ISSN 1990-4789, Journal of Applied and Industrial Mathematics, 2016, Vol. 10, No. 3, pp. 341–348. © Pleiades Publishing, Ltd., 2016.
Original Russian Text © V.P. Il'ev, S.D. Il'eva, A.A. Navrotskaya, 2016, published in Diskretnyi Analiz i Issledovanie Operatsii, 2016, Vol. 23, No. 3, pp. 5–20.

Graph Clustering with a Constraint on Cluster Sizes

V. P. Il'ev[1,2*], S. D. Il'eva[2], and A. A. Navrotskaya[1,2]**

[1] *Sobolev Institute of Mathematics, pr. Akad. Koptyuga 4, Novosibirsk, 630090 Russia*

[2] *Dostoevsky Omsk State University, pr. Mira 55-A, Omsk, 644077 Russia*

Received December 28, 2015; in final form, March 28, 2016

Abstract—A graph clustering problem is under study (also known as the graph approximation problem) with a constraint on cluster sizes. Some new approximation algorithm is presented for this problem, and performance guarantee of the algorithm is obtained. It is shown that the problem belongs to the class *APX* for every fixed p, where p is the upper bound on the cluster sizes.

DOI: 10.1134/S1990478916030042

Keywords: *clustering, approximation, graph, approximation algorithm, performance guarantee*

INTRODUCTION

Clustering is the problem of grouping a set of objects so that objects in each group (called a *cluster*) are more similar to each other than to those in other clusters. There are several suggestions for the *measure of similarity* between two clusterings. One of the most visual formalizations of clustering some related objects is the graph approximation problem which is a version of graph clustering [17]. In this problem, the structure of relations between the objects is given by an undirected graph whose vertices are in one-to-one correspondence to the objects and whose edges connect similar objects with sufficiently many identical attributes. It is required to divide the original set of objects into some pairwise nonoverlapping subsets (*clusters*) minimizing the number of connections between the clusters as well as the number of missing links in the clusters. The number of clusters can be given, bounded, or undefined. The statements and interpretations of the graph approximation problem can be found in [7, 8, 11, 12, 18–20].

Section 1 addresses the three well-known versions of the graph approximation problem which are the formalizations of clustering some related objects. In this section, we briefly overview of available results on computational complexity and approximability of these problems. In Section 2, a relatively new problem of graph clustering is considered with some bounds on the sizes of clusters. This problem is NP-hard. A polynomially solvable case of the problem is given. In Section 3, we propose an approximation algorithm for the problem in which the cluster sizes are bounded above by a given number $p \geq 2$, with the achievable guaranteed accuracy estimate $\lfloor (p-1)^2/2 \rfloor + 1$. Thus, we show that the problem of graph clustering with constraints on the cluster sizes belongs to the class *APX* for every fixed p.

1. STATEMENTS OF THE PROBLEMS. AN OVERVIEW OF AVAILABLE RESULTS

We consider only *ordinary graphs*; i.e., the graphs without loops and multiple edges. An ordinary graph is called a *cluster graph* if its every connected component is a complete graph [18]. Let $\mathcal{M}(V)$ be the set of all cluster graphs on the set of vertices V, let $\mathcal{M}_k(V)$ be the set of all cluster graphs on the vertex set V having exactly k nonempty connected components, and let $\mathcal{M}_{1,k}(V)$ be the set of all cluster graphs on V having at most k connected components, $2 \leqslant k \leqslant |V|$.

* E-mail: iljev@mail.ru
** E-mail: nawrocki@ya.ru

If $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are ordinary graphs both on the set of vertices V then the *distance* $d(G_1, G_2)$ between them is defined as

$$d(G_1, G_2) = |E_1 \Delta E_2| = |E_1 \setminus E_2| + |E_2 \setminus E_1|;$$

i.e., $d(G_1, G_2)$ is the number of distinct edges in G_1 and G_2.

