ISSN 2949-5598

ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ

Том 31 № 4 2024

Новосибирск
Издательство Института математики

ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Октябрь–декабрь 2024. Т. 31, № 4. С. 40–57

УДК 519.8 DOI: 10.33048/daio.2024.31.802

ПРИБЛИЖЁННЫЕ АЛГОРИТМЫ ДЛЯ ЗАДАЧ КЛАСТЕРИЗАЦИИ НА ГРАФАХ С КЛАСТЕРАМИ НЕБОЛЬШОГО РАЗМЕРА

В. П. Ильев[1,2,a], *С. Д. Ильева*[2], *А. В. Кононов*[1,b]

[1] Институт математики им. С. Л. Соболева,
пр. Акад. Коптюга, 4, 630090 Новосибирск, Россия
[2] Омский гос. университет им. Ф. М. Достоевского,
пр. Мира, 55-А, 644077 Омск, Россия
E-mail: [a] iljev@mail.ru, [b] alvenko@math.nsc.ru

Аннотация. В задачах кластеризации на графах требуется разбить множество вершин заданного графа на попарно непересекающиеся подмножества (называемые кластерами) так, чтобы минимизировать число рёбер между кластерами и число отсутствующих рёбер внутри кластеров. Мы рассматриваем вариант задачи, в котором размеры кластеров ограничены сверху натуральным числом s. Задача NP-трудна для любого фиксированного $s \geqslant 3$. Для этого варианта задачи предложены полиномиальные приближённые алгоритмы с гарантированными оценками точности, равными $5/3$ в случае $s = 3$ и 2 в случае $s = 4$. Доказано также, что при $s = 3$ задача APX-трудна. Ил. 5, библиогр. 20.

Ключевые слова: граф, кластеризация, NP-трудная задача, приближённый алгоритм, гарантированная оценка точности.

Введение

В задачах кластеризации требуется разбить заданное множество объектов на попарно непересекающиеся подмножества (кластеры) так, чтобы объекты в каждом кластере были более похожи друг на друга, чем на объекты из других кластеров.

Одной из наиболее наглядных формализаций задач кластеризации являются задачи кластеризации на графах [1]. В 1960–70-х гг. такие задачи изучались под именем задач аппроксимации графов. В этих задачах отношение сходства объектов задаётся посредством неориентированного графа, вершины которого взаимно однозначно соответствуют объектам, а рёбра соединяют похожие объекты. Цель состоит в том, чтобы разбить

множество вершин графа на попарно непересекающиеся подмножества (называемые кластерами) так, чтобы минимизировать число рёбер между кластерами и число отсутствующих рёбер внутри кластеров. Количество кластеров может быть задано, ограничено или заранее не определено. Различные варианты и интерпретации задачи аппроксимации графов содержатся в статьях [2–6].

Позже задачи аппроксимации графов неоднократно и независимо переоткрывались и изучались под разными наименованиями (Correlation Clustering, Cluster Editing, Graph Modification Problem и т. д.) [7–9]. В последнее время в зарубежной литературе за невзвешенной версией задачи закрепилось наименование Cluster Editing, в то время как на задачи, в которых рёбрам графа приписаны произвольные веса, обычно ссылаются как на Correlation Clustering [10–12].

Введём необходимые обозначения и определения.

Будем рассматривать только *обыкновенные* графы, т. е. графы без петель и кратных рёбер. Обыкновенный граф называется *кластерным*, если каждая его компонента связности является полным графом [9]. Обозначим через $\mathcal{M}(V)$ семейство всех кластерных графов на множестве вершин V, $\mathcal{M}_k(V)$ — семейство всех кластерных графов на V, имеющих ровно k компонент связности, $\mathcal{M}_{\leqslant k}(V)$ — семейство всех кластерных графов на множестве V, имеющих не более k компонент связности, $2 \leqslant k \leqslant |V|$.

Если $G_1 = (V, E_1)$ и $G_2 = (V, E_2)$ — обыкновенные графы с нумерованными вершинами на одном и том же множестве вершин V, то *расстояние* $d(G_1, G_2)$ между ними определяется как

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

т. е. $d(G_1, G_2)$ — число несовпадающих рёбер в графах G_1 и G_2.

В 1960–80-х гг. изучались следующие варианты задачи аппроксимации графа, которые эквивалентны задачам Cluster Editing (CE).

Задача CE. *Для произвольного графа $G = (V, E)$ найти граф $M^* \in \mathcal{M}(V)$ такой, что*

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

Задача $\mathbf{CE}_k$. *Для произвольного графа $G = (V, E)$ и натурального числа k, $2 \leqslant k \leqslant |V|$, найти граф $M^* \in \mathcal{M}_k(V)$ такой, что*

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

Задача $\mathbf{CE}_{\leqslant k}$. *Для произвольного графа $G = (V, E)$ и натурального числа k, $2 \leqslant k \leqslant |V|$, найти граф $M^* \in \mathcal{M}_{\leqslant k}(V)$ такой, что*

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

Задача CE NP-трудна, задачи CE_k и $\text{CE}_{\leqslant k}$ NP-трудны для любого фиксированного $k \geqslant 2$. Основные результаты о вычислительной сложности и приближённых алгоритмах с гарантированными оценками точности для этих задач содержатся в обзорах [13–15].

В этой работе рассматривается сравнительно новый вариант задачи, в котором размеры кластеров ограничены сверху натуральным числом s. Обозначим через $\mathcal{M}^{\leqslant s}(V)$ семейство всех кластерных графов на множестве вершин V, в которых размер любой компоненты связности не превосходит s, $2 \leqslant s \leqslant |V|$.

Задача $\text{CE}^{\leqslant s}$. *Для произвольного графа $G = (V, E)$ и натурального числа s, $2 \leqslant s \leqslant |V|$, найти граф $M^* \in \mathcal{M}^{\leqslant s}(V)$ такой, что*

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

При доказательстве NP-трудности задачи CE без каких-либо ограничений на число и размеры кластеров Бансал, Блюм и Чаула в [7] фактически доказали, что задача $\text{CE}^{\leqslant 3}$ NP-трудна. В 2011 г. Ильев и Навроцкая [16] доказали, что задача $\text{CE}^{\leqslant s}$ NP-трудна для любого фиксированного $s \geqslant 3$, а задача $\text{CE}^{\leqslant 2}$ полиномиально разрешима. В 2015 г. Пулео и Миленкович [10] предложили 6-приближённый алгоритм для задачи $\text{CE}^{\leqslant s}$, а в 2016 г. Ильев, Ильева и Навроцкая [17] опубликовали для этой задачи приближённый алгоритм, который является 3-приближённым для случая $s = 3$ и 5-приближённым для $s = 4$.

