Analogies between the Crossing Number and the Tangle Crossing Number

Robin Anderson[1] Shuliang Bai[2] Fidel Barrera-Cruz[3]
Éva Czabarka[2,9] Giordano Da Lozzo[4] Natalie L. F. Hobson[5]
Jephian C.-H. Lin[6] Austin Mohr[7] Heather C. Smith[8]
László A. Székely[2,9] Hays Whitlatch[2]

Submitted: Dec 30, 2017; Accepted: Oct 5, 2018; Published: Nov 2, 2018

Abstract

Tanglegrams are special graphs that consist of a pair of rooted binary trees with the same number of leaves, and a perfect matching between the two leaf-sets. These objects are of use in phylogenetics and are represented with straight-line drawings where the leaves of the two plane binary trees are on two parallel lines and only the matching edges can cross. The tangle crossing number of a tanglegram is the minimum number of crossings over all such drawings and is related to biologically relevant quantities, such as the number of times a parasite switched hosts.

Our main results for tanglegrams which parallel known theorems for crossing numbers are as follows. The removal of a single matching edge in a tanglegram with n leaves decreases the tangle crossing number by at most $n-3$, and this is sharp. Additionally, if $\gamma(n)$ is the maximum tangle crossing number of a tanglegram with n leaves, we prove $\frac{1}{2}\binom{n}{2}(1-o(1)) \leqslant \gamma(n) < \frac{1}{2}\binom{n}{2}$. For an arbitrary tanglegram T,

[1]Department of Mathematics, Saint Louis University, St. Louis, MO, U.S.A. (rander43@slu.edu).
[2]Department of Mathematics, University of South Carolina, Columbia, SC, U.S.A. ({sbai,czabarka,szekely,hww}@math.sc.edu).
[3]Sunnyvale, CA, U.S.A. (fidel.barrera@gmail.com).
[4]Department of Computer Science, Roma Tre University, Rome, Italy (dalozzo@dia.uniroma3.it).
[5]Department of Mathematics & Statistics, Sonoma State University, Rohnert Park, CA, U.S.A. (hobsonn@sonoma.edu).
[6]Department of Mathematics, Iowa State University, Ames, IA, U.S.A. (jephianlin@gmail.com).
[7]Department of Mathematics, Nebraska Wesleyan University, Lincoln, NE, U.S.A. (amohr@nebrwesleyan.edu).
[8]Dept. of Mathematics & Computer Science, Davidson College, Davidson, NC, U.S.A. (hcsmith@davidson.edu). Much of Smith's research was completed while affiliated with the Georgia Institute of Technology, Atlanta, GA, U.S.A.
[9]Czabarka and Székely are also Visiting Professors in the Department of Pure and Applied Mathematics, University of Johannesburg, Johannesburg 2006, South Africa

the tangle crossing number, $\mathrm{crt}(T)$, is NP-hard to compute (Fernau et al. 2005). We provide an algorithm which lower bounds $\mathrm{crt}(T)$ and runs in $O(n^4)$ time. To demonstrate the strength of the algorithm, simulations on tanglegrams chosen uniformly at random suggest that the tangle crossing number is at least $0.055n^2$ with high probabilty, which matches the result that the tangle crossing number is $\Theta(n^2)$ with high probability (Czabarka et al. 2017).

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

1 Introduction

A drawing $\mathcal{D}(G)$ of a graph G in the plane is a set of distinct points in the plane, one for each vertex of G, and a collection of simple open arcs, one for each edge of the graph, such that if e is an edge of G with endpoints v and w, then the closure (in the plane) of the arc α representing e consists precisely of α and the two points representing v and w. We further require that no edge–arc intersects any vertex point. The (standard) crossing number $\mathrm{cr}(\mathcal{D}(G))$ of $\mathcal{D}(G)$ is the number of pairs $(x, \{\alpha, \beta\})$, where x is a point of the plane, α, β are arcs of $\mathcal{D}$ representing distinct edges of G such that $x \in \alpha \cap \beta$. The crossing number $\mathrm{cr}(G)$ of a graph G is defined to be the minimum crossing number over all of its drawings.

Tanglegrams are well studied in the phylogenetics and computer science literature [4, 8, 10]. A tanglegram of size n is a triplet containing two rooted binary trees (L and R), each with n leaves, and a fixed perfect matching M between the two sets of leaves. Two tanglegrams $T_1 = (L_1, R_1, M_1)$ and $T_2 = (L_2, R_2, M_2)$ are the same if there is a pair of tree-isomorphisms (ϕ, ψ) from L_1 to L_2 and from R_1 to R_2 that map each pair of matched leaves to a pair of matched leaves. A layout of a tanglegram is a straight-line plane drawing of the trees, the first drawn in the half plane $x \leqslant 0$ with its leaves on the line $x = 0$ and the second in the half plane $x \geqslant 1$ with its leaves on the line $x = 1$, with a straight-line drawing of the matching edges between the leaves. The tangle crossing number $\mathrm{crt}(T)$ of a tanglegram T is the minimum crossing number over all of its layouts, i.e., the minimum number of unordered pairs of crossing edges over all layouts. The tangle crossing number is related to the number of times parasites switch hosts [10] as well as the number of horizontal gene transfers [4].

Though tangle crossing numbers are crossing numbers of a very specific kind of drawing of a very specific class of graphs, a number of analogies are known between tangle crossing numbers and crossing numbers. As with the crossing numbers of general graphs [9], computing the tangle crossing number is NP-hard [8], even when both trees are complete binary trees [3]. Testing whether a graph is planar can be done in polynomial (in fact linear) time [12]. Analogously, testing for tangle crossing number 0 can also be done in linear time [8]. Recently, Czabarka, Székely, and Wagner [6] gave an analogue of Kuratowski's Theorem [13] for tanglegrams, characterizing tangle crossing number 0. Clearly, for a graph G with e edges we have $\mathrm{cr}(G) = O(e^2)$, while for a tanglegram T of size n, $\mathrm{crt}(T) = O(n^2)$. The expected crossing number of an Erdős–Rényi random graph $G \in G(n, p)$ for $p = \frac{c}{n}$ for any $c > 1$ is $\Theta(e^2)$ where $e = p\binom{n}{2}$ is the expected number of

edges [14], and the expected tangle crossing number of a random and uniformly selected tanglegram with n leaves is $\Theta(n^2)$ [5], i.e., both of these quantities are as large as possible in order of magnitude.

