(v) \cup \{v\}$ denote the open and closed neighborhood of v , respectively. The degree of v is $\deg(v) := |N_G(v)|$. Analogously, for a set $X \subseteq V$, we define $N_G[X] = \bigcup_{x \in X} N_G[x]$ and $N_G(X) = N_G[X] \setminus X$. Moreover, we define $N_G^\cap(X) := \bigcap_{x \in X} N_G(x)$ and $N_G^\cap[X] := N_G^\cap(X) \cup X$. We omit subscripts when there is no ambiguity. The *subgraph of G induced by $U \subseteq V$* is the graph $G[U] = (U, E_U)$, with $E_U = \{\{x, y\} \in E \mid x, y \in U\}$. Given a vertex x , we denote the graph $G[V \setminus \{x\}]$ by $G - x$. A graph is *complete* if every pair of vertices are adjacent. A complete subgraph is called a *clique*. If a vertex set K induces a clique, we also say that K is a clique. A clique K is a *critical clique* if $N_G^\cap[K] = N_G[K]$ and K is inclusion-wise maximal under this property. A *vertex cover* is a vertex subset $S \subseteq V$ such that every edge has at least one endpoint in S .

Fuzzy Graphs. A fuzzy graph $G = (V, E^+, E^0)$ consists of a vertex set V , a set of (real) edges E^+ and a set of fuzzy edges E^0 with $E^+ \cap E^0 = \emptyset$. We call a pair of vertices $\{u, v\} \in \binom{V}{2}$ a *nonedge*, if it is neither a real nor a fuzzy edge, and denote the set of nonedges with $E^- := \binom{V}{2} \setminus (E^+ \cup E^0)$. Similar to standard graphs, we define for a vertex $v \in V$ the real, fuzzy and nonneighborhood as $N^+(v) := \{u \mid \{u, v\} \in E^+\}$, $N^0(v) := \{u \mid \{u, v\} \in E^0\}$ and $N^-(v) := \{u \mid \{u, v\} \in E^-\}$, respectively, and also $N^+[v] := N^+(v) \cup \{v\}$ as well as $N^+[X] = \bigcup_{x \in X} N^+[x]$ for a set $X \subseteq V$. We define $G^{(+)} := (V, E^+)$, $G^{(0)} := (V, E^0)$ and $G^{(-)} := (V, E^-)$, that is, $G^{(+)}$ contains all vertices of G and its edges are the real edges of G , and similarly $G^{(0)}$ and $G^{(-)}$ contain only the fuzzy/nonedges of G as edges. For vertex subsets $A \subseteq V$ and $B \subseteq V$ we define $E^+(A, B)$ as the set of real edges with one endpoint in A and the other endpoint in B and analogously $E^-(A, B)$ and $E^0(A, B)$. If one of the sets, for example A , consists of a single vertex v , we also write $E^+(v, B)$, $E^-(v, B)$ and $E^0(v, B)$. For a vertex subset $C \subseteq V$ we write $E^+(C) := E^+(C, C)$, $E^0(C) := E^0(C, C)$ and $E^-(C) := E^-(C, C)$ for the edges, fuzzy edges and nonedges with both endpoints in C , respectively. We call a vertex set K a *fuzzy clique*, if K is a clique in $G^{(0)}$. When speaking of (critical) cliques, neighbors or connected components in a fuzzy graph, we consider the graph $G^{(+)}$ unless mentioned otherwise.

Problem Definition and Basic Observations. A clustering is a partition of the vertices of V and we call each set of the partition a cluster. For a given clustering $\mathcal{C}$ we denote with $\mathcal{C}(v)$ the cluster that contains the vertex v . We call a vertex pair $\{u, v\}$ a positive mistake if $\{u, v\} \in E^+$ and $u \notin \mathcal{C}(v)$ and a negative mistake if $\{u, v\} \in E^-$ and $u \in \mathcal{C}(v)$. For a weight function $\omega : \binom{V}{2} \rightarrow \mathbb{Z}_{\geq 0}$ and set of vertex pairs F , we let $\omega(F) := \sum_{e \in F} \omega(e)$ denote the total weight of the vertex pairs in F . Note that a vertex pair $\{u, v\}$ has weight 0 if and only if $\{u, v\}$ is a fuzzy edge. The *cost* of a clustering $\mathcal{C}$ for a weight function ω is given by the total weights of the positive and negative mistakes, that is $\text{cost}(\mathcal{C}) = \sum_{C \in \mathcal{C}} \omega(E^-(C)) + \sum_{C_1, C_2 \in \mathcal{C}} \omega(E^+(C_1, C_2))$.

With this we can define our main problem.

CORRELATION CLUSTERING

Input: A fuzzy graph (V, E^+, E^0) , a weight function $\omega : \binom{V}{2} \rightarrow \mathbb{Z}_{\geq 0}$ and a nonnegative integer k .

Question: Is there a clustering $\mathcal{C}$ of V such that $\text{cost}(\mathcal{C}) \leq k$?

We call a clustering $\mathcal{C}$ a *solution* if $\text{cost}(\mathcal{C}) \leq k$ and *optimal* if $\text{cost}(\mathcal{C}) \leq \text{cost}(\mathcal{C}')$ for all clusterings $\mathcal{C}'$. We will also take the alternative graph modification view that is made by CLUSTER EDITING: obtain a clustering by deleting real edges $\{u, v\}$ or inserting nonedges $\{u, v\}$ at the cost of $\omega(u, v)$. For some of our reduction rules we change a fuzzy edge into a real edge or a nonedge. In both cases, we set the new weight of the real edge/nonedge to 1 unless stated otherwise. A fuzzy graph G has a clustering with cost 0 if and only if all vertices in the connected components of $G^{(+)}$ are in G connected via a real or a fuzzy edge; otherwise, at least one real edge or nonedge of G must be modified.

If two vertices u and v have more than k common neighbors, then separating u and v needs more than k edge deletions. This implies the following.

► **Observation 2.1.** *Let u and v be vertices with $|N^+(u) \cap N^+(v)| > k$. Then, u and v are contained in the same cluster in any solution.*

► **Corollary 2.2.** *Let K be a clique of size $k + 2$. If $\mathcal{C}$ is a clustering with $\text{cost}(\mathcal{C}) \leq k$, then all vertices of $N_{G^{(+)}}^\cap[K]$ are in the same cluster of $\mathcal{C}$.*

The next observation considers the opposite case, when neighborhoods are too different.

► **Observation 2.3.** *Let u and v be vertices with $|N^+(u) \cap N^-(v)| > k$. Then, u and v are not contained in the same cluster in any solution.*

If there is a cluster that is not connected by real edges, then we could split it up into several clusters without increasing the cost of the clustering, leading to the following observation.

► **Observation 2.4.** *There is an optimal clustering $\mathcal{C}$ such that for each cluster $C \in \mathcal{C}$ all vertices in C are connected in $G^{(+)}$.*

Parameterized Complexity. A parameterized problem is a language $L \subseteq \Sigma^* \times \mathbb{N}$, where Σ is a finite alphabet. For an instance $(I, k) \in \Sigma^* \times \mathbb{N}$, we refer to k as the parameter. An algorithm for a parameterized problem L is an FPT-algorithm if there is a computable function f such that for every instance (I, k) the algorithm decides in $f(k) \cdot |I|^{\mathcal{O}(1)}$ time whether (I, k) is a yes-instance of L . A *kernel* for L is an algorithm that computes for each instance (I, k) in polynomial time an equivalent instance (I', k') with $(|I'| + k') \leq g(k)$ for a computable function g ; we call g the *size* of the kernel. A kernel of size g is a *polynomial kernel* if $g(k) \in k^{\mathcal{O}(1)}$. A parameterized problem has an FPT-algorithm if and only if it admits a (not necessarily polynomial) problem kernel.

A *polynomial parameter transformation* is a reduction from a parameterized problem A to a parameterized problem B that transforms each instance (I, k) of A in polynomial time into an equivalent instance (I', k') of B with $k' \in k^{\mathcal{O}(1)}$. If $k' = k$, we say the polynomial parameter transformation preserves the parameter.

Moreover, if there is a polynomial parameter transformation from A to B , where the unparameterized versions of A and B are NP-hard and in NP, respectively, then A admits a polynomial kernel if B admits a polynomial kernel.