В настоящей работе предлагаются алгоритмы приближённого решения для случаев $s = 3$ и $s = 4$ с улучшенными гарантированными оценками точности. В 2023 г. эти результаты докладывались на международных конференциях MOTOR 2023 и OPTIMA 2023 [18, 19].

В разд. 1 приведены и доказаны необходимые вспомогательные утверждения. В разд. 2 предложен полиномиальный $\frac{5}{3}$-приближённый алгоритм для задачи $\text{CE}^{\leqslant 3}$, а также доказано, что эта задача APX-трудна. В разд. 3 рассматривается аналогичный полиномиальный 2-приближённый алгоритм для задачи $\text{CE}^{\leqslant 4}$. Гарантированные оценки точности обоих алгоритмов достижимы.

1. Вспомогательные утверждения

Далее будем использовать следующие термины и обозначения. Полный граф на r вершинах будем называть *r-кликой* и обозначать K_r, а обозначение Q_4 будем использовать для 4-вершинного графа с пятью рёбрами. Для произвольного графа $G = (V, E)$ и его подграфа $H = (U, D)$ обозначим через $\delta(H)$ разрез, определённый множеством вершин графа H, т. е. множество рёбер, ведущих из U в $V \setminus U$.

Докажем некоторые свойства оптимальных решений для задач $\text{CE}^{\leqslant 3}$ и $\text{CE}^{\leqslant 4}$.

Утверждение 1. *Пусть $G = (V, E)$ — произвольный граф и H — его подграф. Если $H = K_3$ и $|\delta(H)| \leqslant 2$, то существует оптимальное решение задачи $CE^{\leqslant 3}$ на графе G, которое содержит H как компоненту связности.*

Доказательство. Действительно, пусть вершины 3-клики H принадлежат разным компонентам оптимального решения $M^* \in \mathcal{M}^{\leqslant 3}(V)$. Тогда не менее двух рёбер клики H вносят вклад в $d(G, M^*)$. Удалим из графа M^* все вершины H вместе со всеми инцидентными им рёбрами и добавим вместо них клику H. Получим новый кластерный граф $M \in \mathcal{M}^{\leqslant 3}(V)$, причём $d(G, M) \leqslant d(G, M^*)$. Утверждение 1 доказано.

Аналогично доказывается

Утверждение 2. *Пусть $G = (V, E)$ — произвольный граф и H — его подграф. Если $H = K_2$ и $|\delta(H)| \leqslant 1$, то существует оптимальное решение задачи $CE^{\leqslant 3}$ на графе G, которое содержит H как компоненту связности.*

Из утверждений 1 и 2 следует, что можно произвести предварительную обработку (препроцессинг) данного графа $G = (V, E)$, последовательно находя подграфы, удовлетворяющие условиям утверждений 1 и 2, и добавляя их в решение задачи $\text{CE}^{\leqslant 3}$ с одновременным их удалением из G вместе со всеми рёбрами, инцидентными их вершинам.

Свойство 1. *Пусть G — граф, полученный из данного графа после препроцессинга. Если $H \subseteq G$ и $H = K_3$, то $|\delta(H)| \geqslant 3$.*

Свойство 2. *Пусть G — граф, полученный из данного графа после препроцессинга. Если $H \subseteq G$ и $H = K_2$, то $|\delta(H)| \geqslant 2$.*

Утверждение 3. *В задаче $CE^{\leqslant 3}$ существует оптимальное решение, являющееся подграфом данного графа.*

Утверждение 4. *В задаче $CE^{\leqslant 4}$ существует оптимальное решение, любая клика которого является кликой данного графа, за исключением, быть может, некоторых клик K_4, которые получены из Q_4-подграфов данного графа добавлением одного ребра.*

Доказательство. Докажем утверждение 4, утверждение 3 доказывается аналогично. Очевидно, что если клика $H = K_2$ входит в M^*, то $H \subseteq G$. Предположим, что клика $H = K_3 \subseteq M^*$ получена из некоторого 3-вершинного подграфа $F \subseteq G$ добавлением одного ребра e. Тогда вместо добавления e можем удалить из F другое ребро и получить новый кластерный граф $M' \in \mathcal{M}^{\leqslant 4}(V)$ с тем же значением целевой функции: $d(G, M') = d(G, M^*)$. Наконец, предположим, что клика $H = K_4 \subseteq M^*$

получена из некоторого 4-вершинного подграфа $F \subseteq G$ добавлением не менее двух рёбер. Тогда вместо их добавления можем удалить из F не более двух рёбер и получить кластерный граф $M' \in \mathcal{M}^{\leqslant 4}(V)$ такой, что $d(G, M') \leqslant d(G, M^*)$. Утверждение 4 доказано.

2. Задача CE$^{\leqslant 3}$

Опишем $\frac{5}{3}$-приближённый алгоритм для задачи CE$^{\leqslant 3}$ и докажем его гарантированную оценку точности. Будет также доказано, что задача CE$^{\leqslant 3}$ APX-трудна.

Пусть $\mathcal{F}$ — некоторое семейство графов. Множество попарно непересекающихся по вершинам подграфов произвольного графа G, каждый из которых изоморфен некоторому графу из $\mathcal{F}$, называется *$\mathcal{F}$-упаковкой* в G. Максимальная по включению $\mathcal{F}$-упаковка называется *максимальной*.

Обозначим через $\mathcal{K}_r$ максимальную $\{K_r\}$-упаковку в G и подграф графа G, равный объединению всех клик этой упаковки.

Рассмотрим алгоритм приближённого решения задачи CE$^{\leqslant 3}$. Пусть граф G получен из данного графа после препроцессинга, т. е. для него выполнены свойства 1 и 2.

Алгоритм 1. Приближённый алгоритм для задачи CE$^{\leqslant 3}$

Вход: граф $G = (V, E)$, обладающий свойствами 1 и 2.
Выход: кластерный граф $M \in \mathcal{M}^{\leqslant 3}(V)$.

1: $G' \leftarrow G$;
2: найти максимальную $\{K_3\}$-упаковку $\mathcal{K}_3$ в G';
3: $G' \leftarrow G' - \mathcal{K}_3$; ▷ Удалить все вершины $\mathcal{K}_3$ вместе со всеми инцидентными рёбрами.
4: найти наибольшее паросочетание $\mathcal{K}_2$ в G';
5: $\mathcal{K}_1 \leftarrow G' - \mathcal{K}_2$; ▷ $\mathcal{K}_1$ — множество изолированных вершин графа G'.
6: $M \leftarrow \mathcal{K}_3 \cup \mathcal{K}_2 \cup \mathcal{K}_1$;

Теорема 1. *Пусть $G = (V, E)$ — произвольный граф. Тогда*

$$\frac{d(G, M)}{d(G, M^*)} \leqslant \frac{5}{3}\,, \tag{1}$$

где M^* — оптимальное решение задачи CE$^{\leqslant 3}$ на графе G, M — кластерный граф, построенный алгоритмом 1.