We continue the study of the tangle crossing number with results which parallel results for graph crossing numbers. Hliněný and Salazar [11] studied the crossing number of 1-edge planar graphs (i.e., graphs in which there exists an edge whose removal results in a planar graph). For each $k \geqslant 1$, they define a 1-edge planar graph G_k with $2k+4$ vertices, $6k+7$ edges, and crossing number k. We find that the behavior is quite similar for the tangle crossing number. First we establish an upper bound for $\mathrm{crt}(T) - \mathrm{crt}(T-e)$ given any tanglegram T and any matching edge e. Then for each $n \geqslant 4$, we define a tanglegram of size n with tangle crossing number $n-3$ for which there is a single matching edge whose removal yields a planar subtanglegram. In summary, we prove the following theorem in Section 3:

Theorem 1. *For any tanglegram T of size $n \geqslant 3$ and any matching edge e in T, let $T-e$ be the tanglegram induced by deleting the endpoints of e and suppressing the resulting degree two vertices (i.e., replace each degree two vertex v by an edge between the neighbors of v). Then $\mathrm{crt}(T) - \mathrm{crt}(T-e) \leqslant n-3$. This inequality is best possible, even when $T-e$ is planar.*

We then examine the largest tangle crossing number of a tanglegram of size n (an analogue of the crossing number of the complete graph on n vertices). It is well known (e.g. by the crossing lemma or by the counting method) that the crossing number of the complete graph K_n is $\Theta(n^4) = \Theta\left(\binom{n}{2}^2\right)$. We prove the following result in Section 4:

Theorem 2. *For any tanglegram T of size n, $\mathrm{crt}(T) < \frac{1}{2}\binom{n}{2}$. If $\gamma(n)$ is the maximum tangle crossing number among all tanglegrams of size n, then*

$$\lim_{n\to\infty} \frac{\gamma(n)}{\binom{n}{2}} = \frac{1}{2}.$$

Interestingly, the structure of a size n tanglegram with maximum tangle crossing number remains unknown.

We conclude with a polynomial time algorithm for computing lower bounds on the tangle crossing number in Section 5. Drawing random tanglegrams of size n from a uniform distribution, we give computational evidence that these lower bounds are $\Theta(n^2)$ with high probability, thus matching the result of Czabarka, Székely, Wagner [5] that such a tanglegram has tangle crossing number $\Theta(n^2)$ with high probability.

2 Preliminaries

Before delving into the proofs of our main theorems, we need to establish some terminology and more formal definitions. A rooted binary tree is a tree in which one vertex is designated as the root and each vertex has either 0 or 2 children. A vertex with 0 children

is a *leaf*. A vertex with 2 children is called an *internal vertex*. Thinking of the root as a common ancestor to all other vertices, the notions of *descendant*, *parent*, *children* and *sibling* become clear. If B is a rooted binary tree, a subset of the leaves of B induces a binary subtree B', obtained from the smallest subtree of B which contains the considered subset of leaves, by suppressing all degree 2 vertices and choosing as the root of B' the vertex which was closest to the root of B. For any internal vertex v of B, the subtree induced by the leaves which are descendants of v is a *clade* of B at v, see Figure 1 for an example. If the two children of v are leaves, then the corresponding clade is called a *cherry*.

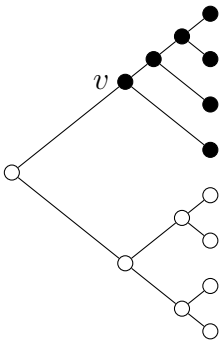

Figure 1: The clade at v is the subtree induced by the black leaves, namely the subtree induced by the vertices in black.

A *tanglegram layout* is a straight-line drawing in the plane of two rooted binary trees, L and R, each with n leaves and a perfect matching M between their leaves (each leaf of L paired with a unique leaf of R) having the following properties:

- A plane drawing of L appears in the half plane $x \leqslant 0$ with only the leaves of L on the line $x = 0$.
- A plane drawing of R appears in the half plane $x \geqslant 1$ with only the leaves of R on the line $x = 1$.
- The matching is represented by a (straight-line) drawing of edges connecting each leaf of L with the appropriate leaf of R.

The crossing number of such a layout is precisely the number of unordered pairs of matching edges which cross. As there are n matching edges, the crossing number is clearly at most $\binom{n}{2}$.

To transform one layout into another, we define a *switch*. First observe that a layout induces a total order on the leaves of L by the y-coordinate of the leaves on the line $x = 0$. Now each internal vertex v of L has two children v_1 and v_2. To make a switch at v, redraw the tree L so that in the new layout, the order of leaves ℓ and ℓ' is reversed if and only if one was a leaf in the clade at v_1 and the other was a leaf in the clade at v_2. The resulting tanglegram layout displays the new drawing of L, an unchanged drawing of

R, and the same matching edges drawn as straight-lines and connecting the appropriate pairs of leaves, for example see Figure 2. Switch operations at internal vertices of R are defined analogously. Observe that the switch operation defines an equivalence relation on the set of tanglegram layouts and each equivalence class will be called a *tanglegram*, denoted by the triple (L, R, M).