In the 1960s–1980s the following three graph approximation problems were under study which can be considered as different formalizations of the graph clustering problem [4, 8, 9, 19, 20]:

Problem A. Given an ordinary graph $G = (V, E)$, find a graph $M^* \in \mathcal{M}(V)$ such that

$$d(G, M^*) = \min_{M \in \mathcal{M}(V)} d(G, M).$$

Problem $\mathbf{A_k}$. Given an ordinary graph $G = (V, E)$ and an integer k, $2 \leqslant k \leqslant |V|$, find a graph $M^* \in \mathcal{M}_k(V)$ such that

$$d(G, M^*) = \min_{M \in \mathcal{M}_k(V)} d(G, M).$$

Problem $\mathbf{A_{1,k}}$. Given an ordinary graph $G = (V, E)$ and an integer k, $2 \leqslant k \leqslant |V|$, find a graph $M^* \in \mathcal{M}_{1,k}(V)$ such that

$$d(G, M^*) = \min_{M \in \mathcal{M}_{1,k}(V)} d(G, M).$$

In the sequel, the graph approximation problem was repeatedly and independently rediscovered and studied under various names (for example, Correlation Clustering [11] and Cluster Editing [12, 18]).

The first theoretical results related to the graph approximation problem were obtained in the 1960s–1970s. In [20] (1964), Problem **A** was under study for the graphs of some special form. In 1971, Fridman [8] defined the first polynomially solvable case of Problem **A**. He showed that every Problem **A** for a graph without triangles reduces to the construction of maximum matching in it.

In [16] (1986), it was shown that Problem **A** is NP-hard, but this article remained unnoticed. In [11] and independently in [18] (2004), the NP-hardness of Problem **A** was proved. In [18], it was also proved that Problem $\mathbf{A_k}$ is NP-hard for every fixed $k \geqslant 2$, and a more simple proof of this result was published in [15] (2006). In the same year, Ageev, Il'ev, Kononov, and Talevnin [1] independently proved that Problems $\mathbf{A}_2$ and $\mathbf{A_{1,2}}$ are NP-hard already for cubic graphs. Therefore, we have that all above-mentioned graph approximation problems are NP-hard, including Problem $\mathbf{A}_{1,k}$.

Thus, in [1, 11, 15, 16, 18] the following is proved:

Theorem 1. *Problem* **A** *is* NP*-hard. Problems* $\mathbf{A_k}$ *and* $\mathbf{A_{1,k}}$ *are* NP*-hard for every fixed* $k \geqslant 2$ *and* NP*-hard for cubic graphs if* $k = 2$.

In [11], a simple 3-approximation algorithm for Problem $\mathbf{A}_{1,2}$ was proposed. In [1], the existence was proved of a randomized polynomial time approximation scheme for Problems $\mathbf{A}_{1,2}$, and in [15], a randomized polynomial time approximation scheme was proposed for Problem $\mathbf{A}_k$ (for every fixed $k \geqslant 2$). Pointing out that the complexity of the polynomial time approximate scheme of [15] deprives it of practical perspectives of using, Coleman, Saunderson and Wirth [14], 2008, proposed some 2-approximation algorithm for Problem $\mathbf{A}_{1,2}$, applying a local search procedure to a feasible solution obtained by the 3-approximation algorithm of [11]. For Problem $\mathbf{A_2}$ an approximate algorithm was proposed in [3] with the tight performance guarantee $3 - 6/|V|$.

As regards to Problem **A**, it was shown in 2005 [13] that Problem **A** is *APX*-hard, and a 4-approximation algorithm was developed. In 2008, some 2.5-approximation algorithm for Problem **A** was presented in [10].

2. THE CLUSTERING PROBLEM WITH RESTRICTION ON THE CLUSTER SIZES

In contrast to Section 1, where the restrictions were imposed on the number of clusters, we now discuss the problem of clustering the interconnected objects with bounded cluster sizes.

Let $\mathcal{M}^{1,p}(V)$ be the set of all cluster graphs on V such that the size of each connected component is at most some integer p, $2 \le p \le |V|$. We say that the *cluster graph belongs to* $\mathcal{M}^{p}(V)$ if the size of each of its connected components is equal to p.

