On the Number of 4-Edge Paths in Graphs With Given Edge Density

DÁNIEL T. NAGY[†]

Eötvös Loránd University, Egyetem tér 1-3, Budapest 1053, Hungary
(e-mail: dani.t.nagy@gmail.com)

We investigate the number of 4-edge paths in graphs with a given number of vertices and edges, proving an asymptotically sharp upper bound on this number. The extremal construction is the quasi-star or the quasi-clique graph, depending on the edge density. An easy lower bound is also proved. This answer resembles the classic theorem of Ahlswede and Katona about the maximal number of 2-edge paths, and a recent theorem of Kenyon, Radin, Ren and Sadun about k-edge stars.

2010 *Mathematics subject classification*: Primary 05C35

1. Introduction

The aim of this paper is to asymptotically determine the maximal and minimal number of 4-edge paths in graphs with a given number of vertices and edges.

The first result of this kind is due to Ahlswede and Katona [1], who described the graphs with a given number of vertices and edges containing the maximal number of 2-edge paths. To state this result, we need some simple definitions.

Definition. The *quasi-clique* C_n^e is a graph with n vertices and e edges, defined as follows. Take the unique representation

$$e = \binom{a}{b} + b \quad 0 \leqslant b < a,$$

connect the first a vertices to each other, and connect the $(a+1)$th vertex to the first b vertices

Definition. The *quasi-star* S_n^e is a graph with n vertices and e edges, defined as follows. Take the unique representation

$$\binom{n}{2} - e = \binom{p}{2} + q, \quad 0 \leqslant q < p,$$

connect the first $n - p - 1$ vertices to every vertex, and connect the $(n-p)$th vertex to the first $n - q$ vertices.

It is easy to see that S_n^e is isomorphic to the complement of $S_n^{\binom{n}{2}-e}$.

Definition. The numbers of 2-edge paths in C_n^e and S_n^e are denoted by $C(n,e)$ and $S(n,e)$ respectively, while the numbers of k-edge stars in these graphs are denoted by $C_k(n,e)$ and $S_k(n,e)$ respectively.

Theorem 1.1 (Ahlswede and Katona 1978 [1]). *Let G be a simple graph with n vertices and e edges. Then the number of 2-edge paths in G is at most* $\max(C(n,e),\ S(n,e))$.
Furthermore,

$$\max(C(n,e),\ S(n,e)) = \begin{cases} S(n,e) & \text{if } 0 \leqslant e \leqslant \frac{1}{2}\binom{n}{2} - \frac{n}{2}, \\ C(n,e) & \text{if } \frac{1}{2}\binom{n}{2} + \frac{n}{2} \leqslant e \leqslant \binom{n}{2}. \end{cases}$$

Roughly speaking, this theorem states that if the edge density is smaller than $1/2$, then the quasi-star is the extremal example, while for higher edge densities the quasi-clique becomes extremal. (The transition between the two cases happens in a non-trivial way.)

Recently, Kenyon, Radin, Ren and Sadun [7] proved a similar result for k-edge stars, using the notion of graphons. Translating the result back to language of graphs, we get the following theorem.

Theorem 1.2 (Kenyon, Radin, Ren and Sadun [7]). *Let G be a simple graph with n vertices and e edges, and let $2 \leqslant k \leqslant 30$. Then the number of k-edge stars in G is at most*

$$\max(C_k(n,e),\ S_k(n,e))(1 + O(e^{-1/2})).$$

The theorem is conjectured to hold for all values of k. (This conjecture would follow from the solution of a certain complicated maximization problem.) Similarly to the case of the 2-edge path, $C_k(n,e) < S_k(n,e)$ if the edge density is small, and $S_k(n,e) < C_k(n,e)$ if it is greater. The point of transition depends on k.

Now let us discuss three theorems with just one fixed parameter: the number of edges. (Here, n is not fixed.) We will start with a general theorem of Alon [2].

Theorem 1.3 (Alon [2]). *Let $N(G,H)$ denote the number of subgraphs of G that are isomorphic to H. Assume that H is a simple graph that has a spanning subgraph which is the*

vertex-disjoint union of edges and cycles. Then

$$N(G,H) \leqslant (1 + O(e^{-1/2}))N(C_n^e, H).$$

This means that for these graphs H, the asymptotically extremal example is always the quasi-clique. Note that this theorem can be applied in the case of fixed n and e, since the extremal example provided by it is the quasi-clique. (No matter how many vertices we are given, we just have to construct a quasi-clique of e edges.)

Also note that this theorem provides upper bounds for all graphs with a perfect matching (*e.g.* all paths with an odd number of edges) and Hamiltonian graphs (*e.g.* complete graphs). For the number of the cliques K_r, the asymptotically best lower bound was proved by Reiher [11]. (The cases $r = 3$ and $r = 4$ were solved by Razborov [10] and Nikiforov [9].)

The problem of finding the maximal number of 4-edge paths in graphs with e edges (and an unlimited number of vertices) was solved by Bollobás and Sarkar [5].

Theorem 1.4 (Bollobás and Sarkar [5]). *The number of* 4-*edge paths among graphs with e edges is maximized by the graph that is obtained by taking the complete bipartite graph $K_{2,\lceil e/2 \rceil}$, and deleting an edge if e is odd.*

Bollobás and Sarkar [4] also proved asymptotic results for $2k$-edge paths. The extremal example in this case is the complete bipartite graph with $\lfloor (9k^2 + 3k + 5)/18 \rfloor$ vertices on one side. For $(2k+1)$-edge paths, the asymptotically extremal example is the quasi-clique. This follows from Theorem 1.3, and is also proved in [4].

Alon [3] had a conjecture for star-forests (vertex-disjoint union of stars), which was partially verified by Füredi [6].

Conjecture 1.5 (Alon [3]). *Let H be a star-forest. For any $e > 0$, the graph maximizing the number of subgraphs isomorphic to H among graphs with e edges is a star-forest.*

Theorem 1.6 (Füredi [6]). *Let H be a star-forest consisting of components with $a_1, a_2, \ldots, a_t$ edges. Assume that $a_i > \log_2(t+1)$ for all $1 \leqslant i \leqslant t$. Let e be sufficiently large. Then the graph maximizing the number of subgraphs isomorphic to H among those with e edges is a star-forest with t components.*

Considering the above results, investigating the number of the 4-edge paths seems to be the 'natural' choice in the case of fixed (n, e). In this paper, an asymptotic upper bound will be given for this quantity. As in the case of k-edge stars, the asymptotically extremal graphs are the quasi-stars and the quasi-cliques. We will also prove an easy asymptotic lower bound.

Theorem 1.7. *Let G be a simple graph with n vertices and e edges. Let $c = 2e/n^2$ (so $0 \leqslant c \leqslant 1$). Let N denote the number of* 4-*edge-paths in G. Then*

$$\frac{1}{2}c^4 n^5 (1 - O(n^{-1})) \leqslant N \leqslant \frac{1}{2} \max((1 - \sqrt{1-c})^2((c+1)\sqrt{1-c} + c),\ c^{5/2})n^5.$$

In Proposition 3.1 we will present constructions to show that these bounds are asymptotically best possible.

