ARTICLE IN PRESS
Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Approximate association via dissociation[☆]

Jie You [a,b], Jianxin Wang [a], Yixin Cao [b,*]

[a] School of Information Science and Engineering, Central South University, Changsha, China
[b] Department of Computing, Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received 27 October 2015
Received in revised form 25 August 2016
Accepted 7 November 2016
Available online xxxx

Keywords:
Association set (cluster vertex deletion)
Dissociation set
Cluster graph
Modular decomposition
Triangle-free graph

ABSTRACT

A vertex set X of a graph G is an association set if each component of $G - X$ is a clique, and a dissociation set if each of these cliques has only one or two vertices. We observe some special structures and show that if none of them exists, then the minimum association set problem can be reduced to the minimum weighted dissociation set problem. This yields the first nontrivial approximation algorithm for the association set problem, with approximation ratio 2.5. The reduction is based on a combinatorial study of modular decomposition of graphs free of these special structures. Further, a novel algorithmic use of modular decomposition enables us to implement our algorithm in $O(mn + n^2)$ time.

1. Introduction

A *cluster graph* comprises a family of disjoint cliques, of which each can be viewed as an *association*. Cluster graphs have thus served as an important model in the study of clustering objects based on their pairwise similarities, particularly in computational biology and machine learning [3]. If we represent each object with a vertex, and add an edge between two objects that are similar, we would expect a cluster graph. If this fails, a natural problem is then to find and exclude a minimum number of vertices such that the rest forms a cluster graph; we call it the *association set* problem. This problem has recently received significant interest from the community of parameterized computation, where it is more commonly called *cluster vertex deletion* [15,4]. The cardinality of a minimum association set of a graph is also known as its *distance to clusters*. It is one of the few meaningful structural parameters for dense graphs [10,9], in contrast with a multitude of structural parameters for sparse graphs, thereby providing another motivation for this line of research. For example, Bruhn et al. [5] recently showed that the boxicity problem (of deciding the minimum d such that a graph G can be represented as an intersection graph of axis-aligned boxes in the d-dimension Euclidean space) is fixed-parameter tractable parameterized by the distance to clusters.

The association set problem belongs to the family of vertex deletion problems studied by Yannakakis et al. [17,19]. The task in these problems is to delete the minimum number of vertices from a graph so that the remaining subgraph satisfies a certain hereditary property; recall that a graph property is *hereditary* if it is closed under taking induced subgraphs [17]. It is known that a hereditary property can be characterized by a (possibly infinite) set of forbidden induced subgraphs. In our case, the property is "being a cluster graph", and the forbidden induced subgraphs are P_3's (i.e., paths on three vertices). A trivial approximation algorithm of ratio 3 can be derived as follows. We search for induced P_3's, and we delete all its three

☆ Supported in part by the Hong Kong Research Grants Council (RGC) under grant 152261/16E and the National Natural Science Foundation of China (NSFC) under grants 61572414 and 61420106009.
* Corresponding author.
E-mail address: yixin.cao@polyu.edu.hk (Y. Cao).

http://dx.doi.org/10.1016/j.dam.2016.11.007

vertices if one is found. This trivial upper bound is hitherto the best known. Indeed, it is a simple application of Lund and Yannakakis's observation [19], which applies to all graph classes having finite forbidden induced subgraphs.

Our main result is the first nontrivial approximation algorithm for the association set problem, with ratio 2.5. As usual, n and m denote the numbers of vertices and edges respectively in the input graph.

Theorem 1.1. *There is an $O(mn + n^2)$-time approximation algorithm of ratio 2.5 for the association set problem.*