Problem $\mathbf{A^{1,p}}$. Given an n-vertex graph $G = (V, E)$ and an integer p, find $M^* \in \mathcal{M}^{1,p}(V)$ such that

$$d(G, M^*) = \min_{M \in \mathcal{M}^{1,p}(V)} d(G, M).$$

Problem $\mathbf{A^{p}}$. Given a graph $G = (V, E)$ such that $|V| = pq$, where p and q are positive integers, find $M^* \in \mathcal{M}^{p}(V)$ such that

$$d(G, M^*) = \min_{M \in \mathcal{M}^{p}(V)} d(G, M).$$

In [11], in the proof of NP-hardness of Problem $\mathbf{A}$ without any constraints on the number and sizes of clusters, it is actually shown that Problem $\mathbf{A^{1,3}}$ NP-hard. In [5], the following is proved:

Theorem 2 [5]. *Problems $\mathbf{A}^{1,p}$ and $\mathbf{A}^{p}$ are* NP*-hard for every fixed $p \ge 3$.*

By the Turing reduction, the NP-complete problem of *partition into isomorphic subgraphs*, called TG12 in [2], is reduced to Problems $\mathbf{A^{1,p}}$ and $\mathbf{A^{p}}$.

In [5] the cases are also considered where the optimal solutions of Problems $\mathbf{A^{1,p}}$ and $\mathbf{A^{p}}$ can be found in polynomial time:

Theorem 3 [5]. *Problems $\mathbf{A^{1,2}}$ and $\mathbf{A^{2}}$ are polynomially solvable. Problems $\mathbf{A^{1,3}}$ on graphs without triangles are polynomially solvable.*

Show now that the latter result can be generalized to the case of an arbitrary p.

Note first that, for every optimal solution $M_A \in \mathcal{M}(V)$ of Problem $\mathbf{A}$ on the graph $G = (V, E)$ and for every optimal solution $M_{A^{1,p}} \in \mathcal{M}^{1,p}(V)$ of Problem $\mathbf{A^{1,p}}$ ($p \geqslant 2$) on G, we have

$$d(G, M_A) \leqslant d(G, M_{A^{1,p}}). \tag{1}$$

Further, for Problem $\mathbf{A}$ the following is proved by Fridman:

Lemma 1 [9]. *If a graph G contains no triangles then one of the optimal solutions of Problem $\mathbf{A}$ on G is an arbitrary graph whose edges form a maximum matching of G.*

Using Lemma 1 and taking (1) into account, we see that for every $p \geqslant 2$ one of the optimal solutions of Problem $\mathbf{A}^{1,p}$ on G, where G has no triangles, is an arbitrary graph whose edges form some maximum matching of G. Therefore, since the maximum matching in every graph can be found in polynomial time, we obtain

Theorem 4. *Problem $\mathbf{A^{1,p}}$ on the graphs without triangles is polynomially solvable for all $p \geqslant 2$.*

3. AN APPROXIMATION ALGORITHM FOR PROBLEM $\mathbf{A}^{1,p}$

For Problem $\mathbf{A^{1,3}}$ some polynomial time approximation algorithm with tight performance guarantee $3 - 6/|V|$ is proposed in [6]. In this section, we offer an approximation algorithm with the performance guarantee $\lfloor (p-1)^2/2 \rfloor + 1$ for Problem $\mathbf{A}^{1,p}$, $p \geqslant 3$.

Let $(V_1, \ldots, V_s)$ stand for a partition of V; i.e., $V = V_1 \cup \cdots \cup V_s$ and $V_i \cap V_j = \varnothing$ for every $i, j \in \{1, \ldots, s\}$, $i \neq j$. Let $M(V_1, \ldots, V_s)$ be a cluster graph of class $\mathcal{M}_s(V)$ in which V_i is the vertex set of the ith cluster, $i \in \{1, \ldots, s\}$.

3.1. The Case of a Connected Graph

Firstly, consider the case that $G = (V, E)$ is a connected graph.

Let $M^* = M(V_1, \ldots, V_l) \in \mathcal{M}^{1,p}(V)$ be an optimal solution of Problem $\mathbf{A^{1,p}}$ on the connected graph G, $1 \leqslant l \leqslant |V|$ and $p \geqslant 3$. Put $n_i = |V_i|$ for each $i \in \{1, \ldots, l\}$.

