

О ЗАДАЧЕ КЛАСТЕРИЗАЦИИ ГРАФА С ОГРАНИЧЕНИЕМ НА РАЗМЕРЫ КЛАСТЕРОВ *)

В. П. Ильев^{1,2}, С. Д. Ильева², А. А. Навроцкая^{1,2}

¹Институт математики им. С. Л. Соболева СО РАН,
пр. Акад. Коптюга, 4, 630090 Новосибирск, Россия

² Омский гос. университет им. Ф. М. Достоевского,
пр. Мира, 55-А, 644077 Омск, Россия

e-mail: iljev@mail.ru, nawrocki@ya.ru

Аннотация. Изучается версия задачи кластеризации графа (известной также как задача аппроксимации графа), в которой размеры кластеров ограничены сверху заданным числом p . Для этой задачи предложен новый приближённый алгоритм с достижимой гарантированной оценкой точности. Тем самым показано, что задача кластеризации графа с ограничением на размеры кластеров принадлежит классу APX для любого фиксированного p . Ил. 2, библиогр. 20.

Ключевые слова: кластеризация, аппроксимация, граф, приближённый алгоритм, гарантированная оценка точности.

Введение

В задаче кластеризации требуется разбить заданное множество объектов на несколько подмножеств (*кластеров*) только на основе сходства объектов друг с другом. Мера сходства определяется по-разному в разных задачах. Одной из наиболее наглядных формализаций задач кластеризации взаимосвязанных объектов является задача аппроксимации графа, которая представляет собой один из вариантов задачи кластеризации графа [17]. В этой задаче структура взаимосвязей объектов задается посредством неориентированного графа, вершины которого взаимно однозначно соответствуют объектам, а рёбра соединяют похожие объекты, обладающие достаточным количеством одинаковых признаков. Требуется разбить множество исходных объектов на попарно не пересекающиеся группы (кластеры) так, чтобы минимизировать число связей

*) Исследование первого и третьего авторов (разд. 1, 2, 3.1) выполнено при финансовой поддержке Российского научного фонда (проект 15-11-10009).

между кластерами и число недостающих связей внутри кластеров. Количество кластеров может быть задано, ограничено или заранее не определено. Постановки и различные интерпретации задачи аппроксимации графа можно найти в [7, 8, 11, 12, 18–20].

В разд. 1 рассматриваются три известных варианта задачи аппроксимации графа, являющейся формализацией задач кластеризации взаимосвязанных объектов. Приводится краткий обзор результатов по вычислительной сложности и аппроксимируемости этих задач. В разд. 2 рассматривается сравнительно новая постановка задачи кластеризации графа, в которой заданы ограничения на размеры кластеров. Эта задача NP-трудна. Выделен полиномиально разрешимый случай задачи. В разд. 3 предложен алгоритм приближённого решения задачи, в которой размеры кластеров ограничены сверху заданным числом $p \geq 2$, с достижимой гарантированной оценкой точности $\lfloor \frac{(p-1)^2}{2} \rfloor + 1$. Тем самым показано, что задача кластеризации графа с ограничением на размеры кластеров принадлежит классу APX для любого фиксированного p .

1. Постановки задач и краткий обзор известных результатов

Будем рассматривать только обыкновенные графы, т. е. графы без петель и кратных рёбер. Обыкновенный граф называется *кластерным графом*, если каждая его компонента связности является полным графом [18]. Обозначим через $\mathcal{M}(V)$ множество всех кластерных графов на множестве вершин V , $\mathcal{M}_k(V)$ — множество всех кластерных графов на множестве вершин V , имеющих ровно k непустых компонент связности, $\mathcal{M}_{1,k}(V)$ — множество всех кластерных графов на множестве V , имеющих не более k компонент связности, $2 \leq k \leq |V|$.

Если $G_1 = (V, E_1)$ и $G_2 = (V, E_2)$ — обыкновенные графы 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-е гг. в литературе изучались следующие три варианта задачи аппроксимации графа, которые можно рассматривать как различные формализации задачи кластеризации графа [4, 8, 9, 19, 20].

