Czechoslovak Mathematical Journal, 65 (140) (2015), 829–836

EDIT DISTANCE MEASURE FOR GRAPHS

Tomasz Dzido, Gdańsk, Krzysztof Krzywdziński, Poznań

(Received August 20, 2014)

Abstract. In this paper, we investigate a measure of similarity of graphs similar to the Ramsey number. We present values and bounds for $g(n,l)$, the biggest number k guaranteeing that there exist l graphs on n vertices, each two having edit distance at least k. By edit distance of two graphs G, F we mean the number of edges needed to be added to or deleted from graph G to obtain graph F. This new extremal number $g(n,l)$ is closely linked to the edit distance of graphs. Using probabilistic methods we show that $g(n,l)$ is close to $\frac{1}{2}\binom{n}{2}$ for small values of $l > 2$. We also present some exact values for small n and lower bounds for very large l close to the number of non-isomorphic graphs of n vertices.

Keywords: extremal graph problem; similarity of graphs

MSC 2010: 05C35, 05C75

1. Introduction

In this paper we describe a new measure of similarity of graphs which is similar to the Ramsey number. We are interested in the biggest number k guaranteeing that there exist l graphs on n vertices, each two having edit distance at least k. We define the edit distance of two graphs G, F to be the number of edges needed to be added to or deleted from the graph G to obtain the graph F and we denote it by $s(G,F)$. More formally, let $\triangle$ be the symmetric difference of the labeled edge sets. For two graphs G and F on the vertex set $[n]$, $s(G,F)$ is the smallest number of switches of edges and non-edges of G that are necessary to turn it into a graph isomorphic to F, i.e.,

$$s(G,F) = \min\left\{|E'|\colon\ E' \subseteq \binom{[n]}{2},\ (V(G), E(G) \triangle E') \cong F\right\}.$$

This research was partially supported by the Polish National Science Centre grant number 2011/02/A/ST6/00201.

Let $\mathcal{P}$ be a family of all graphs on a vertex set $[n] = \{1, \ldots, n\}$. We study the number

$$g(n,l) = \max\{k \in \mathbb{N} \colon \exists_{S \subseteq \mathcal{P}, |S|=l} \forall_{G,F \in S, G \neq F} \colon s(G,F) \geqslant k\}.$$

The edit distance problem began with the following question of Chen, Eulenstein, Fernández-Baca and Sanderson [6]: given a bipartite graph G, how many edge-deletions plus edge-additions are necessary to ensure that G has no copy of M as an induced subgraph. The study of the edit distance in graphs was originated independently by Axenovich, Kézdy and Martin [4] and Alon and Stav [3]. Since then, there has been a great deal of study on the edit distance itself and on the so-called edit distance function (see for example [1], [5]).

To the best of our knowledge the problem of finding $g(n,l)$ has never been invastigated before. A similar problem was introduced by Chang et al. in [7]. They invastigate the following problem: given integers n, e, e', what is the largest number $g(n,e,e')$ such that any two n vertex graphs G and H, with e and e' edges respectively, must have a common subgraph with at least $g(n,e,e')$ edges. Another Ramsey like similarity measure was introduced by Dzido and Krzywdzinski in [9]. The authors consider the problem of finding the smallest n to ensure at least one k-similar pair in any family of l graphs on n vertices. In [9] two graphs G and H, having the same number of vertices n, are k-similar if they contain a common induced subgraph of order k.

In the next sections, we will present some values of and bounds on $g(n,l)$.

2. Main results

In this paper all graphs are undirected, finite and contain neither loops nor multiple edges. Let G be a graph. $\overline{G}$ is the complement of G. By K_n we denote n-vertex clique and N_n is an independent set of size n.

First we have the following observation showing that for $l > 1$ the function $g(n,l)$ is well defined.

Observation 2.1. $g(n,2) = \binom{n}{2}$ and $g(n,l) \leqslant \binom{n}{2}$ for $l > 2$.

Proof. Observe that two graphs with n vertices have edit distance at most $\binom{n}{2}$. So $g(n,2) \leqslant \binom{n}{2}$. On the other hand $s(K_n, N_n) = \binom{n}{2}$, so $g(n,2) \geqslant \binom{n}{2}$.