2. Proof of the main result

Let N' denote the number of the sequences $(v_0, v_1, v_2, v_3, v_4)$ where v_i are (not necessarily different) vertices of G and $v_{i-1} v_i \in E(G)$ for $i = 1, \ldots, 4$. Here, we count every 4-edge path twice (there are two directions). We also count some walks of length 4 with repeated vertices. However, the number of such walks is only $O(n^4)$. Therefore $2N \leqslant N' \leqslant 2N + O(n^4)$, so it suffices to prove

$$c^4 \leqslant \frac{N'}{n^5} \leqslant \max((1 - \sqrt{1-c})^2((c+1)\sqrt{1-c}+c),\ c^{5/2}). \tag{2.1}$$

Let us note that N'/n^5 is often referred to as the homomorphism density of the 4-edge path P^4 in G, and denoted by $t(P^4, G)$. (See [8] for an overview in the topic of graph homomorphisms.)

First, we prove the lower bound, which is much easier. If we want to select a 4-edge walk, we can start by choosing v_2, then v_1 and v_3 (we have to pick them from $N(v_2)$), and finally v_0 and v_4 ($\deg(v_1)$ and $\deg(v_3)$ possibilities). Thus we can write N' as in (2.2) and then estimate it twice over via the Cauchy–Schwarz inequality

$$\sum_{i=1}^{m} x_i^2 \geqslant \frac{1}{n} \left(\sum_{i=1}^{m} x_i \right)^2$$

for all real numbers, obtaining

$$
\begin{aligned}
N' &= \sum_{v_2 \in V(G)} \left(\sum_{v_i \in N(v_2)} \deg(v_i) \right)^2 \geqslant \frac{1}{n} \left(\sum_{v_2 \in V(G)} \left(\sum_{v_i \in N(v_2)} \deg(v_i) \right) \right)^2 \\
&= \frac{1}{n} \left(\sum_{v_i \in V(G)} \deg(v_i)^2 \right)^2 \geqslant \frac{1}{n} \left(\frac{1}{n} \left(\sum_{v_i \in V(G)} \deg(v_i) \right)^2 \right)^2 \\
&= \frac{1}{n} \left(\frac{1}{n} \left(cn^2 \right)^2 \right)^2 = c^4 n^5. \tag{2.2}
\end{aligned}
$$

Now we move on to the proof of the upper bound. Let $\text{codeg}(v, w)$ denote the number of common neighbours of the vertices v and w. Note that

$$N' = \sum_{v_1, v_3 \in V(G)} \deg(v_1) \deg(v_3) \text{codeg}(v_1, v_3),$$

since after fixing v_1 and v_3 we have $\deg(v_1)$ candidates for v_0, $\deg(v_3)$ candidates for v_4, and $\text{codeg}(v_1, v_3)$ candidates for v_2. Obviously, $\text{codeg}(v_1, v_3) \leqslant \min(\deg(v_1), \deg(v_3))$, and therefore

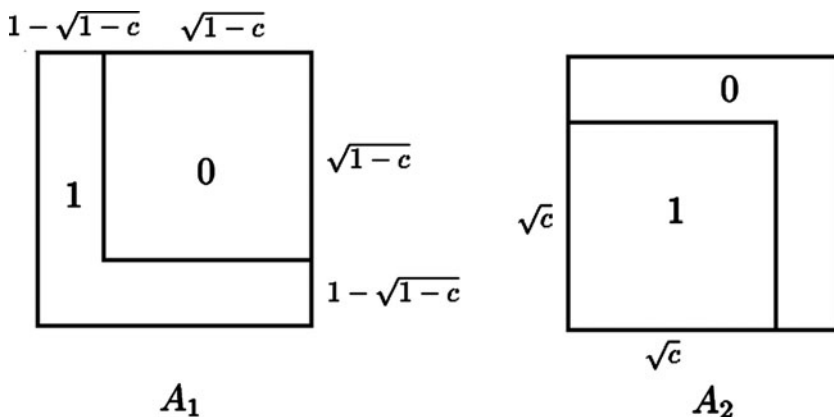$$N' \leqslant \sum_{v_1, v_3 \in V(G)} \deg(v_1) \deg(v_3) \min(\deg(v_1), \deg(v_3)).$$

Figure 1. The functions $A_1(c)$ and $A_2(c)$.

Definition. Let G be a simple graph with n vertices labelled $w_1, w_2, \ldots, w_n$. Then $A_G : [0,1)^2 \to [0,1]$ is the function that is 1 on all rectangles

$$\left[\frac{i-1}{n}, \frac{i}{n}\right) \times \left[\frac{j-1}{n}, \frac{j}{n}\right)$$

where $w_i w_j \in E(G)$, and 0 elsewhere.

Definition. Let $A : [0,1)^2 \to [0,1]$ be an integrable function satisfying $A(x,y) = A(y,x)$ for all $0 \leqslant x,y < 1$. Then for all $0 \leqslant x < 1$ let

$$\ell(x) = \int_0^1 A(x,y) \, dy$$

and let

$$S(A) = \int_0^1 \int_0^1 \ell(x)\ell(y) \min(\ell(x), \ell(y)) \, dx \, dy.$$

Note that A_G satisfies $\int_0^1 \int_0^1 A_G(x,y) \, dx \, dy = c$ and $A_G(x,y) = A_G(y,x)$. If $x \in [(i-1)/n, i/n)$, then $\ell(x) = \deg(w_i)/n$, so

$$S(A_G) = \frac{1}{n^5} \sum_{1 \leqslant i,j \leqslant n} \deg(w_i) \deg(w_j) \min(\deg(w_i), \deg(w_j)) \geqslant \frac{N'}{n^5}. \tag{2.3}$$

Definition. Let $0 \leqslant c \leqslant 1$. Then let $A_1(c) : [0,1)^2 \to [0,1]$ be the function satisfying $A_1(x,y) = 1$ if $\min(x,y) < 1 - \sqrt{1-c}$ and $A_1(x,y) = 0$ otherwise. Let $A_2(c) : [0,1)^2 \to [0,1]$ be the function satisfying $A_2(x,y) = 1$ if $\max(x,y) < \sqrt{c}$ and $A_2(x,y) = 0$ otherwise. It is easy to see that $\int_0^1 \int_0^1 A_i(x,y) \, dx \, dy = c$ holds for $i = 1, 2$: see Figure 1.

To complete the proof of Theorem 1.7, we need to prove an upper bound for N'. Inequality (2.3) shows us that this can be achieved by finding an upper bound for $S(A_G)$. In Theorem 2.1,

we will prove that for a fixed c, the value of S is maximized when $A_G = A_1(c)$ or $A_G = A_2(c)$. An elementary calculation gives that

$$S(A_1(c)) = (1 - \sqrt{1-c})^2((c+1)\sqrt{1-c}+c) \quad \text{and} \quad S(A_2(c)) = c^{5/2}.$$

Therefore

$$\frac{N'}{n^5} \leqslant S(A_G) \leqslant \max(S(A_1(c)), S(A_2(c))) = \max((1-\sqrt{1-c})^2((c+1)\sqrt{1-c}+c), \ c^{5/2}),$$

which as shown above (see (2.1)) implies Theorem 1.7.