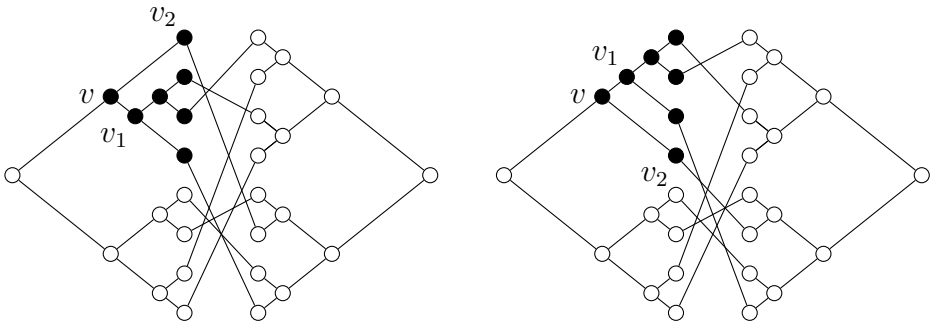

Figure 2: Two tanglegram layouts. Note that one layout is obtained from the other after a switch at vertex v. Thus both tanglegram layouts belong to the same tanglegram.

Let $T = (L, R, M)$ be a tanglegram. The *size* of T is the size of the matching M (also the number of leaves in L and the number of leaves in R). The *tangle crossing number* of T, denoted $\mathrm{crt}(T)$, is the minimum number of pairs of edges that cross, among all layouts of T. If T has size n then one can easily deduce that $\mathrm{crt}(T) \leqslant \binom{n}{2}$. A tanglegram is *planar* provided $\mathrm{crt}(T) = 0$.

Given a tanglegram $T = (L, R, M)$, a subset M' of M induces a *subtanglegram* $T' = (L', R', M')$ where L' is the subtree of L induced by leaves of L which are endpoints of edges in M' and R' is defined similarly.

We let $\gamma(n)$ to denote the maximum tangle crossing number among all tanglegrams of size n. In addition, we utilize the now standard notation $[n]$ for the set $\{1, 2, \ldots, n\}$.

3 Subtanglegrams of size $n - 1$

In a tanglegram of size n, the tangle crossing number is at most $\frac{1}{2}\binom{n}{2}$ (Theorem 2). Given a tanglegram with tangle crossing number close to this upper bound, on average, each matching edge crosses one fourth of all the other matching edges. We explore the *maximum* number of crossings a single edge could contribute to the overall tangle crossing number. Phrased another way, for any tanglegram T of size n and subtanglegram T' of size $n - 1$, we determine the maximum value of $\mathrm{crt}(T) - \mathrm{crt}(T')$. The result is given in Theorem 3, an upper bound which Theorem 6 shows to be tight, even for tanglegrams with T' planar. These two theorems together complete the proof of Theorem 1.

Throughout this section, given a tanglegram $T = (L, R, M)$ and $e \in M$, we use $T - e$ to denote the subtanglegram of T induced by edges in $M - e$.

Theorem 3. *If $T = (L, R, M)$ is a tanglegram of size $n \geqslant 3$ and e is any matching edge of T, then*

$$\operatorname{crt}(T) - \operatorname{crt}(T - e) \leqslant n - 3.$$

Proof. We will proceed by induction on n. For the case $n = 3$ we argue as follows. Each planar drawing of a tree determines a permutation its leaves; and if a tanglegram layout is planar, then the order of the leaves of R is uniquely determined by the order of the leaves of L and the matching. There are 4 planar drawings of L and 4 planar drawings of R. Given a planar drawing of L and a matching, we have 4 possible orders on the leaves of R that could potentially give a planar layout provided R can be drawn in a planar way giving that order. As there are 6 permutations of the leaves of R and 4 of these are planar, only 2 permutations of the leaves of R are non-planar, therefore there are at least $4 - 2 = 2$ planar layouts for the tanglegram.

Let $n \geqslant 4$ and suppose that in every tanglegram of size $n - 1$, each edge contributes at most $(n - 1) - 3$ to the tangle crossing number. Fix a tanglegram $T = (L, R, M)$ of size n, and let $e \in M$ be an arbitrary matching edge of T. Say e has endpoints u in L and v in R. Fix an optimal layout D' of $T - e = (L_u, R_v, M - e)$ with the fewest number of crossings.

In L, let $w_{L'}$ be the parent of u and let L' be the clade at the other child of $w_{L'}$. (Similarly, define $w_{R'}$ and R'.) There are two planar drawings of L whose subdrawings of L_u agree with the drawing of L_u in D', one with u immediately above the leaves of L' and one with u immediately below the leaves of L'. The ordering of the leaves of L_u in each of these drawings of L is exactly the same as the ordering of the leaves in the drawing of L_u in D'. Further, one of these drawings of L can be obtained from the other by making a switch at $w_{L'}$. A similar claim can be made about R, v, R_v, $w_{R'}$ and R'. Figure 3 uses dashed lines to indicate the two potential positions of u and for v in a drawing of T.

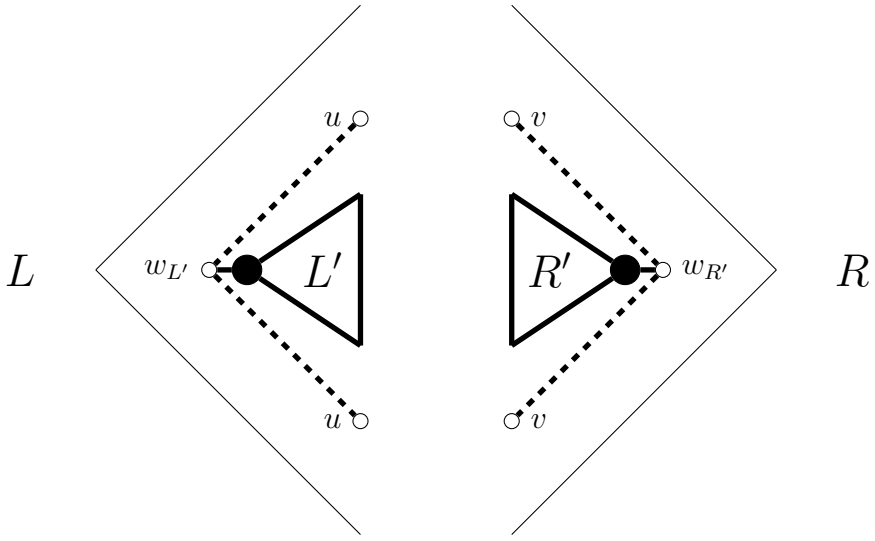

Figure 3: An illustration of the potential positions of u and v in the proof of Theorem 3.

