A More Relaxed Model for Graph-Based Data Clustering: s-Plex Editing

Jiong Guo*, Christian Komusiewicz**, Rolf Niedermeier, and Johannes Uhlmann***

Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany
{jiong.guo,c.komus,rolf.niedermeier,johannes.uhlmann}@uni-jena.de

Abstract. We introduce the s-Plex Editing problem generalizing the well-studied Cluster Editing problem, both being NP-hard and both being motivated by graph-based data clustering. Instead of transforming a given graph by a minimum number of edge modifications into a disjoint union of cliques (Cluster Editing), the task in the case of s-Plex Editing is now to transform a graph into a disjoint union of so-called s-plexes. Herein, an s-plex denotes a vertex set inducing a (sub)graph where every vertex has edges to all but at most s vertices in the s-plex. Cliques are 1-plexes. The advantage of s-plexes for $s \geq 2$ is that they allow to model a more relaxed cluster notion (s-plexes instead of cliques), which better reflects inaccuracies of the input data. We develop a provably efficient and effective preprocessing based on data reduction (yielding a so-called problem kernel), a forbidden subgraph characterization of s-plex cluster graphs, and a depth-bounded search tree which is used to find optimal edge modification sets. Altogether, this yields efficient algorithms in case of moderate numbers of edge modifications.

1 Introduction

The purpose of a clustering algorithm is to group together a set of (many) objects into a relatively small number of clusters such that the elements inside a cluster are highly similar to each other whereas elements from different clusters have low or no similarity. There are numerous approaches to clustering and "there is no clustering algorithm that can be universally used to solve all problems" [16]. To solve data clustering, one prominent line of attack is to use graph theory based methods [14]. In this line, extending and complementing previous work on cluster graph modification problems, we introduce the new edge modification problem s-Plex Editing.

In the context of graph-based clustering, data items are represented as vertices and there is an edge between two vertices iff the interrelation between the

* Partially supported by the DFG, Emmy Noether research group PIAF, NI 369/4, and research project DARE, GU 1023/1.
** Supported by a PhD fellowship of the Carl-Zeiss-Stiftung.
*** Supported by the DFG, research project PABI, NI 369/7.

A. Goldberg and Y. Zhou (Eds.): AAIM 2009, LNCS 5564, pp. 226–239, 2009.

two corresponding items exceeds some threshold value. Clustering with respect to such a graph then means to partition the vertices into sets where each set induces a dense subgraph (that is, a *cluster*) of the input graph whereas there are no edges between the vertices of different clusters. In this scenario, the algorithmic task then typically is to transform the given graph into a so-called cluster graph by a minimum number of graph modification operations [14]. Herein, a *cluster graph* is a graph where all connected components form clusters and a graph modification is to insert or delete an edge. One of the most prominent problems in this context is the NP-hard Cluster Editing problem (also known as Correlation Clustering) [14, 2], where, given a graph G and an integer $k \geq 0$, one wants to transform G into a graph whose connected components all are cliques, using at most k edge insertions and deletions. In this work, with the NP-hard s-Plex Editing problem, we study a more relaxed and often presumably more realistic variant of Cluster Editing: Whereas in the case of Cluster Editing the clusters shall be cliques, in the case of s-Plex Editing we only demand them to be s-plexes. A vertex subset $S \subseteq V$ of a graph $G = (V, E)$ is called *s-plex* if the minimum vertex degree in the induced subgraph $G[S]$ is at least $|S| - s$. Note that a clique is nothing but a 1-plex. Replacing cliques by s-plexes for some integer $s \geq 2$ allows one to reflect the fact that most real-world data are somewhat "spurious" and so the demand for cliques may be overly restrictive in defining what a cluster shall be (also see [5] concerning criticism of the overly restrictive nature of the clique concept).

Problem formulation. In the following, we call a graph an *s-plex cluster graph* if all its connected components are s-plexes.

s-Plex Editing
Input: An undirected graph $G = (V, E)$ and an integer $k \geq 0$.
Question: Can G be modified by up to k edge deletions and insertions into an s-plex cluster graph?

Indeed, seen as an optimization problem, the goal is to minimize the number of edge editing operations. Note that 1-Plex Editing is the same as Cluster Editing. Compared to Cluster Editing, s-Plex Editing with $s \geq 2$ is a more flexible tool for graph-based data clustering: For increasing s, the number of edge modifications should decrease. This important advantage of s-Plex Editing reflects the observation that fewer edge modifications mean that we introduce fewer "errors" into our final cluster solution, because the computed s-plex cluster graph is closer to the original data. This is in accordance with the natural hypothesis that the less one perturbs the input graph the more robust and plausible the achieved clustering is (maximum parsimony principle, also see Böcker et al. [3] for making this point in terms of Cluster Editing). Figure 1 presents a simple example comparing Cluster Editing (that is, 1-Plex Editing) with 2-Plex Editing and 3-Plex Editing in terms of the (number of) necessary editing operations.

Previous work and motivation. The s-plex concept was introduced in 1978 by Seidman and Foster [13] in the context of social network analysis. Recently, a