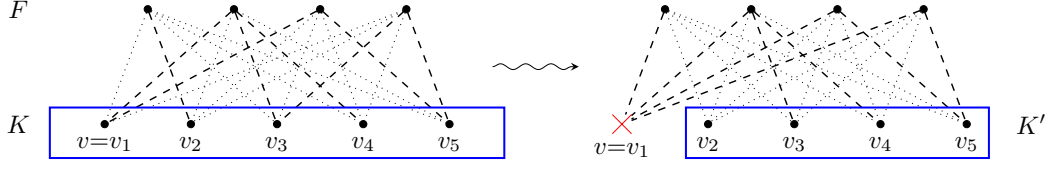
For more details about the relevant concepts of parameterized complexity, such as kernelization and kernelization lower bounds, we refer the reader to the standard literature [12, 9].

3 Fuzzy Degeneracy Kernel

In the following, we provide a kernel for parameterization by k plus the degeneracy of $G^{(0)}$, the fuzzy edge graph of $G = (V, E^+, E^0)$. We will make use of the following definition of degeneracy via vertex orderings. For an ordering $\sigma = (v_1, \dots, v_t)$ of a vertex set $C \subseteq V$ with $|C| = t$, let $D_\sigma[i, j]$ with $i \leq j$ denote the set of vertices $\{v_i, \dots, v_j\}$. Moreover, let $\deg_\sigma(v_i) = |N(v_i) \cap D_\sigma[i, n]|$ denote the *ordering degree* of v_i with regards to the ordering σ .

A graph $G = (V, E)$ has *degeneracy* d , if d is the smallest number for which there is an ordering $\sigma = (v_1, \dots, v_n)$ of V , such that each vertex v_i has at most d neighbors v_j with $j > i$, that is, $\deg_\sigma(v_i) \leq d$. We call such an ordering a *degeneracy ordering* for G . For a fuzzy graph G , we say that a degeneracy ordering of $G^{(0)}$ is a *fuzzy degeneracy ordering* of G and let d denote the *fuzzy degeneracy* of G . For a vertex subset $C \subseteq V$, let $\sigma(C)$ be the subsequence of a fuzzy degeneracy ordering σ that contains exactly the vertices of C . We call such an ordering a *fuzzy degeneracy ordering* for C .

With these definitions at hand, we may now proceed to describe the kernelization. The first rule is the known rule that removes isolated clusters that do not cause any costs.



■ **Figure 1** An application of Rule 3. Fuzzy edges are dashed, nonedges are dotted. After swapping some fuzzy edges and nonedges, v has only fuzzy neighbors in F and can thus be safely deleted.

► **Rule 1.** If G contains a connected component with vertices C such that each pair of vertices $u, v \in C$ forms a real edge or a fuzzy edge, then remove C from G .

The following reduction rule makes vertices with many common neighbors adjacent. It corresponds to a classic reduction rule for CLUSTER EDITING [15] and its correctness is directly implied by Observation 2.1.

► **Rule 2.** Let $\{u, v\}$ be a nonedge or fuzzy edge in G . If $|N^+(u) \cap N^+(v)| > k$, make $\{u, v\}$ a real edge and set $k := k - \omega(u, v)$.

This rule is already sufficient to show that large clusters must contain large real cliques.

► **Lemma 3.1.** Let G be reduced with respect to Rule 2 and let $\mathcal{C}$ be an optimal clustering. Let $C \in \mathcal{C}$ be a cluster of size $t > 2k + 2d$ and let $\sigma(C) = (v_1, v_2, \dots, v_t)$ be a fuzzy degeneracy ordering of C . Then $D_{\sigma(C)}[1, t - 2d - 2k - 1]$ induces a real clique.

Proof. Each vertex in $X := D_{\sigma(C)}[1, t - 2d - 2k - 1]$ has at most d fuzzy neighbors in the last $2k + 2d + 1$ vertices of the fuzzy degeneracy ordering, that is $Y := D_{\sigma(C)}[t - 2d - 2k, t]$.

Let u and v be arbitrary vertices of X . Then u and v have in total at most $2d$ fuzzy neighbors in Y and at most k nonneighbors in Y . Since $|Y| = 2k + 2d + 1$, Y contains at least $k + 1$ common real neighbors of u and v . Thus, u and v are neighbors in G by Rule 2. ◀

Finding such large cliques with the aim of shrinking them seems hard, but as we will see, it is sufficient to shrink critical cliques. This is done by the next rule.

► **Rule 3.** Let K be a critical real clique of size at least $k + 3$ and let $v \in K$ be the vertex in K that appears first in some fuzzy degeneracy ordering σ of G . Let F be the set of fuzzy neighbors of K and for each $u \in F$, let v_u be the first fuzzy neighbor of u in K according to σ . For each $u \in F$ with $v_u \neq v$ make $\{u, v_u\}$ a nonedge with $\omega(u, v_u) := \omega(u, v)$ and make $\{u, v\}$ a fuzzy edge; then delete v from G .

► **Lemma 3.2.** Rule 3 is correct and can be exhaustively applied in $\mathcal{O}(n^3)$ time.

Proof. Let $\mathcal{C}$ be an optimal clustering for G . By Corollary 2.2, since K is a critical clique of size at least $k + 3$, $\mathcal{C}$ contains a cluster C_K with all vertices of $N_{G(+)}^{\cap}[K] = N_G^+[K]$. Since $|K| \geq k + 3$, the only vertices of $V \setminus N_G^+[K]$ in C_K are from F . Let G' be the fuzzy graph obtained after applying Rule 3. Note that $K' := K \setminus \{v\}$ is a critical clique of size at least $k + 2$ in G' , so in an optimal clustering for G' the cluster $C_{K'}$ containing $N_{G'}^+[K']$ also additionally contains only vertices of F . Thus, it suffices to show that for each $u \in F$, the cost of adding u to the cluster containing $N_{G'}^+[K']$ in G' is the same as the cost of adding u to the cluster containing $N_G^+[K]$ in G . Let u be an arbitrary vertex of F . If $v_u = v$, then $N_{G'}^-(u) \cap N_{G'}^+[K'] = N_G^-(u) \cap N_G^+[K]$, so the cost of adding u is the same for G and G' . Otherwise, if $v_u \neq v$, then $N_{G'}^-(u) \cap N_{G'}^+[K'] = (N_G^-(u) \cap N_G^+[K]) \setminus \{v\} \cup \{v_u\}$, so the cost of adding u is also the same for G and G' .

One can compute the critical cliques from the modular decomposition in linear time [8, 28]. We then consider the critical cliques one by one. For each critical clique, F and v_u for all vertices u can be computed in $\mathcal{O}(n)$ time. Since each application of Rule 3 removes a vertex, the rule can be applied at most n times. Thus, exhaustive application takes $\mathcal{O}(n^3)$ time. ◀

► **Lemma 3.3.** *Let $(G = (V, E^+, E^0), k)$ be an instance of CORRELATION CLUSTERING that is reduced with respect to Rules 1, 2, and 3. If $|V| > 30k^3d$, then (G, k) is a no-instance.*

Proof. Let (G, k) be a yes-instance and let $\mathcal{C}$ be a solution for (G, k) . Let C be a cluster of size at least $3k + 2d + 3$ and let $\sigma(C) = (v_1, v_2, \dots, v_t)$ be a fuzzy degeneracy ordering of C . We show that $|C| \leq 14k^2d$. Let $X := D_{\sigma(C)}[1, t - 2d - 2k - 1]$ be the vertices of C except for the last $2k + 2d + 1$ vertices according to $\sigma(C)$. According to Lemma 3.1 the vertices of X form a clique. Let $C_K \subseteq C$ be a maximal real clique in G that contains X and potentially parts of $Y := C \setminus X$ and let $C_F := C \setminus C_K$. Note that $C_F \subseteq Y$ and thus $|C_F| \leq 2k + 2d + 1$. We can partition C_K as follows (see also Figure 2).

- Let $C_{KF} \subseteq C_K$ be the vertices in C_K that have neighbors in C_F ,
- let $\tilde{C}_K \subseteq C_K \setminus C_{KF}$ be the vertices in $C_K \setminus C_{KF}$ that have neighbors outside of C , and
- let $C_K^* = C_K \setminus (C_{KF} \cup \tilde{C}_K)$ be the remaining vertices in C_K .