We claim that there is drawing D of T using one of these two drawings of L and one of these two drawings of R in which the matching edge e crosses at most $n-3$ edges. This is sufficient to complete the proof as the number of crossings between two edges of $M-e$ in D is exactly $\mathrm{crt}(T-e)$ (because the underlying drawing of $T-e$ remained unchanged) which implies $\mathrm{crt}(T) \leqslant \mathrm{crt}(D) \leqslant \mathrm{crt}(T-e)+(n-3)$ as desired.

First observe that L' and R' each have at least one leaf. There are two cases to consider: (1) L' and R' each have exactly one leaf and they are matched in $M-e$, or (2) there is a leaf in L' and a leaf in R' which are not matched with one another.

For the first case, let f be the edge matching the single leaf in L' with the single leaf in R'. Consider the drawing of T with u above L' and v above R' so that e is above f; see Figure 4(a). Suppose, for contradiction, that e participates in strictly more than $n-3$ crossings in this drawing of T. As e does not cross itself or f, matching edge e must cross every other edge in M. Since there are no leaves between the left endpoints of e and f and no leaves between the right endpoints of e and f, it follows that f also participates in $n-2$ crossings in this drawing of T. As the drawing of $T-e$ was optimal, we see that f contributes $n-2$ to the tangle crossing number of tanglegram $T-e$ which had size $n-1$. However, by the induction hypothesis, each edge in $T-e$ contributes at most $(n-1)-3$ crossings to $\mathrm{crt}(T-e)$, a contradiction.

For the second case, let $u_{L'}$ be a leaf in L' and $v_{R'}$ be a leaf in R' which are not matched to each other. We say $u_{L'}$ (respectively, $v_{R'}$) is "matched upward" if the leaf to which it is matched is at least as high as the lowest leaf of R' (respectively, L'). The leaf $u_{L'}$ (respectively, $v_{R'}$) is "matched downward" if the leaf to which it is matched is no higher than the highest leaf of R' (respectively, L').

Let f_1 and f_2 be the matching edges, one with endpoint $u_{L'}$ and the other with endpoint $v_{R'}$. If $u_{L'}$ and $v_{R'}$ are both matched upward (respectively, downward), draw the vertex u below (respectively, above) L' and the vertex v below (respectively, above) R'; see Figure 4(b). On the other hand, if $u_{L'}$ is matched to a leaf higher (lower) than the leaves of R' and $v_{R'}$ is matched to a leaf lower (higher) than the leaves of L', then draw u directly below (above) the leaves of L' and v directly above (below) the leaves of R'; see Figure 4(c). In each of these cases, the edge e crosses neither f_1 nor f_2, and therefore crosses at most $n-3$ other edges, from which $\mathrm{crt}(T)-\mathrm{crt}(T-e) \leqslant n-3$ follows. □

Now we prove that the inequality in Theorem 3 is best possible. To do so, we present an infinite family of tanglegrams $\{P_n : n \geqslant 4\}$ such that P_n has size n, tangle crossing number $n-3$, and there exists a matching edge e such that $\mathrm{crt}(P_n-e)=0$. We say P_n is *1-edge tangle planar* as P_n is not planar but there is a matching edge e such that the subtanglegram P_n-e is planar. The two binary trees in P_n are rooted caterpillars.

Definition 4. The *rooted caterpillar* C_n of size n is the unique rooted binary tree with n leaves such that there are two leaves of distance $n-1$ from the root and for each $i \in [n-2]$ there is one leaf of distance i from the root. (See Figure 5 for an example.)

Definition 5. For each $n \geqslant 4$, we define the *caterpillar tanglegram* $P_n = (L_n, R_n, M_n)$ as follows: L_n and R_n are copies of the rooted caterpillar C_n. We label the leaves of L_n

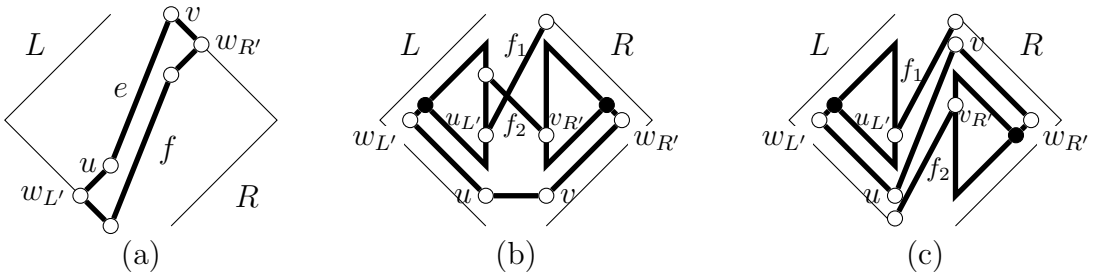

Figure 4: An illustration of the possible relations between L' and R' in the proof of Theorem 3: (a) L' and R' each have exactly one leaf and they are matched in $M - e$. (b) $u_{L'}$ an d $v_{R'}$ are not matched to each other and are both matched upward. (c) $u_{L'}$ is matched to a leaf higher than the leaves of R', and $v_{R'}$ is matched to a leaf lower than the leaves of L'.

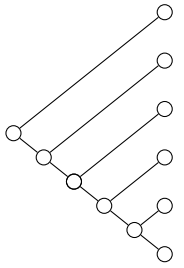

Figure 5: The caterpillar C_6.

as u_i, where i is the leaf's distance from the root. Since there are precisely two leaves at distance $n-1$, we arbitrarily label one of these u_n instead. Similarly, the leaves of R_n are labeled using v_i. Finally, we construct the matching $M_n = \{u_i v_{n-i} \mid i \in [n-1]\} \cup \{u_n v_n\}$. (See Figure 6 for an example.)

