

УДК 510.52

DOI 10.17223/20710410/60/10

**О ГЕНЕРИЧЕСКОЙ СЛОЖНОСТИ ПРОБЛЕМЫ КЛАСТЕРИЗАЦИИ
ГРАФОВ С ОГРАНИЧЕНИЯМИ НА РАЗМЕР КЛАСТЕРОВ¹**

А. Н. Рыбалов

*Институт математики им. С. Л. Соболева СО РАН, г. Омск, Россия***E-mail:** alexander.rybalov@gmail.com

Изучается генерическая сложность проблемы кластеризации графов с ограничением p на размеры кластеров при $p \geq 3$. В этой задаче структура взаимосвязей объектов задаётся с помощью графа, вершины которого соответствуют объектам, а рёбра соединяют похожие объекты. Требуется разбить множество объектов на попарно непересекающиеся группы (кластеры) ограниченного числом p размера так, чтобы минимизировать число связей между кластерами и число недостающих связей внутри кластеров. Доказывается, что при условии $P \neq NP$ и $P = BPP$ для этой проблемы не существует полиномиального сильно генерического алгоритма. Сильно генерический алгоритм решает проблему не на всём множестве входов, а на подмножестве, последовательность частот которого при увеличении размера экспоненциально быстро сходится к 1.

Ключевые слова: *генерическая сложность, кластеризация графа.*

**ON GENERIC COMPLEXITY OF THE GRAPH CLUSTERING
PROBLEM WITH BOUNDED CLUSTERS**

A. N. Rybalov

Sobolev Institute of Mathematics, Omsk, Russia

In this paper, we study the generic complexity of the graph clustering problem with bounded sizes of clusters. In this problem, the structure of object relations is presented as a graph: vertices correspond to objects, and edges connect similar objects. It is required to divide a set of objects into disjoint groups (clusters) of bounded sizes to minimize the number of connections between clusters and the number of missing links within clusters. It is proved that, under the condition $P \neq NP$ and $P = BPP$, for this problem there is no polynomial strongly generic algorithm. A strongly generic algorithm solves a problem not on the entire set of inputs, but on a subset whose frequency sequence converges exponentially to 1 as the size increases. To prove this result, we use the method of generic amplification, which allows to construct generically hard problems from the problems hard in the classical sense. The main component of this method is the cloning technique, which combines the inputs of a problem together into sufficiently large sets of equivalent inputs. Equivalence is understood in the sense that the problem is solved similarly for them.

Keywords: *generic complexity, graph clustering.*

¹Работа поддержана грантом РНФ № 22-11-20019.

Введение

Одной из важных проблем машинного обучения является проблема кластеризации графов. В этой задаче структура взаимосвязей объектов задаётся с помощью графа, вершины которого соответствуют объектам, а рёбра соединяют похожие объекты. Требуется разбить множество объектов на попарно непересекающиеся группы (кластеры) так, чтобы минимизировать число связей между кластерами и число недостающих связей внутри кластеров. В работах [1–7] доказана NP-трудность проблемы кластеризации графа для различных её постановок. Кроме того, В. П. Ильев и А. А. Навроцкая доказали [8], что проблема кластеризации графов, в которой размер (то есть число вершин) каждого кластера ограничен фиксированным числом p , является NP-трудной при $p \geq 3$ и разрешимой за полиномиальное время при $p = 2$.

В работах [9, 10] исследована генерическая сложность общей проблемы кластеризации графов и проблемы кластеризации графов с ограничением на число кластеров. В рамках генерического подхода [11] алгоритмическая проблема изучается не на всём множестве входов, а на некотором естественном подмножестве «почти всех» входов. Здесь генерический алгоритм должен выдавать правильный ответ. Остальные «редкие» входы игнорируются, на них генерический алгоритм выдаёт ответ «не знаю». Понятие «почти все» уточняется введением асимптотической плотности на множестве входных данных. Отметим, что похожий подход для изучения проблем оптимизации предложен ранее в [12].

