Discrete Optimization 8 (2011) 2–17

Contents lists available at ScienceDirect

Discrete Optimization

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

Graph-based data clustering with overlaps☆

Michael R. Fellows [a], Jiong Guo [b], Christian Komusiewicz [c,*], Rolf Niedermeier [c], Johannes Uhlmann [c]

[a] PC Research Unit, Office of DVC (Research), Charles Darwin University, Darwin, Northern Territory 0909, Australia

[b] Universität des Saarlandes, Campus E 1.4, D-66123 Saarbrücken, Germany

[c] Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

ARTICLE INFO

Article history:
Received 3 November 2009
Received in revised form 25 August 2010
Accepted 10 September 2010
Available online 5 October 2010

Keywords:
Cluster graph modification problems
Forbidden subgraph characterization
NP-hardness
Fixed-parameter tractability
W[1]-hardness
Kernelization

ABSTRACT

We introduce overlap cluster graph modification problems where, other than in most previous works, the clusters of the target graph may overlap. More precisely, the studied graph problems ask for a minimum number of edge modifications such that the resulting graph consists of clusters (that is, maximal cliques) that may overlap up to a certain amount specified by the overlap number *s*. In the case of *s-vertex-overlap*, each vertex may be part of at most *s* maximal cliques; *s-edge-overlap* is analogously defined in terms of edges. We provide a complexity dichotomy (polynomial-time solvable versus NP-hard) for the underlying edge modification problems, develop forbidden subgraph characterizations of "cluster graphs with overlaps", and study the parameterized complexity in terms of the number of allowed edge modifications, achieving fixed-parameter tractability (in case of constant *s*-values) and parameterized hardness (in case of unbounded *s*-values).

1. Introduction

Graph-based data clustering is an important tool in exploratory data analysis [1–3]. The applications range from bioinformatics [4,5] to image processing [6]. The formulation as a graph-theoretic problem relies on the notion of a *similarity graph*, where vertices represent data items and an edge between two vertices expresses high similarity between the corresponding data items. Then, the computational task is to group vertices into clusters, where a *cluster* is nothing but a dense subgraph (typically, a clique). Following Ben-Dor et al. [4], Shamir et al. [2] initiated a study of graph-based data clustering in terms of *graph modification* problems. Here, the task is to modify (add or delete) as few edges of an input graph as possible to obtain a *cluster graph*, that is, a *vertex-disjoint* union of cliques. The corresponding problem is referred to as Cluster Editing. Numerous recent publications build on this concept of cluster graphs [7–15]. To uncover the *overlapping* community structure of complex networks in nature and society [16], however, the concept of cluster graphs so far fails to model that clusters may overlap. Consequently, it has been criticized explicitly for this lack of overlaps [11]. In this work, we introduce a graph-theoretic relaxation of the concept of cluster graphs by allowing, to a certain extent, overlaps between the clusters (which are cliques). We distinguish between "vertex-overlaps" and "edge-overlaps" and provide a thorough study of the corresponding cluster graph modification problems.

☆ An extended abstract of this paper appeared in the proceedings of the 15th International Computing and Combinatorics Conference (COCOON'09), volume 5609 in LNCS, pages 516–526, Springer, 2009. Main work was done while all authors were in Jena.

* Corresponding author.

E-mail addresses: michael.fellows@cdu.edu.au (M.R. Fellows), jguo@mmci.uni-saarland.de (J. Guo), c.komus@uni-jena.de (C. Komusiewicz), rolf.niedermeier@uni-jena.de (R. Niedermeier), johannes.uhlmann@uni-jena.de (J. Uhlmann).

1572-5286/$ – see front matter
doi:10.1016/j.disopt.2010.09.006

Table 1
Classical computational complexity of graph-based data clustering with overlaps. Herein, "NPh" means that the respective problem is NP-hard and "P" means that the problem can be solved in polynomial time.

	s-vertex-overlap	s-edge-overlap
Editing	NPh for $s \geq 1$	NPh for $s \geq 1$
Deletion	NPh for $s \geq 1$	NPh for $s \geq 1$
Addition	P for $s = 1$, NPh for $s \geq 2$	P for $s = 1$, NPh for $s \geq 2$

The two core concepts we introduce are *s-vertex-overlap* and *s-edge-overlap*, where in the first case we demand that every vertex in the cluster graph is contained in at most s maximal cliques and in the second case we demand that every edge is contained in at most s maximal cliques. By definition, 1-vertex-overlap means that the cluster graph is a vertex-disjoint union of cliques (that is, there is no overlap of the clusters and, thus, the corresponding graph modification problem is Cluster Editing). Based on these definitions, we study a number of edge modification problems (addition, deletion, editing) in terms of the two overlap concepts, generalizing and extending previous work that focussed on non-overlapping clusters.

