The largest crossing number of tanglegrams

Éva Czabarka[a] Junsheng Liu[a,b] László A. Székely[a]

Submitted: July 10, 2023; Accepted: Feb 17, 2024; Published: Mar 8, 2024

Abstract

A tanglegram $\mathcal{T}$ consists of two rooted binary trees with the same number of leaves, and a perfect matching between the two leaf sets. In a layout, the tanglegram is drawn with the leaves on two parallel lines, the trees on either side of the strip created by these lines are drawn as plane trees, and the edges of the perfect matching are straight line segments inside the strip. The tanglegram crossing number $\mathrm{crt}(\mathcal{T})$ of $\mathcal{T}$ is the smallest number of crossings of pairs of matching edges, over all possible layouts of $\mathcal{T}$. The size of the tanglegram is the number of matching edges, say n. An earlier paper showed that the maximum of the tanglegram crossing number of size n tanglegrams is $< \frac{1}{2}\binom{n}{2}$; but is at least $\frac{1}{2}\binom{n}{2} - \frac{n^{3/2}}{2}$ (and at least $\frac{1}{2}\binom{n}{2} - \frac{n^{3/2}-n}{2}$ for infinitely many n). Here we improve these bounds: the maximum crossing number of a size n tanglegram is at most $\frac{1}{2}\binom{n}{2} - \frac{n}{4}$, but for infinitely many n, at least $\frac{1}{2}\binom{n}{2} - \frac{n\log_2 n}{4}$. The problem shows analogy with the Unbalancing Lights Problem of Gale and Berlekamp.

Mathematics Subject Classifications: 05C10, 05C05, 05C62, 92B10

1 Introduction

A *binary tree* has a root vertex assumed to be a common ancestor of all other vertices, and each vertex either has two children or no children. A vertex with no children is a *leaf*, and a vertex with two children is an *internal vertex*. Note that this definition allows a single-vertex tree that is considered as both root and leaf to be a rooted binary tree. In an *ordered binary tree* an order of the two children is specified, for every vertex that has children.

A *plane binary tree* is a drawn ordered binary tree, without edge crossings, where the left-right order of subtrees in the drawing coincides with the order. The edges are drawn in straight line segments. It is easy to draw a plane binary tree in such a way that all the leaves are on a line, and all other vertices are in the same open half-plane.

[a] Department of Mathematics, University of South, Carolina, Columbia, SC, USA (czabarka@math.sc.edu, szekely@math.sc.edu).

[b] Computer Science and Engineering Department, Washington University in St Louis. (junsheng@wustl.edu).

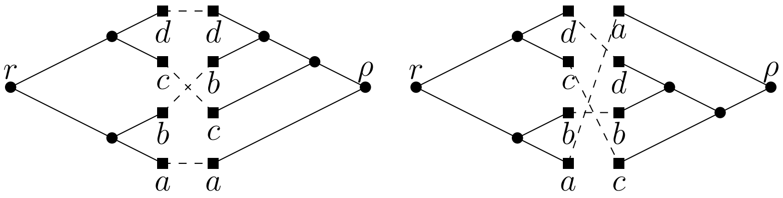
https://doi.org/10.37236/12194

A *tanglegram* $\mathcal{T} = (L, R, \sigma)$ is a graph that consists of a left binary tree L, a right binary tree R with the same number of leaves as L, and a perfect matching σ between the leaves of L and R. Two tanglegrams are considered identical if there is a graph isomorphism between them fixing the root r of R and the root ρ of L. The *size* of a tanglegram is the number of leaves in L (or R). An *abstract tanglegram layout* of the tanglegram (L, R, σ) is given by turning the unordered trees L and R into *ordered trees.* Given an abstract tanglegram layout, an actual *tanglegram layout* consists of a left plane binary tree isomorphic (keeping order as well) to L with root r drawn in the half-plane $x \leqslant 0$, having its leaves on the line $x = 0$, a right plane binary tree isomorphic (keeping order as well) to R with root ρ, drawn in the half-plane $x \geqslant 1$, having its leaves on the line $x = 1$, and a perfect matching σ between their leaves drawn in straight line segments. (Isomorphism of ordered trees (plane trees) keeps the root and the order.)

Our main concern about tanglegram layouts is the number of crossings between the matching edges. As it is determined by the abstract tanglegram layout, it is sufficient to focus on the abstract tanglegram layout to count crossings.

A *switch* on the abstract tanglegram layout (L, R, σ) is the following operation: select an internal vertex v of one of the two trees L and R and change the order of its two children.