Per definition, vertices in C_K^* only have fuzzy neighbors and nonneighbors outside of C and thus C_K^* is a critical clique with $N^+[C_K^*] = C_K$. Observe that $|\tilde{C}_K| \leq k$ since every real edge with an endpoint outside of C has to be deleted to obtain the cluster C .

Moreover, we have $|C_{KF}| \leq (k + 1) \cdot (2k + 2d + 1)$: Otherwise, since $|C_F| \leq 2k + 2d + 1$, per pigeonhole principle there must be some vertex $v \in C_F$ that has at least $k + 1$ neighbors in C_{KF} . Then, since C_K is a clique of size at least $k + 2$, for each $u \in C_K$ there are $k + 1$ common neighbors of u and v in C_{KF} . Thus, according to Rule 2, $\{u, v\} \in E^+$ and therefore $C_K \cup \{v\}$ is a clique. This is a contradiction to the maximality of C_K .

Finally, we have $|C_K^*| \leq k + 2$, since C_K^* is a critical clique and G is reduced with respect to Rule 3. Thus, we can bound the size of C :

$$\begin{aligned} |C| &= |C_F| + |C_{KF}| + |\tilde{C}_K| + |C_K^*| \\ &\leq (2k + 2d + 1) + (k + 1) \cdot (2k + 2d + 1) + k + (k + 2) \\ &= 2k^2 + 2kd + 7k + 4d + 4 \\ &\leq 15k^2d. \end{aligned}$$

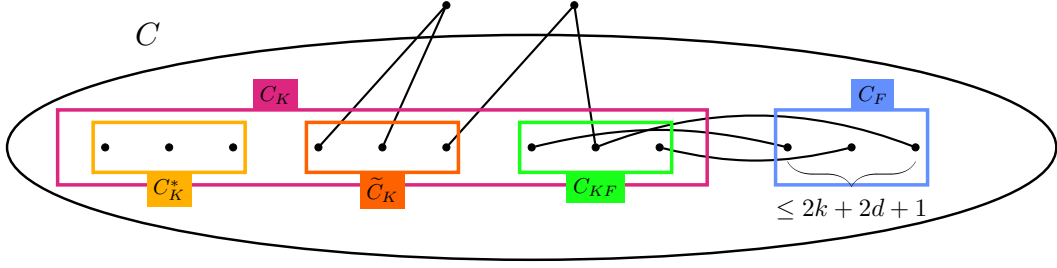
Since G is reduced with respect to Rule 1, each cluster of $\mathcal{C}$ contains a vertex that is incident to a modified edge in G , so $\mathcal{C}$ contains at most $2k$ clusters, each of which contains at most $15k^2d$ vertices. Thus, G contains at most $30k^3d$ vertices. ◀

Lemma 3.3 directly implies Theorem 1.1.

► **Theorem 1.1.** *CORRELATION CLUSTERING parameterized by $k + d$, where d is the fuzzy degeneracy of the input fuzzy graph, admits a kernel with $30k^3d$ vertices.*

4 Fuzzy Closure Kernel

We now exploit the closure of the graph $G^{(0)}$ induced by the fuzzy edges. Recall that a graph H is c -closed for some integer c , if each pair of nonneighbors u and v has less than c common neighbors. The *closure* of a graph H is the smallest number c for which H is c -closed. We call the closure of $G^{(0)}$ the *fuzzy closure* of G and say a fuzzy graph G is *fuzzy c -closed*,



■ **Figure 2** Structure of a cluster C after application of all reduction rules. C_K is a clique and C_F contains a subset of the last $2k + 2d + 1$ vertices according to a given fuzzy degeneracy ordering.

if $G^{(0)}$ is c -closed. For a given value c , we call a pair of vertices u and v *bad*, if they have at least c common fuzzy neighbors, but $\{u, v\}$ is not a fuzzy edge. Thus a fuzzy graph G has fuzzy closure at most c if and only if G does not have any bad pair for c .

We present a polynomial kernel for the parameter $k + c$, where c is the fuzzy closure of the input fuzzy graph G . Note that if $\{u, v\}$ is a fuzzy edge, then making $\{u, v\}$ a real or nonedge may change the fuzzy closure of the graph. Hence, the first rule is a special case of Rule 2 that is only applied to nonedges.

► **Rule 4.** Let $\{u, v\}$ be a nonedge in G . If $|N^+(u) \cap N^+(v)| > k$, make $\{u, v\}$ a real edge and set $k := k - \omega(u, v)$.

We can also state a similar rule for real edges that is directly implied by Observation 2.3.

► **Rule 5.** Let $\{u, v\}$ be a real edge in G . If $|N^+(u) \cap N^-(v)| > k$, make $\{u, v\}$ a nonedge and set $k := k - \omega(u, v)$.

If u and v have less than c common fuzzy neighbors, then they can not be a bad pair and we can apply a similar reduction rule for a fuzzy edge $\{u, v\}$.

► **Rule 6.** Let $\{u, v\}$ be a fuzzy edge in G such that $|N^0(u) \cap N^0(v)| < c$.

■ If $|N^+(u) \cap N^+(v)| > k$, make $\{u, v\}$ a real edge.

■ If $|N^+(u) \cap N^-(v)| > k$, make $\{u, v\}$ a nonedge.

► **Lemma 4.1.** Rule 6 is correct and can be exhaustively applied in $\mathcal{O}(n^3)$ time. If G is fuzzy c -closed, then the resulting graph is also fuzzy c -closed.

Proof. According to Observation 2.1, if $|N^+(u) \cap N^+(v)| > k$, then any solution puts u and v in the same cluster, so we can safely make $\{u, v\}$ a real edge. Let G' be the fuzzy graph obtained by the application of the rule. Since $|N^0(u) \cap N^0(v)| < c$, the vertices $\{u, v\}$ are not a bad pair in G' . For every other pair of vertices $\{x, y\} \neq \{u, v\}$, the number of common fuzzy neighbors does not increase by making $\{u, v\}$ a real edge, so G' is still fuzzy c -closed. We can exhaustively apply the rule by checking the common fuzzy/real neighborhood for each pair of vertices and updating the edges accordingly in $\mathcal{O}(n^3)$ time. The correctness of the case $|N^+(u) \cap N^-(v)| > k$ can be shown analogously. ◀

We will now show that we can modify the instance so that afterwards each vertex has a bounded number of real neighbors or a bounded number of fuzzy neighbors, which we then use to bound the size of clusters in a solution.

► **Lemma 4.2.** Let v be a vertex with more than $2kc$ real neighbors and at least c fuzzy neighbors. Then v has a fuzzy neighbor u with $|N^+(v) \cap N^+(u)| > k$ or $|N^+(v) \cap N^-(u)| > k$.

Proof. Let $X \subseteq N^0(v)$ be a set of c fuzzy neighbors of v . Each real neighbor $w \in N^+(v)$ of v has at least one real or nonneighbor in X , since otherwise G would not be fuzzy c -closed. Since $|N^+(v)| > 2kc$ and $|X| = c$, per pigeonhole principle there must be a vertex $u \in X$ with $|N^+(v) \cap N^+(u)| > k$ or $|N^+(v) \cap N^-(u)| > k$. $\blacktriangleleft$

► **Rule 7.** Let v be a vertex with more than $2kc$ real neighbors and at least c fuzzy neighbors. Then, find a vertex $u \in N^0(v)$ such that $|N^+(v) \cap N^+(u)| > k$ and make $\{u, v\}$ a real edge or find a vertex $u \in N^0(v)$ such that $|N^+(v) \cap N^-(u)| > k$ and make $\{u, v\}$ a nonedge.

► **Lemma 4.3.** Rule 7 is correct and can be exhaustively applied in $\mathcal{O}(n^3)$ time.

Proof. In the first case, v and u have to be in the same cluster, since otherwise we would have to delete more than k edges to separate them, thus we can safely make the fuzzy edge $\{u, v\}$ a real edge. Analogously, in the second case v and u cannot be in the same cluster with at most k edge modifications, so we can safely make $\{u, v\}$ a nonedge.

Checking whether a vertex has sufficiently many real and fuzzy neighbors as well as checking for each fuzzy neighbor the real/nonneighborhood can be done in $\mathcal{O}(n^3)$ time. $\blacktriangleleft$

We now show that after exhaustively applying Rule 7 we can assume that each vertex in G has at most $2kc$ real neighbors or less than c fuzzy neighbors, and the resulting fuzzy graph is still fuzzy c -closed.