Задача А. Дан обыкновенный граф $G = (V, E)$. Требуется найти граф $M^* \in \mathcal{M}(V)$ такой, что

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

Задача \mathbf{A}_k . Даны обыкновенный граф $G = (V, E)$ и целое число k , $2 \leq k \leq |V|$. Требуется найти граф $M^* \in \mathcal{M}_k(V)$ такой, что

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

Задача $\mathbf{A}_{1,k}$. Даны обыкновенный граф $G = (V, E)$ и целое число k , $2 \leq k \leq |V|$. Требуется найти граф $M^* \in \mathcal{M}_{1,k}(V)$ такой, что

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

В дальнейшем задачи аппроксимации графов неоднократно переоткрывались и независимо изучались под разными названиями (Correlation Clustering [11], Cluster Editing [12, 18]).

Первые теоретические результаты, относящиеся к задачам аппроксимации графов, получены в 1960–70-е гг. В 1964 г. в [20] исследована задача \mathbf{A} для графов специального вида. В 1971 г. Фридман [8] выделил первый полиномиально разрешимый случай задачи аппроксимации графа \mathbf{A} . Он показал, что задача \mathbf{A} для любого графа без треугольников сводится к построению в нём наибольшего паросочетания.

В 1986 г. в [16] было показано, что задача \mathbf{A} NP-трудна, однако эта работа осталась незамеченной. В 2004 г. в [11] и независимо в [18] доказана NP-трудность задачи \mathbf{A} . В [18] доказано также, что задача \mathbf{A}_k NP-трудна при любом фиксированном $k \geq 2$, а в 2006 г. в [15] опубликовано более простое доказательство этого результата. В том же году независимо А. А. Агеев, В. П. Ильев, А. В. Кононов и А. С. Талевнин в [1] доказали, что задачи \mathbf{A}_2 и $\mathbf{A}_{1,2}$ NP-трудны уже на кубических графах, откуда вывели, что все упомянутые ранее варианты задачи аппроксимации графа NP-трудны, включая и задачу $\mathbf{A}_{1,k}$.

Таким образом, в [1, 11, 15, 16, 18] доказана

Теорема 1. *Задача \mathbf{A} NP-трудна. Задачи \mathbf{A}_k и $\mathbf{A}_{1,k}$ NP-трудны при любом фиксированном $k \geq 2$, причём для $k = 2$ они NP-трудны на кубических графах.*

В [11] предложен простой 3-приближённый алгоритм для задачи $\mathbf{A}_{1,2}$. В [1] доказано существование рандомизированной полиномиальной приближённой схемы для задачи $\mathbf{A}_{1,2}$, а в [15] предложена рандомизированная полиномиальная приближённая схема для задачи \mathbf{A}_k (для любого фиксированного $k \geq 2$). Указав, что сложность полиномиальной приближённой схемы из [15] лишает её перспективы практического использования, Коулман, Сондерсон и Уирт [14] в 2008 г. предложили 2-приближённый алгоритм для задачи $\mathbf{A}_{1,2}$, применив процедуру локального

поиска к допустимому решению, полученному с помощью 3-приближённого алгоритма из [11]. Для задачи **A₂** в [3] предложен приближённый алгоритм с достижимой гарантированной оценкой точности $3 - 6/|V|$.

Что касается задачи **A**, в 2005 г. в [13] показано, она *APX*-трудна, и для неё разработан 4-приближённый алгоритм. В 2008 г. в [10] представлен 2,5-приближённый алгоритм для задачи **A**.

2. Задача кластеризации с ограничением на размеры кластеров

В отличие от разд. 1, где ограничения накладывались на количество кластеров, в этом разделе рассматриваются задачи кластеризации взаимосвязанных объектов с кластерами ограниченного размера.

Обозначим через $\mathcal{M}^{1,p}(V)$ множество всех кластерных графов на V , в которых размер каждой компоненты связности не превышает целого числа p , $2 \leq p \leq |V|$. Будем говорить, что кластерный граф принадлежит множеству $\mathcal{M}^p(V)$, если размер каждой его компоненты связности равен p .