Theorem 6. *For each $n \geqslant 4$, the caterpillar tanglegram P_n is 1-edge tangle planar and has tangle crossing number $n-3$.*

Proof. Note that the tanglegram $P_n - u_n v_n$ is clearly a planar tanglegram (see Figure 6), so P_n is 1-edge tangle planar. The same drawing demonstrates that $\mathrm{crt}(P_n) \leqslant n-3$. It remains to show that $\mathrm{crt}(P_n) \geqslant n-3$. Suppose, for contradiction, that there is some k for which $\mathrm{crt}(P_k) < k-3$. Furthermore, let k be the least index witnessing this strict inequality.

Since P_3 has a planar drawing, $k \geqslant 4$. There are two cases for a fixed optimal drawing of P_k: at least one matching edge in the set $S = \{u_i v_{k-i} \mid 1 \leqslant i \leqslant k-1\}$ is part of a crossing or else none of them are.

In the former case, say the edge $u_j v_{k-j} \in S$ is part of a crossing. The subtanglegram induced by $M_k - u_j v_{k-j}$ is isomorphic to P_{k-1} and has tangle crossing number at most $\mathrm{crt}(P_k) - 1$. It follows that $\mathrm{crt}(P_{k-1}) \leqslant \mathrm{crt}(P_k) - 1 < (k-1) - 3$, which contradicts the minimality of k.

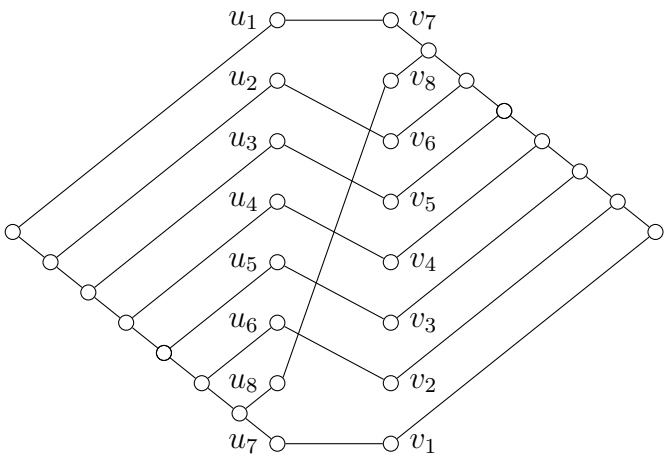

Figure 6: The caterpillar tanglegram P_8.

In the latter case, the subtanglegram $T = (L_T, R_T, M_T)$ induced by S has a planar layout in P_k. Any planar drawing of L_T must have u_1 at the top or at the bottom. Flipping the tanglegram if necessary, we may assume u_1 is at the top of the layout of T. Since the drawing of T is planar, v_{k-1} is also at the top. Similarly, v_1 and u_{k-1} have to be at the bottom in the drawing of T. It follows that every edge in $S' = \{u_i v_{k-i} \mid 2 \leqslant i \leqslant k-2\}$ is in between $u_1 v_{k-1}$ and $u_{k-1} v_1$. Now consider the possible positions of u_k and v_k in P_k. Since u_k and v_k has to be next to u_{k-1} (at the bottom) and v_{k-1} (at the top), respectively, the edge $u_k v_k$ is forced to pass through all edges in S'. Therefore, the tangle crossing number is at least $|S'| = n - 3$. □

4 Maximizing the crossing number

While a single edge in a tanglegram of size n can contribute up to $n-3$ to the tangle crossing number, not all matching edges can realize this many crossings in a drawing which minimizes the tangle crossing number. The aim of this section is to better understand the maximum tangle crossing number among tanglegrams of the same size. To prove Theorem 2, we begin with the first part:

Theorem 7. *If T is a tanglegram of size n then* $\operatorname{crt}(T) < \frac{1}{2}\binom{n}{2}$. *Consequently,*

$$\limsup_{n\to\infty} \frac{\gamma(n)}{\binom{n}{2}} \leqslant \frac{1}{2}.$$

Proof. Let $T = (L, R, M)$ be a tanglegram. Suppose $\operatorname{crt}(T) = k$ and let D be a tanglegram layout of T having k crossings. By making a switch at every internal vertex in R, we obtain a new layout D' of T. Note that in D', the plane drawing of R can be viewed as a reflection of the drawing of R in D across the line $y = 0$, while the plane drawing of L is the same in both D and D'. For any unordered pair of edges $\{e, f\}$ in M, e and f cross in D if

and only if they do not cross in D'. This implies that D' has exactly $\binom{n}{2} - k$ crossings. Since $\mathrm{crt}(T) = k$, every layout has at least k crossings. Consequently, $\binom{n}{2} - k \geqslant k$ and $\mathrm{crt}(T) = k \leqslant \frac{1}{2}\binom{n}{2}$.

Suppose that, contrary to our statement, $\mathrm{crt}(T) = \frac{1}{2}\binom{n}{2}$. It follows from our proof so far that any layout of T has $\frac{1}{2}\binom{n}{2}$ crossings, and for any unordered pair $\{e, f\}$ of matching edges there is a layout in which they cross. Let C be a cherry of R with leaves ℓ_1 and ℓ_2 incident with matching edges $e, f \in M$. As noted above, e and f must cross in some layout D of T. From D, we create a new layout D'' by making a switch at the parent of ℓ_1 and ℓ_2. The number of crossings in D'' is $\frac{1}{2}\binom{n}{2} - 1$, a contradiction. □

To complete the proof of Theorem 2, we prove

$$\liminf_{n\to\infty} \frac{\gamma(n)}{\binom{n}{2}} \geqslant 1/2$$

by constructing for each $n \geqslant 4$ a family $\mathcal{T}_n$ of tanglegrams of size n such that for any $\varepsilon > 0$ and large enough n, for all $T \in \mathcal{T}_n$ $\mathrm{crt}(T)/\binom{n}{2} \geqslant \frac{1}{2} - \varepsilon$.