It is easy to see that two abstract tanglegram layouts represent the same tanglegram if and only if a sequence of switches moves one abstract layout into the other. (A switch on a tanglegram layout illustrated in Figure 1.) Hence tanglegrams of a given size partition the set of all abstract tanglegram layouts of the same size, or equivalently a tanglegram can be seen as an equivalence class of abstract tanglegram layouts. Note that interchanging L and R is not allowed, as it may result in a different tanglegram.



original layout

after switching at ρ

Figure 1: Result of a switch operation.

The crossing number of a tanglegram layout is the number of pairs of matching edges that cross, which is determined by the abstract tanglegram layout.

It is desirable to draw a tanglegram with the *least possible number of crossings*, which is known as the Tanglegram Layout Problem [4, 8]. The *(tanglegram) crossing number* $\text{crt}(\mathcal{T})$ of a tanglegram $\mathcal{T}$ is defined as the minimum number of crossings among its layouts. The Tanglegram Layout Problem problem is NP-hard [2, 4], but is Fixed Parameter Tractable [2, 1]. It does not allow constant factor approximation under the Unique Game Conjecture [2]. Tanglegrams play a major role in phylogenetics, especially in the theory of cospeciation [6]. For example, the first binary tree is the phylogenetic tree of the hosts, the second binary tree is the phylogenetic tree of their parasites (e.g., gopher and louse), and the matching connects the host with its parasite [5]. The tanglegram crossing number has been related to the number of times parasites switched hosts [5], or, working with gene trees instead of phylogenetic trees, to the number of horizontal gene transfers ([3], pp. 204–206). Besides phylogenetics, tanglegrams are also well-studied objects in computer science.

Let M_n denote $\max_{\mathcal{T}} \operatorname{crt}(\mathcal{T})$ among size n tanglegrams. It is easy to see that for any tanglegram, the expected number of crossings in a random layout of any fixed labeled tanglegram of size n is $\frac{1}{2}\binom{n}{2}$. (For details, see Section 3.) Therefore, $M_n \leqslant \frac{1}{2}\binom{n}{2}$. An earlier paper [10] made a slight improvement showing that equality cannot happen: $M_n < \frac{1}{2}\binom{n}{2}$; and also showed that for every n,

$$\frac{1}{2}\binom{n}{2} - \frac{n^{3/2}}{2} \leqslant M_n,$$

and for $n = k^2$

$$\frac{1}{2}\binom{n}{2} - \frac{n^{3/2} - n}{2} \leqslant M_n.$$

The goal of this paper is to find more proper separation as $M_n \leqslant \frac{1}{2}\binom{n}{2} - \frac{n}{4}$, and for each $n = 2^k$,

$$M_n \geqslant \frac{1}{2}\binom{n}{2} - \frac{n \log_2 n}{4}.$$

In Section 2 we provide a construction for the lower bound. In Section 3 we relate the number of crossings in different layouts of the tanglegram. In Section 5 we relate the largest crossing number problem to the Unbalancing Lights Problem of Gale and Berlekamp, and show the separation from $\frac{1}{2}\binom{n}{2}$. In Section 4 we derive some technical results that we need for the proof.

2 A construction for tanglegrams with large crossing number

We established a better lower bound on M_n for $n = 2^k$ only:

Theorem 1. *For every $i \geqslant 1$, there exists a tanglegram of size 2^i, which has tanglegram crossing number $\frac{1}{2}\binom{2^i}{2} - i2^{i-2}$ exactly.*

Let $X = \{0, 1\}$ and let X^i be the set of binary strings of length i, i.e., words over the alphabet X, which make the binary representations of the non-negative integers that are less than 2^i. Given a string $\vec{x} = x_1 x_2 \dots x_i$, we will denote by $\overleftarrow{x}$ the string obtained by reversing x, i.e., $\overleftarrow{x} = x_i x_{i-1} \dots x_1$.

For every $i \in \mathbb{N}$ we will define a tanglegram $\mathcal{T}_i = (R^{(i)}, L^{(i)}, \sigma_i)$ of size 2^i by the following procedure:

Both $L^{(i)}$ and $R^{(i)}$ are the rooted complete binary trees of height i. We label the vertices of $L^{(i)}$ (resp. of $R^{(i)}$) as follows: The set of vertices at distance j (which we call the j^{th} layer) from the root are labeled as $u_{\vec{x}}$ (resp. $w_{\vec{x}}$) where $\vec{x}$ is an element of X^j.