Since $g(n,l) \geqslant g(n,l')$ for $l < l'$, we immediately obtain that for $l > 2$ we have $g(n,l) \leqslant \binom{n}{2}$. □

Now we give the value of $g(n,3)$. We use following lemma:

Lemma 2.2. *Let G and F be arbitrary n-vertex graphs, then*

$$s(G,F) \leqslant \frac{e(G)\big(\binom{n}{2} - e(F)\big) + e(F)\big(\binom{n}{2} - e(G)\big)}{\binom{n}{2}}.$$

Proof. For two graphs H_1 and H_2 on the vertex set $[n]$, let $s_{\mathrm{ind}}(H_1, H_2)$ be the Hamming distance between H_1 and H_2, i.e., the number of two-element subsets of $[n]$ that are edges of exactly one of the two graphs. Let π be a random permutation of $[n]$ and G and F two graphs on vertex set $[n]$. We denote by $\pi(G)$ the (random) graph obtained by applying the permutation π on the vertex set of G. Then clearly for every two-element subset of $[n]$, the probability that after the permutation it is an edge of F is $e(F)/\binom{n}{2}$, and similarly the probability that it is a non-edge of F is $1 - e(F)/\binom{n}{2}$. Hence,

$$\mathbb{E}(s_{\mathrm{ind}}(\pi(G), F)) = \frac{e(G)\big(\binom{n}{2} - e(F)\big) + e(F)\big(\binom{n}{2} - e(G)\big)}{\binom{n}{2}},$$

and the lemma follows. □

Theorem 2.3. *Let $n \geqslant 2$. Then*

$$g(n,3) = \left\lfloor \frac{1}{2}\binom{n}{2} \right\rfloor.$$

Proof. Let $\alpha = \left\lfloor \frac{1}{2}\binom{n}{2} \right\rfloor$. For the lower bound $g(n,3) \geqslant \alpha$, consider the graphs $G_1 = K_n$, $G_2 = N_n$ and any graph G_3 with α edges. Then any two graphs among G_1, G_2, G_3 have edit distance at least α.

For the upper bound $g(n,3) \leqslant \alpha$, suppose that there are three n-vertex graphs G_1, G_2, G_3 such that any two of them have edit distance at least $\alpha + 1$. First suppose that $e(G_1) \leqslant \frac{1}{2}\binom{n}{2}/a$ and $e(G_2) \leqslant \frac{1}{2}\binom{n}{2}/b$ where $a, b \geqslant 1$. Then by Lemma 2.2 we have

$$s(G_1, G_2) \leqslant \binom{n}{2}\Big(\frac{1}{2a} + \frac{1}{2a} - \frac{1}{2ab}\Big) \leqslant \frac{1}{2}\binom{n}{2}.$$

This is contradiction with the assumption that $s(G_1, G_2) \geqslant \alpha + 1 = \left\lfloor \frac{1}{2}\binom{n}{2} \right\rfloor + 1$.

Now suppose that among G_1, G_2, G_3 there are two graphs with more than $\frac{1}{2}\binom{n}{2}$ edges. Without loss of generality we assume that $|e(G_1)| > \frac{1}{2}\binom{n}{2}$ and $|e(G_2)| > \frac{1}{2}\binom{n}{2}$. So we have that $|e(\overline{G_1})| \leqslant \frac{1}{2}\binom{n}{2}$, $|e(\overline{G_2})| \leqslant \frac{1}{2}\binom{n}{2}$ and also $s(G_1, G_2) = s(\overline{G_1}, \overline{G_2}) \leqslant \frac{1}{2}\binom{n}{2}$. But this is contradiction with the assumption that $s(G_1, G_2) \geqslant \alpha + 1$. □

Theorem 2.4. *Let $l \in \{3,4,5,6\}$. Then*

$$\frac{n^2}{4} - \frac{n}{2} \leqslant g(n,l) \leqslant \left\lfloor \frac{1}{2}\binom{n}{2} \right\rfloor.$$

Proof. We construct a class of 6 graphs in which any two graphs have a large edit distance. Consider the family of graphs of order n, say $G_1, G_2, \ldots, G_6$, as follows: $G_1 = K_n$, $G_2 = N_n$, $G_3 = K_{\lceil (n-1)/2 \rceil, \lfloor (n+1)/2 \rfloor}$, $G_4 = \overline{G_3} = K_{\lceil (n-1)/2 \rceil} \cup K_{\lfloor (n+1)/2 \rfloor}$, $G_5 = K_{\lceil n/\sqrt{2} \rceil} \cup N_{n-\lceil n/\sqrt{2} \rceil}$ and $G_6 = \overline{G_5}$. We show that for every $1 \leqslant i < j \leqslant 6$, $s(G_i, G_j) \geqslant n^2/4 - n/2$.