We begin by constructing a family $\mathcal{T}_{k^2}$ for each integer $k \geqslant 2$. Any $T \in \mathcal{T}_{k^2}$ is the result of the following procedure: Take an arbitrary $(2k+2)$-tuple of rooted binary trees $(L_0, \ldots, L_k, R_0, \ldots, R_k)$, each of size k. Label the k leaves of L_0 with labels $\{1, 2, \ldots, k\}$ arbitrarily. For each $i \in [k]$, identify the root of L_i with leaf i in L_0 and assign labels $\{v_{ij} : j \in [k]\}$ to the k leaves of L_i. The result is the rooted binary tree L with k^2 leaves. Similarly, R is built from $(R_0, R_1, \ldots, R_k)$ with leaf labels $\{w_{ij} : i, j \in [k]\}$. The matching is defined as $M = \{v_{ij}w_{ji} : i, j \in [k]\}$.

Figure 7 shows a tanglegram in $\mathcal{T}_9$. The binary trees L_1, L_2, L_3 are marked by dashed rectangles. The tree L_0 is the subtree of L consisting of the roots of L_1, L_2, L_3, and their ancestors. Note that the trees L_0, L_1, L_2, and L_3 need not be isomorphic. They are only isomorphic here because there is only one binary tree, up to isomorphism, with 3 leaves. Further, for any pair of trees from $\{L_1, L_2, L_3\}$ and two trees from $\{R_1, R_2, R_3\}$, there is at least one pair of edges which cross.

With a well-defined set of tanglegrams $\mathcal{T}_{k^2}$ for each $k \geqslant 2$, we now define $\mathcal{T}_n$ for any integer n. Fix n and choose k such that $k^2 \leqslant n < (k+1)^2$. Let $\mathcal{T}_n$ be the set of tanglegrams of size n such that $T \in \mathcal{T}_n$ if and only if there is a tanglegram $T' \in \mathcal{T}_{k^2}$ with T' a subtanglegram of T. Figure 8 shows a tanglegram in $\mathcal{T}_5$. The tanglegram with bold edges is a subtanglegram in $\mathcal{T}_4$.

Theorem 8.

$$\liminf_{n\to\infty} \frac{\gamma(n)}{\binom{n}{2}} \geqslant \frac{1}{2}.$$

Proof. First we show that for each $k \geqslant 2$ and $T \in \mathcal{T}_{k^2}$, $\mathrm{crt}(T) \geqslant \binom{k}{2}^2$. Observe that for each $i \in [k]$, L_i is a clade of L at one of the leaves of L_0. Therefore in any tanglegram layout of (L, R, M) all the leaves of L_i appear forming a vertical consecutive block, for each $i \in [k]$. A similar assertion holds for the leaves of R_i, $i \in [k]$. For any $\{i, j\}, \{a, b\} \subseteq [k]$, there are 4 edges with both endpoints in the clades L_i, L_j, R_a, and R_b. Because the leaves

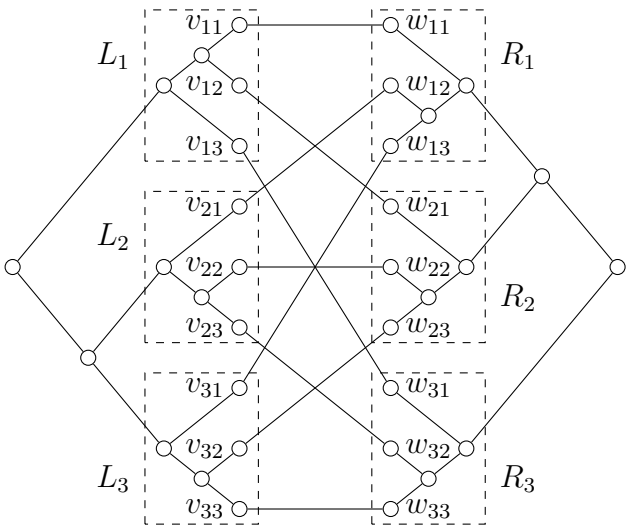

Figure 7: A tanglegram in $\mathcal{T}_9$.

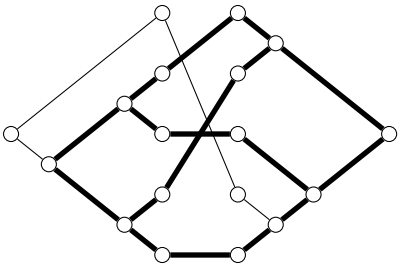

Figure 8: A tanglegram in $\mathcal{T}_5$.

in a single clade form a vertical consecutive block in any layout, either the edges $v_{ia}w_{ai}$ and $v_{jb}w_{bj}$ or the edges $v_{ja}w_{aj}$ and $v_{ib}w_{bi}$ form a crossing. As a result, $\operatorname{crt}(T) \geqslant \binom{k}{2}^2$.

As the tangle crossing number of each tanglegram in $\mathcal{T}_{k^2}$ is at least $\binom{k}{2}^2$, the tangle crossing number of each tanglegram in $\mathcal{T}_n$ with $n > k^2$ is also at least $\binom{k}{2}^2$.

Let $n \geqslant 4$ and $k = \lfloor \sqrt{n} \rfloor$, so $k^2 \leqslant n < (k+1)^2$. Observe that for each tanglegram $T \in \mathcal{T}_n$, $\operatorname{crt}(T) \geqslant \binom{k}{2}^2$. Therefore

$$\begin{aligned}\frac{\gamma(n)}{\binom{n}{2}} \geqslant \frac{\max\limits_{T\in\mathcal{T}_n} \operatorname{crt}(T)}{\binom{n}{2}} \geqslant \frac{\binom{k}{2}^2}{\binom{(k+1)^2}{2}} &= \frac{1}{2}\left(1-\frac{2}{k+2}\right)\left(1-\frac{2}{k+1}\right)^2 \\ &\geqslant \frac{1}{2}\left(1-\frac{2}{\sqrt{n}+1}\right)\left(1-\frac{2}{\sqrt{n}}\right)^2.\end{aligned}$$

As a result,

$$\liminf_{n\to\infty} \frac{\gamma(n)}{\binom{n}{2}} \geqslant \liminf_{n\to\infty} \frac{1}{2}\left(1-\frac{2}{\sqrt{n}+1}\right)\left(1-\frac{2}{\sqrt{n}}\right)^2 = \frac{1}{2}. \qquad \square$$

Theorems 7 and 8 complete the proof of Theorem 2.