The root of $L^{(i)}$ is labeled as u_ϵ, and the root of $R^{(i)}$ is labeled as v_ϵ, where ϵ is the empty string. The labels of children of $u_{\vec{x}}$ (resp. $w_{\vec{x}}$) are created by suffixes: $u_{\vec{x}0}$ and $u_{\vec{x}1}$ (resp. $w_{\vec{x}0}$ and $w_{\vec{x}1}$). The matching is $\sigma_i = \{u_{\vec{x}} w_{\overleftarrow{x}} : \vec{x} \in X^i\}$, see Fig. 2.

For $t \in X$, let $L_t^{(i)}$ (resp. $R_t^{(i)}$) denote the subtree of $L^{(i)}$ (resp. $R^{(i)}$) rooted at u_t (resp. w_t).

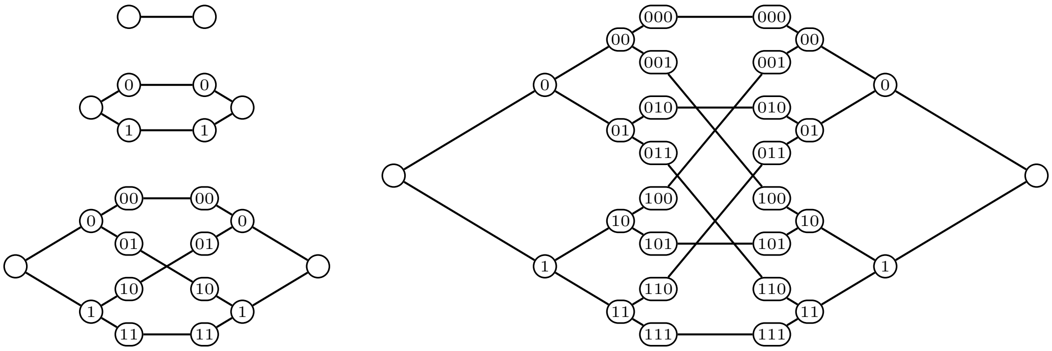


Figure 2: The tanglegrams $\mathcal{T}_i$ for $i \in \{0, 1, 2, 3\}$. The vertices are labeled with their indices as in the text and the tanglegrams are shown with a crossing-optimal layout.

Lemma 2. *For every $i \geqslant 2$, $\operatorname{crt}(\mathcal{T}_i) \geqslant 2\operatorname{crt}(\mathcal{T}_{i-1}) + \binom{2^{i-1}}{2}$.*

Proof. Let $\mathcal{T} = (L, R, \sigma)$ be an arbitrary tanglegram, and v be a non-leaf vertex of one of the trees L, R. Let Z be the set of leaves in the tree where v lives (L or R) that are descendants of v. Note that in any layout of $\mathcal{T}$, the elements of Z appear consecutively in the sequence of leaves. Moveover, if both children of v are leaves, and the matching edges incident upon these children cross in the layout, then switching the order of these children in the layout eliminates this crossing and decreases the crossing number, so the original layout was not optimal.

Assume $i \geqslant 2$, and let $\mathcal{D}$ be an optimal layout of $\mathcal{T}_i$. Let A be the number of crossings in $\mathcal{D}$ between edges incident upon leaves $u_{\vec{x}}$ and $u_{\vec{y}}$ where the first digits of $\vec{x}$ and $\vec{y}$ are the same and let B be the the number of crossings in $\mathcal{D}$ between edges incident upon leaves $u_{\vec{x}}$ and $u_{\vec{y}}$ where the first digit of $\vec{x}$ and $\vec{y}$ differ. Obviously, $\operatorname{crt}(\mathcal{T}_i) = A + B$. As for each $t \in X$ the matching edges incident upon leaves of $L_t^{(i)}$ induce a $\mathcal{T}_{i-1}$ with a sublayout in $\mathcal{D}$ with at least $\operatorname{crt}(\mathcal{T}_{i-1})$ crossings, $A \geqslant 2\operatorname{crt}(\mathcal{T}_{i-1})$, so it is enough to show that $B \geqslant \binom{2^{i-1}}{2}$.

Let $t, s \in X$ be chosen such that u_t lies above u_s in the layout $\mathcal{D}$. Let $\vec{x}, \vec{y} \in X^{i-1}$ be different words. Clearly, $w_{\vec{x}0}, w_{\vec{x}1}, w_{\vec{y}0}, w_{\vec{y}1}$ are distinct leaves of $R^{(i)}$, and $u_{0\overleftarrow{x}}, u_{1\overleftarrow{x}}, u_{0\overleftarrow{y}}, u_{1\overleftarrow{y}}$ are distinct leaves of $L^{(i)}$. Also, the leaves $w_{\vec{x}0}, w_{\vec{x}1}$ as well as $w_{\vec{y}0}, w_{\vec{y}1}$ are consecutive in any layout, including D.

