УДК 510.52

# О ГЕНЕРИЧЕСКОЙ СЛОЖНОСТИ ПРОБЛЕМЫ КЛАСТЕРИЗАЦИИ ГРАФОВ[1]

А. Н. Рыбалов

*Институт математики им. С. Л. Соболева СО РАН, г. Омск, Россия*

Генерический подход к алгоритмическим проблемам предложен Мясниковым, Капович, Шуппом и Шпильрайном в 2003 г. В рамках этого подхода рассматривается поведение алгоритмов на множествах почти всех входов. В данной работе изучается генерическая сложность проблемы 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. Доказывается, что при условии $P \neq NP$ и $P = BPP$ для проблемы кластеризации графов 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.*

DOI 10.17223/20710410/46/6

# ON GENERIC COMPLEXITY OF THE GRAPH CLUSTERING PROBLEM

A. N. Rybalov

*Sobolev Institute of Mathematics, Omsk, Russia*

**E-mail:** alexander.rybalov@gmail.com

Generic-case approach to algorithmic problems was suggested by Miasnikov, Kapovich, Schupp and Shpilrain in 2003. This approach studies behavior of algorithms on typical (almost all) inputs and ignores the rest of inputs. In this paper, we study the generic complexity of the problem of clustering graphs. In this problem the structure of relations of objects 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) 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 the graph clustering problem there is no polynomial strongly generic algorithm. A strongly generic algorithm solves a problem not on the whole set of inputs, but on its subset, in which the sequence of frequencies of inputs converges exponentially fast to 1 with increasing its size.

**Keywords:** *generic complexity, graph clustering.*

[1]Работа поддержана грантом РНФ № 18-71-10028.

## Введение

Теория генерической вычислимости и сложности вычислений предложена в 2003 г. [1]. В рамках этого подхода алгоритмическая проблема рассматривается не на всём множестве входов, а на некотором подмножестве «почти всех» входов. Такие входы образуют так называемое генерическое множество. Понятие «почти все» формализуется введением естественной меры на множестве входных данных. С точки зрения практики алгоритмы, решающие быстро проблему на генерическом множестве, так же хороши, как и быстрые алгоритмы для всех входов. Классическим примером такого алгоритма является симплекс-метод — он за полиномиальное время решает задачу линейного программирования для большинства входных данных, но имеет экспоненциальную сложность в худшем случае. Более того, может так оказаться, что проблема трудноразрешима или вообще неразрешима в классическом смысле, но легкоразрешима на генерическом множестве. Отметим, что похожий подход для изучения проблем оптимизации был предложен ранее Э. Х. Гимади, Н. И. Глебовым и В. А. Перепелицей [2].

Одной из важных проблем машинного обучения является проблема кластеризации графов. В этой задаче структура взаимосвязей объектов задается с помощью графа, вершины которого соответствуют объектам, а ребра 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 [3 – 9] доказана NP-трудность проблемы кластеризации графа для различных её постановок. Таким образом, при условии $\mathrm{P} \neq \mathrm{NP}$ полиномиального алгоритма для решения этой задачи не существует. А при условии совпадения классов P и BPP (класс проблем, решаемых за полиномиальное время вероятностными алгоритмами) для неё не существует и полиномиальных вероятностных алгоритмов. Имеются серьёзные доводы в пользу равенства $\mathrm{P} = \mathrm{BPP}$. В частности, доказано [10], что это равенство следует из весьма правдоподобных гипотез о вычислительной сложности некоторых трудных проблем.

Данная работа посвящена изучению генерической сложности задачи кластеризации графов. Доказывается, что при условии $\mathrm{P} \neq \mathrm{NP}$ и $\mathrm{P} = \mathrm{BPP}$ для проблемы кластеризации графов 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.

## 1. Генерические алгоритмы

Пусть $I$ — некоторое множество входов, $I_n$ — подмножество входов размера $n$. Для подмножества $S \subseteq I$ определим последовательность

$$\rho_n(S) = \frac{|S_n|}{|I_n|}, \ n = 1, 2, 3, \ldots,$$

где $S_n = S \cap I_n$ — множество входов из $S$ размера $n$. Заметим, что $\rho_n(S)$ — это вероятность попасть в $S$ при случайной и равновероятной генерации входов из $I_n$. *Асимптотической плотностью* $S$ назовём предел

$$\rho(S) = \overline{\lim_{n\to\infty}} \rho_n(S).$$

Верхний предел здесь нужен потому, что часто при кодировании входных данных не для каждого $n$ существуют коды размера $n$. Множество $S$ называется *пренебрежи-*

*мым*, если $\rho(S) = 0$, и *сильно пренебрежимым*, если последовательность $\rho_n(S)$ экспоненциально быстро сходится к 0, т. е. существуют константы $\sigma$, $0 < \sigma < 1$, и $C > 0$ такие, что для любого $n$ имеет место $\rho_n(S) < C\sigma^n$.

Алгоритм $\mathcal{A}$ с множеством входов $I$ и множеством выходов $J \cup \{?\}$ $(? \notin J)$ называется (*сильно*) *генерическим*, если

1) $\mathcal{A}$ останавливается на всех входах из $I$;
2) множество $\{x \in I : \mathcal{A}(x) = ?\}$ является (сильно) пренебрежимым.