Lemma 2. *For each* $i \in \{1, \ldots, l\}$

$$\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor \leqslant \left\lfloor \frac{(p-1)^2}{2} \right\rfloor.$$

Proof. If n_i is even then

$$\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor = \frac{n_i(n_i-1)-n_i}{2} = \frac{n_i(n_i-2)}{2} = \left\lfloor \frac{(n_i-1)^2}{2} \right\rfloor.$$

So,

$$n_i(n_i-1)/2 - \lfloor n_i/2 \rfloor = \lfloor (n_i-1)^2/2 \rfloor, \qquad i \in \{1, \ldots, l\}.$$

Since $M^* \in \mathcal{M}^{1,p}(V)$, we have $n_i \leqslant p$ for all $i \in \{1, \ldots, l\}$ and, therefore,

$$n_i(n_i-1)/2 - \lfloor (n_i)/2 \rfloor \leqslant \lfloor (p-1)^2/2 \rfloor.$$

This completes the proof. □

Consider a cluster graph M' on the vertex set V built with the graph structure of $G = (V, E)$ and $M^* = (V_1, \ldots, V_l)$ by the following rule:

Rule 1. For each $i \in \{1, \ldots, l\}$, we divide V_i into $\lfloor n_i/2 \rfloor$ pairs in an arbitrary manner. For every pair of vertices $u, v \in V_i$, we define either one or two connected components of the graph M': if $uv \in E$ then $\{u, v\}$ is a bimodal component of M'; if $uv \notin E$ then $\{u\}$ and $\{v\}$ is the trivial components of M'. If n_i is odd then the vertex $w \in V_i$, remaining single, forms a trivial component $\{w\}$ of M'.

Note, that the edges of G corresponding to some two-element cluster graph M' form a matching in G (because $V_i \cap V_j = \varnothing$ for every $i, j \in \{1, \ldots, l\}$, $i \neq j$).

Let us estimate the distance between M' and G using $d(G, M^*)$:

Lemma 3. *Given a connected graph* G,

$$d(G, M') \leqslant d(G, M^*) + \sum_{i=1}^{l} \left(\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor \right), \tag{2}$$

where $M^* = M(V_1, \ldots, V_l) \in \mathcal{M}^{1,p}(V)$ *is an optimal solution of Problem* $\mathbf{A}^{1,p}$ *on* G, $M' \in \mathcal{M}^{1,2}(V)$ *is the cluster graph built by Rule* 1, *and* $n_i = |V_i|$ *with* $i \in \{1, \ldots, l\}$.

Proof. Let E_1 be a subset of edges of G whose ends are in different clusters of M^*:

$$E_1 = \{uv \in E \mid u \in V_i,\ v \in V_j,\ i \neq j\}.$$

By the definition of the distance,

$$d(G, M^*) \geqslant |E_1|. \tag{3}$$

We now estimate the distance between M' and G. By construction, M' is the subgraph of G. Consequently, $D(G, M')$ is equal to the number of edges in G whose endpoints belong to different clusters of M'. Obviously, this set includes the edges from E_1 and the edges within the clusters of M^* that are not included in the two-element clusters of M'.

For $i \in \{1, \dots, l\}$, the number of edges of G in the ith cluster of M^* does not exceed $n_i(n_i-1)/2$. The number of edges of G in the ith cluster of the graph M^*, which are not included in the two-element clusters of M', is at most $n_i(n_i-1)/2 - \lfloor n_i/2 \rfloor$. Hence,

$$d(G, M') \leqslant |E_1| + \sum_{i=1}^{l} \left(\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor \right).$$

Owing to (3), from here we have the required inequality (2). The proof of Lemma 3 is complete. □

Let $\mathcal{G}^* = \mathcal{G}^*(V_1, \dots, V_l)$ be a family of connected graphs consisting of l cliques on the sets $V_1, \dots, V_l$ arbitrarily connected by $l-1$ bridges. It is obvious that $d(G, M^*) = l-1$ for every $G \in \mathcal{G}^*$ and $d(G, M^*) \geqslant l$ for each connected graph $G \notin \mathcal{G}^*$.

Thus, for every connected graph G we have

$$d(G, M^*) \geqslant l - 1. \tag{4}$$

Lemma 4. *Let G be a connected graph. If $G \notin \mathcal{G}^*$ then*

$$d(G, M') \leqslant d(G, M^*) \left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1 \right),$$

where $M^* = M(V_1, \dots, V_l)$ is an optimal solution of Problem $\mathbf{A}^{1,p}$ on G, and M is a cluster graph constructed by Rule 1.