► **Lemma 4.4.** Let G' be the resulting fuzzy graph after exhaustively applying Rule 7. Then, for each vertex $v \in V$ we have $|N_{G'}^0(v)| < c$ or $|N_{G'}^+(v)| \leq 2kc$ and G' is fuzzy c -closed.

Proof. Assume towards a contradiction that there is a vertex $v \in V$ in G' with more than $2kc$ real neighbors and at least c fuzzy neighbors. Then, according to Lemma 4.2 there is a vertex $u \in N^0(v)$ with $|N^+(v) \cap N^+(u)| > k$ or $|N^+(v) \cap N^-(u)| > k$ and thus Rule 7 can be applied, leading to a contradiction.

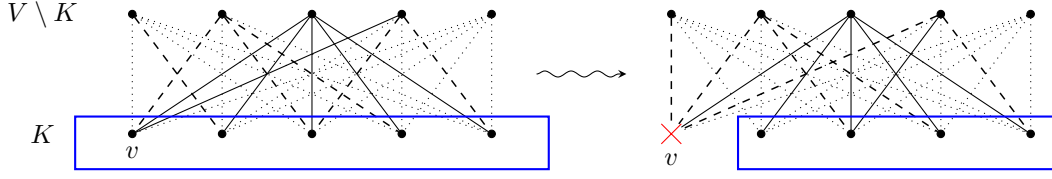
Thus, it remains to show that G' is also fuzzy c -closed. Let v be a vertex in G with more than $2kc$ real neighbors and at least c fuzzy neighbors to which Rule 7 is applied. Note that applying Rule 7 once can only create a bad pair $\{u, v\}$ that is formed by a fuzzy edge $\{u, v\}$ that was made into a real edge or nonedge, since G is c -closed and for each other pair of vertices the common fuzzy neighborhood either stays the same or gets smaller. As long as v still has at least c fuzzy neighbors, Rule 7 can be applied again, where after each application a fuzzy neighbor of v is then made a real or a nonneighbor, so the number of fuzzy neighbors of v decreases with each application of Rule 7. Hence, after exhaustively applying Rule 7, v has less than c fuzzy neighbors and can thus not be in a bad pair anymore. $\blacktriangleleft$

In the following we assume G is reduced with regards to Rule 7 and define V_+ to be the set of vertices with less than c fuzzy neighbors and $V_0 := V \setminus V_+$. Note that each vertex in V_0 has at most $2kc$ real neighbors. Next, we show that if a cluster is large, then we can partition it into a real clique and a fuzzy clique.

► **Lemma 4.5.** Let $\mathcal{C}$ be a solution for (G, k) and let $C \in \mathcal{C}$ be a cluster. If $|C| > 4kc + k + c + 2$, then $C \cap V_0$ is a clique in $G^{(0)}$ and $C \cap V_+$ is a clique in $G^{(+)}$.

Proof. Let u and v be vertices of $C \cap V_0$. By definition of V_0 , each of u and v has at most $2kc$ real neighbors in C . Moreover, since C is a cluster of a solution, u and v have in total at most k nonneighbors in C . Thus, more than c vertices of C are common fuzzy neighbors of u and v . Since G has fuzzy closure c , the vertices u and v must form a fuzzy edge.

17:10 Clustering with Locally Bounded Ignorance



■ **Figure 3** An application of Rule 8. Fuzzy edges are dashed, nonedges are dotted. After swapping some real edges and nonedges with fuzzy edges, each real neighbor/nonneighbor of v in $V \setminus K$ is also a real neighbor/nonneighbor of every other vertex in K and thus v can be safely deleted.

Now let u and v be arbitrary vertices of $C \cap V_+$. Per definition of V_+ , each of u and v has less than c fuzzy neighbors in C . Moreover, since C is a cluster, u and v have in total at most k nonneighbors in C . Since $|C| > 4kc + k + c + 2 > 2c + 2k + 2$, more than k vertices of C are common real neighbors of u and v . Thus, according to Rule 4 and Rule 6, the vertices u and v must form a real edge. ◀

We can bound the size of a clique in G with the following reduction rule.

► **Rule 8.** Let K be a real clique of size at least $3kc + 2$ and let v be an arbitrary fixed vertex of K . Then for each vertex $u \in V \setminus K$ with $\{u, v\} \notin E^0$ do the following:

- Let $w \in K$ be a vertex such that $\{u, w\} \in E^0$, if such a vertex exists.
- If $\{u, v\}$ is a real edge, make $\{u, w\}$ a real edge with $\omega(u, w) := \omega(u, v)$.
- If $\{u, v\}$ is a nonedge, make $\{u, w\}$ a nonedge with $\omega(u, w) := \omega(u, v)$.

Afterwards, delete v .

► **Lemma 4.6.** Rule 8 is correct and can be exhaustively applied in $\mathcal{O}(n^4)$ time. If G is fuzzy c -closed, then the resulting graph is also fuzzy c -closed.

Proof. Let G' be the fuzzy graph after applying Rule 8 with vertex set $V' := V \setminus \{v\}$. We show that (G, k) is a yes-instance if and only if (G', k) is a yes-instance. First, suppose that (G, k) is a yes-instance and $\mathcal{C}$ is a clustering for G with $\text{cost}(\mathcal{C}) \leq k$. According to Corollary 2.2 all vertices of K are contained in the same cluster, which we denote by C_K . Now consider the clustering $\mathcal{C}' := (\mathcal{C} \setminus \{C_K\}) \cup \{C_K \setminus \{v\}\}$ for G' that is identical to $\mathcal{C}$ except that C_K is replaced by $C'_K := C_K \setminus \{v\}$. Note that if for each $u \in V \setminus K$ we have $\{u, v\} \in E_G^0$, then G' is an induced subgraph of G and thus $\text{cost}(\mathcal{C}') \leq \text{cost}(\mathcal{C}) \leq k$. Otherwise, let $U \subseteq V \setminus K$ be the vertices that are not fuzzy neighbors of v . Consider some fuzzy edge $\{u, w\} \in E_G^0$ with $u \in U$, $w \in K$ that becomes a real/nonedge in G' and contributes to $\text{cost}(\mathcal{C}')$. Then, $\{u, v\}$ must also be a real/nonedge in G and equally contribute to $\text{cost}(\mathcal{C})$, so we get $\text{cost}(\mathcal{C}') \leq \text{cost}(\mathcal{C}) \leq k$ again.

Conversely, suppose that (G', k) is a yes-instance and $\mathcal{C}'$ is a clustering for G' with $\text{cost}(\mathcal{C}') \leq k$. Note that, since $|K \setminus \{v\}| \geq 3kc + 1$, again according to Corollary 2.2 all vertices of $K \setminus \{v\}$ are contained in the same cluster, which we denote by C'_K . Consider the clustering $\mathcal{C} := (\mathcal{C}' \setminus \{C'_K\}) \cup \{C'_K \cup \{v\}\}$ for G that is identical to $\mathcal{C}'$ except that C'_K is replaced by $C_K := C'_K \cup \{v\}$. We have

$$\text{cost}(\mathcal{C}) - \text{cost}(\mathcal{C}') = \omega(E_G^+(C_K, V \setminus C_K)) + \omega(E_G^-(C_K)) - \omega(E_{G'}^+(C'_K, V' \setminus C'_K)) - \omega(E_{G'}^-(C'_K))$$

and it remains to show that $\text{cost}(\mathcal{C}) - \text{cost}(\mathcal{C}') \leq 0$. Clearly, the contribution of edges between two vertices of $K \setminus \{v\}$ is the same for both $\text{cost}(\mathcal{C})$ and $\text{cost}(\mathcal{C}')$. The same is true for edges between two vertices of $V \setminus K$. Now consider an arbitrary vertex $u \in V \setminus K$. If $\{u, v\} \in E_G^0$, then $\{u, v\}$ does not contribute to neither $\text{cost}(\mathcal{C})$ nor $\text{cost}(\mathcal{C}')$ and for each $w \in K$ the contribution of $\{u, w\}$ to $\text{cost}(\mathcal{C})$ and $\text{cost}(\mathcal{C}')$ is the same.