В исследованиях по генерической вычислимости и сложности вычислений можно выделить два основных направления. Первое связано с построением генерических (полиномиальных) алгоритмов для алгоритмических проблем, которые являются неразрешимыми или трудноразрешимыми в классическом смысле. Второе направление концентрируется на поиске алгоритмических проблем, которые остаются неразрешимыми или трудноразрешимыми и в генерическом смысле. Это второе направление имеет важное значение для современной криптографии, так как для построения криптосистем с открытым ключом не годятся алгоритмические проблемы, являющиеся легко разрешимыми для почти всех входов.

Данная работа посвящена изучению генерической сложности задачи кластеризации графов с ограничением p на размеры кластеров при $p \geq 3$. Доказывается, что при условии $P \neq NP$ и $P = BPP$ для этой проблемы не существует полиномиально сильно генерического алгоритма. Отсюда, в частности, следует, что теоретически существует эффективная процедура генерации трудных входов данной проблемы, что указывает на потенциальную возможность её использования в криптографии. Здесь равенство $P = BPP$ означает, что любой вероятностный полиномиальный алгоритм можно эффективно дерандомизировать, то есть построить эквивалентный ему полиномиальный алгоритм, не использующий генератор случайных чисел. Эта гипотеза пока не доказана, однако имеются веские основания в пользу этого равенства [13].

1. Предварительные сведения

Определения сильно генерического полиномиального алгоритма можно найти в работах [9, 10]. С понятием вероятностного алгоритма можно познакомиться, например, в [14].

Здесь и далее будем рассматривать неориентированные графы без петель и кратных рёбер. Граф называется *кластерным*, если каждая его компонента связности является полным графом. Обозначим через $\mathcal{M}^{\leq p}(V)$ множество всех кластерных графов на множестве вершин V , у которых каждая компонента связности имеет не более p вершин.

Если $G_1 = (V, E_1)$ и $G_2 = (V, E_2)$ — графы на одном и том же множестве вершин V , то расстояние $\rho(G_1, G_2)$ между ними есть число несовпадающих рёбер в графах G_1 и G_2 , то есть

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

Проблема p -кластеризации графа состоит в следующем. Задан граф $G = (V, E)$. Найти такой граф $M^* \in \mathcal{M}^{\leq p}(V)$, что

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

Лемма 1. Пусть G_1 и G_2 — два графа с непересекающимися множествами вершин и M^* — кластерный граф, являющийся решением проблемы p -кластеризации для графа $G_1 \cup G_2$. Тогда $M^* = M_1^* \cup M_2^*$, где M_i^* — решение проблемы p -кластеризации для графа G_i , $i = 1, 2$.

Доказательство. Пусть $G_1 = (V_1, E_1)$ и $G_2 = (V_2, E_2)$. Допустим, что существует кластерный граф M , являющийся решением проблемы p -кластеризации для графа $G_1 \cup G_2$, такой, что

$$M = C_1 \cup C_2 \cup \dots \cup C_m,$$

где C_i , $i = 1, \dots, m$, — непересекающиеся полные компоненты связности с не более чем p вершинами, причём среди них есть компонента $C_k = K(V_c)$, которая имеет непустое пересечение как с графом G_1 , так и с графом G_2 . Заменим в кластерном графе M его компоненту C_k на два полных графа $C_{k,i} = K(V_c \cap V_i)$, $i = 1, 2$. Очевидно, что число вершин каждого из этих графов не превосходит числа вершин $K(V_c)$, то есть оно не больше p . Вершины первого лежат в графе G_1 , а второго — в G_2 . Обозначим через M' получившийся кластерный граф. Теперь заметим, что

$$\rho(G_1 \cup G_2, M') < \rho(G_1 \cup G_2, M),$$

так как в новых компонентах $C_{k,i}$, $i = 1, 2$, присутствуют все рёбра старой компоненты C_k , обе вершины которых лежат либо в G_1 , либо в G_2 , и отсутствуют все рёбра старой компоненты C_k , одна вершина которых лежит в G_1 , а другая — в G_2 . При этом последних рёбер нет и в графе $G_1 \cup G_2$.