We may assume without loss of generality that $u_{t\overleftarrow{x}}$ is above $u_{t\overleftarrow{y}}$ in $\mathcal{D}$. As u_t lies above u_s, both $u_{s\overleftarrow{x}}, u_{s\overleftarrow{y}}$ lie below $u_{t\overleftarrow{y}}$. If the pair $w_{\vec{x}0}, w_{\vec{x}1}$ lies above the pair $w_{\vec{y}0}, w_{\vec{y}1}$, then the matching edges incident upon $u_{s\overleftarrow{y}}$ and $u_{t\overleftarrow{x}}$ cross; otherwise the matching edges incident upon $u_{t\overleftarrow{y}}$ and $u_{s\overleftarrow{x}}$ cross. This shows that for any $\vec{x}, \vec{y} \in X^{i-1}$, if $\vec{x} \neq \vec{y}$, then for some k, ℓ such that $\{k, \ell\} = \{0, 1\}$ we have that the matching edges incident upon $u_{k\vec{x}}$ and $u_{\ell\vec{y}}$ cross in $\mathcal{D}$. Therefore we have $B \geqslant \binom{2^{i-1}}{2}$. □

Lemma 3. *Let $\mathcal{D}_i^\star$ be the layout of $\mathcal{T}_i$ in which the leaf labels from top to bottom appear in the order of the integers corresponding to the binary words, both in $L^{(i)}$ and $R^{(i)}$. (See* Fig. 2 *for this layout.) Let* $\mathrm{cr}(\mathcal{D}_i^\star)$ *denote the number of crossings in this layout. Then, for all $i \in \mathbb{N}$, we have*

$$\mathrm{cr}(\mathcal{D}_i^\star) = \mathrm{crt}(\mathcal{T}_i) = \frac{1}{2}\binom{2^i}{2} - i2^{i-2}.$$

Proof. Set $\omega_i = \mathrm{cr}(\mathcal{D}_i^\star)$. We will show the statement by induction on i, with base cases $i \in \{0, 1\}$.

$\mathcal{T}_0$ and $\mathcal{T}_1$ are the unique planar tanglegrams of size 1 and 2 respectively, $\frac{1}{2}\binom{2^0}{2} - 0 \cdot 2^{0-2} = 0 = \mathrm{crt}(\mathcal{T}_0)$ and $\frac{1}{2}\binom{2^1}{2} - 1 \cdot 2^{1-2} = 0 = \mathrm{crt}(\mathcal{T}_1)$. Since for $i \in \{0,1\}$ we have $\vec{x} = \overleftarrow{x}$ for any $\vec{x} \in X^i$, $\mathcal{D}_i^\star$ is a planar layout, so $\omega_0 = \omega_1 = 0$. Thus, the statement is true for $i \in \{0, 1\}$.

Assume now $i > 1$, and consider the layout $\mathcal{D}_i^\star$. For each $t \in X$, the matching edges incident upon a leaf of L_t^i induce a drawing of a subtanglegram of $\mathcal{T}_{i-1}$ that is isomorphic to $\mathcal{D}_{i-1}^\star$, contributing exactly $2\omega_{i-1}$ crossings. We want to count the number of crossings in $\mathcal{D}_i^\star$ between matching edges whose left-endpoints are $u_{0\overleftarrow{x}}, u_{1\overleftarrow{y}}$, where $\vec{x}, \vec{y} \in X^{i-1}$. The edges cross precisely when $\vec{y}1 < \vec{x}0$, which is equivalent with $\vec{y} < \vec{x}$ (where we consider the words as binary representations of numbers). So we have exactly one such crossings for each unordered pair $\vec{x}, \vec{y}$ from X^{i-1}. By the induction hypothesis and Lemma 2 we have

$$\mathrm{crt}(\mathcal{T}_i) \leqslant \omega_i = 2\omega_{i-1} + \binom{2^{i-1}}{2} = 2\,\mathrm{crt}(\mathcal{T}_{i-1}) + \binom{2^{i-1}}{2} \leqslant \mathrm{crt}(\mathcal{T}_i),$$

which gives $\mathrm{crt}(\mathcal{T}_i) = \omega_i$. Also, by the induction hypothesis

$$\begin{aligned}\omega_i &= 2\omega_{i-1} + \binom{2^{i-1}}{2} = 2\left(\frac{1}{2}\binom{2^{i-1}}{2} - (i-1)2^{i-3}\right) + \binom{2^{i-1}}{2} \\ &= 2^{i-1}(2^{i-1}-1) - (i-1)2^{i-2} = \frac{1}{2}\binom{2^i}{2} - i2^{i-2}. \quad \square\end{aligned}$$

Unfortunately, we know that this construction is not the best possible. For size 8, the tanglegram on Fig. 3 is shown with an optimal drawing and has one more crossings than our construction on Fig. 2, and, replacing $\mathcal{T}_3$ with this tanglegram in our construction, it is easy to see that $\mathrm{crt}(\mathcal{T}_i) < M_{2^i}$ for each $i > 2$, but we have not managed to improve on our lower bound by a term that is not of order $o(n \log(n))$.