5 Lower bound of the tangle crossing number

Computing the exact value of $\mathrm{crt}(T)$ for an arbitrary tanglegram T is NP-hard [8], so algorithms for approximating tangle crossing numbers are needed. Let $T = (L, R, M)$ be an arbitrary tanglegram of size n. In this section, we present an algorithm, which produces a lower bound for $\mathrm{crt}(T)$ and runs in $O(n^4)$ time. Czabarka, Székely, and Wagner [5] proved that a random tanglegram with n leaves has tangle crossing number $\Theta(n^2)$ with high probability. To demonstrate the strength of the lower bounds output by our algorithm, we run it on a series of tanglegrams chosen uniformly at random, finding that the lower bounds from our algorithm are approximately $0.055n^2$ which matches the "with high probability" result of $\Omega(n^2)$ in [5].

The algorithm runs in two phases. First it partitions the leaves of each tree into clades. In the second phase the clades are used to compute the lower bound for $\mathrm{crt}(T)$. Now we describe the algorithm for partitioning the leaves of a given tree into clades, given a restriction on their size. Note that we use this algorithm independently for L and R.

Algorithm 1 Partition leaves into clades.

Input: A binary tree B and a number $C > 1$.

Output: A partition of the leaves of B into clades of size at most C.

1: Label each vertex v in B with the number of leaves in the clade of B at v.
2: Let $\{v_i\}_{i=1}^{k}$ be the set of vertices such that the label at v_i is at most C and whose parent has label greater than C.
3: **return** $\{V_i\}_{i=1}^{k}$, where V_i is the set of leaves in the clade of B at v_i.

Algorithm 1 can be implemented in $O(n)$ time. This follows from noting that step 1 requires a post-order traversal of B and each of steps 2 and 3 requires a pre-order traversal of B. Let $W = \{v_i\}_{i=1}^{k}$ be the set of vertices from step 2. Note that if $v_i, v_j \in W$, then v_i is not an ancestor of v_j and vice versa. It is easy to see that a consequence of this property is that the collection $\{V_i\}_{i=1}^{k}$ from step 3 is a partition of the leaves of B into clades. Algorithm 2 below computes the lower bound for the tangle crossing number.

Algorithm 2 Tangle crossing number lower bound.

Input: Tanglegram $T = (L, R, M)$, and numbers $C_L, C_R > 1$.

Output: A lower bound for $\mathrm{crt}(T)$.

1: Use Algorithm 1 to obtain $\{U_i\}_{i=1}^{\ell}$ for L and C_L.
2: Use Algorithm 1 to obtain $\{V_i\}_{i=1}^{r}$ for R and C_R.
3: For each U_i and V_j, let $M_{i,j}$ be the set of matching edges with one endpoint in U_i and one in V_j.
4: **return** $\sum_{i_1,i_2\subseteq[\ell]} \sum_{j_1,j_2\subseteq[r]} \min\{|M_{i_1,j_1}||M_{i_2,j_2}|, |M_{i_1,j_2}||M_{i_2,j_1}|\}$

Note that Algorithm 2 runs in $O(n^4)$. This follows since steps 1 and 2 take $O(n)$ time, step 3 takes $O(n^2)$ time, and step 4 takes $O(n^4)$ time. To prove correctness, suppose U_a and U_b are clades in L and suppose V_c and V_d are clades in R. Because these are clades, any layout of T will have either the M_{ac} edges cross the M_{bd} edges or the M_{ad} edges cross the M_{bc} edges. As a result, these 4 clades will contribute at least $\min\{|M_{ac}||M_{bd}|, |M_{ad}||M_{bc}|\}$ to the tangle crossing number. Thus, as done in step 4, summing these minimums over all $\binom{\ell}{2}$ pairs of clades from L and $\binom{r}{2}$ pairs of clades from R, we obtain a lower bound on $\mathrm{crt}(T)$.

One may notice that Algorithm 2 depends on the choice of C_L and C_R. When $n = k^2$, the choice of $C_L = C_R = \sqrt{n}$ is optimal for the tanglegrams in $\mathcal{T}_{k^2}$ from Section 4 described for the proof of Theorem 8. For each tree in these tanglegrams, Algorithm 1 finds the k clades with k leaves that were used to build these trees. With this clade partition, $M_{i,j} = 1$ for all $i, j \in [k]$. So the tangle crossing number is at least $\binom{k}{2}^2$ by Algorithm 2. It is not hard to find tanglegrams in $\mathcal{T}_{k^2}$ with tangle crossing number exactly $\binom{k}{2}^2$. Thus the output of Algorithm 2 for the family of tanglegrams $\mathcal{T}_{k^2}$ is tight.

We ran simulations for different choices of C_L and C_R with random tanglegrams drawn from a uniform distribution. Figure 9 shows the average lower bounds when $C_L, C_R \in \{4, \sqrt{n}, \frac{n}{2}\}$. For each $n \in \{10, \ldots, 100\}$, we picked 100 tanglegrams of size n uniformly at random. The random sampling algorithm is a SageMath [7] implementation of Algorithm 3 from [2, p. 253]. The source code for our implementation is available at [1]. Based on the simulations, it appears that $C_L = C_R = \frac{n}{2}$ yields better lower bounds.