Полученное противоречие показывает, что для кластерного графа M^* , являющегося решением проблемы p -кластеризации для графа $G_1 \cup G_2$, имеет место $M^* = M_1^* \cup M_2^*$, причём M_i^* — решение проблемы p -кластеризации для графа G_i , $i = 1, 2$, так как иначе кластерный граф M_i^* можно заменить на более подходящий, уменьшив тем самым расстояние $\rho(G_1 \cup G_2, M^*)$. ■

2. Основной результат

Будем использовать представление графов с помощью матриц смежности. Так как графы неориентированные, для кодирования графа с n вершинами достаточно верхней части матрицы, состоящей из $n(n-1)/2$ бит. Таким образом, будем считать, что размер графа с n вершинами равен $n(n-1)/2$. Обозначим через $\mathcal{G}$ множество всех графов с таким представлением.

Теорема 1. Не существует сильно генерического полиномиального алгоритма для решения проблемы p -кластеризации графов при $p \geq 3$, если $P \neq NP$ и $P = BPP$.

Доказательство. Допустим, что существует сильно генерический полиномиальный алгоритм $\mathcal{A}$, решающий проблему p -кластеризации графа. Тогда существует

вероятностный полиномиальный алгоритм $\mathcal{B}$, разрешающий эту проблему на всём множестве входов. На графе G с n вершинами (размера $n(n-1)/2$) алгоритм $\mathcal{B}$ работает следующим образом:

- 1) генерирует случайный граф H с $n^2 - n$ вершинами;
- 2) запускает алгоритм $\mathcal{A}$ на графе $G \cup H$;
- 3) если $\mathcal{A}(G \cup H) \neq ?$, то по решению проблемы p -кластеризации для графа $G \cup H$, согласно лемме 1, выдаёт решение проблемы p -кластеризации для графа G ;
- 4) если $\mathcal{A}(G \cup H) = ?$, то выдаёт ответ K_n .

Заметим, что полиномиальный вероятностный алгоритм $\mathcal{B}$ выдаёт правильный ответ на шаге 3, а на шаге 4 может выдать неправильный ответ. Нужно доказать, что вероятность того, что ответ выдаётся на шаге 4, меньше $1/3$.

Граф $G \cup H$ имеет n^2 вершин, то есть его размер равен $m = (n^4 - n^2)/2$. Вероятность того, что для случайного графа $G \cup H$ имеет место $\mathcal{A}(G \cup H) = ?$, не больше

$$\frac{|\{G \in \mathcal{G} : \mathcal{A}(G) = ?\}_m|}{|\{G \cup H : H \in \mathcal{G}\}_m|} = \frac{|\{G \in \mathcal{G} : \mathcal{A}(G) = ?\}_m|}{|\mathcal{G}_m|} \frac{|\mathcal{G}_m|}{|\{G \cup H : H \in \mathcal{G}\}_m|}.$$

Так как множество $\{G \in \mathcal{G} : \mathcal{A}(G) = ?\}$ сильно пренебрежимое, то существует константа $\alpha > 0$, такая, что

$$\frac{|\{G \in \mathcal{G} : \mathcal{A}(G) = ?\}_m|}{|\mathcal{G}_m|} < \frac{1}{2^{\alpha m}} = \frac{1}{2^{\alpha(n^4 - n^2)/2}}$$

для любого n .