Доказательство. Заметим, что если неравенство (1) будет доказано для графа, полученного из G после препроцессинга, то оно будет верно

и для графа G, поэтому далее без ограничения общности будем считать, что G — граф, полученный из данного графа после препроцессинга.

Из утверждения 3 следует, что существует оптимальное решение M^* задачи CE$^{\leqslant 3}$ на графе G, которое содержит только K_2- и K_3-подграфы графа G. Пусть $\mathcal{O}$ — семейство всех клик оптимального решения M^*, а $\mathcal{E}_i$ — семейство всех i-клик в $\mathcal{O}$, $i = 2, 3$. Обозначим через $N(\mathcal{O})$ и $N(\mathcal{A})$ число рёбер в $\mathcal{E}_2 \cup \mathcal{E}_3$ и $\mathcal{K}_2 \cup \mathcal{K}_3$ соответственно. Очевидно, что $N(\mathcal{O}) = 3|\mathcal{E}_3| + |\mathcal{E}_2|$, а из свойств 1 и 2 следует, что $d(G, M^*) \geqslant \frac{3|\mathcal{E}_3|+2|\mathcal{E}_2|}{2}$.

В [20] доказано, что

$$N(\mathcal{A}) \geqslant 2|\mathcal{E}_3| + |\mathcal{E}_2|. \tag{2}$$

Для полноты изложения приведём здесь собственное доказательство этого неравенства.

Инъективно сопоставим (назначим) каждой клике $C \in \mathcal{E}_2$ не менее одного ребра, а каждой клике $C \in \mathcal{E}_3$ — не менее двух рёбер из $\mathcal{K}_2 \cup \mathcal{K}_3$. Рассмотрим произвольную клику $K \in \mathcal{K}_3$ и три её вершины u, v, w.

(а) Пусть $C \in \mathcal{E}_3$ и $C = K$. Сопоставим клике C все три ребра клики K.

(б) Пусть $C \in \mathcal{E}_3 \cup \mathcal{E}_2$ и вершины u, v принадлежат C, а $w \notin C$. Сопоставим рёбра uv и uw клике C.

(в) Пусть $C \in \mathcal{E}_3 \cup \mathcal{E}_2$ и $\{w\} = K \cap C$. Назначим ребро vw клике C.

В результате получим, что если клика $C \in \mathcal{E}_3 \cup \mathcal{E}_2$ имеет i общих вершин с K, то i рёбер клики K сопоставлены C. При этом каждое такое ребро связано только с одной кликой $C \in \mathcal{E}_3 \cup \mathcal{E}_2$.

Далее, рассмотрим граф $G' = G - \mathcal{K}_3$. В G' уже нет треугольников. Обозначим через E^* множество рёбер в $G' \cap (\mathcal{E}_3 \cup \mathcal{E}_2)$. Тогда E^* — некоторое паросочетание в G'.

(г) Пусть $\mathcal{C} = \{C \in \mathcal{E}_3 \cup \mathcal{E}_2 \mid C \cap E^* \neq \varnothing\}$. Так как $\mathcal{K}_2$ — это наибольшее паросочетание в G', то $|E^*| \leqslant |\mathcal{K}_2|$. Следовательно, можем сопоставить каждой компоненте $C \in \mathcal{C}$ одно ребро из $\mathcal{K}_2$; если $C \in \mathcal{E}_3$, то это будет уже второе ребро, назначенное клике C.

Таким образом, не менее двух рёбер из $\mathcal{K}_2 \cup \mathcal{K}_3$ сопоставлены каждой клике $C \in \mathcal{E}_3$, и как минимум одно ребро из $\mathcal{K}_2 \cup \mathcal{K}_3$ каждой клике $C \in \mathcal{E}_2$. Тем самым неравенство (2) доказано.

Из (2) следует, что $d(G, M) \leqslant d(G, M^*) + |\mathcal{E}_3|$ и

$$\frac{d(G, M)}{d(G, M^*)} \leqslant 1 + \frac{|\mathcal{E}_3|}{d(G, M^*)} \leqslant 1 + \frac{2|\mathcal{E}_3|}{3|\mathcal{E}_3| + 2|\mathcal{E}_2|} \leqslant \frac{5}{3}.$$

Теорема 1 доказана.

Замечание 1. Оценка (1) достижима.

Теорема 2. *Задача CE$^{\leqslant 3}$ APX-трудна.*

Доказательство. Докажем, что задача CE$^{\leqslant 3}$ APX-трудна на графах максимальной степени 4.

В [20] доказано, что задача о $\{K_2, K_3\}$-упаковке APX-трудна на графах максимальной степени 4. $\{K_2, K_3\}$-упаковка в графе G — это множество попарно непересекающихся по вершинам подграфов G, каждый из которых изоморфен K_2 или K_3. Задача о $\{K_2, K_3\}$-упаковке состоит в отыскании максимальной по числу рёбер $\{K_2, K_3\}$-упаковки.

Пусть G — вход задач о $\{K_2, K_3\}$-упаковке и CE$^{\leqslant 3}$. Заметим, что любое допустимое решение задачи о $\{K_2, K_3\}$-упаковке является также допустимым решением задачи CE$^{\leqslant 3}$ (не считая изолированных вершин). Пусть $f(\mathrm{Sol})$ и $g(\mathrm{Sol})$ — значения целевых функций на допустимом решении Sol задач о $\{K_2, K_3\}$-упаковке и CE$^{\leqslant 3}$ соответственно. Очевидно, что $f(\mathrm{Sol}) = m - g(\mathrm{Sol})$, где m — число рёбер графа G. Значит, если Opt — оптимальное решение задачи о $\{K_2, K_3\}$-упаковке, то Opt есть также оптимальное решение задачи CE$^{\leqslant 3}$ и $f(\mathrm{Opt}) = m - g(\mathrm{Opt})$. Кроме того,

$$f(\mathrm{Opt}) - f(\mathrm{Sol}) = m - g(\mathrm{Opt}) - m + g(\mathrm{Sol}) = g(\mathrm{Sol}) - g(\mathrm{Opt}).$$

Пусть G — граф максимальной степени 4. Заметим, что $f(\mathrm{Opt}) \geqslant m/7$. В самом деле, обозначим через z число рёбер некоторого максимального паросочетания M в G. Каждое ребро паросочетания M смежно (т. е. имеет общую вершину) не более чем с 6 рёбрами графа G. Поскольку любое ребро графа G либо принадлежит Opt, либо смежно с некоторым ребром из Opt, то $m \leqslant 7z \leqslant 7f(\mathrm{Opt})$.

