Journal of Computer and System Sciences 80 (2014) 1430–1447

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Tight bounds for parameterized complexity of Cluster Editing with a small number of clusters ☆

Fedor V. Fomin [a], Stefan Kratsch [b], Marcin Pilipczuk [c], Michał Pilipczuk [a,*], Yngve Villanger [a]

[a] Department of Informatics, University of Bergen, Norway
[b] Technical University Berlin, Germany
[c] Institute of Informatics, University of Warsaw, Poland

ARTICLE INFO

Article history:
Received 3 May 2013
Received in revised form 3 January 2014
Accepted 8 April 2014
Available online 16 April 2014

Keywords:
Cluster editing
Correlation clustering
Parameterized complexity
Subexponential-time algorithms
Exponential-time hypothesis

ABSTRACT

In the Cluster Editing problem, also known as Correlation Clustering, we are given an undirected n-vertex graph G and a positive integer k. The task is to decide if G can be transformed into a cluster graph, i.e., a disjoint union of cliques, by changing at most k adjacencies, i.e. by adding/deleting at most k edges. We give a subexponential-time parameterized algorithm that in time $2^{\mathcal{O}(\sqrt{pk})} + n^{\mathcal{O}(1)}$ decides whether G can be transformed into a cluster graph with exactly p cliques by changing at most k adjacencies. Our algorithmic findings are complemented by the following tight lower bound on the asymptotic behavior of our algorithm. We show that unless ETH fails, for any constant $0 < \sigma \leq 1$, there is $p = \Theta(k^\sigma)$ such that there is no algorithm deciding in time $2^{o(\sqrt{pk})} \cdot n^{\mathcal{O}(1)}$ whether G can be transformed into a cluster graph with at most p cliques by changing at most k adjacencies.

1. Introduction

Cluster editing, also known as *clustering with qualitative information* or *correlation clustering*, is the problem to cluster objects based only on the qualitative information concerning similarity between their pairs. For every pair of objects we have a binary indication whether they are similar or not. The task is to find a partition of the objects into clusters minimizing the number of similarities between different clusters and non-similarities inside of clusters. The problem was introduced by Ben-Dor, Shamir, and Yakhini [6] motivated by problems from computational biology, and, independently, by Bansal, Blum, and Chawla [5], motivated by machine learning problems concerning document clustering according to similarities. The correlation version of clustering was studied intensively, including [1,3,4,15,16,29,39].

☆ A preliminary version of this work [26] has appeared in the Proceedings of the 30th International Symposium on Theoretical Aspects of Computer Science, STACS 2013. The first and the fourth authors have received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement No. 267959. This work has been done while the second author was at the Utrecht University, the Netherlands, and supported by the Dutch Research Foundation (NWO). The third author is supported by the National Science Centre grant N206 567140 and the Foundation For Polish Science.

* Corresponding author.
E-mail addresses: fomin@ii.uib.no (F.V. Fomin), stefan.kratsch@tu-berlin.de (S. Kratsch), malcin@mimuw.edu.pl (M. Pilipczuk), michal.pilipczuk@ii.uib.no (M. Pilipczuk), yngve.villanger@ii.uib.no (Y. Villanger).

http://dx.doi.org/10.1016/j.jcss.2014.04.015
0022-0000/

The graph-theoretic formulation of the problem is the following. A graph K is a *cluster graph* if every connected component of K is a complete graph. Let $G = (V, E)$ be a graph; then $F \subseteq V \times V$ is called a *cluster editing set* for G if $G \triangle F = (V, E \triangle F)$ is a cluster graph. Here $E \triangle F$ is the symmetric difference between E and F. In the optimization version of the problem the task is to find a cluster editing set of the minimum size. Constant factor approximation algorithms for this problem were obtained in [1,5,15]. On the negative side, the problem is known to be NP-complete [39] and, as was shown by Charikar, Guruswami, and Wirth [15], also APX-hard.

Giotis and Guruswami [29] initiated the study of clustering when the maximum number of clusters that we are allowed to use is stipulated to be a fixed constant p. As observed by them, this type of clustering is well-motivated in settings where the number of clusters might be an external constraint that has to be met. It appeared that p-clustering variants posed new and non-trivial challenges. In particular, in spite of the APX-hardness of the general case, Giotis and Guruswami [29] gave a PTAS for this version of the problem.