It should be noted that there is some $c_0 \approx 0.0865$ such that $S(A_1(c)) \geqslant S(A_2(c))$ in $[0, c_0]$ and $S(A_1(c)) \leqslant S(A_2(c))$ in $[c_0, 1]$.

The theorem below summarizes what is needed to complete the proof of the upper bound in Theorem 1.7. It is actually more general than needed, in that it allows the function A to take any value in the interval $[0, 1]$, rather than just 0 and 1. However, the problem turns out to be easier to settle in this more general setting.

Theorem 2.1. *Let $0 \leqslant c \leqslant 1$ and $K \in \mathbb{N}^+$ be fixed numbers. Assume that $A := [0,1)^2 \to [0,1]$ is a function satisfying $\int_0^1 \int_0^1 A(x,y)\,\mathrm{d}x\,\mathrm{d}y = c$ and $A(x,y) = A(y,x)$ for all $0 \leqslant x,y < 1$, and that there are some numbers $0 = q_0 < q_1 < q_2 < \cdots < q_K = 1$ such that A is constant on $[q_{i-1}, q_i) \times [q_{j-1}, q_j)$ for all $1 \leqslant i, j \leqslant K$. Then*

$$S(A) \leqslant \max(S(A_1(c)), S(A_2(c))).$$

Proof. We will use the following notation: $I_i = [q_{i-1}, q_i)$, $t_i = q_i - q_{i-1} = |I_i|$, $\ell_i = \ell(x)$ for any $x \in I_i$, and $A_{i,j}$ is the value of A in the rectangle $I_i \times I_j$. We will refer to sets of the form $[0,1) \times I_i$ and $I_i \times [0,1)$ as *rows* and *columns* respectively.

The function S is continuous on a compact set defined by the conditions, so its maximum is attained for some A. Let A be a function maximizing S, and let

$$\begin{aligned}
T(A) &= \int_0^1 \int_0^1 \int_0^1 \int_0^1 |A(x_1, y_1) - A(x_2, y_2)|\,\mathrm{d}x_1\,\mathrm{d}y_1\,\mathrm{d}x_2\,\mathrm{d}y_2 \\
&= \sum_{1 \leqslant a_1, a_2, b_1, b_2 \leqslant K} t_{a_1} t_{b_1} t_{a_2} t_{b_2} |A_{a_1, b_1} - A_{a_2, b_2}|.
\end{aligned}$$

By a similar compactness argument, the minimum of T is also attained for some A (among those that maximize S). We now show that such an A cannot have four rectangles $I_{i_1} \times I_{j_1}$, $I_{i_1} \times I_{j_2}$, $I_{i_2} \times I_{j_1}$ and $I_{i_2} \times I_{j_2}$ satisfying $A_{i_1, j_1} < A_{i_2, j_1}$ and $A_{i_1, j_2} > A_{i_2, j_2}$.

For some $\varepsilon > 0$, replace the values A_{i_1, j_1}, A_{i_2, j_1}, A_{i_1, j_2} and A_{i_2, j_2} by

$$A_{i_1, j_1} + \frac{\varepsilon}{t_{i_1} t_{j_1}}, \quad A_{i_2, j_1} - \frac{\varepsilon}{t_{i_2} t_{j_1}}, \quad A_{i_1, j_2} - \frac{\varepsilon}{t_{i_1} t_{j_2}}, \quad A_{i_2, j_2} + \frac{\varepsilon}{t_{i_2} t_{j_2}}$$

respectively. By choosing sufficiently small ε, the value of A remains greater in $I_{i_2} \times I_{j_1}$ and $I_{i_1} \times I_{j_2}$ than in $I_{i_1} \times I_{j_1}$ and $I_{i_2} \times I_{j_2}$ respectively. Note that such a change does not change the values $\ell(x)$, and therefore does not change $S(A)$. (To see this, take a line that intersects two of the four rectangles where the value of A changes. It increases in one of them while decreasing in the other. This results in a 0 net change in the integral of A over that line, since if one of the

rectangles intersects the line in a segment λ times as long as the other one, then its area is λ times greater, so the change in the value of A is λ times smaller.)

Now we show that the value $T(A)$ decreases during this transformation. $T(A)$ is the sum of differences between the values $A_{i,j}$, weighted with the areas of these rectangles. Assume that the value of A is greater in r_1 than in r_2 for two rectangles r_1 and r_2. If we decrease the value of A in a rectangle r_1 by $\varepsilon/\mathrm{Area}(r_1)$, and increase it in r_2 by $\varepsilon/\mathrm{Area}(r_2)$ for sufficiently small ε, then $T(A)$ decreases. To see this, note that

$$\mathrm{Area}(r_1)\,\mathrm{Area}(r_2)|A_{r_1}-A_{r_2}| > \mathrm{Area}(r_1)\,\mathrm{Area}(r_2)\left|A_{r_1}-\frac{\varepsilon}{\mathrm{Area}(r_1)}-\left(A_{r_2}+\frac{\varepsilon}{\mathrm{Area}(r_2)}\right)\right|,$$

and for any rectangle $r_3 \notin \{r_1, r_2\}$,

$$\mathrm{Area}(r_1)\,\mathrm{Area}(r_3)|A_{r_1}-A_{r_3}| + \mathrm{Area}(r_2)\,\mathrm{Area}(r_3)|A_{r_2}-A_{r_3}|$$
$$\geqslant \mathrm{Area}(r_1)\,\mathrm{Area}(r_3)\left|A_{r_1}-\frac{\varepsilon}{\mathrm{Area}(r_1)}-A_{r_3}\right| + \mathrm{Area}(r_2)\,\mathrm{Area}(r_3)\left|A_{r_2}+\frac{\varepsilon}{\mathrm{Area}(r_2)}-A_{r_3}\right|.$$

Applying this to $(r_1,r_2)=(I_{i_2}\times I_{j_1}, I_{i_1}\times I_{j_1})$ and $(r_1,r_2)=(I_{i_1}\times I_{j_2}, I_{i_2}\times I_{j_2})$, the desired result follows.

The symmetry of A can be ruined by this transformation, but replacing $A(x,y)$ with

$$\frac{A(x,y)+A(y,x)}{2} \quad \text{for all } 0\leqslant x,y\leqslant 1$$

fixes this while not increasing $T(A)$ and not changing $S(A)$. (The fact that $T(A)$ does not increase can be verified by the above calculation dealing with the decrease in A for a high-valued rectangle and its increase for a lower-valued one.)

Rearrange the intervals I_i so that $\ell_1 \geqslant \ell_2 \geqslant \cdots \geqslant \ell_K$. The property we just proved for the rectangles implies that, for any four rectangles of the form $I_{i_1}\times I_{j_1}$, $I_{i_1}\times I_{j_2}$, $I_{i_2}\times I_{j_1}$ and $I_{i_2}\times I_{j_2}$, we have

$$A_{i_1,j_1} < A_{i_2,j_1} \Rightarrow A_{i_1,j_2} \leqslant A_{i_2,j_2}.$$

Now we prove that A is decreasing in both variables. (Since $A(x,y)=A(y,x)$, it is enough to show that for one variable.) Assume to the contrary that for some $i_1 < i_2$ and j we have $A_{i_1,j} < A_{i_2,j}$. Then for all $1\leqslant p\leqslant K$ we have $A_{i_1,p}\leqslant A_{i_2,p}$. This results in $\ell_{i_1}<\ell_{i_2}$, a contradiction.