Задача $\mathbf{A}^{1,p}$. Даны n -вершинный граф $G = (V, E)$ и целое число p , $2 \leq p \leq n$. Найти граф $M^* \in \mathcal{M}^{1,p}(V)$ такой, что

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

Задача \mathbf{A}^p . Дан граф $G = (V, E)$ такой, что $|V| = pq$, где p, q — целые положительные числа. Найти граф $M^* \in \mathcal{M}^p(V)$ такой, что

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

В [11] при доказательстве NP-трудности задачи **A**, в которой нет никаких ограничений на число и размеры кластеров, фактически показано, что задача **A^{1,3}** NP-трудна. В [5] доказано, что для любого фиксированного $p \geq 3$ задачи **A^{1,p}** и **A^p** NP-трудны.

Теорема 2 [5]. *Задачи $\mathbf{A}^{1,p}$ и \mathbf{A}^p NP-трудны для любого фиксированного $p \geq 3$.*

К задачам **A^{1,p}** и **A^p** сводится по Тьюрингу NP-полная задача Разбиение на изоморфные подграфы, содержащаяся в [2] под номером ТГ12.

В [5] рассмотрены также случаи, когда оптимальные решения задач **A^{1,p}** и **A^p** можно найти за полиномиальное время.

Теорема 3 [5]. Задачи $\mathbf{A}^{1,2}$ и \mathbf{A}^2 полиномиально разрешимы. Задача $\mathbf{A}^{1,3}$ на графах, не содержащих треугольников, полиномиально разрешима.

Покажем, что последний результат можно обобщить на случай произвольного p .

Заметим вначале, что для любого оптимального решения $M_A \in \mathcal{M}(V)$ задачи \mathbf{A} на графе $G = (V, E)$ и для любого оптимального решения $M_{A^{1,p}} \in \mathcal{M}^{1,p}(V)$ задачи $\mathbf{A}^{1,p}$ ($p \geq 2$) на графе G выполнено неравенство

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

Далее, для задачи \mathbf{A} Фридманом доказана

Лемма 1 [9]. Если граф G не содержит треугольников, то одним из оптимальных решений задачи \mathbf{A} на графе G является произвольный граф, рёбра которого образуют наибольшее паросочетание графа G .

Из леммы 1 с учётом неравенства (1) получаем, что для любого $p \geq 2$ одним из оптимальных решений задачи $\mathbf{A}^{1,p}$ на графе G , не содержащем треугольников, является любой граф, рёбра которого образуют наибольшее паросочетание графа G . Отсюда, учитывая, что наибольшее паросочетание в любом графе можно найти за полиномиальное время, немедленно получаем следующее утверждение.

Теорема 4. Задача $\mathbf{A}^{1,p}$ на графах, не содержащих треугольников, полиномиально разрешима для любого $p \geq 2$.

3. Приближённый алгоритм для задачи $\mathbf{A}^{1,p}$

Для задачи $\mathbf{A}^{1,3}$ в [6] предложен полиномиальный приближённый алгоритм с достижимой гарантированной оценкой точности $3 - 6/|V|$. В этом разделе предложим приближённый алгоритм с гарантированной оценкой точности $\lfloor \frac{(p-1)^2}{2} \rfloor + 1$ для задачи $\mathbf{A}^{1,p}$, $p \geq 3$.

Пусть (V_1, \dots, V_s) — разбиение множества V , т. е. $V = V_1 \cup \dots \cup V_s$ и $V_i \cap V_j = \emptyset$ для любых $i, j \in \{1, \dots, s\}$, $i \neq j$, $M(V_1, \dots, V_s)$ — кластерный граф из класса $\mathcal{M}_s(V)$, в котором V_i — множество вершин i -го кластера, $i \in \{1, \dots, s\}$.

3.1. Случай связного графа. Вначале рассмотрим случай, когда граф $G = (V, E)$ связный.

Обозначим через $M^* = M(V_1, \dots, V_l) \in \mathcal{M}^{1,p}(V)$ оптимальное решение задачи $\mathbf{A}^{1,p}$ на связном графе G , $1 \leq l \leq |V|$, $p \geq 3$. Для всех $i \in \{1, \dots, l\}$ положим $n_i = |V_i|$.