3 Crossings in different layouts of the same tanglegram

In this section we consider a fixed layout $\mathcal{D}_0$ of the tanglegram, the state (crossing and noncrossing) of pairs of matching edges in this layout, and compute the crossing number of any layout using these states plus switches on internal vertices of the left-and right-tree of the tanglegram. We will use this expression to establist an upper bound on the maximum tanglegram crossing number.

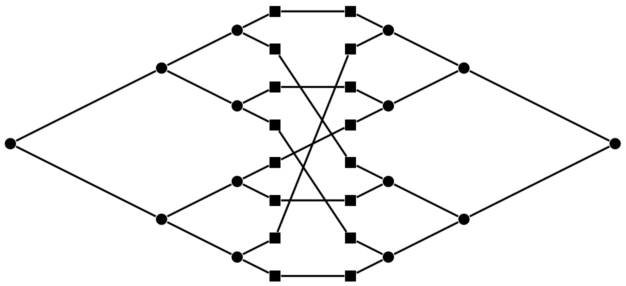


Figure 3: A tanglegram of size 8 with tanglegram crossing number 9. This is the maximum tanglegram crossing number for size 8, found by brute force search.

Let us be given a tanglegram $\mathcal{T} = (L, R, \sigma)$ of size n. Vertices of R form a partially ordered set, for the following order: if r is the root of R, then $x \leqslant y$ if y is a vertex of the unique rx path in $\mathcal{T}$. This partial order is a semilattice, in which the least upper bound of vertices u and v is denoted by $\mathrm{lca}_R(u, v)$ (lca stands for least common ancestor in phylogenetics). Similar arguments apply for the tree L, where the notation will be lca_L. For matching edhes e, f, $\mathrm{lca}_R(e, f)$ (resp. $\mathrm{lca}_L(e, f)$) will denote the lca of the two leaves of R adjacent to the edges e and f (resp. the lca of the two leaves of L adjacent to the edges e and f).

Consider a fixed layout $\mathcal{D}_0$ of $\mathcal{T}$. Assume that a layout $\mathcal{D}$ is obtained from $\mathcal{D}_0$ by making a switch in certain internal (non-leaf) vertices of R and certain internal (non-leaf) vertices of L. Note that changing the order of switches has no effect on $\mathcal{D}$. Also note that each of R, L has exactly $n - 1$ internal vertices. We denote the set of internal vertices by $\mathrm{int}(R)$ and $\mathrm{int}(L)$.

Define α (β) on $\mathrm{int}(R)$ ($\mathrm{int}(L)$) as 1 if no switch takes place in the vertex, and -1 if a switch takes place in the vertex. Fixing $\mathcal{D}_0$, the combinatorially different layouts $\mathcal{D}$ are in one-to-one correspondence with the pairs (α, β) of ± 1 valued functions.

Consider now two matching edges, e, f of $\mathcal{T}$. Let $x = \mathrm{lca}_R(e, f)$ and $u = \mathrm{lca}_L(e, f)$. Define the *crossing status* of matching edges e, f in layout $\mathcal{D}$ as

$$\chi_{\mathcal{D}}(e, f) = \begin{cases} -1 & \text{if } e \text{ crosses } f \text{ in } \mathcal{D}; \\ 1 & \text{otherwise.} \end{cases}$$

Observe that $\chi_{\mathcal{D}}(e, f) = \chi_{\mathcal{D}_0}(e, f)$ if and only if $\alpha(x)\beta(u) = 1$. Therefore,

$$\chi_{\mathcal{D}}(e, f) = \alpha(\mathrm{lca}_R(e, f))\beta(\mathrm{lca}_L(e, f))\chi_{\mathcal{D}_0}(e, f).$$

Counting the number of crossings in a layout $\mathcal{D}$, we have

$$\mathrm{cr}(\mathcal{D}) = \sum_{\{e,f\}} \frac{1 + \chi_{\mathcal{D}}(e, f)}{2}.$$