It is a simple observation that

$$s(G_i, G_j) \geqslant \frac{n^2}{4} - \frac{n}{2} \quad \text{for } i \in \{1,2\} \text{ and } i < j \leqslant 6.$$

Let us now evaluate $s(G_3, G_5)$. Let $\{A, B\}$ be the bipartition of G_3. Define C and D to be the subsets of $V(G_5)$ such that C contains all vertices of the clique and D contains the remaining vertices of $V(G_5)$. Any isomorphism f between G_3 and G_5 maps in some way the set A to C and D. Let us divide the sets C and D into 4 subsets as follows:

(1) $ac = |f(A) \cap C|$;
(2) $ad = |f(A) \cap D|$ (naturally $ac + ad = |A| = \lceil (n-1)/2 \rceil$ and hence $ad = \lceil (n-1)/2 \rceil - ac$);
(3) $bc = |C| - ac = \lceil n/\sqrt{2} \rceil - ac$;
(4) $bd = |D| - ad = n - \lceil n/\sqrt{2} \rceil - (\lceil (n-1)/2 \rceil - ac)$.

Thus

$$s_f(G_3, G_5) = \frac{ac(ac-1)}{2} + \frac{bc(bc-1)}{2} + ad \cdot bd + ac \cdot bd + ad \cdot bc,$$

where $0 \leqslant ac \leqslant \lceil n/\sqrt{2} \rceil$.

We obtain that

$$\begin{aligned} s_f(G_3, G_5) &= \frac{ac(ac-1)}{2} + \frac{\bigl(\lceil n/\sqrt{2} \rceil - ac\bigr)\bigl(\lceil n/\sqrt{2} \rceil - ac - 1\bigr)}{2} \\ &\quad + \left\lceil \frac{n-1}{2} \right\rceil \left(n - \left\lceil \frac{n}{\sqrt{2}} \right\rceil - \left(\left\lceil \frac{n-1}{2} \right\rceil - ac\right)\right) \\ &\quad + \left(\left\lceil \frac{n-1}{2} \right\rceil - ac\right)\left(\left\lceil \frac{n}{\sqrt{2}} \right\rceil - ac\right) \\ &= 2ac^2 - 2ac\left\lceil \frac{n}{\sqrt{2}} \right\rceil + \frac{1}{2}\left\lceil \frac{n}{\sqrt{2}} \right\rceil^2 - \frac{1}{2}\left\lceil \frac{n}{\sqrt{2}} \right\rceil + n\left\lceil \frac{n-1}{2} \right\rceil - \left\lceil \frac{n-1}{2} \right\rceil^2. \end{aligned}$$

One can calculate that the last formula achieves the minimum value for $ac = \lceil n/\sqrt{2}\rceil/2$. The minimum value is $n\lceil(n-1)/2\rceil - (\lceil(n-1)/2\rceil)^2 - \lceil n/\sqrt{2}\rceil/2$ which is greater than $n^2/4 - n/2$.

Similarly, in the case $s_f(G_3, G_6)$ let us divide the sets C and D into 4 subsets as follows:

(1) $bd = |f(B) \cap D|$;
(2) $bc = |f(B) \cap C|$ (naturally $bc + bd = |B| = \lfloor(n+1)/2\rfloor$ and hence $bc = \lfloor(n+1)/2\rfloor - bd$);
(3) $ad = |D| - bd = n - \lceil n/\sqrt{2}\rceil - bd$;
(4) $ac = |C| - bc = \lceil n/\sqrt{2}\rceil - \lfloor(n+1)/2\rfloor + bd$.

Thus

$$s_f(G_3, G_6) = \frac{ad(ad-1)}{2} + \frac{bd(bd-1)}{2} + ac \cdot ad + bc \cdot bd + ac \cdot bc,$$

where $0 \leqslant bd \leqslant \lfloor(n+1)/2\rfloor$.

One can calculate that the last formula achieves the minimum value for $bd = \lfloor(n+1)/2\rfloor$. The minimum value is $n^2/2 - n/2 + (\lfloor(n+1)/2\rfloor)^2 - n\lfloor(n+1)/2\rfloor$ which is greater or equal to $n^2/4 - n/2$.