Proof. As already noted, for every connected graph G we have $d(G, M^*) = l-1$ in (4) if and only if $G \in \mathcal{G}^*$. Since by hypothesis of the lemma $G \notin \mathcal{G}^*$, we have $d(G, M^*) \geqslant l$.

Owing to Lemmas 3 and 2 together with inequality $l \leqslant d(G, M^*)$, we infer

$$\begin{aligned} d(G, M') &\leqslant d(G, M^*) + \sum_{i=1}^{l} \left(\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor \right) \leqslant d(G, M^*) + l \left\lfloor \frac{(p-1)^2}{2} \right\rfloor \\ &\leqslant d(G, M^*) + d(G, M^*) \left\lfloor \frac{(p-1)^2}{2} \right\rfloor = d(G, M^*) \left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1 \right). \end{aligned}$$

This completes the proof of Lemma 4. □

Consider the approximation algorithm for Problem $\mathbf{A}^{1,p}$ on a connected graph G:

Algorithm A_1:

Input: a connected graph $G = (V, E)$.

Step 1. Remove from G all bridges and denote the resultant graph by M_1. Go to Step 2.

Step 2. Construct a cluster graph $M_2 \in \mathcal{M}^{1,2}(V)$: Find a maximum matching in G; the found matching forms the bimodal components of the cluster of M_2, and the vertices that are not included in the matching form a trivial component. Go to Step 3.

Step 3. If $M_1 \in \mathcal{M}^{1,p}(V)$ and $d(G, M_1) \leqslant d(G, M_2)$ then put $M = M_1$; otherwise, $M = M_2$.

End.

Note, that the number of bimodal connected components in M_2 built at Step 2 by Algorithm A_1 coincides with the number of edges in the maximum matching in G. Since the edges corresponding to the bimodal components of M' constructed by Rule 1 also form a matching in G (not necessarily the largest), we have

$$d(G, M_2) \leqslant d(G, M'). \tag{5}$$

Theorem 5. *Consider a connected graph G. Then*

$$d(G,M) \leqslant d(G,M^*)\left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1\right), \tag{6}$$

where $M \in \mathcal{M}^{1,p}(V)$ is a cluster graph constructing by Algorithms A_1, and M^* is an optimal solution of Problem $\mathbf{A}^{1,p}$ on G.

Proof. Let start with the case that $G \notin \mathcal{G}^*$. By (5) and Lemma 4, we obtain the required inequalities

$$d(G,M) \leqslant d(G,M_2) \leqslant d(G,M') \leqslant d(G,M^*)\left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1\right).$$

Let now $G \in \mathcal{G}^*$; i.e., G consists of l cliques on the sets $V_1,\dots,V_l$ connected by bridges. The two cases are possible:

Case 1: $n_i \geqslant 3$ for every $i \in \{1,\dots,l\}$. In this case, the graph M_1, built at Step 1 of Algorithm A_1, coincides with M^* and belongs to $\mathcal{M}^{1,p}(V)$. Consequently, we have

$$d(G,M) = d(G,M_1) = d(G,M^*) < d(G,M^*)\left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1\right).$$

Case 2: $n_j \leqslant 2$ for some $j \in \{1,\dots,l\}$. In this case, M_1, built at Step 1 of Algorithm A_1, may fail to coincide with M^*. Without loss of generality, we assume that $j = l$. Then $n_l(n_l-1)/2 - \lfloor n_l/2 \rfloor = 0$ and, by Lemma 3,

$$d(G,M') \leqslant d(G,M^*) + \sum_{i=1}^{l}\left(\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor\right) = d(G,M^*) + \sum_{i=1}^{l-1}\left(\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor\right).$$

By Lemma 2, for every $i \in \{1,\dots,l\}$ we have

$$\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor \leqslant \left\lfloor \frac{(p-1)^2}{2} \right\rfloor;$$

therefore,

$$d(G,M') \leqslant d(G,M^*) + (l-1)\left\lfloor \frac{(p-1)^2}{2} \right\rfloor \leqslant d(G,M^*)\left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1\right).$$

Note that the last inequality holds because of (4). Hence, owing to (5), we obtain

$$d(G,M) \leqslant d(G,M_2) \leqslant d(G,M') \leqslant d(G,M^*)\left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1\right).$$

Theorem 5 is proved. □

3.2. The General Case

In this section we show that (6) remains valid for a disconnected graph. We describe the approximation algorithm for Problem $\mathbf{A^{1,p}}$ on an arbitrary graph G:

Algorithm A_2:

Input: An arbitrary graph $G = (V,E)$, where $G_i = (U_i,E_i)$ is the ith connected component of G, $i = 1,\dots,k$ for some $k \in \{1,\dots,|V|\}$.