Лемма 2. Для любого $i \in \{1, \dots, l\}$ справедливо неравенство

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

ДОКАЗАТЕЛЬСТВО. Если n_i чётно, то

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

Если n_i нечётно, то

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

Итак, $n_i(n_i - 1)/2 - \lfloor n_i/2 \rfloor = \lfloor (n_i - 1)^2/2 \rfloor$ для любого $i \in \{1, \dots, l\}$. Так как $M^* \in \mathcal{M}^{1,p}(V)$, имеем $n_i \leq p$ для любого $i \in \{1, \dots, l\}$, откуда получаем

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

Лемма 2 доказана.

Рассмотрим кластерный граф M' на множестве вершин V , построенный с учётом структуры графов $G = (V, E)$ и $M^* = (V_1, \dots, V_l)$ по следующему правилу.

Правило 1. Для каждого $i \in \{1, \dots, l\}$ разобьём V_i на $\lfloor n_i/2 \rfloor$ пар произвольным образом. Для каждой такой пары вершин $u, v \in V_i$ определим либо одну, либо две компоненты связности графа M' : если $uv \in E$, то $\{u, v\}$ — двухвершинная компонента графа M' ; если $uv \notin E$, то $\{u\}$, $\{v\}$ — тривиальные компоненты графа M' . Если n_i нечётно, то вершина $w \in V_i$, оставшаяся без пары, образует тривиальную компоненту $\{w\}$ графа M' .

Заметим, что рёбра графа G , соответствующие двухэлементным кластерам графа M' , образуют паросочетание в графе G (так как $V_i \cap V_j = \emptyset$ для любых $i, j \in \{1, \dots, l\}$, $i \neq j$).

Оценим расстояние от графа M' до графа G через $d(G, M^*)$.

Лемма 3. Для любого связного графа G

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

где $M^* = M(V_1, \dots, V_l) \in \mathcal{M}^{1,p}(V)$ — оптимальное решение задачи $\mathbf{A}^{1,p}$ на графе G , $M' \in \mathcal{M}^{1,2}(V)$ — кластерный граф, построенный по правилу 1, $n_i = |V_i|$, $i \in \{1, \dots, l\}$.

ДОКАЗАТЕЛЬСТВО. Обозначим через E_1 подмножество рёбер графа G , концы которых находятся в разных кластерах графа M^* : $E_1 = \{uv \in E \mid u \in V_i, v \in V_j, i \neq j\}$. По определению расстояния

$$d(G, M^*) \geq |E_1|. \quad (3)$$

Оценим теперь расстояние от графа M' до графа G . По построению M' является подграфом графа G . Следовательно, $d(G, M')$ равно количеству рёбер графа G , концы которых находятся в разных кластерах графа M' . Очевидно, что это рёбра множества E_1 и рёбра внутри кластеров графа M^* , не вошедшие в двухэлементные кластеры графа M' .

Для любого $i \in \{1, \dots, l\}$ количество рёбер графа G внутри i -го кластера графа M^* не превосходит $\frac{n_i(n_i-1)}{2}$. Количество рёбер графа G внутри i -го кластера графа M^* , которые не вошли в двухэлементные кластеры графа M' , не превосходит $\frac{n_i(n_i-1)}{2} - \left\lfloor \frac{n_i}{2} \right\rfloor$. Таким образом,

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

откуда с учётом (3) вытекает требуемое неравенство (2). Лемма 3 доказана.

Обозначим через $\mathcal{G}^* = \mathcal{G}^*(V_1, \dots, V_l)$ семейство связных графов, состоящих из l клик на множествах V_1, \dots, V_l , произвольным образом соединённых $l-1$ мостами. Очевидно, что $d(G, M^*) = l-1$ для всякого $G \in \mathcal{G}^*$ и $d(G, M^*) \geq l$ для любого связного графа $G \notin \mathcal{G}^*$.