Thus, now assume that $\{u, v\} \notin E_G^0$. If $|N_G^0(u) \cap K| \geq 1$, then there is a vertex $w \in K$ with $\{u, w\} \in E_G^0$ as stated in Rule 8. If $\{u, v\}$ is a real edge in G , then $\{u, w\}$ becomes a real edge in G' . Thus, if $\{u, v\}$ contributes to $\omega(E_G^+(C_K, V \setminus C_K))$, then $\{u, w\}$ contributes to $\omega(E_{G'}^+(C'_K, V' \setminus C'_K))$. Similarly, if $\{u, v\}$ is a nonedge in G , then $\{u, w\}$ becomes a nonedge in G' and if $\{u, v\}$ contributes to $\omega(E_G^-(C_K))$, then $\{u, w\}$ contributes to $\omega(E_{G'}^-(C'_K))$. For each $z \in K \setminus \{w\}$ the contribution of $\{u, z\}$ to $\text{cost}(\mathcal{C})$ and $\text{cost}(\mathcal{C}')$ is the same.

Otherwise, we either have $|N_G^+(u) \cap K| > k+1$ or $|N_G^-(u) \cap K| > k+1$. If $|N_G^+(u) \cap K| > k+1$, then u and each vertex in K have more than k common neighbors in K , so according to Rule 4 and Rule 6 for each $w \in K$ we have that $\{u, w\}$ is a real edge, which implies $u \in C_K$ and thus the contribution of edges in $E_G^+(u, K)$ to $\omega(E_G^+(C_K, V \setminus C_K))$ is zero. Similarly, if $|N_G^-(u) \cap K| > k+1$, then according to Rule 5 and Rule 6 for each $w \in K$ we have that $\{u, w\}$ is a nonedge, which implies $u \notin C_K$ and the contribution of edges in $E_G^-(u, K)$ to $\omega(E_G^-(C_K))$ is zero. All in all we get $\text{cost}(\mathcal{C}) - \text{cost}(\mathcal{C}') \leq 0$.

Note that, since G is fuzzy c -closed, applying Rule 8 once could potentially only create a bad pair $\{u, w\}$ that was made a real edge or a nonedge for a vertex $w \in K$. However, since K is a real clique of size $3kc+2$, each vertex $w \in K$ is in V_+ , has less than c fuzzy neighbors and thus cannot be part of a bad pair.

In $\mathcal{O}(n^3)$ time we can find a clique of size $3kc+2$ or verify that such a clique does not exist: Let K be a clique of size $3kc+2$ and let u be a vertex in K . Clearly, only real neighbors of u with at least $3kc$ common neighbors with u can possibly be part of K . Now, let $w \in N^+(u)$ with $|N^+(u) \cap N^+(w)| \geq 3kc$ be such a neighbor of u and assume towards a contradiction that $w \notin K$ and $K \cup \{w\}$ is not a clique. Since $w \in V_+$, we have $|N^0(w) \cap K| < c$. Moreover, $|N^-(w) \cap K| < k+1$, since otherwise $\{u, w\}$ would be a nonedge according to Rule 5. This implies $|N^+(w) \cap K| > k+1$ and thus w has more than k common neighbors with each vertex in K , so $K \cup \{w\}$ must also be a clique according to Rule 4. Thus, we can iterate over all vertices and for a fixed $u \in V$ we check whether u together with each neighbor v with at least $3kc$ common neighbors forms a clique, which can be done in $\mathcal{O}(n^3)$ time.

Given a clique of size at least $3kc+2$, we can then apply Rule 8 in $\mathcal{O}(n^2)$ time by iterating through $V \setminus K$ and K . Since each application of Rule 8 removes a vertex, it can be applied at most n times and thus also exhaustively applied in $\mathcal{O}(n^4)$ time. $\blacktriangleleft$

The final reduction rule will help bounding the size of fuzzy cliques inside clusters.

► **Rule 9.** Let K be a real clique in G such that $X := N_{G(+)}^\cap(K) \cap \{v \in V \mid N^+(v) = K\}$ is a fuzzy clique with $|X| > 2k + c + 1$. Then, remove an arbitrary vertex $v \in X$ from G .

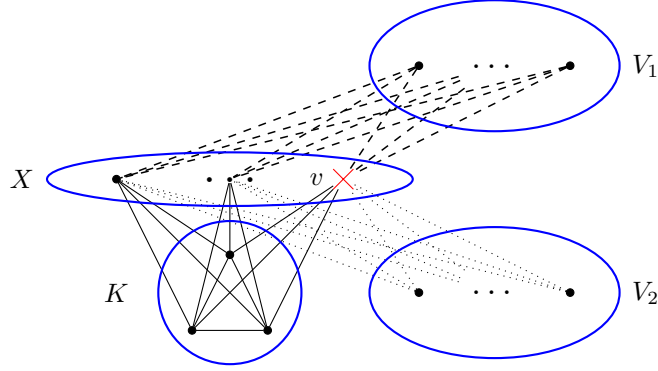
► **Lemma 4.7.** Rule 9 is correct and can be exhaustively applied in $\mathcal{O}(n^4)$ time.

Proof. Let $\mathcal{C}$ be a solution, so $\text{cost}(\mathcal{C}) \leq k$. Note that since $|X| > k$ each pair of vertices in K has at least $k+1$ common neighbors in X , so all vertices of K are part of the same cluster C_K in $\mathcal{C}$ according to Observation 2.1. Moreover, since $|X| > 2k + c$, there is a subset X_K of size at least $k + c$ that also belongs to the same cluster C_K .

We first show that we can assume $X_K = X$: Suppose that there is some vertex $u \in X \setminus X_K$ that is not in the cluster C_K , but in some other cluster C_u . Consider the clustering $\mathcal{C}' := (\mathcal{C} \setminus \{C_K, C_u\}) \cup \{C_K \cup \{u\}, C_u \setminus \{u\}\}$ where u is assigned to C_K instead. Since u is adjacent to each vertex in K , but has no real neighbors outside of K , and X is a fuzzy clique we have

$$\begin{aligned} \text{cost}(\mathcal{C}) - \text{cost}(\mathcal{C}') &= \omega(E^+(u, C_K)) + \omega(E^-(u, C_u \setminus \{u\})) - \omega(E^+(u, C_u \setminus \{u\})) - \omega(E^-(u, C_K)) \\ &\geq |K| + \omega(E^-(u, C_u \setminus \{u\})) - \omega(E^-(u, C_K \setminus \{K \cup X\})) \\ &\geq |K| - \omega(E^-(u, C_K \setminus \{K \cup X\})) \end{aligned}$$

We show that $E^-(u, C_K \setminus \{K \cup X\}) = \emptyset$, which implies $\text{cost}(\mathcal{C}) - \text{cost}(\mathcal{C}') > 0$. Consider



■ **Figure 4** An illustration of Rule 9. Fuzzy edges are dashed, nonedges are dotted. Only relevant edges are drawn. Vertices in V_1 are fuzzy neighbors of all vertices in X , vertices in V_2 are not part of the same cluster as $K \cup X$.

some vertex $w \in V \setminus (K \cup X)$. Per definition of X , there is no real edge between w and vertices in X . Thus, we can distinguish between two cases:

1. If $|N^0(w) \cap X| > c$, then for each $x \in X$ we have that w and x have more than c common fuzzy neighbors in X , which implies $N^0(w) \cap X = X$.
2. Otherwise, we have $|N^-(w) \cap X| > 2k$ and thus $|N^-(w) \cap X_K| > k$. This implies that w is not in the cluster C_K .

Let V_1 be the set of vertices $w \in V \setminus (K \cup X)$ with $|N^0(w) \cap X| > c$ and let V_2 be the set of vertices $w \in V \setminus (K \cup X)$ with $|N^-(w) \cap X_K| > k$. Since each vertex in V_1 is a fuzzy neighbor of u and each vertex in V_2 is not part of C_K , we have $E^-(u, C_K \setminus \{K \cup X\}) = \emptyset$ and thus $\text{cost}(\mathcal{C}) - \text{cost}(\mathcal{C}') > 0$. Hence, $\mathcal{C}'$ is also a solution with $\text{cost}(\mathcal{C}') < k$ and we can assume $X_K = X$.