Предположим, что A — некоторый α-приближённый алгоритм для задачи CE$^{\leqslant 3}$, и пусть Sol — приближённое решение, найденное этим алгоритмом. По аналогии с утверждением 3 можно доказать, что тогда существует допустимое решение задачи CE$^{\leqslant 3}$ с тем же значением целевой функции, которое является подграфом графа G. Без ограничения общности будем считать, что Sol — именно такое решение. Тогда Sol является допустимым решением задачи о $\{K_2, K_3\}$-упаковке, откуда получаем

$$\begin{aligned} f(\mathrm{Sol}) = f(\mathrm{Opt}) + g(\mathrm{Opt}) - g(\mathrm{Sol}) &\geqslant m - \alpha g(\mathrm{Opt}) = \\ &= m - \alpha(m - f(\mathrm{Opt})) = \alpha f(\mathrm{Opt}) - (\alpha - 1)m \geqslant \\ &\geqslant \alpha f(\mathrm{Opt}) - 7(\alpha - 1)f(\mathrm{Opt}) = (7 - 6\alpha)f(\mathrm{Opt}). \end{aligned}$$

Таким образом, если существует α-приближённый алгоритм для задачи CE$^{\leqslant 3}$, то существует и $(7 - 6\alpha)$-приближённый алгоритм для задачи о $\{K_2, K_3\}$-упаковке. В частности, если бы для любого $\varepsilon > 0$ существовал $(1+\varepsilon)$-приближённый алгоритм для задачи CE$^{\leqslant 3}$, то существовал бы $(1 - 6\varepsilon)$-приближённый алгоритм для задачи о $\{K_2, K_3\}$-упаковке. Поскольку задача о $\{K_2, K_3\}$-упаковке APX-трудна на графах максимальной степени 4, задача CE$^{\leqslant 3}$ также APX-трудна. Теорема 2 доказана.

3. Задача CE$^{\leqslant 4}$

Опишем 2-приближённый алгоритм для задачи CE$^{\leqslant 4}$ и докажем его гарантированную оценку точности.

Рассмотрим алгоритм приближённого решения задачи CE$^{\leqslant 4}$.

Алгоритм 2. Приближённый алгоритм для задачи CE$^{\leqslant 4}$

Вход: произвольный граф $G = (V, E_G)$.
Выход: кластерный граф $M = (V, E_M) \in \mathcal{M}^{\leqslant 4}(V)$.

1: $G' \leftarrow G$;
2: найти максимальную $\{K_4\}$-упаковку $\mathcal{K}_4$ в G';
3: $G' \leftarrow G' - \mathcal{K}_4$; ▷ Удалить все вершины $\mathcal{K}_4$ вместе со всеми инцидентными рёбрами.
4: найти максимальную $\{K_3\}$-упаковку $\mathcal{K}_3$ в G';
5: $G' \leftarrow G' - \mathcal{K}_3$; ▷ Удалить все вершины $\mathcal{K}_3$ вместе со всеми инцидентными рёбрами.
6: найти максимальное паросочетание $\mathcal{K}_2$ в G';
7: $\mathcal{K}_1 \leftarrow G' - \mathcal{K}_2$; ▷ $\mathcal{K}_1$ — множество изолированных вершин графа G'.
8: $M \leftarrow \mathcal{K}_4 \cup \mathcal{K}_3 \cup \mathcal{K}_2 \cup \mathcal{K}_1$;

Далее будем рассматривать только те оптимальные решения задачи CE$^{\leqslant 4}$, что удовлетворяют утверждению 4. Пусть G — произвольный граф, M^* — оптимальное решение задачи CE$^{\leqslant 4}$ на G. Введём следующие обозначения:

- E_1 — множество рёбер графа G, которые не входят в M^*, но помещены в M алгоритмом 2 (т. е. $E_1 = (E_G \cap E_M) \setminus E_{M^*}$);
- $E_2 = E_{M^*} \setminus E_G$;
- E^* — множество рёбер графа G, которые не входят в M, но входят в M^* (т. е. $E^* = (E_G \cap E_{M^*}) \setminus E_M$).

Лемма 1. *Имеет место неравенство* $|E^*| \leqslant 2(|E_1| + |E_2|)$.

Доказательство. Рассмотрим мысленную процедуру пометки рёбер множества E^*, шаги которой соответствуют шагам алгоритма 2. Процедура присваивает метку каждому ребру из E^*, которое удаляется из текущего графа G' на очередном шаге алгоритма 2. Пометка рёбер осуществляется с помощью рёбер множеств E_1 и E_2. Каждое ребро $e \in E_1 \cup E_2$ даёт свои метки вида $e*$ не более чем двум смежным с e непомеченным рёбрам множества E^*. Изначально рёбра E^* не помечены.

Покажем сначала, что в множестве E_1 достаточно рёбер для пометки рёбер всех клик K_4, K_3 и K_2 графа M^*, являющихся подграфами

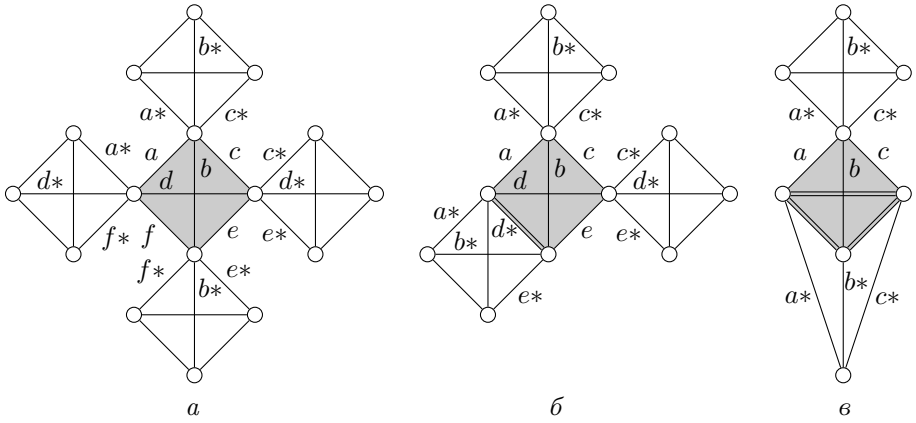

Рис. 1

графа G, и только некоторые рёбра клик $K_4 \subseteq M^*$, полученных из подграфов $Q_4 \subset G$, могут остаться непомеченными с использованием рёбер множества E_1 в момент удаления из текущего графа G'.

