Extremal Distances for Subtree Transfer Operations in Binary Trees

Ross Atkins and Colin McDiarmid

Abstract. Three standard subtree transfer operations for binary trees, used in particular for phylogenetic trees, are: tree bisection and reconnection (TBR), subtree prune and regraft (SPR), and rooted subtree prune and regraft (rSPR). We show that for a pair of leaf-labelled binary trees with n leaves, the maximum number of such moves required to transform one into the other is $n - \Theta(\sqrt{n})$, extending a result of Ding, Grünewald, and Humphries, and this holds also if one of the trees is fixed arbitrarily. If the pair is chosen uniformly at random, then the expected number of moves required is $n - \Theta(n^{2/3})$. These results may be phrased in terms of agreement forests: we also give extensions for more than two binary trees.

Mathematics Subject Classification. 05C05, 92D15.

Keywords. Phylogenetic tree, Subtree prune and regraft, Tree bisection and reconnection, Binary tree, Agreement forest, Tree rearrangement, Extremal distance, Average distance.

1. Introduction

A standard way to transform one binary tree into another is by performing ‘subtree-prune-and-regraft’ (SPR) moves (definitions are given below). These operations are of particular interest for phylogenetic trees. For a pair of binary trees with n labelled leaves, it was shown by Allen and Steel [2] in 2001 that the maximum number of SPR moves required to change one into the other, $D_{SPR}(n)$, is between $\frac{n}{2} - o(n)$ and $n - 3$ (for $n \geq 3$). Martin and Thatte [11] show the existence of a common (unrooted) subtree of size $\Omega(\sqrt{\log n})$, which brings the upper bound for $D_{SPR}(n)$ down to $n - \Omega(\sqrt{\log n})$.

Ding, Grünewald, and Humphries [6] show that $D_{SPR}(n)$ is actually $n - \Theta(\sqrt{n})$. Our Theorem 1.9 is an extended version of this result for the rooted as well as the unrooted cases. It says also that if one of the trees is fixed arbitrarily and the other is chosen to maximise the number of SPR moves

required to turn one into the other, then still $n - \Theta(\sqrt{n})$ moves are required. We prove the last result by showing that for a fixed pair of binary trees, we can label the leaves in such a way that $n - \Theta(\sqrt{n})$ moves are required to turn one into the other. This result is set in a more general context of agreement forests for sets of $k \geq 2$ trees.

We also consider the case when two trees are chosen independently, with an exchangeable distribution (for example, the uniform or the Yule–Harding distribution). We show that, if X is the number of moves required to transform one into the other, then $\mathbb{E}[X] = n - \Theta(n^{2/3})$ and indeed $X = n - \Theta(n^{2/3})$ with high probability; see Corollary 1.11. This implies that the expected size of a largest common (unrooted) subtree of two random trees is $\mathcal{O}(n^{2/3})$, where we measure size by the number of leaves. It is known, see [3] and [5], that this expected size is $\Omega(n^{1/8})$ in the uniform case and $\Omega(n^{0.344})$ in the Yule–Harding case; and it is $\mathcal{O}(n^{1/2})$ in both cases. Our results generalise the upper bound to random sets of $k \geq 2$ trees with n leaves in the following way. The expected size of a minimum agreement forest of such a random set of k trees is $n - \Theta(n^{2/(k+1)})$, and the expected size of a maximum agreement subtree is $\mathcal{O}(n^{2/(k+1)})$; see Eqs. (5.2) and (5.3).

Subtree transfer operations are subtly different when acting on rooted trees as opposed to unrooted trees. Some papers have considered SPR and rSPR moves as acting on different types of binary trees. We give a unified treatment here, and consider TBR, SPR, and rSPR to be three different classes of moves which act on the same class of trees. We work on the class $\mathcal{B}(n)$ of unrooted trees defined below, where always one of the leaf labels is 0. The leaf labelled 0 acts as the root of the tree for *rooted*-subtree-prune-and-regraft (rSPR) moves, and it behaves in exactly the same manner as the other leaves for SPR and TBR moves.¹

In the subsections below, we define binary trees, the three subtree transfer operations TBR, SPR, and rSPR, and the corresponding metrics they induce over the class of binary trees, and present our main theorems.