С другой стороны, так как граф H имеет $n^2 - n$ вершин, то

$$|\{G \cup H : H \in \mathcal{G}\}_m| = |\{H : H \in \mathcal{G}\}_{((n^2 - n)^2 - (n^2 - n))/2}| = 2^{(n^4 - 2n^3 + n)/2}.$$

Отсюда

$$\frac{|\mathcal{G}_m|}{|\{G \cup H : H \in \mathcal{G}\}_m|} = \frac{2^{(n^4 - n^2)/2}}{2^{(n^4 - 2n^3 + n)/2}} = 2^{(2n^3 - n^2 + n)/2}.$$

Поэтому искомая вероятность не больше

$$\frac{2^{(2n^3 - n^2 + n)/2}}{2^{\alpha(n^4 - n^2)/2}} < \frac{1}{3}$$

при больших n .

Теперь покажем, что при условиях $P \neq NP$ и $P = BPP$ не существует полиномиального вероятностного алгоритма для решения проблемы p -кластеризации графов при $p \geq 3$. Действительно, пусть такой алгоритм существует. Так как проблема p -кластеризации графов при $p \geq 3$ является NP-трудной [8], существует полиномиально эквивалентная ей NP-полная проблема распознавания A . Из полиномиального вероятностного алгоритма для проблемы p -кластеризации графов легко получается полиномиальный вероятностный алгоритм для решения проблемы A . А так как $P = BPP$, то существует и детерминированный полиномиальный алгоритм для A , откуда $P = NP$. Противоречие. ■

Заключение

В работе изучена генерическая сложность проблемы кластеризации графов с ограничением p на размеры кластеров при $p \geq 3$. Доказано, что при условии $P \neq NP$ и $P = BPP$ для этой проблемы не существует полиномиального сильно генерического алгоритма. Этот результат показывает некоторый потенциал применения данной алгоритмической проблемы в криптографии, устанавливая, что существует возможность эффективно генерировать случайные трудные входы для этой задачи.

Автор выражает благодарность рецензенту за полезные замечания и предложения по улучшению текста статьи.

ЛИТЕРАТУРА

1. *Křivánek M. and Morávek J.* NP-hard problems in hierarchical-tree clustering // *Acta Informatica*. 1986. V. 23. P. 311–323.
2. *Bansal N., Blum A., and Chawla S.* Correlation clustering // *Machine Learning*. 2004. V. 56. P. 89–113.
3. *Shamir R., Sharan R., and Tsur D.* Cluster graph modification problems // *Discrete Appl. Math.* 2004. V. 144. No. 1–2. P. 173–182.
4. Агеев А. А., Ильев В. П., Кононов А. В., Талевнин А. С. Вычислительная сложность задачи аппроксимации графов // *Дискретный анализ и исследование операций*. Сер. 1. 2006. Т. 13. № 1. С. 3–11.
5. Ильев В. П., Ильева С. Д. О задачах кластеризации графов // *Вестник Омского университета*. 2016. № 2. С. 16–18.
6. Ильев А. В., Ильев В. П. Об одной задаче кластеризации графа с частичным обучением // *Прикладная дискретная математика*. 2018. № 42. С. 66–75.
7. Талевнин А. С. О сложности задачи аппроксимации графов // *Вестник Омского университета*. 2004. № 4. С. 22–24.
8. Ильев В. П., Навроцкая А. А. Вычислительная сложность задачи аппроксимации графами с компонентами связности ограниченного размера // *Прикладная дискретная математика*. 2011. № 3(13). С. 80–84.
9. Рыбалов А. Н. О генерической сложности проблемы кластеризации графов // *Прикладная дискретная математика*. 2019. № 46. С. 72–77.
10. Рыбалов А. Н. О генерической сложности ограниченной проблемы кластеризации графов // *Прикладная дискретная математика*. 2022. № 57. С. 91–97.
11. *Karovich I., Miasnikov A., Schupp P., and Shpilrain V.* Generic-case complexity, decision problems in group theory and random walks // *J. Algebra*. 2003. V. 264. No. 2. P. 665–694.
12. Гимади Э. Х., Глебов Н. И., Перепелица В. А. Алгоритмы с оценками для задач дискретной оптимизации // *Проблемы кибернетики*. 1975. Т. 31. С. 35–42.
13. *Impagliazzo R. and Wigderson A.* $P = BPP$ unless E has subexponential circuits: Derandomizing the XOR Lemma // *Proc. 29th STOC*. El Paso: ACM, 1997. P. 220–229.
14. Вьялый М., Китаев А., Шень А. Классические и квантовые вычисления. М.: МЦНМО, ЧеРо, 1999. 192 с.