1. Заметим, что клика K_4, найденная алгоритмом 2 на шаге 2, может иметь 0, 1, 3 или 6 общих рёбер с кликами графа M^*. Тем самым только 6, 5 или 3 ребра любой клики $K_4 \subseteq M$ могут принадлежать множеству E_1. Рассмотрим эти случаи с иллюстрацией на рис. 1: серые тетраэдры — клики K_4, найденные алгоритмом 2; белые тетраэдры — клики графа M^* (некоторые из них могут быть заменены кликами K_3, K_2 или K_1). На всех рисунках двойной линией обозначены общие рёбра графов M и M^*.

(а) Пусть K_4 — клика, найденная алгоритмом 2 на шаге 2, все шесть рёбер a, b, c, d, e, f которой принадлежат множеству E_1. Тогда в графе G' не более 12 непомеченных рёбер множества E^* смежны с этими шестью рёбрами, они получат не более 12 меток вида $a*, b*, c*, d*, e*, f*$ (рис. 1а).

(б) Пусть K_4 — клика, найденная алгоритмом 2 на шаге 2, в которой пять рёбер a, b, c, d, e принадлежат множеству E_1. Тогда в графе G' не более 10 непомеченных рёбер множества E^* смежны с этими пятью рёбрами, они получат не более 10 меток вида $a*, b*, c*, d*, e*$ (рис. 1б).

(в) Пусть K_4 — клика, найденная алгоритмом 2 на шаге 2, три ребра a, b, c которой принадлежат множеству E_1. Тогда в графе G' не более 6 непомеченных рёбер множества E^* смежны с этими тремя рёбрами, они получат не более 6 меток вида $a*, b*, c*$ (рис. 1в).

2. Заметим, что клика K_3, найденная алгоритмом 2 на шаге 4, может иметь 0, 1 или 3 общих ребра с кликами графа M^*, поэтому только 3

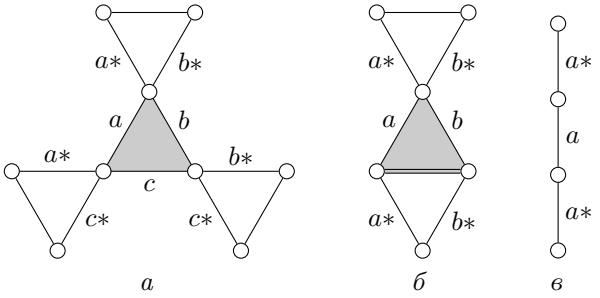

Рис. 2

или 2 ребра каждой клики $K_3 \subseteq M$ принадлежат множеству E_1. Рассмотрим эти случаи с иллюстрацией на рис. 2а, 2б: серые треугольники — клики, найденные алгоритмом 2, белые треугольники — клики графа M^*, в котором часть клик K_3 могут быть заменены кликами K_2 или K_1, а часть клик K_3 могут содержаться в кликах $K_4 \subset M^*$ в качестве подграфов.

(а) Пусть K_3 — клика, найденная алгоритмом 2 на шаге 4, все три ребра a, b, c которой принадлежат множеству E_1. Тогда в графе G' не более 6 непомеченных рёбер множества E^* смежны с этими тремя рёбрами, они получат не более 6 меток вида $a*, b*, c*$ (рис. 2а).

(б) Пусть K_3 — клика, найденная алгоритмом 2 на шаге 4, только два ребра a, b которой принадлежат множеству E_1. Тогда в графе G' не более 4 непомеченных рёбер множества E^* смежны с этими двумя рёбрами, они получат не более 4 меток вида $a*, b*$ (рис. 2б).

3. Заметим, что если ребро e клики K_2, найденной алгоритмом 2 на шаге 6, принадлежит множеству E_1, то не более двух рёбер множества E^* смежны с e в текущем графе G', они получат метки $a*$ (рис. 2в).

Перечислим теперь случаи, когда рёбер E_1 может оказаться недостаточно для пометки рёбер из E^*, удаляемых из текущего графа G' на очередном шаге алгоритма 2. В каждом из этих случаев рёбра из E^*, оставшиеся непомеченными рёбрами из E_1, принадлежат некоторому подграфу $Q_4 \subset G$ с недостающим ребром $x \in E_2 = E_{M^*} \setminus E_G$. Эти случаи иллюстрируются рис. 3 и 4, на которых серым обозначены клики, найденные алгоритмом 2, а пунктирной линией — ребро x.

Случай 1. На шаге 2 алгоритм 2 находит клику $K_4 \subset G'$, три ребра a, b, c которой инцидентны вершине 1 подграфа Q_4. Тогда $a, b, c \in E_1$, и они дают свои метки $a*, b*, c*$ рёбрам 12, 13 и 14 соответственно (возможно, это их вторые метки, а первые метки $a*, b*, c*$ были даны некоторым рёбрам, инцидентным другим концам рёбер a, b, c) (рис. 3а).

После удаления K_4 и всех инцидентных ей рёбер из G' следующий граф G' будет содержать рёбра $23, 34 \in Q_4$. Заметим, что по меньшей мере одно из этих рёбер должно принадлежать E^*. Возможны два случая.

(а) Только одно из рёбер 23, 34 принадлежит E^*. Без ограничения общности предположим, что $23 \in E^*$. Это означает, что алгоритм 2 включил в M некоторую клику, содержащую ребро 34. Если это клика K_4 или K_3, то она также содержит некоторое ребро d, инцидентное вершине 3. Заметим, что $d \in E_1$, и оно даёт свою метку $d*$ ребру 23.

Единственная ситуация, когда ребро 23 в момент его удаления из текущего графа G' может остаться без метки, полученной от рёбер множества E_1, следующая. Алгоритм 2 на шаге 6 находит клику $K_2 = \langle 34 \rangle$, и таким образом ребро 34 удаляется из графа G' вместе со всеми инцидентными рёбрами. Однако в этом случае $34 \notin E_1$, следовательно, ребро 23 остаётся непомеченным с помощью рёбер множества E_1. Во избежание такой ситуации процедура приписывает ребру 23 метку $x*$, где $x = 24 \in E_2$ (рис. 3б).