Step 1. For every G_i build a graph $M_i \in \mathcal{M}^{1,p}(U_i)$, $i = 1,\dots,k$, by Algorithm A_1. Go to Step 2.

Step 2. Put $M = \bigcup_{i=1}^{k} M_i$.

End.

It is obvious that the cluster graph M built by Algorithm A_2 belongs to $\mathcal{M}^{1,p}(V)$.

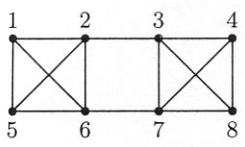

Fig. 1. The graph G_4.

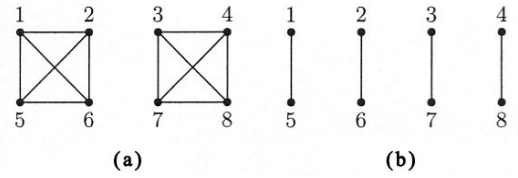

Fig. 2. The graphs M^* (a) and M (b).

Theorem 6. *Consider a graph G. Then*

$$d(G,M) \leqslant d(G,M^*)\big(\lfloor (p-1)^2/2 \rfloor + 1\big), \tag{7}$$

where $M \in \mathcal{M}^{1,p}(V)$ is the cluster graph built by Algorithm A_2, and M^* is the optimal solution of Problem $\mathbf{A}^{1,p}$ on G.

Proof. If G is a connected graph then (7) is valid by Theorem 5.

Let $G = (V,E)$ be a disconnected graph, and let $G_i = (U_i, E_i)$ denote the ith connected component of G, $i = 1,\dots,k$ for some k. By Theorem 5, the estimate (6) holds for each connected component; i.e.,

$$d(G_i,M_i) \leqslant d(G_i,M_i^*)\big(\lfloor (p-1)^2/2 \rfloor + 1\big),$$

where $M_i \in \mathcal{M}^{1,p}(U_i)$ is the cluster graph constructed by Algorithm A_2, and M_i^* is the optimal solution of Problem $\mathbf{A}^{\mathbf{1},\mathbf{p}}$ on G_i. Then

$$\begin{aligned} d(G,M) = \sum_{i=1}^{k} d(G_i,M_i) &\leqslant \sum_{i=1}^{k} d(G_i,M_i^*)\left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1\right) \\ &= \left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1\right) \sum_{i=1}^{k} d(G_i,M_i^*) = \left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1\right) d(G,M^*). \end{aligned}$$

Theorem 6 is proved. □

Corollary. *Problem $\mathbf{A}^{1,p}$ belongs to APX for every fixed $p \geqslant 3$.*

By the following assertion, estimate (7) is tight for even values of p:

Remark. For every even $p \geqslant 4$, there exists a graph G_p such that

$$d(G_p,M) = d(G_p,M^*)\big(\lfloor (p-1)^2/2 \rfloor + 1\big). \tag{8}$$

The graph G_p has $2p$ vertices and consists of the two cliques K_p connected by the two edges. Fig. 1 shows an example for the case $p = 4$.

As we see in Fig. 2, $d(G_4,M^*) = 2$, $d(G_4,M) = 10$, and equality (8) holds.

The cluster graph M for G_p, found by Algorithm A_2, is a maximum matching in association of the cliques K_p, and hence

$$d(G_p,M) = 2\big(p(p-1)/2 - p/2\big) + 2 = p^2 - 2p + 2.$$

Since $d(G_p,M^*) = 2$ and for even p we have

$$\lfloor (p-1)^2/2 \rfloor + 1 = \big(p^2 - 2p + 2\big)/2,$$

(8) holds for all even $p \geqslant 4$.

ACKNOWLEDGMENTS

The first and third authors (Sections 1, 2, and 3.1) were supported by the Russian Science Foundation (project no. 15–11–10009).