Now, let G' be the fuzzy graph obtained from G after removing some arbitrary $v \in X$ from G and let $X' = X \setminus \{v\}$. We show that G has a clustering $\mathcal{C}$ with $\text{cost}(\mathcal{C}) \leq k$ if and only if G' has a clustering $\mathcal{C}'$ with $\text{cost}(\mathcal{C}') \leq k$.

First, suppose $\mathcal{C}$ is a clustering for G with $\text{cost}(\mathcal{C}) \leq k$. Since G' is an induced subgraph of G , we have $\text{cost}(\mathcal{C}') \leq \text{cost}(\mathcal{C}) \leq k$.

Second, suppose that $\mathcal{C}'$ is a clustering for G' with $\text{cost}(\mathcal{C}') \leq k$. With the same arguments as above, since $|X'| = |X| - 1$, we can see that all the vertices $K \cup X'$ are part of the same cluster, which we denote by C'_K . Consider the clustering $\mathcal{C}$ for G that is identical to $\mathcal{C}'$ except that the cluster C'_K is replaced by $C_K = C'_K \cup \{v\}$. We have

$$\text{cost}(\mathcal{C}) - \text{cost}(\mathcal{C}') = \omega(E^-(v, C'_K)) + \omega(E^+(v, V \setminus C'_K)).$$

Note that v has no real edges to vertices outside of $K \cup X'$ per definition of X , so $E^+(v, V \setminus C'_K) = \emptyset$. Moreover, v has no nonedges to vertices of $K \cup X' \subseteq C'_K$. For each vertex $u \in V_1$ we have $\{u, v\} \in E^0$. Furthermore, each vertex $u \in V_2$ is not part of the cluster C'_K due to the same arguments as above. Therefore, we also have $E^-(v, C'_K) = \emptyset$ and thus $\text{cost}(\mathcal{C}) = \text{cost}(\mathcal{C}') \leq k$.

Concerning the running time, we consider each vertex $v \in V$, determine $K = N^+(v)$ and $X = N_{G(+)}^0(K)$ and then check if K and X satisfy the requirements of the rule in $\mathcal{O}(n^3)$ time. Since each rule application removes a vertex, we can exhaustively apply the rule in $\mathcal{O}(n^4)$ time. ◀

We can now bound the size of a cluster in a solution.

► **Lemma 4.8.** *Let (G, k) be a yes-instance that is reduced with respect to Rule 1 and Rules 4 to 9. Let $\mathcal{C}$ be a solution for (G, k) and let $C \in \mathcal{C}$ be a cluster. Then, $|C| \leq 3kc^2 + 3kc + 3k + 3c + 3$.*

Proof. Since G is reduced with respect to Rule 7 we can partition C into $C_1 := C \cap V_0$ and $C_2 := C \cap V_+$. Assume towards a contradiction that $|C| > 3kc^2 + 3kc + 3k + 3c + 3$. Then C_1 is a fuzzy clique and C_2 is a real clique according to Lemma 4.5. In particular, C does not contain a nonedge. Since G is reduced with respect to Rule 8 we can assume $|C_2| < 3kc + 2$. Moreover, we can bound the number of vertices in C_1 that have a fuzzy neighbor in C_2 by $|C_2| \cdot c < 3kc^2 + 2c$. Note that we can assume that C_2 is not empty according to Observation 2.4. Now, let $C'_1 \subseteq C_1$ be the vertices in C_1 that have only real neighbors in C_2 . Since $\mathcal{C}$ is a solution, at most k vertices of C'_1 have real neighbors outside of C . Finally, let $X \subseteq C'_1$ be the vertices in C'_1 that have only real neighbors in C_2 and no real neighbors outside of C . Note that since X is a fuzzy clique, if $|X| > 2k + c + 1$ we could apply Rule 9 to X . Since G is reduced with respect to Rule 9, we thus have $|X| \leq 2k + c + 1$. Overall, we get

$$\begin{aligned} |C| &= |C_1| + |C_2| \\ &\leq (3kc^2 + 2c) + k + (2k + c + 1) + (3kc + 2) \\ &= 3kc^2 + 3kc + 3k + 3c + 3. \end{aligned}$$

◀

With the previous lemma we now can bound the size of a yes-instance.

► **Lemma 4.9.** *Let $I := (G, k)$ be an instance of CORRELATION CLUSTERING that is reduced with respect to Rule 1 and Rules 4 to 9. If $|V(G)| > 30k^2c^2$, then I is a no-instance.*

Proof. Let I be a yes-instance and let $\mathcal{C}$ be a solution. Since I is reduced with respect to Rule 1, each cluster in $\mathcal{C}$ contributes to $\text{cost}(\mathcal{C})$. Since each deletion of a real edge can create at most 2 clusters, $\mathcal{C}$ contains at most $2k$ clusters. According to Lemma 4.8, if I is a yes-instance, then each cluster has a size of at most $3kc^2 + 3kc + 3k + 3c + 3$. Hence, either I is a no-instance or G contains at most $6k^2c^2 + 6k^2c + 6k^2 + 6kc + 6k \leq 30k^2c^2$ vertices. ◀

The previous lemma immediately gives us Theorem 1.2.

► **Theorem 1.2.** *CORRELATION CLUSTERING parameterized by $k + c$, where c is the fuzzy closure of the input fuzzy graph, admits a kernel with $30k^2c^2$ vertices.*

5 Hardness for Restricted Real Edges and Nonedges

In this section we assume that all real and nonedges have unit weight.

We first adapt a reduction of Demaine et al. [11] from EDGE MULTICUT to CORRELATION CLUSTERING to show hardness for the case where the nonedges form a matching.

EDGE MULTICUT

Input: A graph $G = (V, E)$, a set of terminal pairs $T = \{(s_1, t_1), \dots, (s_r, t_r)\} \subseteq \binom{V}{2}$ and a nonnegative integer k .

Question: Is there a set of edges $S \subseteq E$ such that $|S| \leq k$ and no terminal pair (s_i, t_i) is connected in $G - S$?

Construction. Let $(G = (V, E), T, k)$ with $T = \{(s_1, t_1), \dots, (s_r, t_r)\}$ be an instance of EDGE MULTICUT with maximum degree Δ . We construct an instance $(G' = (V', E^+, E^0), k)$ of CORRELATION CLUSTERING as follows. We initialize G' with $V' = V$ and $E^+ = E, E^0 = \binom{V}{2} \setminus E$. For each terminal vertex $v \in V(T)$ let T_v denote the set of terminals $u \in V(T)$ that are in a terminal pair together with v and let $t_v^1, \dots, t_v^{|T_v|}$ be an arbitrary fixed ordering of T_v . For each terminal vertex $v \in V(T)$ we add the following gadget Q_v consisting of a chain of $|T_v|$ cliques of size $\Delta + 1$: For each $i \in \{1, \dots, |T_v|\}$ we have a clique Q_v^i of size $\Delta + 1$ associated with the vertex t_v^i , where each vertex of Q_v^1 is a real neighbor of v and for each $i \in \{1, \dots, |T_v| - 1\}$ every vertex in Q_v^i is a real neighbor of every vertex in Q_v^{i+1} . Note that $Q_v \cup \{v\}$ is $\Delta + 1$ -edge-connected, that is, after deleting up to Δ edges, every vertex in $Q_v \cup \{v\}$ is still connected to every other vertex in $Q_v \cup \{v\}$. For each terminal pair $(u, v) \in T$ each vertex in Q_v^i is a nonneighbor of exactly one vertex in Q_u^j and vice versa, where $u = t_v^i$ and $v = t_u^j$. In other words, $E^-(Q_v^i, Q_u^j)$ is a matching of size $\Delta + 1$ in G'^{-} . All other edges adjacent to some vertex in a clique Q_v^i are fuzzy edges.

► **Lemma 5.1.** (G, k) is a yes-instance of EDGE MULTICUT if and only if (G', k) is a yes-instance of CORRELATION CLUSTERING.

Proof. First, let (G, k) be a yes-instance with solution set S and let $C_1, \dots, C_\ell$ be the connected components of $G - S$. Consider the clustering $\mathcal{C} = \{C'_1, \dots, C'_\ell\}$, where each C'_i consists of the vertices in C_i together with all vertices of Q_v for each terminal vertex $v \in V(T)$ in C_i . Note that the only real edges between clusters in $\mathcal{C}$ are the edges in S . Moreover, since S is a solution for EDGE MULTICUT, no connected component in $G - S$ contains both vertices of a terminal pair, so no cluster in $\mathcal{C}$ contains a nonedge. Thus, $\text{cost}(\mathcal{C}) = |S| \leq k$ and (G', k) is a yes-instance.