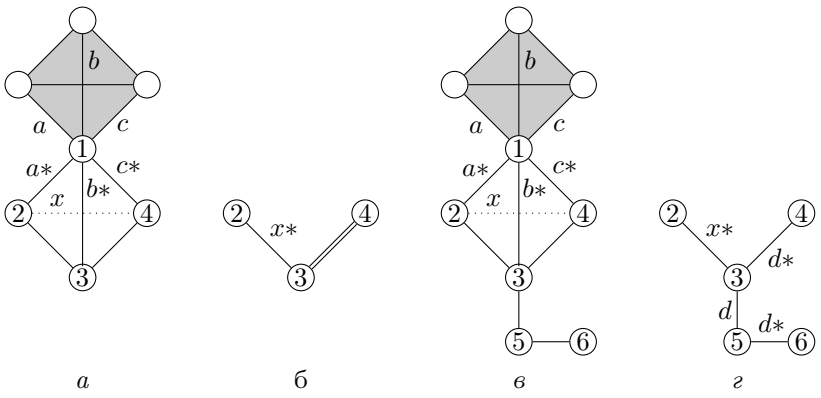

Рис. 3

(б) Оба ребра 23, 34 принадлежат E^*. Это означает, что алгоритм 2 включил в M некоторые рёбра, инцидентные одной, двум или трём вершинам множества $\{2, 3, 4\}$. Эти рёбра принадлежат E_1 и дают свои метки обоим рёбрам 23, 34, кроме единственной ситуации, когда алгоритм 2 на шаге 6 находит некоторую клику $K_2 = \langle 35 \rangle$, где $5 \notin \{2, 4\}$, и кроме того, в G' существует ребро $56 \in E^*$ (рис. 3в). Тогда $d = 35 \in E_1$ и d даёт свои метки $d*$ ребру 56 и одному из рёбер 23 и 34, скажем 34. В этой ситуации ребро 23 остаётся непомеченным с помощью рёбер множества E_1. Чтобы избежать такой ситуации, процедура назначает ребру 23 метку $x*$, где $x = 24 \in E_2$ (рис. 3г).

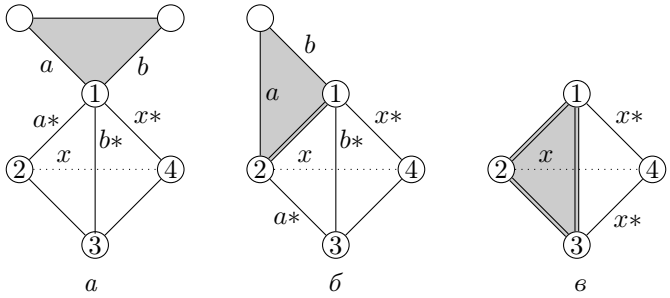

Рис. 4

Случай 2 отличается от случая 1 тем, что алгоритм 2 находит на шаге 4 в графе G' клику K_3 вместо K_4 (рис. 4а). Здесь, как и в случае 1, рёбра 12 и 13 получают метки $a*$ и $b*$, но ребро 14 не может быть помечено с помощью рёбер множества E_1. Таким образом, процедура присваивает ребру 14 метку $x*$, где $x = 24 \in E_2$.

И точно так же, чтобы избежать нежелательных ситуаций в следующем графе G', перечисленных в пп. (а), (б) случая 1, ребро 23 получает метку $x*$, где $x = 24 \in E_2$.

Случай 3 отличается от случая 2 тем, что клика K_3, найденная алгоритмом 2 на шаге 4, имеет общее ребро с Q_4, скажем 12 (рис. 4б). В этом случае, как и ранее, ребро 14 получает метку $x*$, где $x = 24 \in E_2$.

Случай 4. На шаге 4 алгоритм 2 находит клику K_3 с рёбрами 12, 13, 23 и удаляет её из графа G' вместе со всеми инцидентными рёбрами, но ещё до этого удаления процедура присваивает рёбрам 14 и 34 метки $x*$, где $x = 24 \in E_2$ (рис. 4в).

По завершении работы алгоритма 2 граф G' становится пустым. Поскольку все рёбра множества E^*, которые удалялись из текущего графа G', получили метки, а каждое ребро $e \in E_1$ или E_2 дало свои метки не более чем двум смежным с e непомеченным рёбрам множества E^*, то $|E^*| \leqslant 2(|E_1| + |E_2|)$. Лемма 1 доказана.

Пример работы алгоритма 2 представлен на рис. 5. Этот пример показывает достижимость следующей оценки.

Теорема 3. *Пусть $G = (V, E_G)$ — произвольный граф. Тогда*

$$\frac{d(G, M)}{d(G, M^*)} \leqslant 2, \tag{3}$$

где $M^* = (V, E_{M^*})$ — оптимальное решение задачи $CE^{\leqslant 4}$ на графе G, $M = (V, E_M)$ — кластерный граф, построенный алгоритмом 2.

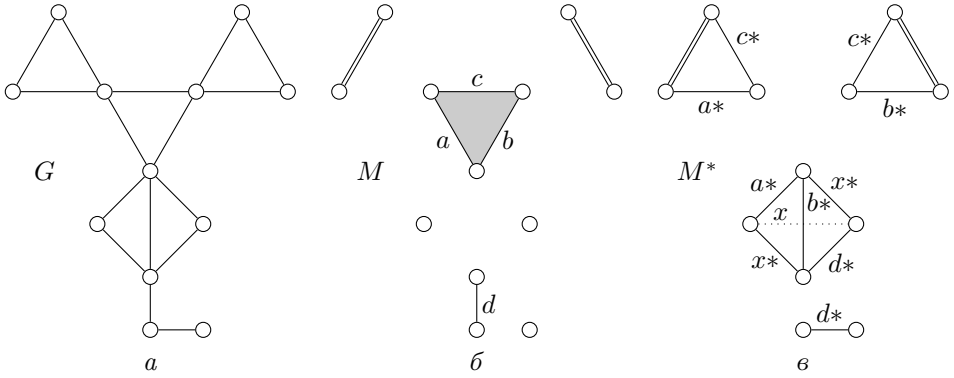

Рис. 5. Пример работы алгоритма 2: графы G, M и M^*

Доказательство. По определению

$$d(G, M^*) = |E_G \setminus E_{M^*}| + |E_{M^*} \setminus E_G|.$$

Представим разность $E_G \setminus E_{M^*}$ в виде $E_G \setminus E_{M^*} = E_0 \cup E_1$, где E_0 — множество рёбер графа G, которые не входят в M^* и не помещены в M алгоритмом 2, а E_1 — множество рёбер графа G, которые не входят в M^*, но помещены в M алгоритмом 2. Положим $E_2 = E_{M^*} \setminus E_G$. Тогда

$$d(G, M^*) = |E_0| + |E_1| + |E_2|. \tag{4}$$

По построению граф M является подграфом графа G, следовательно, $d(G, M) = |E_G \setminus E_M|$. Запишем разность $E_G \setminus E_M$ в следующем виде: $E_G \setminus E_M = E_0 \cup E^*$, где E^* — множество рёбер графа G, не помещённых в M, но входящих в M^*. Тогда

$$d(G, M) = |E_0| + |E^*|. \tag{5}$$

Следовательно, в силу (4), (5) и леммы 1 получим

$$\frac{d(G, M)}{d(G, M^*)} = \frac{|E_0| + |E^*|}{|E_0| + |E_1| + |E_2|} \leqslant \frac{|E^*|}{|E_1| + |E_2|} \leqslant \frac{2(|E_1| + |E_2|)}{|E_1| + |E_2|} = 2.$$

Теорема 3 доказана.