Генерический алгоритм $\mathcal{A}$ вычисляет функцию $f : I \to J$, если для всех $x \in I$ выполнено

$$(\mathcal{A}(x) = y \in J) \Rightarrow (f(x) = y).$$

Ситуация $\mathcal{A}(x) = ?$ означает, что $\mathcal{A}$ не может вычислить функцию $f$ на аргументе $x$. Но условие 2 гарантирует, что $\mathcal{A}$ корректно вычисляет $f$ на почти всех входах (входах из генерического множества). Различие между генерически разрешимыми проблемами и сильно генерически разрешимыми проблемами поясняется в работе [11].

## 2. Проблема кластеризации графа

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 $\mathcal{M}(V)$ множество всех кластерных графов на множестве вершин $V$. 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

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

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

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

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

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

где $C_i$, $i = 1, \ldots, m$, — непересекающиеся полные компоненты связности, 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 $C_{k,i} = K(V_c \cap V_i)$, $i = 1, 2$. Вершины первого лежат в графе $G_1$, а второго — в $G_2$. Обозначим через $M'$ получившийся кластерный граф. 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 $G_2$. Последних рёбер нет и в графе $G_1 \cup G_2$.

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

## 3. Основной результат

Для изучения генерической сложности проблемы кластеризации графов будем использовать представление графов с помощью матриц смежности. Поскольку 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 $n(n-1)/2$.

**Теорема 1.** Если существует сильно генерический полиномиальный алгоритм, решающий проблему кластеризации графа, то существует вероятностный полиномиальный алгоритм, разрешающий эту проблему на всём множестве входов.

***Доказательство.*** Допустим, что существует сильно генерический полиномиальный алгоритм $\mathcal{A}$, решающий проблему кластеризации графа. Построим вероятностный полиномиальный алгоритм $\mathcal{B}$, разрешающий эту проблему на всём множестве входов, который на графе $G$ с $n$ вершинами (размера $n(n-1)/2$) 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 кластеризации для графа $G \cup H$, согласно лемме 1, выдаёт решение проблемы кластеризации для графа $G$.
4) Если $\mathcal{A}(G \cup H) = ?$, то выдаёт ответ $K_n$.

Заметим, что 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 $1/2$.

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

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

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

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

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

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

при больших $n$. ■

Непосредственным следствием теоремы 1 является следующее утверждение.

**Теорема 2.** Если P $\neq$ NP и P = BPP, то не существует сильно генерического полиномиального алгоритма для решения проблемы кластеризации графов.

## ЛИТЕРАТУРА


1. *Kapovich 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.
2. *Гимади Э.Х., Глебов Н.И., Перепелица В.А.* Алгоритмы с оценками для задач дискретной оптимизации // Проблемы кибернетики. 1975. Т. 31. С. 35–42.
3. *Kŕivanek M. and Morávek J.* NP-hard problems in hierarchical-tree clustering // Acta Informatica. 1986. V. 23. P. 311–323.
4. *Bansal N., Blum A., and Chawla S.* Correlation clustering // Machine Learning. 2004. V. 56. P. 89–113.
5. *Shamir R., Sharan R., and Tsur D.* Cluster graph modification problems // Discrete Appl. Math. 2004. V. 144. No. 1–2. P. 173–182.
6. *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.
7. *Ильев В. П., Ильева С. Д.* О задачах кластеризации графов // Вестник Омского университета. 2016. № 2. С. 16–18.
8. *Ильев А. В., Ильев В. П.* Об одной задаче кластеризации графа с частичным обучением // Прикладная дискретная математика. 2018. № 42. С. 66–75.
9. *Талевнин А. С.* О сложности задачи аппроксимации графов // Вестник Омского университета. 2004. № 4. С. 22–24.
10. *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.
11. *Рыбалов А. Н.* О генерической сложности проблемы общезначимости булевых формул // Прикладная дискретная математика. 2016. № 2(32). С. 119–126.


## REFERENCES


1. *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.
2. *Gimadi E.H., Glebov N.I., and Perepelitsa V.A.* Algoritmy s ocenkami dlya zadach diskretnoi optimizacii [Algorithms with bounds for problems of discrete optimization]. Problemy Kibernetiki, 1975, vol. 31, pp. 35–42. (in Russian)
3. *Kŕivanek M. and Morávek J.* NP-hard problems in hierarchical-tree clustering. Acta Informatica, 1986, vol. 23, pp. 311–323.
4. *Bansal N., Blum A., and Chawla S.* Correlation clustering. Machine Learning, 2004, vol. 56, pp. 89–113.
5. *Shamir R., Sharan R., and Tsur D.* Cluster graph modification problems. Discrete Appl. Math., 2004, vol. 144, no. 1–2, pp. 173–182.
6. *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.

7. *Il'ev V. P. and Il'eva S. D.* O zadachah klasterizacii grafov [On problems of graph clustering]. Vestnik Omskogo Universiteta, 2016, no. 2, pp. 16–18. (in Russian)
8. *Il'ev A. V. and Il'ev V. P.* Ob odnoi zadache klasterizacii grafa s chastichnym obucheniem [On a problem of graph clustering with partial learning]. Prikladnaya Diskretnaya Matematika, 2018, no. 42, pp. 66–75. (in Russian)
9. *Talevnin A. S.* O slozhnosti zadachi approksimacii grafov [On the complexity of the graph approximation problem]. Vestnik Omskogo Universiteta, 2004, no. 4, pp. 22–24. (in Russian)
10. *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.
11. *Rybalov A. N.* O generricheskoy slozhnosti problemy obshcheznachimosti bulevykh formul [On generic complexity of the validity problem for Boolean formulas]. Prikladnaya Diskretnaya Matematika, 2016, no. 2(32), pp. 119–126. (in Russian)