Now, let (G', k) be a yes-instance with an optimal clustering $\mathcal{C}$ with $\text{cost}(\mathcal{C}) \leq k$. We show the following claim:

► **Claim 5.2.** For each $v \in V(T)$ all vertices of $Q_v \cup \{v\}$ are in the same cluster in $\mathcal{C}$ and no cluster in $\mathcal{C}$ contains vertices from gadgets $Q_v \cup \{v\}$ and $Q_u \cup \{u\}$ of a terminal pair $(u, v) \in T$.

Proof. Assume towards a contradiction that for some $v \in V(T)$ there are clusters $C_1, \dots, C_\ell$ that each contain vertices of $Q_v \cup \{v\}$. Consider the new clustering $\mathcal{C}' = (\mathcal{C} \setminus \{C_1, \dots, C_\ell\}) \cup \{Q_v \cup \{v\}, C_1 \setminus (Q_v \cup \{v\}), \dots, C_\ell \setminus (Q_v \cup \{v\})\}$, where $Q_v \cup \{v\}$ becomes its own cluster. Since $Q_v \cup \{v\}$ is $\Delta + 1$ -edge-connected, at least $\Delta + 1$ edges in $E^+(Q_v \cup \{v\})$ have to be deleted to obtain $\mathcal{C}'$, whereas disconnecting $Q_v \cup \{v\}$ from the other clusters in $\mathcal{C}'$ needs $\deg_G(v) \leq \Delta$ edge deletions. Thus, $\text{cost}(\mathcal{C}') < \text{cost}(\mathcal{C})$, a contradiction to the optimality of $\mathcal{C}$.

Now, assume towards a contradiction that some cluster $C \in \mathcal{C}$ contains the gadgets $Q_v \cup \{v\}$ and $Q_u \cup \{u\}$ for a terminal pair $(u, v) \in T$. Consider the new clustering $\mathcal{C}' = (\mathcal{C} \setminus \{C\}) \cup \{Q_v \cup \{v\}, C \setminus (Q_v \cup \{v\})\}$, where $Q_v \cup \{v\}$ becomes its own cluster. We have

$$\begin{aligned} \text{cost}(\mathcal{C}) - \text{cost}(\mathcal{C}') &= E^-(Q_v \cup \{v\}, C \setminus (Q_v \cup \{v\})) - E^+(Q_v \cup \{v\}, C \setminus (Q_v \cup \{v\})) \\ &\geq E^-(Q_v \cup \{v\}, Q_u \cup \{u\}) - \deg_G(v) \\ &\geq (\Delta + 1) - \Delta > 0 \end{aligned}$$

and thus a contradiction to the optimality of $\mathcal{C}$. ◁

By the above claim, only real edges between clusters in $\mathcal{C}$ contribute to $\text{cost}(\mathcal{C})$, since per construction nonedges are only present between vertices of gadgets Q_u and Q_v for some terminal pair (u, v) . Thus, if we consider the set $S = \{e \in E^+(C_1, C_2) \mid C_1, C_2 \in \mathcal{C}, C_1 \neq C_2\}$ of real edges between clusters in $\mathcal{C}$, we have $|S| \leq k$. Since for each terminal pair (u, v) the

gadgets $Q_u \cup \{u\}$ and $Q_v \cup \{v\}$ are in different clusters of $\mathcal{C}$, after deleting the edges of S in G , no connected component in $G - S$ contains both vertices of a terminal pair and thus S is a solution for (G, k) . $\blacktriangleleft$

We now describe a reduction from CORRELATION CLUSTERING where the nonedges induce a matching to CORRELATION CLUSTERING with the additional constraint that the closure and the degeneracy of $G^{(+)}$ are at most 2.

Construction. Let $(G = (V, E^+, E^0), k)$ be an instance of CORRELATION CLUSTERING where the nonedges induce a matching. We construct a new instance $(G' = (V', \tilde{E}^+, \tilde{E}^0), k)$ of CORRELATION CLUSTERING as follows. We subdivide each real edge $\{u, v\}$ by adding a new vertex x_{uv} and making $\{u, x_{uv}\}$ as well as $\{v, x_{uv}\}$ real edges. Moreover, all other edges incident with x_{uv} as well as the edge $\{u, v\}$ become fuzzy edges. Note that the set of nonedges did not change and thus the nonedges in G' also induce a matching. Furthermore, since no vertex pair in G' has more than one common real neighbor and after deleting each new vertex x_{uv} with real degree 2 all other vertices have no real neighbors anymore, G'^{+} has closure 2 and degeneracy 2.

► **Lemma 5.3.** *(G, k) is a yes-instance of CORRELATION CLUSTERING if and only if (G', k) is a yes-instance of CORRELATION CLUSTERING.*

Proof. Let (G, k) be a yes-instance with a clustering $\mathcal{C}$ of cost at most k . Consider the clustering $\mathcal{C}'$ for G' where for each cluster $C \in \mathcal{C}$ we have a cluster $C' \in \mathcal{C}'$ such that $C \subseteq C'$ and for each new vertex x_{uv} we have $x_{uv} \in C'(u)$ or $x_{uv} \in C'(v)$ arbitrarily. Note that the contribution of nonedges to $\text{cost}(\mathcal{C}')$ is the same as the contribution of nonedges to $\text{cost}(\mathcal{C})$. Now, if some real edge $\{u, x_{uv}\}$ contributes to $\text{cost}(\mathcal{C}')$, then this implies that $x_{uv} \in C'(v)$ and $C'(u) \neq C'(v)$, so in particular $C(u) \neq C(v)$ and $\{u, v\}$ contributes to $\text{cost}(\mathcal{C})$. Thus, $\text{cost}(\mathcal{C}') \leq \text{cost}(\mathcal{C}) = k$.

Conversely, let (G', k) be a yes-instance with a clustering $\mathcal{C}'$ of cost at most k . Note that we can assume that each vertex x_{uv} is in the cluster $C'(u)$ or $C'(v)$. Otherwise, since x_{uv} has only fuzzy neighbors except for u and v , we could put x_{uv} in the cluster $C'(u)$ instead and thus reduce the cost of the clustering by not deleting $\{u, x_{uv}\}$. Consider the clustering $\mathcal{C}$ for G where for each cluster $C' \in \mathcal{C}'$ we have a cluster $C \in \mathcal{C}$ with $C = C' \cap V$. Again, the contribution of nonedges to $\text{cost}(\mathcal{C})$ is the same as the contribution of nonedges to $\text{cost}(\mathcal{C}')$. Moreover, if some real edge $\{u, v\}$ contributes to $\text{cost}(\mathcal{C})$, then this implies $C(u) \neq C(v)$ and thus $C'(u) \neq C'(v)$. In particular, we have either $x_{uv} \notin C'(u)$ or $x_{uv} \notin C'(v)$ and thus either $\{u, x_{uv}\}$ or $\{v, x_{uv}\}$ contributes to $\text{cost}(\mathcal{C}')$. Hence, $\text{cost}(\mathcal{C}) \leq \text{cost}(\mathcal{C}') = k$. $\blacktriangleleft$

Since EDGE MULTICUT is NP-hard and does not admit a polynomial kernel for the parameter k , the reductions above which preserve the parameter value imply the following.

► **Theorem 1.3.** *CORRELATION CLUSTERING is NP-hard and does not admit a polynomial kernel for the solution size k , unless $\text{NP} \subseteq \text{coNP/poly}$, even if the nonedges form a matching and the graph consisting of the real edges has degeneracy 2 and closure 2.*

6 Conclusion

Motivated by the fact that CORRELATION CLUSTERING presumably does not admit a polynomial problem kernel for the solution cost parameter k , we continued the investigation of the influence of structural parameters of the fuzzy edge graph on the problem complexity. We showed that combining k with the degeneracy d or closure c of the fuzzy edge graph

leads to polynomial kernels, while cases where the real edges and nonedges have a rather restrictive local structure remain hard. A natural direction for further research would be to improve the sizes of the kernels and to find FPT-algorithms with good running times for the parameters $k + d$ and $k + c$. While our results do not rule out polynomial kernels for k together with the maximum degree Δ , the presented reduction from EDGE MULTICUT to CORRELATION CLUSTERING with bounded degeneracy and closure increases the maximum degree only by a constant factor. Thus, showing that EDGE MULTICUT does not admit a polynomial kernel for $k + \Delta$ would also transfer the result to CORRELATION CLUSTERING. We leave the (non)existence of such a kernel as an open problem.