REFERENCES

1. A. A. Ageev, V. P. Il'ev, A. V. Kononov, and A. S. Talevnin, "Computational Complexity of a Graph Approximation Problem," Diskretn. Anal. Issled. Oper. Ser. 1, **13** (1), 3–11 (2006) [J. Appl. Indust. Math. **1** (1), 1–8 (2007)].
2. M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979; Mir, Moscow, 1982).
3. V. P. Il'ev, S. D. Il'eva, and A. A. Navrotskaya, "Approximation Algorithms for Graph Approximation Problems," Diskretn. Anal. Issled. Oper. **18** (1), 41–60 (2011) [J. Appl. Indust. Math. **5** (4), 569–581 (2011)].
4. V. P. Il'ev and G. Sh. Fridman, "On the Problem of Approximation by Graphs with a Fixed Number of Components," Dokl. Akad. Nauk SSSR **264** (3), 533–538 (1982) [Sov. Math. Dokl. **25** (3), 666–670 (1982)].
5. V. P. Il'ev and A. A. Navrotskaya, "Computational Complexity of the Problem of Approximation by Graphs with Connected Components of bounded Size," Prikl. Diskretn. Mat., No. 3, 80–84 (2011).
6. V. P. Il'ev and A. A. Navrotskaya, "An Approximate and Exact Solution to a Variant of the Problem of Clustering Interconnected Objects," in *Proceedings of the XI International Asian School-Seminar on Optimization Problems for Complex Systems, Cholpon-Ata, Kyrghyzstan, July 27—August 7, 2015* (Inst. Vychisl. Mat. Mat. Geofiz., Novosibirsk, 2015), pp. 278–283.
7. A. A. Lyapunov, "On the Structure and Evolution of Control Systems in Connection with the Theory of Classification," *Problems of Cybernetics*, Vol. 27 (Fizmatgiz, Moscow, 1973), pp. 7–18.
8. G. Sh. Fridman, "A Graph Approximation Problem," in Upravlyaemye Sistemy (Izd. Inst. Mat., Novosibirsk), **8**, 73–75 (1971).
9. G. Sh. Fridman, "Investigation of a Classifying Problem on Graphs," in *Methods of Modelling and Data Processing* (Nauka, Novosibirsk, 1976), pp. 147–177.
10. N. Ailon, M. Charikar, and A. Newman, "Aggregating Inconsistent Information: Ranking and Clustering," J. ACM **55** (5), 1–27 (2008).
11. N. Bansal, A. Blum, and S. Chawla, "Correlation Clustering," Mach. Learn. **56** (1–3), 89–113 (2004).
12. A. Ben-Dor, R. Shamir, and Z. Yakhimi, "Clustering Gene Expression Patterns," J. Comput. Biol. **6** (3–4), 281–297 (1999).
13. M. Charikar, V. Guruswami, and A. Wirth, "Clustering with Qualitative Information," J. Comput. Syst. Sci. **71** (3), 360–383 (2005).
14. T. Coleman, J. Saunderson, and A. Wirth, "A local-search 2-approximation for 2-correlation-clustering," in *Lecture Notes in Computer Sciences*, Vol. 5193: *Algorithms—ESA 2008 (Proceedings of the 16th Annual European Symposium on Algorithms, Karlsruhe, Germany, Sept. 15–17, 2008)* (Springer, Heidelberg, 2008), pp. 308–319.
15. I. Giotis and V. Guruswami, "Correlation Clustering with a Fixed Number of Clusters," Theory Comput. **2**, 249–266 (2006).
16. M. Křivánek and J. Morávek, "NP-Hard Problems in Hierarchical-Tree Clustering," Acta Inform. **23**, 311–323 (1986).
17. S. E. Schaeffer, "Graph Clustering," Comput. Sci. Rev. **1** (1), 27–64 (2007).
18. R. Shamir, R. Sharan, and D. Tsur, "Cluster Graph Modification Problems," Discrete Appl. Math. **144** (1–2), 173–182, (2004).
19. I. Tomescu, "Minimal Reduction of a Graph to a Union of Cliques," Discrete Math. **10** (1), 173–179 (1974).
20. C. T. Zahn, Jr., "Approximating Symmetric Relations by Equivalence Relations," J. Soc. Ind. Appl. Math. **12** (4), 840–847 (1964).