This decreasing property implies that if $\ell_i=\ell_{i+1}$ then A is identical in $I_i\times[0,1)$ and $I_{i+1}\times[0,1)$, so we can merge all such intervals and assume that $\ell_1 > \ell_2 > \cdots > \ell_k$ for some $k\leqslant K$.

Note that $S(A)$ can be expressed as

$$S(A) = \sum_{i=1}^{k}\sum_{j=1}^{k} t_i t_j \ell_i \ell_j \min(\ell_i,\ell_j). \tag{2.4}$$

Consider an A that meets the theorem's requirements, maximizes $S(A)$ and is decreasing in both variables. We state that there cannot be two rectangles in the same row (or column) where the value of A is neither 0 nor 1. Assume that for some $1\leqslant a<b\leqslant k$ and $1\leqslant p\leqslant k$ we have $0<A_{a,p}<1$ and $0<A_{b,p}<1$. Pick some $\varepsilon\in\mathbb{R}$ and change $A_{a,p}$ and $A_{p,a}$ to $A_{a,p}+\varepsilon/(t_a t_p)$ while changing $A_{b,p}$ and $A_{p,b}$ to $A_{b,p}-\varepsilon/(t_b t_p)$. This transformation changes only two ℓ values: ℓ_a becomes $\ell_a+\varepsilon/t_a$ and ℓ_b becomes $\ell_b-\varepsilon/t_b$. If $|\varepsilon|$ is small enough then $0<A_{a,p}+\varepsilon/(t_a t_p), A_{b,p}-$
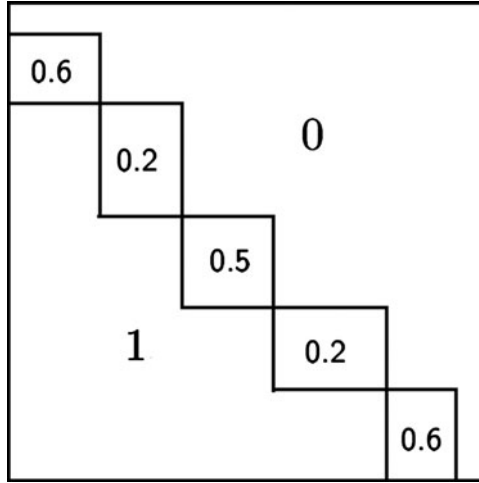
Figure 2. Example of a function considered at this point in the proof.

$\varepsilon/(t_b t_p) < 1$ and the order of the ℓ values is preserved. Now we show that $S(A)$ is a strictly convex function of ε in a neighbourhood of 0. Consider the k^2 terms in the expression (2.4). The terms including other terms than a and b are obviously convex functions of ε, since ε appears at a power of at most 2 in them, and it has a positive coefficient when it has power 2. So the terms of $S(A)$ corresponding to pairs of indices other than (a,a), (a,b), (b,a) and (b,b) are convex functions of ε. All we have to prove is that the sum of the terms corresponding to these four pairs is strictly convex at $\varepsilon = 0$. This sum can be written as

$$t_a^2\left(\ell_a + \frac{\varepsilon}{t_a}\right)^3 + t_b^2\left(\ell_b - \frac{\varepsilon}{t_b}\right)^3 + 2t_a t_b\left(\ell_a + \frac{\varepsilon}{t_a}\right)\left(\ell_b - \frac{\varepsilon}{t_b}\right)^2.$$

Differentiating twice with respect to ε and substituting $\varepsilon = 0$, we get

$$6\ell_a - 2\ell_b + \frac{4t_a\ell_a}{t_b}.$$

This is positive, because $a < b$ implies $\ell_a > \ell_b$. Therefore $S(A)$ is a strictly convex function of ε in a neighbourhood of 0, so cannot have a maximum at 0. This proves that the A under investigation has at most one rectangle in every row and column with a value different from 0 or 1, as depicted in Figure 2.

Now we will prove that in fact there are no such rectangles at all. Since A is decreasing in both variables, each row (and column) starts with some 1-valued rectangles; thus it might include a single rectangle with value between 0 and 1, in which case it contains 0-valued rectangles. (Of course, a row or column need not necessarily contain all three types of rectangle.) If A has a single rectangle of size $t \times t$ with non-zero value, then A's value is c/t^2 there. (So $c/t^2 \leqslant 1$.) Then $S(A) = t^2 \cdot (t \cdot c/t^2)^3 = c^3/t \leqslant c^{5/2}$. If there are multiple non-zero-valued rectangles, then $A_{1,1} = 1$.

Assume that $0 < \lambda = A_{1,j} = A_{j,1} < 1$ for some j. (Then $j \in \{k-1, k\}$.) We will show that it is possible to modify A to increase $S(A)$, so this case is not possible. (From now on we will modify the lengths of the intervals too, not just the value of A in the rectangles.) Divide the interval I_j

into two intervals $I_{j'}$ and $I_{j''}$ of length $t_{j'} = \lambda \cdot t_j$ and $t_{j''} = (1-\lambda) \cdot t_j$ respectively. Then divide $I_1 \times I_j$ into two rectangles of size $t_1 \times t_{j'}$ and $t_1 \times t_{j''}$, and set A to be 1 and 0 respectively in them. Modify A in $I_j \times I_1$ similarly to keep A symmetric.

After this modification, we will get $\ell_{j'} = \ell_j / \lambda$ and $\ell_{j''} = 0$. This means that the terms with j'' can be ignored in (2.4). The only terms to change in (2.4) are t_j becoming $t_{j'} = \lambda \cdot t_j$ and ℓ_j becoming $\ell_{j'} = \ell_j / \lambda$. Since the power of ℓ_j is not smaller than the power t_j in any term of (2.4), and greater than it in $t_j^2 \cdot \ell_j^3$, the value of $S(A)$ increases by this modification. So we can assume that A takes only 0 and 1 values in $I_1 \times [0,1)$ and $[0,1) \times I_1$. To show that non-integer values are not possible in the other places, we need a technical lemma.

Note that the variables t_i and ℓ_i appearing in the following lemma should be considered real numbers with no connection to any function $A : [0,1)^2 \to [0,1]$, but the same notations are used, since the lemma will be applied in such settings.

Lemma 2.2. *Let $t_1, t_2, \ldots, t_k$ be positive reals and let $\ell_1 > \ell_2 > \cdots > \ell_k \geqslant 0$. Assume that there is a neighbourhood H of t_β such that for some $\alpha \in \{\beta - 1, \beta + 1\}$ and $x \in H$ we can replace the numbers $t_\alpha, t_\beta, \ell_\beta$ by $t_\alpha(x) = t_\alpha + t_\beta - x$, $t_\beta(x) = x$ and $\ell_\beta(x) = \ell_\alpha + (t_\beta/x)(\ell_\beta - \ell_\alpha)$ respectively, without changing the non-negativity of the variables and preserving the order of the ℓ. The other variables are left unchanged: $t_i(x) = t_i$, if $i \notin \{\alpha, \beta\}$ and $\ell_i(x) = \ell_i$, if $i \neq \beta$. (Roughly speaking, this transformation preserves the sum $c = \sum_{i=1}^{k} t_i(x)\ell_i(x)$, while changing only three of the values: two neighbouring t and the ℓ corresponding to one of them.) Then the function*

$$S(x) = \sum_{i=1}^{k} \sum_{j=1}^{k} t_i(x) t_j(x) \ell_i(x) \ell_j(x) \min(\ell_i(x), \ell_j(x)) \tag{2.5}$$

is strictly convex at $x = t_\beta$. Therefore it has no maximum there.