Таким образом, для любого связного графа G имеем

$$d(G, M^*) \geq l-1. \quad (4)$$

Лемма 4. Пусть граф G связан. Если $G \notin \mathcal{G}^*$, то

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

где $M^* = M(V_1, \dots, V_l)$ — оптимальное решение задачи $\mathbf{A}^{1,p}$ на графе G , M' — кластерный граф, построенный по правилу 1.

ДОКАЗАТЕЛЬСТВО. Как уже было отмечено, для любого связного графа G равенство $d(G, M^*) = l - 1$ в (4) имеет место тогда и только тогда, когда $G \in \mathcal{G}^*$. Так как по условию леммы $G \notin \mathcal{G}^*$, получаем $d(G, M^*) \geq l$.

Пользуясь леммами 3, 2 и неравенством $l \leq d(G, M^*)$, выводим

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

Лемма 4 доказана.

Рассмотрим алгоритм приближённого решения задачи $\mathbf{A}^{1,p}$ на связном графе G .

АЛГОРИТМ A_1

ВХОД: связный граф $G = (V, E)$.

ШАГ 1. Удалим из графа G все мосты, полученный граф обозначим через M_1 . Переходим на шаг 2.

ШАГ 2. Строим кластерный граф $M_2 \in \mathcal{M}^{1,2}(V)$: находим наибольшее паросочетание в G ; найденное паросочетание образует двухвершинные компоненты связности кластерного графа M_2 , а вершины, не вошедшие в паросочетание, образуют тривиальные компоненты. Переходим на шаг 3.

ШАГ 3. Если $M_1 \in \mathcal{M}^{1,p}(V)$ и $d(G, M_1) \leq d(G, M_2)$, то полагаем $M = M_1$, иначе $M = M_2$.

КОНЕЦ.

Заметим, что число двухвершинных компонент связности в графе M_2 , построенном на шаге 2 алгоритма A_1 , совпадает с числом рёбер в наибольшем паросочетании в G . Так как рёбра, соответствующие двухвершинным компонентам графа M' , построенного по правилу 1, тоже образуют паросочетание в G (не обязательно наибольшее), получим

$$d(G, M_2) \leq d(G, M'). \quad (5)$$

Теорема 5. Для любого связного графа G имеет место следующая оценка:

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

где $M \in \mathcal{M}^{1,p}(V)$ — кластерный граф, построенный алгоритмом A_1 , M^* — оптимальное решение задачи $\mathbf{A}^{1,p}$ на графе G .

ДОКАЗАТЕЛЬСТВО. Сначала рассмотрим случай, когда $G \notin \mathcal{G}^*$. Тогда из (5) и леммы 4 получаем

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

что и требовалось.

Пусть теперь $G \in \mathcal{G}^*$, т. е. граф G состоит из l клик на множествах V_1, \dots, V_l , соединённых мостами. Возможны два случая.

СЛУЧАЙ 1: $n_i \geq 3$ для всех $i \in \{1, \dots, l\}$. В этом случае граф M_1 , построенный на шаге 1 алгоритма A_1 , совпадает с графом M^* и принадлежит классу $\mathcal{M}^{1,p}(V)$. Следовательно,

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

что и требовалось.

СЛУЧАЙ 2: $n_j \leq 2$ для некоторого $j \in \{1, \dots, l\}$. В этом случае граф M_1 , построенный на шаге 1 алгоритма A_1 , может не совпадать с M^* . Без ограничения общности считаем, что $j = l$. В этом случае $\frac{n_l(n_l-1)}{2} - \lfloor \frac{n_l}{2} \rfloor = 0$ и по лемме 3

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

В силу леммы 2 для любого $i \in \{1, \dots, l\}$ верно неравенство

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

поэтому

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

(последнее неравенство выполнено в силу (4)). Отсюда с учётом (5) получаем

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

Теорема 5 доказана.

3.2. Общий случай. В данном пункте будет показано, что оценка (6) остаётся верной и для несвязного графа.

Опишем алгоритм приближённого решения задачи $\mathbf{A}^{1,p}$ на произвольном графе G .

АЛГОРИТМ A_2

ВХОД: произвольный граф $G = (V, E)$, $G_i = (U_i, E_i)$ — i -я компонента связности графа G , $i = 1, \dots, k$, для некоторого $k \in \{1, \dots, |V|\}$.