We have

$$\begin{aligned}
\operatorname{cr}(\mathcal{D}) &= \sum_{\{e,f\}} \frac{1+\alpha(\mathrm{lca}_R(e,f))\beta(\mathrm{lca}_L(e,f))\chi_{\mathcal{D}_0}(e,f)}{2} && (1)\\
&= \frac{1}{2}\binom{n}{2} + \frac{1}{2}\sum_{x\in \mathrm{int}(R)}\sum_{u\in \mathrm{int}(L)} \alpha(x)\beta(u) \sum_{\substack{\{e,f\}: x=\mathrm{lca}_R(e,f)\\ u=\mathrm{lca}_L(e,f)}} \chi_{\mathcal{D}_0}(e,f). && (2)
\end{aligned}$$

To justify the claim in Section 1 on the expected number of crossings in a random layout of a labeled tanglegram, select randomly and independently the α and β values to transform the fixed drawing $\mathcal{D}_0$ into the random drawing $\mathcal{D}$. The displayed formula above implies $\mathbb{E}[\operatorname{cr}(\mathcal{D})] = \frac{1}{2}\binom{n}{2}$.

4 Tools

We will establish a technical lemma in this section. For a rooted binary tree T, let $L(T)$ denote the set of its leaves. and let $A(T)$ be the set of internal vertices that have a leaf neighbor. For set $\psi(x) = 1$ if $x \in A(T)$ and the number of leaves that are descendants of x is even, 0 otherwise, and let $\psi_T = \sum_{x\in V(T)} \psi(x)$. Set $h(1) = 0$ and for $n \geqslant 2$ let $h(n) = \min_{T:|L(T)|=n} \psi_T$. A tree T with n leaves is called a *realizer*, if $\psi_T = h(n)$.

Lemma 4. *For any $n \geqslant 2$, $h(n) = \lfloor \frac{n}{4} \rfloor + 1$. In words, in any rooted binary tree with n leaves, at least $\lfloor \frac{n}{4} \rfloor + 1$ vertices have a leaf neighbor and an even number of leaf descendants.*

Proof. Note that $h(1) = 0 = \lfloor \frac{1}{4} \rfloor$. Let $n \geqslant 2$ and write $n = 4q + r$ where $q = \lfloor \frac{n}{4} \rfloor$ and $0 \leqslant r < q$. We will show that $h(n) = q + 1$ by induction on n.

For $n \in \{2, 3\}$, there is only one rooted binary tree on n vertices. Clearly, $h(2) = 1$ and $h(3) = 1$, so the claim is true.

Let $n \geqslant 4$ (i.e. $q \geqslant 1$) and assume that the statement is true for all trees with n' leaves, where $2 \leqslant n' < n$.

Take a tree T on n vertices, and let T_1, T_2 be the subtrees rooted at the two neighbors of the root. Without loss of generality $1 \leqslant k = |L(T_1)| \leqslant |L(T_2)| = n - k \leqslant n - 1$. Set $q' = \lfloor \frac{k}{4} \rfloor$, $r' = k - 4q'$, $q'' = \lfloor \frac{n-k}{4} \rfloor$ and $r'' = (n - k) - 4q''$.

When $0 \leqslant r' \leqslant r$ we get $r'' = r - r'$, $q = q' + q''$: $\psi_T \geqslant \psi_{T_1} + \psi_{T_2} \geqslant q' + (q - q') + 1 = q + 1$. The first inequality is an equality iff (n is odd or (n is even and $k \neq 1$)), and the second inequality is an equality iff $k = 1$ and T_2 is a realizer. Thus, for an odd n, $h(n) \leqslant q + 1$ is obtained by choosing a tree T such that T_1 is a single vertex and T_2 is a realizer with $n - 1$ leaves.

When $r < r' \leqslant 3$ we get $r'' = 4 + r - r'$, $q'' = q - q' - 1$, $\psi_T \geqslant \psi_{T_1} + \psi_{T_2} \geqslant q' + (q - q' - 1) + 1 = q$, the first inequality is an equality iff (n is odd or (n is even and $k \neq 1$)), the second inequality is an equality iff $k = 1$ and T_2 is a realizer. From $r < r' \leqslant 3$ we get that if n is odd, then $r = 1$ and consequently $k \geqslant r' > 1$; so for odd n this gives that $h(n) \geqslant q + 1$, and by the remark at the end of the previous paragraph, $h(n) = q + 1$ for odd n. If n is even then $r \in \{0, 2\}$. When $r = 0$ then equality cannot hold in both places, so $h(4q) \geqslant q+1$. Chosing T with $4q$ leaves such that T_1 is a single vertex and T_2 is a realizer with $4q-1$ leaves, we get $\psi(T) = q+1$, so $h(4q) = q+1$. When $r=2$, $r'=3$, consequently $k>1$. This gives $h(4q+2) \geqslant q+1$. Let T_1 be the tree on 3 leaves and T_2 be a realizer on $4q-1$ leaves, then $\psi(T) = q+1$, so $h(4q+2) = q+1$. □