REFERENCES

1. *Křivánek M. and Morávek J.* NP-hard problems in hierarchical-tree clustering. *Acta Informatica*, 1986, vol. 23, pp. 311–323.
2. *Bansal N., Blum A., and Chawla S.* Correlation clustering. *Machine Learning*, 2004, vol. 56, pp. 89–113.

3. Shamir R., Sharan R., and Tsur D. Cluster graph modification problems. *Discrete Appl. Math.*, 2004, vol. 144, no. 1–2, pp. 173–182.
4. Ageev A. A., Il'ev V. P., Kononov A. V., and Talevnin A. S. Computational complexity of the graph approximation problem. *J. Appl. Ind. Math.*, 2007, vol. 1, no. 1, pp. 1–8.
5. Il'ev V. P. and Il'eva S. D. O zadachakh klasterizatsii grafov [On problems of graph clustering]. *Vestnik Omskogo Universiteta*, 2016, no. 2, pp. 16–18. (in Russian)
6. Il'ev A. V. and Il'ev V. P. Ob odnoy zadache klasterizatsii grafa s chastichnym obucheniem [On a semi-supervised graph clustering problem]. *Prikladnaya Diskretnaya Matematika*, 2018, no. 42, pp. 66–75. (in Russian)
7. Talevnin A. S. O slozhnosti zadachi approksimatsii grafov [On the complexity of the graph approximation problem]. *Vestnik Omskogo Universiteta*, 2004, no. 4, pp. 22–24. (in Russian)
8. Il'ev V. P. and Navrotskaya A. A. Vychislitel'naya slozhnost' zadachi approksimatsii grafami s komponentami svyaznosti ogranichenogo razmera [Computational complexity of the problem of approximation by graphs with connected components of bounded size]. *Prikladnaya Diskretnaya Matematika*, 2011, no. 3(13), pp. 80–84. (in Russian)
9. Rybalov A. N. O genericheskoy slozhnosti problemy klasterizatsii grafov [On generic complexity of the graph clustering problem]. *Prikladnaya Diskretnaya Matematika*, 2019, no. 46, pp. 72–77. (in Russian)
10. Rybalov A. N. O genericheskoy slozhnosti ogranichennoy problemy klasterizatsii grafov [The generic complexity of the bounded problem of graphs clustering]. *Prikladnaya Diskretnaya Matematika*, 2022, no. 57, pp. 91–97. (in Russian)
11. Kapovich I., Miasnikov A., Schupp P., and Shpilrain V. Generic-case complexity, decision problems in group theory and random walks. *J. Algebra*, 2003, vol. 264, no. 2, pp. 665–694.
12. Gimadi E. X., Glebov N. I., and Perepelitsa V. A. Algoritmy s otsenkami dlya zadach diskretnoy optimizatsii [Algorithms with bounds for problems of discrete optimization]. *Problemy Kibernetiki*, 1975, vol. 31, pp. 35–42. (in Russian)
13. Impagliazzo R. and Wigderson A. $P=BPP$ unless E has subexponential circuits: Derandomizing the XOR Lemma. *Proc. 29th STOC, El Paso, ACM*, 1997, pp. 220–229.
14. Vyalyy M., Kitaev A., and Shen' A. Klassicheskie i kvantovye vychisleniya // Moscow, MCCME & CheRo Publ., 1999. 192 p. (in Russian)