ШАГ 1. Для каждого графа G_i строим граф $M_i \in \mathcal{M}^{1,p}(U_i)$, $i = 1, \dots, k$, с помощью алгоритма A_1 . Переходим на шаг 2.

ШАГ 2. Полагаем $M = \bigcup_{i=1}^k M_i$.

КОНЕЦ.

Очевидно, что построенный алгоритмом A_2 кластерный граф M принадлежит классу $\mathcal{M}^{1,p}(V)$.

Теорема 6. Для любого графа G имеет место оценка

$$d(G, M) \leq d(G, M^*) \left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1 \right), \quad (7)$$

где $M \in \mathcal{M}^{1,p}(V)$ — кластерный граф, построенный алгоритмом A_2 , M^* — оптимальное решение задачи $\mathbf{A}^{1,p}$ на графе G .

ДОКАЗАТЕЛЬСТВО. Если G — связный граф, то оценка справедлива в силу теоремы 5.

Пусть $G = (V, E)$ — несвязный граф, $G_i = (U_i, E_i)$ — i -я компонента связности графа G , $i = 1, \dots, k$, для некоторого k . По теореме 5 для каждой компоненты связности верна оценка (6), т. е.

$$d(G_i, M_i) \leq d(G_i, M_i^*) \left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1 \right),$$

где $M_i \in \mathcal{M}^{1,p}(U_i)$ — кластерный граф, построенный алгоритмом A_2 , M_i^* — оптимальное решение задачи $\mathbf{A}^{1,p}$ на графе G_i . Тогда

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

Теорема 6 доказана.

Следствие. Задача $\mathbf{A}^{1,p}$ принадлежит классу APX для любого фиксированного $p \geq 3$.

Как показывает следующее утверждение, оценка (7) достижима для чётных значений p .

Замечание. Для любого чётного $p \geq 4$ существует такой граф G_p , что

$$d(G_p, M) = d(G_p, M^*) \left(\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1 \right). \quad (8)$$

Граф G_p имеет $2p$ вершин и состоит из двух клик K_p , соединённых двумя рёбрами. На рис. 1 приведён пример для случая $p = 4$.

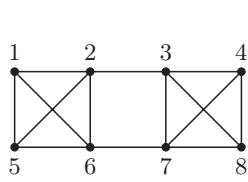
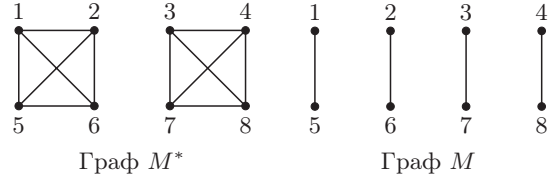


Рис. 1. Граф G_4



Граф M^*

Граф M

Рис. 2. Графы M^* и M

Как видно из рис. 2, $d(G_4, M^*) = 2$, $d(G_4, M) = 10$, и равенство (8) выполнено.

Для графа G_p кластерный граф M , найденный алгоритмом A_2 , представляет собой наибольшее паросочетание в объединении клик K_p , поэтому

$$d(G_p, M) = 2 \left(\frac{p(p-1)}{2} - \frac{p}{2} \right) + 2 = p^2 - 2p + 2.$$

Поскольку $d(G_p, M^*) = 2$ и при чётном p

$$\left\lfloor \frac{(p-1)^2}{2} \right\rfloor + 1 = \frac{p^2 - 2p + 2}{2},$$

равенство (8) выполняется для любого чётного $p \geq 4$.