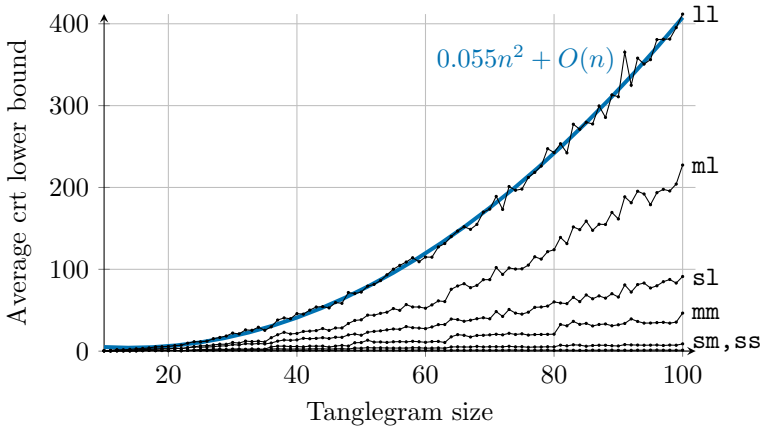

Figure 9: The average lower bound for $\mathrm{crt}(T)$ for different choices of C_L and C_R. The symbols `s`, `m` and `l` represent 4, $\sqrt{n}$, and $n/2$ respectively. The curve labeled `ml` represents the average output with $C_L = \sqrt{n}$ and $C_R = n/2$.

In [5] it is shown that there exists $C > 0$ such that a random tanglegram has tangle crossing number Cn^2 with high probability. Fitting the `ll` curve from Figure 9, the curve corresponding to $C_L = C_R = n/2$, to a quadratic function via least squares yields $0.055n^2 + O(n)$. This suggests that the tangle crossing number of the random tanglegram is at least $0.055n^2$. For the same sample, a plot of the maximum lower bounds is fit by the

curve $0.08n^2 + O(n)$. These two growth rates are to be compared with the upper bound of $0.25n^2$ from Theorem 2.

Another way to view this process is to create an auxiliary bipartite multigraph with a vertex for each clade and the number of edges between two clades is the number of edges which match a vertex of one clade to a vertex of the other clade. We then restrict to straight-line drawings where the vertices of one part remain on the line $x = 0$ and the vertices of the other part lie on the line $x = 1$. The minimum crossing number over all such drawings of this multigraph provides a lower bound on the crossing number of the tanglegram. However, Garey and Johnson [9] proved that even this problem on the auxiliary bipartite multigraph is NP-complete.

6 Open Questions and Further Work

Although the lower bound provided in Section 5 is tight for many small tanglegrams, we don't expect it being close to the real answer all the time, since we are doing a polynomial time approximation to an NP-hard problem. One may notice that the lower bound is dependent on the choice of clades. While we made an arbitrary choice, we are interested in polynomial time algorithms to choose the clades for an optimized lower bound.

In Section 4, we provided a family of tanglegrams with crossing number asymptotically $\frac{1}{2}\binom{n}{2}$. While the tangle crossing number of tanglegrams in $\mathcal{T}_n$ is at least $\binom{\lfloor\sqrt{n}\rfloor}{2}^2$, there are tanglegrams of size n with larger tangle crossing number. Is it perhaps true that $\max\{\mathrm{crt}(T) : T \in \mathcal{T}_n\} = \gamma(n)$, at least for $n = k^2$? We remain interested in the maximum tangle crossing number over all tanglegrams of size n.

Acknowledgements

The authors would like to extend their gratitude to the American Mathematical Society for organizing the Mathematics Research Community workshops where this work began. All authors were supported by the National Science Foundation under Grant Number DMS 1641020. Smith was also supported in part by NSF-DMS grant 1344199 and Székely was also supported by the NSF-DMS grants 1300547 and 1600811. Da Lozzo was supported by the U.S. Defense Advanced Research Projects Agency (DARPA) under agreement no. AFRL FA8750-15-2-0092. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The authors would like to thank the referees for their comments to improve readability and for their suggestion of proof for Theorem 6 that did not rely on computer checks for small tanglegrams.

References

[1] F. Barrera-Cruz and J. C.-H. Lin. Code for generating tanglegrams of size n uniformly. https://github.com/AMS-MRC-tanglegrams/tanglegrams, 2017.

[2] S. C. Billey, M. Konvalinka, and F. A. Matsen IV. On the enumeration of tanglegrams and tangled chains. *J. Combin. Theory Ser. A*, 146:239–263, 2017.

[3] K. Buchin, M. Buchin, J. Byrka, M. Nöllenburg, Y. Okamoto, R. Silveira, and A. Wolff. Drawing (complete) binary tanglegrams - hardness, approximation, fixed-parameter tractability. *Algorithmica*, 62(1-2):309–332, 2012.

[4] A. Burt and R. Trivers. *Genes in Conflict: The Biology of Selfish Genetic Elements*. Belknap Press, Cambridge, MA, 2008.

[5] É. Czabarka, L. A. Székely, and S. Wagner. Inducibility in binary trees and crossings in random tanglegrams. *SIAM J. Disc. Math.*, 31(3):1732–1750, 2017.

[6] Éva Czabarka, László A Székely, and Stephan Wagner. A tanglegram Kuratowski theorem. *Journal of Graph Theory*, 2018.

[7] The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 8.0)*, 2017. http://www.sagemath.org.

[8] H. Fernau, M. Kaufmann, and M. Poths. Comparing trees via crossing minimization. In S. Sarukkai and S. Sen, editors, *Proc. 25th Intern. Conf. Found. Softw. Techn. Theoret. Comput. Sci. (FSTTCS'05), LNCS vol. 3821*, pages 457–469. Springer-Verlag, 2005.

[9] M. R. Garey and D. S. Johnson. Crossing number is NP-complete. *SIAM J. Alg. Disc. Meth.*, 4(3):312–316, 1983.

[10] M. S. Hafner and S. A. Nadler. Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, 332:258–259, 1988.

[11] P. Hliněný and G. Salazar. On the crossing number of almost planar graphs. In M. Kaufmann and D. Wagner, editors, *Graph Drawing. GD 2006. LNCS vol. 4372*, pages 162–173. Springer Berlin Heidelberg, 2007.

[12] J.E. Hopcroft and R.E. Tarjan. Efficient planarity testing. *J. Assoc. Comput. Mach.*, 21(4):549–568, 1974.

[13] K. Kuratowski. Sur le problème des courbes gauches en topologie. *Fund. Math.*, 15:271–283, 1930.

[14] J. Spencer and G. Tóth. Crossing numbers of random graphs. *Random Structures and Algorithms*, 21:347–358, 2002.