5 Unbalancing lights

We will establish our claimed upper bound on the maximum tanglegram crossing number here. Equation (2) gives an expression for the difference $\mathrm{cr}(\mathcal{D}) - \frac{1}{2}\binom{n}{2}$ between the crossing number $\mathrm{cr}(\mathcal{D})$ of a layout $\mathcal{D}$ of a tanglegram and $\frac{1}{2}\binom{n}{2}$ that is similar to the following theorem found in Alon and Spencer's paper [9]:

Theorem 5. *Let $a_{ij} = \pm 1$ for $1 \leqslant i,j \leqslant n$. Then there exists $x_i,\ y_j = \pm 1$, $1 \leqslant i,j \leqslant n$, so that*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_iy_j \geqslant \Big(\sqrt{\frac{2}{\pi}} + o(1)\Big)n^{3/2}. \tag{3}$$

Alon and Spencer [9] contains an amusing interpretation of Theorem 5, which explains the title of this section: "Let an $n \times n$ array of lights be given, each either on ($a_{ij} = +1$) or off ($a_{ij} = -1$). Suppose for each row and each column there is a switch so that if the switch is pulled ($x_i = -1$ for row i and $y_j = -1$ for column j) all of the lights in that line are 'switched': on to off and off to on. Then for any initial configuration it is possible to perform switches so that the number of lights on minus the number of lights off is at least $\Big(\sqrt{\frac{2}{\pi}} + o(1)\Big)n^{3/2}$."

Clearly, redefining all y_j to their negative, one obtains from (3)

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_iy_j \leqslant -\Big(\sqrt{\frac{2}{\pi}} + o(1)\Big)n^{3/2}. \tag{4}$$

If we want to find a drawing $\mathcal{D}$ of the tanglegram $\mathcal{T}$, where the number of crossings $\mathrm{cr}(\mathcal{D})$ in $\mathcal{D}$ is way below $\frac{1}{2}\binom{n}{2}$ in (1)–(2), then the formulation (4) of Theorem 5 is absolutely relevant—except that instead of $a_{ij} = \pm 1$, we have to deal with the following values

$$a_{xu} = \sum_{\substack{\{e,f\}:x=\mathrm{lca}_R(e,f)\\ u=\mathrm{lca}_L(e,f)}} \chi_{\mathcal{D}_0}(e,f),$$

that are computed from the fixed layout $\mathcal{D}_0$ and give $\mathrm{cr}(\mathcal{D}) - \frac{1}{2}\binom{n}{2}$ as

$$\sum_{x\in \mathrm{int}(R)}\sum_{u\in \mathrm{int}(L)} \alpha(x)\beta(u) \sum_{\substack{\{e,f\}:x=\mathrm{lca}_R(e,f)\\ u=\mathrm{lca}_L(e,f)}} \chi_{\mathcal{D}_0}(e,f). \tag{5}$$

The difficulty is that now a_{xu} may take other values than ± 1, in fact, it is difficult to find many non-zero a_{xu} terms. Therefore we were unable to utilize the probabilistic method.

We will call internal vertices of R that satisfies the property in Lemma 4, i.e., that have a leaf neighbor and an even number of leaf descendant *special vertices*. We have the following:

Lemma 6. *If x is a special vertex of the tanglegram $\mathcal{T}$, then*

$$\sum_{u \in \operatorname{int}(L)} a_{xu} \neq 0$$

Proof. Let e be the matching edge incident upon the leaf neighbor of the special vertex x and let $f_1, f_2, \ldots, f_{2k+1}$ be the matching edges at further leaf descendants of x. As we add up an odd number of ± 1 values,

$$\sum_{u \in \operatorname{int}(L)} a_{xu} = \sum_{i=1}^{2k+1} \chi_{\mathcal{D}_0}(e, f_i) \not\equiv 0 \mod 2. \quad \square$$

Now we are ready to prove

Theorem 7. *For any tanglegram $\mathcal{T}$ of size n,*

$$\operatorname{crt}(\mathcal{T}) \leqslant \frac{1}{2}\binom{n}{2} - \frac{n}{4}$$

Proof. Let us be given an arbitrary tanglegram $\mathcal{T}$ of size n. Without loss of generality, we can assume that the fixed layout $\mathcal{D}_0$ chosen realizes the crossing number of $\mathcal{T}$. Then for every $x \in S$, $\sum_{u \in \operatorname{int}(L)} a_{xu} < 0$, as by Lemma 6 this sum is non-zero, and if it was positive, a switch in x would yield a layout with strictly smaller number of crossings.