Proof. Consider the formula (2.5) and select all the terms depending on x. We can ignore the terms where one of the indices is α or β, and the other is greater than $\max(\alpha, \beta)$ because

$$(t_\alpha + t_\beta - x) t_p \ell_\alpha \ell_p^2 + x t_p \left(\ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right) \ell_p^2 = t_\alpha t_p \ell_\alpha \ell_p^2 + t_\beta t_p \ell_\beta \ell_p^2$$

does not depend on x. The sum of the other terms depending on x can be written as $2S_1(x) + S_2(x)$. Here $S_1(x)$ is the sum of the terms corresponding to pairs of indices where one of the elements is α or β and the other is smaller than α and β; $S_2(x)$ denotes the sum of the terms corresponding to the pairs of indices $(\alpha, \alpha), (\alpha, \beta), (\beta, \alpha)$ and (β, β). Hence

$$S_1(x) = \left((t_\alpha + t_\beta - x) \cdot \ell_\alpha^2 + x \cdot \left(\ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right)^2 \right) \cdot \sum_{p=1}^{\min(\alpha, \beta) - 1} t_p \ell_p \quad \text{and}$$

$$S_2(x) = (t_\alpha + t_\beta - x)^2 \cdot \ell_\alpha^3 + x^2 \cdot \left(\ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right)^3$$
$$+ 2(t_\alpha + t_\beta - x) x \ell_\alpha \left(\ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right) \min \left(\ell_\alpha, \ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right).$$

We have to show that $S_1(x)$ and $S_2(x)$ are strictly convex at $x = t_\beta$, and therefore $S(x)$ does not take its maximum there. We will start with $S_1(x)$. We can disregard the constant factor on

the right, as it does not change convexity. The left factor can be expressed as $\lambda_2 x^2 + \lambda_1 x + \lambda_0 + \lambda_{-1} x^{-1}$. Since $\lambda_2 = \ell_\alpha^2 > 0$ and $\lambda_{-1} = t_\beta (\ell_\beta - \ell_\alpha)^2 > 0$, S_1 is strictly convex.

Now we consider $S_2(x)$. First, assume that $\beta = \alpha + 1$, and therefore $\ell_\alpha > \ell_\beta$. In this case we have

$$S_2(x) = (t_\alpha + t_\beta - x)^2 \cdot \ell_\alpha^3 + x^2 \cdot \left(\ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right)^3$$
$$+ 2(t_\alpha + t_\beta - x) x \ell_\alpha \left(\ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right)^2.$$

Differentiating twice by x and setting $x = t_\beta$, we get

$$2 t_\beta^{-1} (\ell_\beta - \ell_\alpha)^2 \left((\ell_\alpha + \ell_\beta) t_\beta + 2 \ell_\alpha t_\alpha \right) > 0.$$

If $\beta = \alpha - 1$, and therefore $\ell_\alpha < \ell_\beta$, we have

$$S_2(x) = (t_\alpha + t_\beta - x)^2 \cdot \ell_\alpha^3 + x^2 \cdot \left(\ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right)^3$$
$$+ 2(t_\alpha + t_\beta - x) x \ell_\alpha^2 \left(\ell_\alpha + \frac{t_\beta}{x} (\ell_\beta - \ell_\alpha) \right).$$

Differentiating twice by x and setting $x = t_\beta$, we get

$$2 (\ell_\beta - \ell_\alpha)^3 > 0.$$

In both cases, $S_2(x)$ is strictly convex at $x = t_\beta$. This concludes the proof of the lemma. $\qquad\square$

Now we can continue the proof of Theorem 2.1. Assume that A is not entirely 0–1-valued. Let $I_p \times [0, 1)$ be the first column containing a rectangle with a value different from 0 and 1. We have already proved that $p \geqslant 2$. Let $I_p \times I_q$ be the unique rectangle in the pth column with $0 < A_{p,q} < 1$. Since $A_{p,q} = A_{q,p}$, we know that $p \leqslant q$. We will show that A admits the type of transformation described in Lemma 2.2, and therefore does not maximize $S(A)$. We will describe transformations in each case that change only two neighbouring t values and the ℓ value corresponding to one of them.

Case 1. First, assume that $p < q < k$ and $A_{p-1,q+1} = 1$. Then the rows $[0, 1) \times I_q$ and $[0, 1) \times I_{q+1}$ differ only in the pth column. We can move the point separating the intervals I_q and I_{q+1}, keeping $A_{p,q+1} = 0$ while adjusting $A_{p,q}$ such that the integral of A over the whole square remains unchanged. We apply these changes to the other side of the main diagonal to keep A symmetric. During this transformation the only ℓ value to change is ℓ_q. So Lemma 2.2 can be applied with $\alpha = q + 1, \beta = q$.

Case 2. Now assume that $p < q = k$ or $p < q < k$ and $A_{p-1,q+1} = 0$. Then the columns $I_{p-1} \times [0, 1)$ and $I_p \times [0, 1)$ differ only in the qth row. We can move the point separating the intervals I_{p-1} and I_p, keeping $A_{p-1,q} = 1$ while adjusting $A_{p,q}$ such that the integral of A over the whole square stays the same. We apply these changes to the other side of the main diagonal to keep A symmetric. During this transformation the only ℓ value to change is ℓ_p. So Lemma 2.2 can be applied with $\alpha = p - 1, \beta = p$.
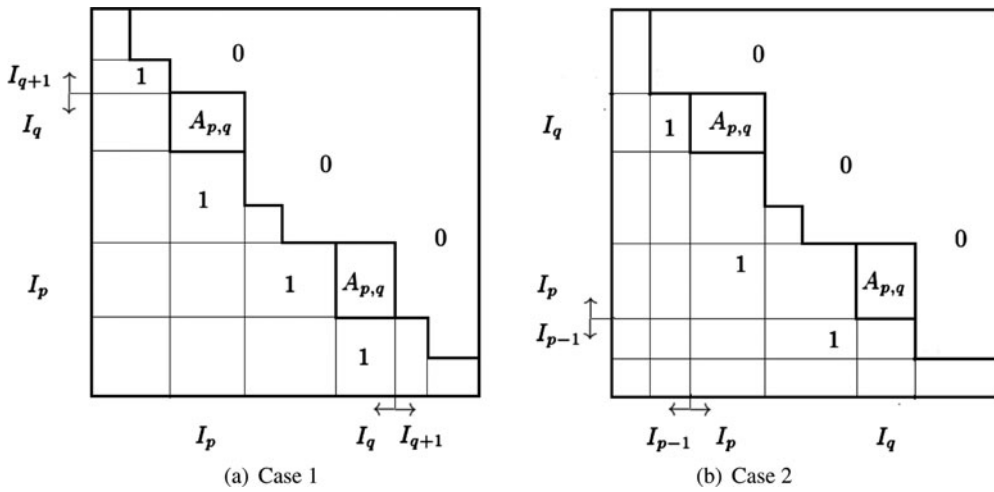
(a) Case 1   (b) Case 2