Заключение

В работе рассматривается вариант задачи кластеризации на графе, в котором размеры кластеров ограничены сверху натуральным числом s. Эта задача NP-трудна для любого фиксированного $s \geqslant 3$. Для указанного варианта задачи предложены полиномиальные приближённые алгоритмы с достижимыми гарантированными оценками точности, равными $\frac{5}{3}$ в случае $s = 3$ и 2 в случае $s = 4$. Доказано также, что для $s = 3$ эта задача APX-трудна.

Финансирование работы

Работа первого автора выполнена в рамках гос. задания Института математики им. С. Л. Соболева (проект № FWNF–2022–0020). Работа второго автора выполнена за счёт бюджета Омского гос. университета им. Ф. М. Достоевского. Работа третьего автора выполнена в рамках гос. задания Института математики им. С. Л. Соболева (проект № FWNF–2022–0019). Дополнительных грантов на проведение или руководство этим исследованием получено не было.

Конфликт интересов

Авторы заявляют, что у них нет конфликта интересов.

Литература

1. **Schaeffer S. E.** Graph clustering // Comput. Sci. Rev. 2005. V. 1, No. 1. P. 27–64.
2. **Фридман Г. Ш.** Одна задача аппроксимации графов // Управляемые системы. Вып. 8. Новосибирск: Изд-во Ин-та математики, 1971. С. 73–75.
3. **Фридман Г. Ш.** Исследование одной задачи классификации на графах // Методы моделирования и обработка информации. Новосибирск: Наука, 1976. С. 147–177.
4. **Tomescu I.** La reduction minimale d'un graphe à une reunion de cliques // Discrete Math. 1974. V. 10, No. 1–2. P. 173–179 [French].
5. **Zahn C. T.** Approximating symmetric relations by equivalence relations // J. Soc. Ind. Appl. Math. 1964. V. 12, No. 4. P. 840–847.
6. **Агеев А. А., Ильев В. П., Кононов А. В., Талевнин А. С.** Вычислительная сложность задачи аппроксимации графов // Дискрет. анализ и исслед. операций. Сер. 1. 2006. Т. 13, № 1. С. 3–11.
7. **Bansal N., Blum A., Chawla S.** Correlation clustering // Mach. Learn. 2004. V. 56, No. 1–3. P. 89–113.
8. **Ben-Dor A., Shamir R., Yakhimi Z.** Clustering gene expression patterns // J. Comput. Biol. 1999. V. 6, No. 3–4. P. 281–297.
9. **Shamir R., Sharan R., Tsur D.** Cluster graph modification problems // Discrete Appl. Math. 2004. V. 144, No. 1–2. P. 173–182.
10. **Puleo G. J., Milenkovic O.** Correlation clustering with constrained cluster sizes and extended weights bounds // SIAM J. Optim. 2015. V. 25, No. 3. P. 1857–1872.
11. **Pandove D., Goel S., Rani R.** Correlation clustering methodologies and their fundamental results // Expert Syst. 2018. V. 35, No. 1. Paper ID e12229. 24 p.
12. **Chawla S., Makarychev K., Schramm T., Yaroslavtsev G.** Near optimal LP rounding algorithm for correlation clustering on complete and complete k-partite graphs // Proc. 47th Annu. ACM Symp. Theory of Computing (Portland, OR, USA, June 14–17, 2015). New York: ACM, 2015. P. 219–228. DOI: 10.1145/2746539.2746604.

13. Il'ev V. P., Il'eva S. D., Kononov A. V. Short survey on graph correlation clustering with minimization criteria // Discrete optimization and operations research. Proc. 9th Int. Conf. (Vladivostok, Russia, Sept. 19–23, 2016). Cham: Springer, 2016. P. 25–36. (Lect. Notes Comput. Sci.; V. 9869). DOI: 10.1007/978-3-319-44914-2_3.

14. Wahid D. F., Hassini E. A literature review on correlation clustering: Cross-disciplinary taxonomy with bibliometric analysis // Oper. Res. Forum. 2022. V. 3. Paper ID 47. 42 p.

15. Bonchi F., García-Soriano D., Gullo F. Correlation clustering. Cham: Springer, 2022. 148 p. DOI: 10.1007/978-3-031-79210-6.

16. Ильев В. П., Навроцкая А. А. Вычислительная сложность задачи аппроксимации графами с компонентами связности ограниченного размера // Прикл. дискрет. математика. 2011. № 3. С. 80–84.

17. Ильев В. П., Ильева С. Д., Навроцкая А. А. О задаче кластеризации графа с ограничением на размеры кластеров // Дискрет. анализ и исслед. операций. 2016. Т. 23, № 3. С. 5–20.

18. Il'ev V. P., Il'eva S. D. Approximation algorithms for graph cluster editing problems with cluster size at most 3 or 4 // Mathematical optimization theory and operations research: Recent trends. Rev. Sel. Pap. 22nd Int. Conf. (Yekaterinburg, Russia, July 2–8, 2023). Cham: Springer, 2023. P. 134–145. (Commun. Comput. Inf. Sci.; V. 1881). DOI: 10.1007/978-3-031-43257-6_11.

19. Kononov A. V., Il'ev V. P. On cluster editing problem with clusters of small sizes // Optimization and applications. Rev. Sel. Pap. 14th Int. Conf. (Petrovac, Montenegro, Sept. 18–22, 2023). Cham: Springer, 2023. P. 316–328. (Lect. Notes Comput. Sci.; V. 14395). DOI: 10.1007/978-3-031-47859-8_23.

20. Chataigner F., Manić G., Wakabayashi Y., Yuster R. Approximation algorithms and hardness results for the clique packing problem // Discrete Appl. Math. 2009. V. 157, No. 7. P. 1396–1406. DOI: 10.1016/j.dam.2008.10.017.