ЛИТЕРАТУРА

1. 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.
2. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982. 416 с.
3. Ильев В. П., Ильева С. Д., Навроцкая А. А. Приближённые алгоритмы для задач аппроксимации графов // Дискрет. анализ и исслед. операций. 2011. Т. 18, № 1. С. 41–60.
4. Ильев В. П., Фридман Г. Ш. К задаче аппроксимации графами с фиксированным числом компонент // Докл. АН СССР. 1982. Т. 264, № 3. С. 533–538.
5. Ильев В. П., Навроцкая А. А. Вычислительная сложность задачи аппроксимации графами с компонентами связности ограниченного размера // Прикл. дискрет. математика. 2011. № 3. С. 80–84.
6. Ильев В. П., Навроцкая А. А. Приближённое и точное решение одного варианта задачи кластеризации взаимосвязанных объектов // Тр. XI междунар. Азиатской школы-семинара «Проблемы оптимизации сложных систем» (Чолпон-Ата, 27 июля–7 августа 2015 г.). Новосибирск: Инст. вычисл. математики и мат. геофизики СО РАН, 2015. С. 278–283.
7. Ляпунов А. А. О строении и эволюции управляющих систем в связи с теорией классификации // Пробл. кибернетики. М.: Наука, 1973. Вып. 27. С. 7–18.
8. Фридман Г. Ш. Одна задача аппроксимации графов // Управляемые системы. 1971. Вып. 8. С. 73–75.
9. Фридман Г. Ш. Исследование одной задачи классификации на графах // Методы моделирования и обработка информации. Новосибирск: Наука, 1976. С. 147–177.
10. Ailon N., Charikar M., Newman A. Aggregating inconsistent information: Ranking and clustering // J. ACM. 2008. Vol. 55, No. 5. P. 1–27.
11. Bansal N., Blum A., Chawla S. Correlation clustering // Mach. Learn. 2004. Vol. 56, No. 1–3. P. 89–113.
12. Ben-Dor A., Shamir R., Yakhimi Z. Clustering gene expression patterns // J. Comput. Biol. 1999. Vol. 6, No. 3–4. P. 281–297.
13. Charikar M., Guruswami V., Wirth A. Clustering with qualitative information // J. Comput. Syst. Sci. 2005. Vol. 71, No. 3. P. 360–383.
14. Coleman T., Saunderson J., Wirth A. A local-search 2-approximation for 2-correlation-clustering // Algorithms — ESA 2008. Proc. 16th Annu. Eur. Symp. Algorithms (Karlsruhe, Germany, Sept. 15–17, 2008). Heidelberg: Springer-Verl., 2008. P. 308–319. (Lect. Notes Comput. Sci., Vol. 5193).
15. Giotis I., Guruswami V. Correlation clustering with a fixed number of clusters // Theory Comput. 2006. Vol. 2, No. 1. P. 249–266.

-
16. **Křivánek M., Morávek J.** NP-hard problems in hierarchical-tree clustering // Acta Inf. 1986. Vol. 23. P. 311–323.
 17. **Schaeffer S. E.** Graph clustering // Comput. Sci. Rev. 2007. Vol. 1, No. 1. P. 27–64.
 18. **Shamir R., Sharan R., Tsur D.** Cluster graph modification problems // Discrete Appl. Math. 2004. Vol. 144, No. 1–2. P. 173–182.
 19. **Tomescu I.** La reduction minimale d'un graphe à une reunion de cliques // Discrete Math. 1974. Vol. 10, No. 1–2. P. 173–179.
 20. **Zahn C. T., Jr.** Approximating symmetric relations by equivalence relations // J. Soc. Ind. Appl. Math. 1964. Vol. 12, No. 4. P. 840–847.

Ильев Виктор Петрович,
Ильева Светлана Диадоровна,
Навроцкая Анна Александровна

Статья поступила
28 декабря 2015 г.
Исправленный вариант —
28 марта 2016 г.

DISKRETNYYI ANALIZ I ISSLEDOVANIE OPERATSII

July–August 2016. Volume 23, No. 3. P. 5–20

UDC 519.8

DOI: 10.17377/daio.2016.23.521

GRAPH CLUSTERING WITH A CONSTRAINT
ON CLUSTER SIZESV. P. Il'ev^{1,2}, S. D. Il'eva², A. A. Navrotskaya^{1,2}¹Sobolev Institute of Mathematics,

4 Acad. Koptyug Ave., 630090 Novosibirsk, Russia

²Omsk State University,

55-A Mir Ave., 644077 Omsk, Russia

e-mail: iljev@mail.ru, nawrocki@ya.ru

Abstract. A graph clustering problem (also known as the graph approximation problem) with a constraint on cluster sizes is studied. A new approximation algorithm is presented for this problem and performance guarantee of this algorithm is obtained. It is shown that the problem belongs to class *APX* for every fixed p , where p is the upper bound on the cluster sizes. Ill. 2, bibliogr. 20.