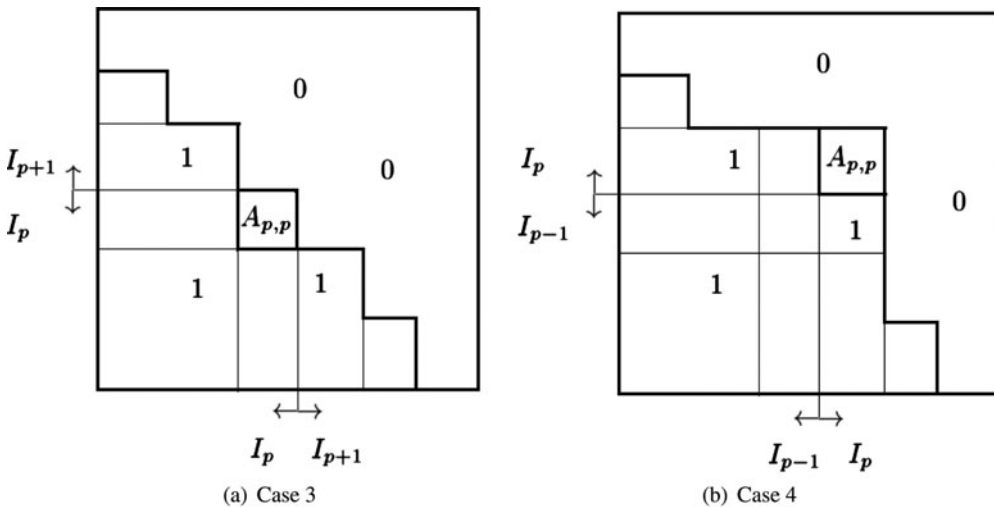*Figure 3.* Illustration of Cases 1 and 2.

(a) Case 3   (b) Case 4

Figure 4. Illustration of Cases 3 and 4.

Case 3. Assume that $p = q < k$ and $A_{p-1,p+1} = 1$. Then the rows $[0,1) \times I_p$ and $[0,1) \times I_{p+1}$ differ only in the pth column. We can move the point separating the intervals I_p and I_{p+1}, keeping $A_{p,p+1} = 0$ while adjusting $A_{p,p}$ such that the integral of A over the whole square stays the same. We apply these changes to the the intervals defining the rows and columns simultaneously to keep A symmetric. During this transformation the only ℓ value to change is ℓ_p. So Lemma 2.2 can be applied with $\alpha = p+1, \beta = p$.

Case 4. Now assume that $p = q = k$ or $p = q < k$ and $A_{p-1,p+1} = 0$. Then the columns $I_{p-1} \times [0,1)$ and $I_p \times [0,1)$ differ only in the pth row. We can move the point separating the intervals I_{p-1} and I_p, keeping $A_{p-1,p} = 1$ while adjusting $A_{p,p}$ such that the integral of A over the whole
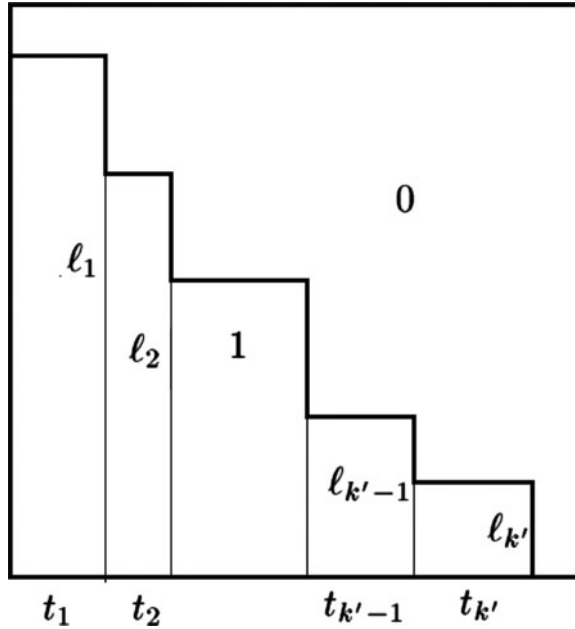
Figure 5. The case $k' \geqslant 4$.

square stays the same. We apply these changes to the the intervals defining the rows and columns simultaneously to keep A symmetric. During this transformation the only ℓ value to change is ℓ_p. So Lemma 2.2 can be applied with $\alpha = p - 1, \beta = p$.

With this, we have covered all the possibilities. From now on, we can assume that $A : [0, 1)^2 \rightarrow \{0, 1\}$.

Since A is 0–1-valued and decreasing in both variables, there exists some $k' \in \{k - 1, k\}$ such that k of the ℓ values are positive and $A_{i,j} = 1$ if and only if $i + j \leqslant k' + 1$. Now we will show that A cannot maximize $S(A)$ if $k' \geqslant 4$.

Assume that $k' \geqslant 4$. It is possible to move the point separating the intervals $I_{k'-1}$ and $I_{k'}$, while keeping $\ell_{k'-1}$ unchanged and adjusting $\ell_{k'}$ to keep the integral over the whole square unchanged. When these changes are applied to the other side of the main diagonal to preserve the symmetry, we see that the point separating I_1 and I_2 moves, ℓ_1 remains unchanged, but ℓ_2 changes. During this transformation, the following values change (see Figure 5):

$$t_{k'} \to x, \quad t_{k'-1} \to t_{k'-1} + t_{k'} - x, \quad \ell_{k'} \to \ell_{k'-1} + \frac{t_{k'}}{x}(\ell_{k'} - \ell_{k'-1}),$$

$$\ell_2 \to \ell_1 - x, \quad t_2 \to \frac{t_2(\ell_1 - \ell_2)}{x}, \quad t_1 \to t_1 + t_2 - \frac{t_2(\ell_1 - \ell_2)}{x}.$$

Now we investigate how $S(A)$ changes during such a transformation. First, apply the changes only to $t_{k'}, t_{k'-1}$ and $\ell_{k'}$. Lemma 2.2 states that S is now a strictly convex function of x. (This is because we changed only two neighbouring t and the ℓ corresponding to one of them, while preserving the sum $\sum_{i=1}^{k'} t_i \ell_i$.) Now apply the changes to t_1, t_2 and ℓ_2 too. Since $t_1 \ell_1 + t_2 \ell_2$ does not change during the transformation, for any $3 \leqslant s \leqslant k'$, the sum of the terms in (2.4) corresponding
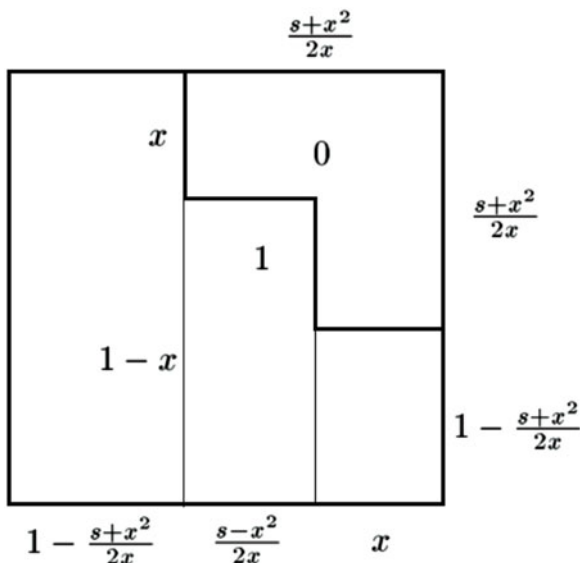
Figure 6.   The case $k' = 3$.