Consider now the following layout $\mathcal{D}_1$: switch in all $u \in \operatorname{int}(L)$. It is easy to see that

$$\operatorname{cr}(\mathcal{D}_1) = \binom{n}{2} - \operatorname{cr}(\mathcal{D}_0).$$

Now switch in layout $\mathcal{D}_1$ at every vertex $x \in S$ to obtain the layout $\mathcal{D}_2$:

$$\begin{aligned}
\operatorname{cr}(\mathcal{D}_2) &= \operatorname{cr}(\mathcal{D}_1) + 2\sum_{x \in S} \sum_{u \in \operatorname{int}(L)} a_{xu} \\
&= \binom{n}{2} - \operatorname{cr}(\mathcal{D}_0) + 2\sum_{x \in S} \sum_{u \in \operatorname{int}(L)} a_{xu} \\
&\leqslant \binom{n}{2} - \operatorname{cr}(\mathcal{D}_0) - 2|S|.
\end{aligned}$$

Hence

$$\operatorname{crt}(\mathcal{T}) = \operatorname{cr}(\mathcal{D}_0) \leqslant \frac{1}{2}\left(\operatorname{cr}(\mathcal{D}_0) + \operatorname{cr}(\mathcal{D}_2)\right) \leqslant \frac{1}{2}\binom{n}{2} - |S| \leqslant \frac{1}{2}\binom{n}{2} - \frac{n}{4}$$

by Lemma 4. $\square$

References

[1] S. Böcker, F. Hüffner, A. Truss, M. Wahlström. A faster fixed-parameter approach to drawing binary tanglegrams, In: Chen, J., Fomin, F.V. (eds) Parameterized and Exact Computation. IWPEC 2009. Lecture Notes in Computer Science, vol 5917. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-11269-0_3.

[2] K. Buchin, M. Buchin, J. Byrka, M. Nöllenburg, Y. Okamoto, R.I. Silveira, A. Wolff, Drawing (Complete) Binary Tanglegrams. Algorithmica 62, 309–332 (2012) doi:10.1007/s00453-010-9456-3

[3] A. Burt and R. Trivers, Genes in Conflict, Belknap Harvard Press, Cambridga MA, 2006.

[4] H. Fernau, M. Kaufmann, and M. Poths, Comparing trees via crossing minimization, In: Sarukkai, S., Sen, S. (eds) FSTTCS 2005: Foundations of Software Technology and Theoretical Computer Science. FSTTCS 2005. Lecture Notes in Computer Science, vol 3821. Springer, Berlin, Heidelberg doi:10.1007/11590156_37

[5] M.S. Hafner and S.A. Nadler, Phylogenetic trees support the coevolution of parasites and their hosts, Nature 332, 258–259 (1988) doi:10.1038/332258a0

[6] R.D.M. Page, (Editor) Tangled Trees. Phylogeny, Cospeciation and Coevolution, University of Chicago Press, Chicago IL, 2002.

[7] C. Scornavacca, F. Zickmann, D.H. Huson, Tanglegrams for rooted phylogenetic trees and networks, Bioinformatics. 27(13)(2011), 248–256 doi:10.1093/bioinformatics/btr210

[8] B. Venkatachalam, J. Apple, K. St. John, and D. Gusfield, Untangling tanglegrams: comparing trees by their drawings, In: Măndoiu, I., Narasimhan, G., Zhang, Y. (eds) Bioinformatics Research and Applications. ISBRA 2009. Lecture Notes in Computer Science, vol 5542. Springer, Berlin, Heidelberg doi:10.1007/978-3-642-01551-9_10

[9] N. Alon and J.H. Spencer, The Probabilistic Method, third edition, (John Wiley and Sons, New York, 2008).

[10] Robin Anderson, Shuliang Bai, Fidel Barrera-Cruz, Éva Czabarka, Giordano Da Lozzo, Natalie L.F. Hobson, Jephian C.-H. Lin, Austin Mohr, Heather C. Smith, László A. Székely, Hays Whitlatch, Analogies between the crossing number and the tangle crossing number, *Electronic J. Comb.* **25**(4) (2018) #P4.24 doi:10.37236/7581