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

REFERENCES

1. 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**, No. 1, 3–11, 2006. Translated in *J. Appl. Ind. Math.*, **1**, No. 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. Translated under the title *Vychislitel'nye mashiny i trudnoreshaemye zadachi*, 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**, No. 1, 41–60, 2011. Translated in *J. Appl. Ind. Math.*, **5**, No. 4, 569–581, 2011.
4. V. P. Il'ev and G. Š. Fridman, On the problem of approximation by graphs with a fixed number of components, *Dokl. Akad. Nauk SSSR*, **264**, No. 3, 533–538, 1982. Translated in *Sov. Math. Dokl.*, **25**, No. 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 one variant of the problem of clustering interconnected objects, in *Tr. XI Mezhdunarodnoi aziatskoi shkoly-seminara "Problemy optimizatsii slozhnykh sistem"* (Proc. XI Int. Asian School-Seminar "Optimization Problems for Complex Systems"), *Cholpon-Ata, Kyrgyzstan, July 27 – Aug. 7, 2015*, pp. 278–283, Inst. Vychisl. Mat. Mat. Geofiz. SO RAN, Novosibirsk, 2015.
7. **A. A. Lyapunov**, On the structure and evolution of control systems in connection with the theory of classification, *Problemy kibernetiki* (Problems of Cybernetics), Vol. 27, pp. 7–18, Fizmatgiz, Moscow, 1973.
8. **G. Š. Fridman**, A graph approximation problem, in *Upravlyaemye sistemy* (Control Systems), Vol. 8, pp. 73–75, Izd. Inst. Mat., Novosibirsk, 1971.
9. **G. Š. Fridman**, Investigation of a classifying problem on graphs, in *Metody modelirovaniya i obrabotka informatsii* (Methods of Modelling and Data Processing), pp. 147–177, Nauka, Novosibirsk, 1976.
10. **N. Ailon**, **M. Charikar**, and **A. Newman**, Aggregating inconsistent information: Ranking and clustering, *J. ACM*, **55**, No. 5, 1–27, 2008.
11. **N. Bansal**, **A. Blum**, and **S. Chawla**, Correlation clustering, *Mach. Learn.*, **56**, No. 1–3, 89–113, 2004.
12. **A. Ben-Dor**, **R. Shamir**, and **Z. Yakhimi**, Clustering gene expression patterns, *J. Comput. Biol.*, **6**, No. 3–4, 281–297, 1999.
13. **M. Charikar**, **V. Guruswami**, and **A. Wirth**, Clustering with qualitative information, *J. Comput. Syst. Sci.*, **71**, No. 3, 360–383, 2005.
14. **T. Coleman**, **J. Saunderson**, and **A. Wirth**, A local-search 2-approximation for 2-correlation-clustering, in *Algorithms — ESA 2008* (Proc. 16th Annual Eur. Symp. Algorithms, Karlsruhe, Germany, Sept. 15–17, 2008), pp. 308–319, Springer, Heidelberg, 2008 (Lect. Notes Comput. Sci., Vol. 5193).
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**, No. 1, 27–64, 2007.
18. **R. Shamir**, **R. Sharan**, and **D. Tsur**, Cluster graph modification problems, *Discrete Appl. Math.*, **144**, No. 1–2, 173–182, 2004.
19. **I. Tomescu**, Minimal reduction of a graph to a union of cliques, *Discrete Math.*, **10**, No. 1, 173–179, 1974 [French].

20. **C. T. Zahn, Jr.**, Approximating symmetric relations by equivalence relations, *J. Soc. Ind. Appl. Math.*, **12**, No. 4, 840–847, 1964.

Victor P. Il'ev,
Svetlana D. Il'eva,
Anna A. Navrotskaya

Received
28 December 2015
Revised
28 March 2016