to the pairs of indices $(1,s)$, $(s,1)$, $(2,s)$ and $(s,2)$, which is

$$2t_1 t_s \ell_1 \ell_s^2 + 2t_2 t_s \ell_2 \ell_s^2 = 2t_s \ell_s^2 (t_1 \ell_1 + t_2 \ell_2),$$

does not change. Therefore it is enough to consider the terms where both indices are 1 or 2:

$$\left(t_1 + t_2 - \frac{t_2(\ell_1 - \ell_2)}{x} \right)^2 \ell_1^3 + \left(\frac{t_2(\ell_1 - \ell_2)}{x} \right)^2 (\ell_1 - x)^3$$

$$+ 2\left(t_1 + t_2 - \frac{t_2(\ell_1 - \ell_2)}{x} \right) \left(\frac{t_2(\ell_1 - \ell_2)}{x} \right) \ell_1 (\ell_1 - x)^2.$$

Differentiating twice by x and setting $x = t_{k'} = \ell_1 - \ell_2$, we get

$$\frac{2t_2^2 \ell_1^2}{t_{k'}} > 0,$$

so the above formula is a strictly convex function of x. As the sum of two strictly convex functions, S is a strictly convex function of x, and therefore A does not maximize $S(A)$.

Now assume that $k' = 3$. We claim that if $A : [0,1)^2 \to \{0,1\}$ is a symmetric function decreasing in both variables with at most three positive ℓ values, then there is another such function B with at most two positive ℓ values such that $\int_0^1 \int_0^1 A(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 \int_0^1 B(x,y)\,\mathrm{d}x\,\mathrm{d}y$ and $S(B) \geqslant S(A)$. If A is such a function then $\ell_1 = t_1 + t_2 + t_3$, $\ell_2 = t_1 + t_2$ and $\ell_3 = t_1$. Without loss of generality we can assume that $t_1 + t_2 + t_3 = \ell_1 = 1$. (Replacing $A(x,y)$ with $A(\lambda x, \lambda y)$ changes $S(A)$ with a factor of λ^5, so the rescaling does not change our problem.) Note that in this case the two parameters $s = 1 - \int_0^1 \int_0^1 A(x,y)\,\mathrm{d}x\,\mathrm{d}y$ and $x = t_3$ are enough to define A (see Figure 6). We have

$$\ell_1 = 1, \quad \ell_2 = 1 - x, \quad t_1 = \ell_3 = 1 - \frac{s+x^2}{2x}, \quad t_2 = \frac{s-x^2}{2x}, \quad t_3 = x.$$

Note that x can take any value from $[1 - \sqrt{1-s}, \sqrt{s}]$. The two endpoints correspond to functions with at most two intervals. (After scaling back, we get a step function with at most two positive ℓ values.) Thus

$$S(A) = f(x) = t_1^2 \ell_1^3 + t_2^2 \ell_2^3 + t_3^2 \ell_3^3 + 2t_1 t_2 \ell_1 \ell_2^2 + 2t_1 t_3 \ell_1 \ell_3^2 + 2t_2 t_3 \ell_2 \ell_3^2$$

$$= \left(1 - \frac{s + x^2}{2x}\right)^2 + \left(\frac{s - x^2}{2x}\right)^2 (1 - x)^3 + x^2 \left(1 - \frac{s + x^2}{2x}\right)^3$$

$$+ 2\left(1 - \frac{s + x^2}{2x}\right)\left(\frac{s - x^2}{2x}\right)(1 - x)^2$$

$$+ 2\left(1 - \frac{s + x^2}{2x}\right)x\left(1 - \frac{s + x^2}{2x}\right)^2 + 2\left(\frac{s - x^2}{2x}\right)x(1 - x)\left(1 - \frac{s + x^2}{2x}\right)^2.$$

We will prove that for a fixed s, $f(x)$ is either increasing in $[1 - \sqrt{1-s}, \sqrt{s}]$ or there exists an $x_0 \in [1 - \sqrt{1-s}, \sqrt{s}]$ such that $f(x)$ is strictly decreasing in $[1 - \sqrt{1-s}, x_0]$ and strictly increasing in $[x_0, \sqrt{s}]$. In both cases, $f(x)$ must take its maximum in one of the endpoints. Differentiating $f(x)$ with respect to x, we need that $f'(x)$ is either positive in $(1 - \sqrt{1-s}, \sqrt{s})$ or it is negative in $(1 - \sqrt{1-s}, x_0)$ and positive in $(x_0, \sqrt{s})$. Since $x > 0$, it is sufficient to prove the same for $f'(x)x$. An elementary calculation shows that $\lim_{x \searrow 0} f'(x)x = -\infty$ and $f'(\sqrt{s})\sqrt{s} = 0$. If we can show that $f'(x)x$ is strictly concave in $[0, \sqrt{s}]$, then the claim will follow. After further calculation,

$$4(f'(x)x)'' = -50x^3 + 96x^2 + (27s - 54)x - 16s - 3s^2(2 - s)\frac{1}{x^3}.$$

We are going to prove that the above formula is negative if $0 < x, s < 1$. Since $s < 1$ it is enough to show that

$$g(x, s) = -50x^3 + 96x^2 + (27s - 54)x - 16s - 3s^2 \frac{1}{x^3} \leqslant 0.$$

This is a polynomial of s of degree 2. For a fixed x, it takes its maximum at $s = (27x^4 - 16x^3)/6$. If $(27x^4 - 16x^3)/6 \geqslant 1$, then $x \geqslant 0.89$ and

$$g(x, s) \leqslant g(x, 1) = -50x^3 + 96x^2 - 27x - 16 - 3x^{-3} \leqslant 0.$$

If $(27x^4 - 16x^3)/6 \leqslant 1$, then $x \leqslant 0.9$ and

$$g(x, s) \leqslant g\left(x, \frac{27x^4 - 16x^3}{6}\right) = \frac{1}{12}(729x^5 - 864x^4 - 344x^3 + 1152x^2 - 648x) \leqslant 0.$$

(Both of the above results follow by elementary calculus.) This concludes the proof of the claim and hence of the case $k' = 3$. We obtained that $S(A)$ is maximized by a function with at most two positive ℓ values.