Since $G_3 = \overline{G_4}$ and $G_5 = \overline{G_6}$, similar calculations lead us to the same result on two cases $s_f(G_4, G_5)$ and $s_f(G_4, G_6)$. The remaining case $s_f(G_5, G_6)$ results from the properties of graphs.

Theorem 2.3 gives also an upper bound $g(n,l) \leqslant \left\lfloor \frac{1}{2}\binom{n}{2} \right\rfloor$. □

The next observation shows that we can find a very large collection of graphs with the pairwise distance close to the extremal value $\frac{1}{2}\binom{n}{2}$.

Theorem 2.5. *For every $\varepsilon > 0$ there exist $\delta > 0$ and n_0 such that for every $n \geqslant n_0$ we have*

$$\Big(\frac{1}{2} - \varepsilon\Big)\binom{n}{2} < g(n, 2^{\delta n^2}) \leqslant \left\lfloor \frac{1}{2}\binom{n}{2} \right\rfloor.$$

Proof. Assume without loss of generality that $\varepsilon < 1/3$.

Let G_1 and G_2 be two independent random graphs drawn according to the distribution of $G(n, 1/2)$ (a graph in which each edge is added independently with probability $1/2$). By Chernoff's inequality (see, e.g., Theorems A.1.11 and A.1.13 in [2])

$$\mathbb{P}\Big(s(G_1, G_2) < \Big(\frac{1}{2} - \varepsilon\Big)\binom{n}{2}\Big) < n!\, 2^{(-\mathrm{e}^2/6)\binom{n}{2}}.$$

Thus, sampling l independent graphs $G_1, \dots, G_l$ with the distribution $G(n, 1/2)$ and applying the union bound over all choices to take two of them, we obtain

$$\mathbb{P}\left(\exists i,\ j \in [l],\ i \neq j\colon\ s(G_i, G_j) < \left(\frac{1}{2} - \varepsilon\right)\binom{n}{2}\right) \\ < \binom{l}{2} n!\, 2^{(-\mathrm{e}^2/6)\binom{n}{2}} < 2^{(2\delta - \mathrm{e}^2/14)n^2} \to 0$$

for $\delta < \mathrm{e}^2/28$ and sufficiently large n. □

At the end of this section we give exact results for small values of n.

Theorem 2.6. *The values of* $g(n, l)$ *for* $n = 3$ *and* $n = 4$ *are as follows:*

l	2	3	4	5	6	7	8	9	10	11	$\geqslant 12$
$n = 3$	3	1	1	0	0	0	0	0	0	0	0
$n = 4$	6	3	2	2	2	1	1	1	1	1	0

Proof. The cases $l = 2$ and $l = 3$ are consequences of Observation 2.1 and Theorem 2.3, respectively. The cases $l \geqslant 4$ for $n = 3$ and $l \geqslant 11$ for $n = 4$ are the consequences of Observation 3.1. The cases $l = 4, 5$ for $n = 4$ follow from Theorem 3.2 and the cases $7 \leqslant l \leqslant 10$ follow from the proof of Theorem 3.5. For the remaining case $l = 6$ for $n = 4$ consider all non-isomorphic graphs of order 4 with even number of edges. □

3. Large l

Let Q_n denote the set of all non-isomorphic graphs on n vertices. In this section we will state various properties of $g(n, l)$ for l close to the number $|Q_n|$. The exact value of the number $|Q_n|$ is not known. To the best of our knowledge, the best known estimate on $|Q_n|$ is $\sqrt{T_n} < |Q_n| < T_n$, where $T_n = 2^{\binom{n}{2}}$, which was proved by de Wet [8].

Simply by the definition of $g(n, l)$ and Q_k we get the following observation.

Observation 3.1. $g(n, |Q_n|) = 1$ and for all $l > |Q_n|$, $g(n, l) = 0$.

Theorem 3.2. *For $l \leqslant |Q_n|/2$ we have $g(n,l) \geqslant 2$.*

Proof. Divide the set $|Q_n|$ into two subsets of all non-isomorphic graphs with odd and even number of edges, respectively. By Pigeonhole Principle, the size of one of these subsets of graphs, say with even number of edges, is at least $|Q_n|/2$. Therefore, if G_1 and G_2 are non-isomorphic graphs with even number of edges, then $s(G_1, G_2) \geqslant 2$ and the proof is complete. □

Now we give estimates on $g(n,l)$ for l close to $|Q_n|$. We use the following definition.