A cluster graph G is called a *p-cluster graph* if it has exactly p connected components or, equivalently, if it is a disjoint union of exactly p cliques. Similarly, a set F is a *p-cluster editing set* of G, if $G \triangle F$ is a p-cluster graph. In parameterized complexity, correlation clustering and its restriction to bounded number of clusters were studied under the names CLUSTER EDITING and p-CLUSTER EDITING, respectively.

CLUSTER EDITING *Parameter: k.*
Input: A graph $G = (V, E)$ and a non-negative integer k.
Question: Is there a cluster editing set for G of size at most k?

p-CLUSTER EDITING *Parameters: p, k.*
Input: A graph $G = (V, E)$ and non-negative integers p and k.
Question: Is there a p-cluster editing set for G of size at most k?

The parameterized version of CLUSTER EDITING, and variants of it, were studied intensively [7,9–12,18,22,30,32,33,35,38]; in particular, there is a recent survey of Böcker and Baumbach [8]. The problem is solvable in time $\mathcal{O}(1.62^k + |V(G)| + |E(G)|)$ [7] and it has a kernel with $2k$ vertices [14,17] (see Section 2 for the definition of a kernel). On the negative side, Komusiewicz and Uhlmann [35] have shown that existence of a *subexponential-time parameterized algorithm*, i.e., with running time $2^{o(k)} \cdot |V(G)|^{\mathcal{O}(1)}$, would contradict the Exponential Time Hypothesis of Impagliazzo et al. [34].

As for the p-CLUSTER EDITING problem, Shamir et al. [39] have shown that the problem is NP-complete for every fixed $p \geq 2$. A kernel with $(p + 2)k + p$ vertices was given by Guo [31].

Our results A new and active direction in parameterized complexity is the pursuit of asymptotically tight bounds on the complexity of problems. In several cases, it is possible to obtain a complete analysis by providing matching lower (complexity) and upper (algorithmic) bounds. The most widely used complexity assumption for such tight lower bounds is the *Exponential Time Hypothesis* (*ETH*), which posits that no subexponential-time algorithms for k-CNF-SAT or CNF-SAT exist [34]. For more information about this "optimality program", we refer to a survey of Lokshtanov et al. [36] and to an appropriate section of the recent survey of Marx [37].

The complexity class SUBEPT defined by Flum and Grohe [23, Chapter 16] comprises all parameterized problems that are solvable in subexponential parameterized time, i.e., in time $2^{o(k)} \cdot n^{\mathcal{O}(1)}$ for inputs of length n and parameter k. Until very recently, the only problems known to be in the class SUBEPT were the problems with additional constraints on the input, like being a planar, H-minor-free, or tournament graph [2,19]. However, recent algorithmic developments indicate that the structure of the class SUBEPT is much more interesting than expected. It appears that some parameterized problems related to chordal graphs, like MINIMUM FILL-IN or CHORDAL GRAPH SANDWICH, are also in SUBEPT [28].

Based on the similarities with problems on tournaments, it had been conjectured by Cao and Chen [13] that CLUSTER EDITING also belongs to SUBEPT. Unfortunately, this conjecture has been recently disproved by Komusiewicz and Uhlmann [35]: CLUSTER EDITING does not admit a $2^{o(k)} \cdot |V(G)|^{\mathcal{O}(1)}$ algorithm unless ETH fails. We remark that in our study we have obtained the same result independently. Our reduction is very similar in principles to the one given by Komusiewicz and Uhlmann, however the graph in the obtained instance of CLUSTER EDITING has maximum degree 5, instead of 6 as is the case in [35]. Consequently, our reduction shows that CLUSTER EDITING cannot be solved in subexponential time even on graphs of maximum degree 5. We believe that this improvement is of minor importance, and hence we refrain from presenting this result in order not to reiterate already published material. An interested reader is invited to the arXiv version [25] of this work for details of our reduction.

It is therefore an interesting question whether stipulating the target number of clusters can lead to a better time complexity, as was the case for the approximation viewpoint. In this work we answer this question in affirmative, and we extend our study to show a tight multivariate analysis of the p-CLUSTER EDITING problem. Our main algorithmic result is the following.