Now we can assume that $k' \leqslant 2$. Then A is completely defined by the parameters

$$\int_0^1 \int_0^1 A(x, y) \, dx \, dy = c \quad \text{and} \quad t_1 = x$$

(see Figure 7). In this setting,

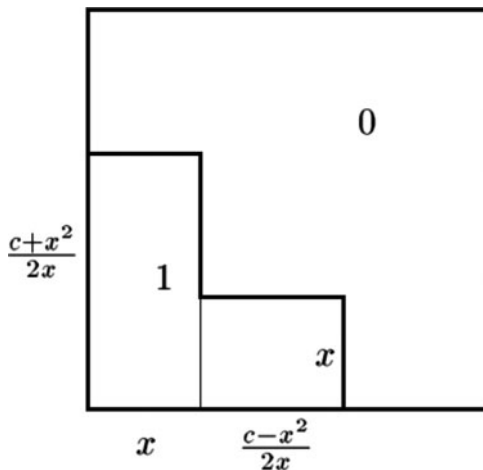$$t_1 = \ell_2 = x, \quad t_2 = \frac{c - x^2}{2x}, \quad \ell_1 = \frac{c + x^2}{2x}.$$

Figure 7. The case $k' = 2$.

Note that $1 - \sqrt{1-c} \leqslant x \leqslant \sqrt{c}$, and $x = 1 - \sqrt{1-c}$ corresponds to $A_1(c)$, while $x = \sqrt{c}$ corresponds to $A_2(c)$ (see Figure 1). Thus

$$S(A) = t_1^2 \ell_1^3 + t_2^2 \ell_2^3 + 2t_1 t_2 \ell_1 \ell_2^2$$
$$= x^2 \left(\frac{c+x^2}{2x} \right)^3 + \left(\frac{c-x^2}{2x} \right)^2 x^3 + 2x \left(\frac{c-x^2}{2x} \right) \left(\frac{c+x^2}{2x} \right) x^2$$
$$= \frac{1}{8} \left(\frac{c^3}{x} + 9c^2 x - cx^3 - x^5 \right).$$

Using the substitution $y = x/\sqrt{c}$ (where $(1 - \sqrt{1-c})/\sqrt{c} \leqslant y \leqslant 1$), we get

$$S(A) = \frac{c^{5/2}}{8} \left(\frac{1}{y} + 9y - y^3 - y^5 \right).$$

We want to show that this function takes its maximum at one of the endpoints of its domain. It suffices to show that there exists a real number $0 < y_0 < 1$ such that the function $y \to 1/y + 9y - y^3 - y^5$ is strictly decreasing in $(0, y_0)$ and strictly increasing in $(y_0, 1)$.

Differentiating once, we obtain $f(y) = -1/y^2 + 9 - 3y^2 - 5y^4$. We need there to be some $0 < y_0 < 1$ such that $f(y) < 0$ if $0 < y < y_0$ and $f(y) > 0$ if $y_0 < y < 1$. Consider the function $g(y) = f(\sqrt{y})y$. It is obvious that g has the desired property if and only f has it. Since $g(y) = -5y^3 - 3y^2 + 9y - 1$, a polynomial of degree 3, this property can be verified for g by elementary calculus. With this, Theorem 2.1 is proved. $\square$

As shown before Theorem 2.1, Theorem 1.7 follows from Theorem 2.1.

3. Remarks and open questions

First, we show that the bounds in Theorem 1.7 are asymptotically sharp.

Proposition 3.1. *Let n and e be fixed positive integers satisfying $e \leqslant \binom{n}{2}$. Let $c = 2e/n^2$. Then there is a simple graph G_1 with n vertices and e edges containing at most $\frac{1}{2}c^4n^5$ 4-edge paths. Additionally, there are simple graphs G_2 and G_3 with n vertices and e edges that contain at least $\frac{1}{2}c^{5/2}n^5(1 - O(n^{-1}))$ and $\frac{1}{2}(1 - \sqrt{1-c})^2((c+1)\sqrt{1-c}+c)n^5(1 - O(n^{-1}))$ 4-edge paths respectively.* $\qquad\square$

Proof. Let G_1 be a graph with n vertices and e edges such that the degrees of any two vertices differ by at most 1. (It is well known that such a graph exists.) Then the degree of any vertex is at most $2e/n + 1 = cn + 1$. So the number of 4-edge paths is at most

$$\frac{1}{2}n(cn+1)(cn)(cn-1)(cn-2) \leqslant \frac{1}{2}n(cn)^4 = \frac{1}{2}c^4n^5.$$

Now we show that we can choose the quasi-clique C_n^e for G_2. C_n^e contains an a-clique, where a is the greatest integer satisfying $\binom{a}{2} \leqslant e$. Therefore $\binom{a+1}{2} \geqslant e$, implying $a \geqslant \sqrt{2e} - 1$. The number of 4-edge paths in this clique is

$$\frac{1}{2}a(a-1)(a-2)(a-3)(a-4) \geqslant \frac{1}{2}(2e)^{5/2}(1 - O(e^{-1/2})) \geqslant \frac{1}{2}c^{5/2}n^5(1 - O(n^{-1})).$$

A similar (but more complicated) calculation gives that we can choose the quasi-star S_n^e for G_3. $\qquad\square$

We conclude the paper with a few open questions.

Question 1. We proved that either the quasi-star or the quasi-clique asymptotically maximizes the number of 4-edge paths in graphs with given edge density. Is it true that this maximum is in fact exactly (not just asymptotically) achieved by either the quasi-star or the quasi-clique?

Theorem 1.1 states that the above is true for 2-edge paths.

Question 2. Is it true for all graphs H that the number of subgraphs isomorphic to H in graphs with given edge density is (asymptotically) maximized by either the quasi-star or the quasi-clique?

It is true for 4-edge paths and k-edge stars, when $2 \leqslant k \leqslant 30$ (see Theorem 1.2). When H is a graph having a spanning subgraph that is a vertex-disjoint union of edges and cycles, only the quasi-clique comes into play (see Theorem 1.3).

Question 3. Is it true that for every graph H, there is a constant $c_H < 1$ such that among graphs with n vertices and edge density $c > c_H$, the number of subgraphs isomorphic to H is (asymptotically) maximized by the quasi-clique?

Acknowledgement

References

[1] Ahlswede, R. and Katona, G. O. H. (1978) Graphs with maximal number of adjacent pairs of edges. *Acta Math. Acad. Sci. Hungar.* **32** 97–120.

[2] Alon, N. (1981) On the number of subgraphs of prescribed type of graphs with a given number of edges. *Israel J. Math.* **38** 116–130.

[3] Alon, N. (1986) On the number of certain subgraphs contained in graphs with a given number of edges. *Israel J. Math.* **53** 97–120.

[4] Bollobás, B. and Sarkar, A. (2001) Paths in graphs. *Studia Sci. Math. Hungar.* **38** 115–137.

[5] Bollobás, B. and Sarkar, A. (2003) Paths of length four. *Discrete Math.* **265** 357–363.

[6] Füredi, Z. (1992) Graphs with maximum number of star-forests. *Studia Sci. Math. Hungar.* **27** 403–407.

[7] Kenyon, R., Radin, C., Ren, K. and Sadun, L. Multipodal structure and phase transitions in large constrained graphs. `arXiv:1405.0599`

[8] Lovász, L. (2012) *Large Networks and Graph Limits*, AMS.

[9] Nikiforov, V. (2011) The number of cliques in graphs of given order and size. *Trans. Amer. Math. Soc.* **363** 1599–1618.

[10] Razborov, A. A. (2008) On the minimal density of triangles in graphs. *Combin. Probab. Comput.* **17** 603–618.

[11] Reiher, C. (2016) The clique density theorem. *Annals of Mathematics* **184** 683–707.