Definition 3.3. Let H_n be the graph for which $V(H_n)$ is the set of all non-isomorphic graphs on n vertices. Two vertices of H_n are connected by an edge if the graphs corresponding to these vertices have edit distance equal to 1.

Observe that H_n is a bipartite graph with partition classes consisting of graphs with odd and even number of edges, respectively. Let us consider a maximum matching M in H_n.

Lemma 3.4. *The cardinality of a maximum matching M in H_n is at least $|Q_{n-1}|/2$.*

Proof. Consider the set S consisting of all non-isomorphic graphs which are a disjoint union of a graph on $n-1$ vertices and one isolated vertex v. It is clear that $|S| = |Q_{n-1}|$. Define a function $f\colon\ S \to S'$ which assigns to each graph $G \in S$ a new graph $G' \in S'$ created by joining v with the maximum degree vertex from $V(G) - v$. Observe that f is surjective since if $G' \in S'$ is isomorphic to $F' \in S'$, then $G \in S$ is isomorphic to $F \in S$. It means that for any graph $G \in S$ we assign exactly one graph $G' \in S'$. The assignment f forms the edges in the graph H_n. Since it may happen that $G' \in S'$ is also an element of S, matching M in H_n has at least $|Q_{n-1}|/2$ edges. □

Theorem 3.5. *For $|Q_n| \geqslant l > |Q_{n-1}|/2 \geqslant |Q_n|(1 - 2^{-n-1})$,*

$$g(n,l) = 1.$$

Proof. Consider a family of l non-isomorphic graphs on n vertices. Let M be a matching in H. If $l > |Q_n| - |M|$, then there are at least two graphs connected by a matching M in H_n and these graphs have edit distance equal to 1. Since $|Q_{n-1}|2^n > |Q_n|$ and by Lemma 3.4 we obtain the result. □

Acknowledgement. The authors are grateful to Professor Roman Glebov whose valuable suggestions resulted in an improved paper.

References

[1] *N. Alon, A. Shapira, B. Sudakov*: Additive approximation for edge-deletion problems. Ann. Math. (2) *170* (2009), 371–411. zbl MR

[2] *N. Alon, J. H. Spencer*: The Probabilistic Method. Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Hoboken, 2008. zbl MR

[3] *N. Alon, U. Stav*: What is the furthest graph from a hereditary property? Random Struct. Algorithms *33* (2008), 87–104. zbl MR

[4] *M. Axenovich, A. Kézdy, R. Martin*: On the editing distance of graphs. J. Graph Theory *58* (2008), 123–138. zbl MR

[5] *J. Balogh, R. Martin*: Edit distance and its computation. Electron. J. Comb. (electronic only) *15* (2008), Research Paper R20, 27 pages. zbl MR

[6] *D. Chen, O. Eulenstein, D. Fernández-Baca, M. Sanderson*: Supertrees by flipping. Computing and Combinatorics (O. H. Ibarra et al., eds.). Proc. of the 8th Annual International Conf., Singapore, 2002, Lecture Notes in Comput. Sci. 2387, Springer, Berlin, 2002, pp. 391–400. zbl MR

[7] *F. R. K. Chung, P. Erdős, R. L. Graham*: Minimal decompositions of graphs into mutually isomorphic subgraphs. Combinatorica *1* (1981), 13–24. zbl MR

[8] *P. O. de Wet*: Constructing a large number of nonisomorphic graphs of order n. Morehead Electronic Journal of Applicable Mathematics *1* (2001), 2 pages.

[9] *T. Dzido, K. Krzywdziński*: On a local similarity of graphs. Discrete Math. *338* (2015), 983–989. MR

Authors' addresses: Tomasz Dzido, Institute of Informatics, University of Gdańsk, ul. Wita Stwosza 57, 80-952 Gdańsk, Poland, e-mail: `tdz@inf.ug.edu.pl`; Krzysztof Krzywdziński, Faculty of Mathematics and Computer Science, Adam Mickiewicz University, ul. Umultowska 87, 61-614 Poznań, Poland, e-mail: `kkrzywd@amu.edu.pl`.