Ильев Виктор Петрович
Ильева Светлана Диадоровна
Кононов Александр Вениаминович

Статья поступила
17 мая 2024 г.
После доработки —
11 июня 2024 г.
Принята к публикации
22 июня 2024 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
/DISCRETE ANALYSIS AND OPERATIONS RESEARCH/
October–December 2024. Vol. 31, No. 4. P. 40–57

UDC 519.8 DOI: 10.33048/daio.2024.31.802

APPROXIMATION ALGORITHMS FOR GRAPH CLUSTERING PROBLEMS WITH CLUSTERS OF BOUNDED SIZE

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

[1] Sobolev Institute of Mathematics,
4 Acad. Koptyug Avenue, 630090 Novosibirsk, Russia
[2] Dostoevsky Omsk State University,
55-A Mir Avenue, 644077 Omsk, Russia
E-mail: [a] iljev@mail.ru, [b] alvenko@math.nsc.ru

Abstract. In the cluster editing problem, one has to partition the set of vertices of a graph into pairwise disjoint subsets (called clusters) minimizing the number of edges between clusters and the number of missing edges within clusters. We consider a version of the problem in which cluster sizes are bounded from above by a positive integer s. This problem is NP-hard for any fixed $s \geqslant 3$. We propose polynomial-time approximation algorithms for this version of the problem. Their performance guarantees equal 5/3 for the case $s = 3$ and 2 for $s = 4$. We also show that the cluster editing problem is APX-hard for the case of $s = 3$. Illustr. 5, bibliogr. 20.

Keywords: graph, clustering, NP-hard problem, approximation algorithm, performance guarantee.

References

1. **S. E. Schaeffer,** Graph clustering, *Comput. Sci. Rev.* **1** (1), 27–64 (2005).
2. **G. Sh. Fridman,** One problem in graph approximation, in *Controlled Systems,* Vol. 8 (Izd. Inst. Mat., Novosibirsk, 1971), pp. 73–75 [Russian].
3. **G. Sh. Fridman,** Investigation of one problem in graph classification, in *Modelling Methods and Information Processing* (Nauka, Novosibirsk, 1976), pp. 147–177 [Russian].
4. **I. Tomescu,** La reduction minimale d'un graphe à une reunion de cliques, *Discrete Math.* **10** (1–2), 173–179 (1974) [French].
5. **C. T. Zahn,** Approximating symmetric relations by equivalence relations, *J. Soc. Ind. Appl. Math.* **12** (4), 840–847 (1964).

English transl.: *Journal of Applied and Industrial Mathematics* **18** (4), 686–696 (2024), DOI: 10.1134/S1990478924040069.

6. **A. A. Ageev, V. P. Il'ev, A. V. Kononov,** and **A. S. Talevnin,** Computational complexity of the graph approximation problem, *Diskretn. Anal. Issled. Oper., Ser. 1,* **13** (1), 3–11 (2006) [Russian] [*J. Appl. Ind. Math.* **1** (1), 1–8 (2023)].
7. **N. Bansal, A. Blum,** and **S. Chawla,** Correlation clustering, *Mach. Learn.* **56** (1–3), 89–113 (2004).
8. **A. Ben-Dor, R. Shamir,** and **Z. Yakhimi,** Clustering gene expression patterns, *J. Comput. Biol.* **6** (3–4), 281–297 (1999).
9. **R. Shamir, R. Sharan,** and **D. Tsur,** Cluster graph modification problems, *Discrete Appl. Math.* **144** (1–2), 173–182 (2004).
10. **G. J. Puleo** and **O. Milenkovic,** Correlation clustering with constrained cluster sizes and extended weights bounds, *SIAM J. Optim.* **25** (3), 1857–1872 (2015).
11. **D. Pandove, S. Goel,** and **R. Rani,** Correlation clustering methodologies and their fundamental results, *Expert Syst.* **35** (1), ID e12229 (2018).
12. **S. Chawla, K. Makarychev, T. Schramm,** and **G. Yaroslavtsev,** Near optimal LP rounding algorithm for correlation clustering on complete and complete k-partite graphs, in *Proc. 47th Annu. ACM Symp. Theory of Computing, Portland, OR, USA, June 14–17, 2015* (ACM, New York, 2015), pp. 219–228, DOI: 10.1145/2746539.2746604.
13. **V. P. Il'ev, S. D. Il'eva,** and **A. V. Kononov,** Short survey on graph correlation clustering with minimization criteria, in *Discrete Optimization and Operations Research* (Proc. 9th Int. Conf., Vladivostok, Russia, Sept. 19–23, 2016) (Springer, Cham, 2016), pp. 25–36 (Lect. Notes Comput. Sci., Vol. 9869), DOI: 10.1007/978-3-319-44914-2_3.
14. **D. F. Wahid** and **E. Hassini,** A literature review on correlation clustering: Cross-disciplinary taxonomy with bibliometric analysis, *Oper. Res. Forum* **3**, ID 47 (2022).
15. **F. Bonchi, D. García-Soriano** and **F. Gullo,** *Correlation Clustering* (Springer, Cham, 2022), DOI: 10.1007/978-3-031-79210-6.
16. **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) [Russian].
17. **V. P. Il'ev, S. D. Il'eva,** and **A. A. Navrotskaya,** Graph clustering with a constraint on cluster sizes, *Diskretn. Anal. Issled. Oper.* **23** (3), 5–20 (2016) [Russian] [*J. Appl. Ind. Math.* **10** (3), 341–348 (2016)].
18. **V. P. Il'ev** and **S. D. Il'eva,** Approximation algorithms for graph cluster editing problems with cluster size at most 3 or 4, in *Mathematical Optimization Theory and Operations Research: Recent Trends* (Rev. Sel. Pap. 22nd Int. Conf., Yekaterinburg, Russia, July 2–8, 2023) (Springer, Cham, 2023), pp. 134–145 (Commun. Comput. Inf. Sci., Vol. 1881), DOI: 10.1007/978-3-031-43257-6.
19. **A. V. Kononov** and **V. P. Il'ev,** On cluster editing problem with clusters of small sizes, in *Optimization and Applications* (Rev. Sel. Pap. 14th Int. Conf., Petrovac, Montenegro, Sept. 18–22, 2023) (Springer, Cham, 2023), pp. 316–328 (Lect. Notes Comput. Sci., Vol. 14395), DOI: 10.1007/978-3-031-47859-8.

20. F. Chataigner, G. Manić, Y. Wakabayashi, and **R. Yuster,** Approximation algorithms and hardness results for the clique packing problem, *Discrete Appl. Math.* **157** (7), 1396–1406 (2009), DOI: 10.1016/j.dam.2008.10.017.

Victor P. Il'ev
Svetlana D. Il'eva
Aleksandr V. Kononov

Received May 17, 2024
Revised June 11, 2024
Accepted June 22, 2024