Our approach is to reduce the association set problem to the weighted dissociation set problem. Given a vertex-weighted graph, the *weighted dissociation set* problem asks for a set of vertices with the minimum weight such that its deletion leaves a graph of maximum degree 1 or 0. This problem was first studied by Yannakakis [26], who proved that its unweighted version is already NP-hard on bipartite graphs. Note that the problem asks for breaking all P_3's. In a triangle-free graph, every P_3 is induced, and thus the weighted version of the association set problem on triangle-free graphs is equivalent to the weighted dissociation set problem. It is easy to observe that for the association set problem, vertices in a twin class (i.e., whose vertices have the same closed neighborhood) are either fully contained in or disjoint from a minimum solution. This observation inspires us to transform the input graph G into a vertex-weighted graph Q by identifying each twin class of G with a vertex of Q whose weight is the cardinality of the corresponding twin class. We further observe that there are five small graphs such that if G has none of them as an induced subgraph, then Q either has a simple structure, hence trivially solvable, or is triangle-free, and can be solved using the ratio-2 approximation algorithm for the weighted dissociation set problem [23,22]. From the obtained solution for Q we can easily retrieve a solution for the original graph G. Since each of these five graphs has at most five vertices, of which at least two need to be deleted to make it free of induced P_3's, the approximation ratio 2.5 follows.

The main idea of this paper appears in the argument justifying the reduction from the (unweighted) association set problem to the weighted dissociation set problem. Indeed, we are able to provide a stronger algorithmic result that implies the aforementioned combinatorial result. We develop an efficient algorithm that detects one of the five graphs in G, solves the problem completely, or determines that Q is already triangle-free. Our principal tool is modular decomposition, of which a similar use was recently invented by the authors [18] in parameterized algorithms. It is worth noting that the basic observation on vertex deletion problems to graph properties with finite forbidden induced subgraphs has been used on both approximation and parameterized algorithms, by Lund and Yannakakis [19] and by Cai [6] respectively.

We would like to remark that for some vertex deletion problem, the trivial bound from Lund and Yannakakis [19] is already the best one can expect. For example, the vertex cover problem, with "being edgeless" the property and P_2 the forbidden induced subgraph, is NP-hard to approximate within ratio $2 - \epsilon$ for any positive constant ϵ, assuming the Unique Games Conjecture (UGC) [16]. Indeed, with the standard reduction [17], the same lower bound applies to our problem as well, and it is the only lower bound we are aware of. A natural question is then to close the gap between the upper and lower bounds. After a preliminary version of this work appeared in arXiv, Fiorini et al. [11] managed to improve the ratio to 7/3. The first part of their algorithm is similar as ours, with more small induced subgraphs taken into consideration, while their analysis, using the so-called "local ratio" technique, is quite different from ours.

Closely related is the cluster editing problem, which allows us to use, instead of vertex deletions, both edge additions and deletions [3]. Approximation algorithms of the cluster editing problem have been intensively studied, and the current best approximation ratio is 2.5 [2,8,1]. The cluster editing problem has a $2k$-vertex kernel [7], while it remains an open problem to find a linear-vertex kernel for the association set (cluster vertex deletion) problem.

2. Preliminaries

This paper will be only concerned with undirected and simple graphs. The vertex set and edge set of a graph G are denoted by $V(G)$ and $E(G)$ respectively. Without loss of generality, we assume throughout the paper that the input graph contains no isolated vertices (vertices of degree 0), hence $n = O(m)$. For $\ell \geqslant 3$, let P_ℓ and C_ℓ denote respectively an induced path and an induced cycle on ℓ vertices. A C_3 is also called a *triangle*. For a given set $\mathcal{F}$ of graphs, a graph G is $\mathcal{F}$-free if it contains no graph in $\mathcal{F}$ as an induced subgraph. When $\mathcal{F}$ consists of a single graph F, we use also F-*free* for short. For each vertex v in $V(G)$, its *neighborhood* and *closed neighborhood* are denoted by $N_G(v)$ and $N_G[v]$ respectively.

The weighted versions of the associated set problem and the dissociation set problem are formally defined as follows.

Associated set	
Input:	A vertex-weighted graph G.
Task:	To find a subset $X \subset V(G)$ of the minimum weight such that every component of $G - X$ is a clique.

Dissociation set	
Input:	A vertex-weighted graph G.
Task:	To find a subset $X \subset V(G)$ of the minimum weight such that every component of $G - X$ is a single vertex or a single edge.

Please cite this article in press as: J. You, et al., Approximate association via dissociation, Discrete Applied Mathematics (2016), http://dx.doi.org/10.1016/j.dam.2016.11.007