Previous work

Perhaps the most extensively studied cluster graph modification problem is the NP-hard Cluster Editing, where one asks for a minimum number of edges to add or delete in order to transform the input graph into a disjoint union of cliques. Cluster Editing has been studied from a theoretical [17,18,10,12–15] as well as a practical side [8,11]. The majority of these works deals with the parameterized complexity of Cluster Editing, having led to efficient search-tree based [18,13] and polynomial-time kernelization [9,12–15] algorithms. One motivation of our work is drawn from these intensive studies, motivated by the practical relevance of Cluster Editing and related problems. As discussed above, however, Cluster Editing forces a sometimes too strict notion of cluster graphs by disallowing any overlap. To the best of our knowledge, relaxed versions of Cluster Editing and the cluster graph concept have been largely unexplored.[1] There are only two approaches studying overlapping cliques in the context of Cluster Editing that we are aware of. One was proposed by Barthélemy and Brucker [21] under the name t-Zahn Clustering, where the aim is to obtain by a minimum number of edge modifications a graph in which each pair of maximal cliques has at most $t - 1$ vertices in common. The base case $t = 1$ is thus equivalent to Cluster Editing. Among other things, Barthélemy and Brucker [21] showed that 2-Zahn Clustering is NP-hard. The model of Barthélemy and Brucker [21] allows, for constant t, for vertices and edges to be in an unbounded number of maximal cliques. In contrast, our model limits the number of maximal cliques that a vertex or clique is contained in, but already for constant s there can be maximal cliques that intersect in an unbounded number of vertices. The second approach was presented by Damaschke [10], who investigated the Twin Graph Editing problem, where the goal is to obtain a so-called *twin graph* (with a further parameter t specified as part of the input) with a minimum number k of edge modifications. A t-twin graph is a graph whose "critical clique graph" has at most t edges, where the critical clique graph is the representation of a graph obtained by keeping for each set of vertices with identical closed neighborhoods exactly one vertex. Roughly speaking, our model expresses a more local property of the target graph. The main result of Damaschke [10] is fixed-parameter tractability with respect to the combined parameter (t, k). We note that already for $s = 2$ our s-vertex-overlap model includes graphs whose twin graphs can have an unbounded number t of edges. Hence, s is not a function of t.

Our results

We provide a thorough study of the computational complexity of clustering with vertex and edge-overlaps, extending previous work on Cluster Editing and closely related problems. In particular, in terms of the overlap number s, we provide a complete complexity dichotomy (polynomial-time solvable versus NP-hard) of the corresponding edge modification problems, most of them turning out to be NP-hard (for an overview, see Table 1 in Section 4). For instance, whereas Cluster Editing restricted to only allowing edge additions (also known as Cluster Addition or 1-Vertex-Overlap Addition) is trivially solvable in polynomial time, 2-Vertex-Overlap Addition turns out to be NP-hard. We also study the parameterized complexity of clustering with overlaps. On the negative side, we show $W[1]$-hardness results with respect to the parameter "number of edge modifications" in the case of an unbounded overlap number s. On the positive side, we prove that the problems become fixed-parameter tractable for the combined parameter (s, k). This result is based on forbidden subgraph characterizations of the underlying overlap cluster graphs, that may be of independent graph-theoretic interest. In particular, it turns out that the "1-edge-overlap cluster graphs" are exactly the diamond-free graphs. Finally, we develop polynomial-time data reduction rules for two special cases. More precisely, we show an $O(k^4)$-vertex problem kernel for 1-Edge-Overlap Deletion and an $O(k^3)$-vertex problem kernel for 2-Vertex-Overlap Deletion, where in both cases k denotes the number of allowed edge deletions. We conclude in Section 7 with a number of open problems.

[1] Two recent exceptions are so-called s-plex cluster graphs [19] and (p, q)-cluster graphs [20].