References

- 1 Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Mach. Learn.*, 56(1-3):89–113, 2004. doi:10.1023/B:MACH.0000033116.57574.95.
- 2 Sebastian Böcker. A golden ratio parameterized algorithm for cluster editing. *J. Discrete Algorithms*, 16:79–89, 2012. doi:10.1016/J.JDA.2012.04.005.
- 3 Sebastian Böcker and Jan Baumbach. Cluster editing. In *The Nature of Computation. Logic, Algorithms, Applications - 9th Conference on Computability in Europe, CiE 2013, Milan, Italy, July 1-5, 2013. Proceedings*, volume 7921 of *Lecture Notes in Computer Science*, pages 33–44. Springer, 2013. doi:10.1007/978-3-642-39053-1_5.
- 4 Sebastian Böcker, Sebastian Briesemeister, Quang Bao Anh Bui, and Anke Truß. Going weighted: Parameterized algorithms for cluster editing. *Theor. Comput. Sci.*, 410(52):5467–5480, 2009. doi:10.1016/J.TCS.2009.05.006.
- 5 Hans L. Bodlaender, Michael R. Fellows, Pinar Heggernes, Federico Mancini, Charis Papadopoulos, and Frances A. Rosamond. Clustering with partial information. *Theor. Comput. Sci.*, 411(7-9):1202–1211, 2010. doi:10.1016/J.TCS.2009.12.016.
- 6 Nicolas Bousquet, Jean Daligault, and Stéphan Thomassé. Multicut is FPT. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 459–468. ACM, 2011. doi:10.1145/1993636.1993698.
- 7 Yixin Cao and Jianer Chen. Cluster editing: Kernelization based on edge cuts. *Algorithmica*, 64(1):152–169, 2012. doi:10.1007/S00453-011-9595-1.
- 8 Alain Cournier and Michel Habib. A new linear algorithm for modular decomposition. In *Trees in Algebra and Programming - CAAP'94, 19th International Colloquium, Edinburgh, UK, April 11-13, 1994, Proceedings*, volume 787 of *Lecture Notes in Computer Science*, pages 68–84. Springer, 1994. doi:10.1007/BFB0017474.
- 9 Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015. doi:10.1007/978-3-319-21275-3.
- 10 Marek Cygan, Stefan Kratsch, Marcin Pilipczuk, Michal Pilipczuk, and Magnus Wahlström. Clique cover and graph separation: New incompressibility results. *ACM Trans. Comput. Theory*, 6(2):6:1–6:19, 2014. doi:10.1145/2594439.
- 11 Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. *Theor. Comput. Sci.*, 361(2-3):172–187, 2006. doi:10.1016/J.TCS.2006.05.008.
- 12 Rodney G. Downey and Michael R. Fellows. *Fundamentals of Parameterized Complexity*. Texts in Computer Science. Springer, 2013. doi:10.1007/978-1-4471-5559-1.
- 13 Jacob Fox, Tim Roughgarden, C. Seshadhri, Fan Wei, and Nicole Wein. Finding cliques in social networks: A new distribution-free model. *SIAM Journal on Computing*, 49(2):448–464, 2020. doi:10.1137/18M1210459.
- 14 Jaroslav Garvardt, Christian Komusiewicz, and Nils Morawietz. When can cluster deletion with bounded weights be solved efficiently? *Discret. Appl. Math.*, 383:169–183, 2026. doi:10.1016/J.DAM.2025.12.028.

- 15 Jens Gramm, Jiong Guo, Falk Hüffner, and Rolf Niedermeier. Graph-modeled data clustering: Exact algorithms for clique generation. *Theory Comput. Syst.*, 38(4):373–392, 2005. doi:10.1007/S00224-004-1178-Y.
- 16 Jiong Guo. A more effective linear kernelization for cluster editing. *Theor. Comput. Sci.*, 410(8-10):718–726, 2009. doi:10.1016/J.TCS.2008.10.021.
- 17 Giuseppe F. Italiano, Athanasios L. Konstantinidis, and Charis Papadopoulos. Structural parameterization of cluster deletion. *Algorithmica*, 87(6):961–981, 2025. doi:10.1007/S00453-025-01303-1.
- 18 Lawqueen Kanesh, Jayakrishnan Madathil, Sanjukta Roy, Abhishek Sahu, and Saket Saurabh. Further exploiting c -closure for FPT algorithms and kernels for domination problems. *SIAM J. Disc. Math.*, 37(4):2626–2669, 2023. doi:10.1137/22M1491721.
- 19 Tomohiro Koana, Christian Komusiewicz, André Nichterlein, and Frank Sommer. Covering Many (Or Few) Edges with k Vertices in Sparse Graphs. In *Proceedings of the 39th International Symposium on Theoretical Aspects of Computer Science, STACS 2022, Marseille, France (Virtual Conference), March 15-18 2022*, volume 219 of *LIPICs*, pages 42:1–42:18, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPICs.STACS.2022.42.
- 20 Tomohiro Koana, Christian Komusiewicz, and Frank Sommer. Exploiting c -closure in kernelization algorithms for graph problems. *SIAM J. Discret. Math.*, 36(4):2798–2821, 2022. doi:10.1137/21M1449476.
- 21 Tomohiro Koana, Christian Komusiewicz, and Frank Sommer. Computing dense and sparse subgraphs of weakly closed graphs. *Algorithmica*, 85(7):2156–2187, 2023. doi:10.1007/S00453-022-01090-Z.
- 22 Tomohiro Koana, Christian Komusiewicz, and Frank Sommer. Essentially tight kernels for (weakly) closed graphs. *Algorithmica*, 85(6):1706–1735, 2023. doi:10.1007/S00453-022-01088-7.
- 23 Tomohiro Koana and André Nichterlein. Detecting and enumerating small induced subgraphs in c -closed graphs. *Disc. Appl. Math.*, 302:198–207, 2021. doi:10.1016/j.dam.2021.06.019.
- 24 Christian Komusiewicz and Johannes Uhlmann. Alternative parameterizations for cluster editing. In *SOFSEM 2011: Theory and Practice of Computer Science - 37th Conference on Current Trends in Theory and Practice of Computer Science, Nový Smokovec, Slovakia, January 22-28, 2011. Proceedings*, volume 6543 of *Lecture Notes in Computer Science*, pages 344–355. Springer, 2011. doi:10.1007/978-3-642-18381-2_29.
- 25 Christian Komusiewicz and Johannes Uhlmann. Cluster editing with locally bounded modifications. *Discret. Appl. Math.*, 160(15):2259–2270, 2012. doi:10.1016/J.DAM.2012.05.019.
- 26 Mirko Krivánek and Jaroslav Morávek. NP-hard problems in hierarchical-tree clustering. *Acta Informatica*, 23(3):311–323, 1986. doi:10.1007/BF00289116.
- 27 Dániel Marx and Igor Razgon. Fixed-parameter tractability of multicut parameterized by the size of the cutset. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 469–478. ACM, 2011. doi:10.1145/1993636.1993699.
- 28 Ross M. McConnell and Jeremy P. Spinrad. Linear-time modular decomposition and efficient transitive orientation of comparability graphs. In *Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. 23-25 January 1994, Arlington, Virginia, USA*, pages 536–545. ACM/SIAM, 1994. URL: <http://dl.acm.org/citation.cfm?id=314464.314641>.
- 29 Ron Shamir, Roded Sharan, and Dekel Tsur. Cluster graph modification problems. *Discrete Applied Mathematics*, 144(1):173–182, 2004. Discrete Mathematics and Data Mining. doi:10.1016/j.dam.2004.01.007.
- 30 René van Bevern, Vincent Froese, and Christian Komusiewicz. Parameterizing edge modification problems above lower bounds. *Theory Comput. Syst.*, 62(3):739–770, 2018. doi:10.1007/S00224-016-9746-5.