1.1. Binary Trees

Definition 1.1. Let S be a non-empty, finite set. If $|S| > 1$, then a binary leaf-labelled tree with leaf-label set S is a finite tree which has 2 types of vertices:

- tree vertices of degree 3, and
- leaf vertices of degree 1, labelled with the set S (there is a bijection between the leaves and S).

If $0 \in S$, the leaf labelled 0 is called the root of the tree. In the trivial case, when $S = \{\alpha\}$ is a singleton, the trivial graph containing only one vertex, labelled α , is considered to be a binary leaf-labelled tree, and its vertex is called a leaf.

¹Sometimes, the root of a rooted-binary-tree is defined to have degree 2. This is equivalent to our definition (see Definition 1.1). Simply add a new leaf, adjacent to the degree-2 root, label it 0 and make it the new root.

Definition 1.2. For any non-empty set S , let $\mathcal{B}(S)$ be the set of all binary leaf-labelled trees, with label set S . For a leaf-labelled tree A , let $L(A)$ denote the label set of A ; so if $A \in \mathcal{B}(S)$, then $L(A) = S$. To simplify notation, for any positive integer n , let $[n]_0 = \{0, 1, 2, \dots, n-1\}$ and let $\mathcal{B}(n)$ be given by

$$\mathcal{B}(n) = \mathcal{B}([n]_0) = \mathcal{B}(\{0, 1, 2, 3, \dots, n-1\}).$$

So, a tree in $\mathcal{B}(n)$ has n leaves and, for $n \geq 2$, it has $n-2$ tree-vertices (giving $2n-2$ vertices in total) and $2n-3$ edges.

For trees T and T' in $\mathcal{B}(S)$, we say T is isomorphic to T' (denoted $T \equiv T'$) if there is a graph isomorphism preserving leaf labels; that is, if there is a bijection $\psi: V(T) \rightarrow V(T')$ such that uv is an edge in T , if and only if $\psi(u)\psi(v)$ is an edge in T' , and for each leaf v of T , v and $\psi(v)$ are labelled with the same element of S . We identify isomorphic trees, as is customary. For non-empty $S' \subseteq L(A)$, let $A|S'$ be the minimal subtree of A that contains all the leaves with labels in S' . Let A/S' be the tree formed by suppressing any vertices of degree 2 from $A|S'$. So,

$$A|S' \text{ is a subgraph of } A \quad \text{and} \quad A/S' \in \mathcal{B}(S').$$

For a given tree $T \in \mathcal{B}(n)$ and a permutation $\pi \in S_{n-1}$, let T^π denote the tree in $\mathcal{B}(n)$ obtained from T by using π to permute the non-root leaf labels.

1.2. Tree Bisection and Reconnection

Definition 1.3. A tree-bisection-and-reconnection (TBR) move on a tree $T \in \mathcal{B}(n)$, for $n \geq 3$, is a two-step process:

- (1) (Bisection step) Delete an edge e from T (this edge is called the bisection edge) and then suppress all degree 2 vertices (the edges created by these suppressions are called new edges) to obtain a pair T_1, T_2 of binary trees with $L(T_1) \sqcup L(T_2) = [n]_0$. (The number of new edges is 2, unless the bisection edge was incident with a leaf. In the latter case, there is only one new edge and either T_1 or T_2 is an isolated vertex.)
- (2) (Reconnection step) Connect T_1 and T_2 by creating an edge between the midpoint of an edge in T_1 and the midpoint of an edge in T_2 . This edge is called the reconnecting edge. (In the case that T_1 or T_2 is an isolated vertex, the reconnecting edge is incident to this vertex.)

See Fig. 1: on the left, the dark vertical edge is the bisection edge, and the diagonal grey edges will form the new edges; and on the right, the dark vertical edge is the reconnection edge, and the horizontal grey edges are the new edges. If $A \in \mathcal{B}(n)$, and B is the result of performing a TBR on A , then $B \in \mathcal{B}(n)$ because B must be connected, acyclic, and all non-leaf vertices have degree 3.

Definition 1.4. Let $\mathcal{TBR} = \mathcal{TBR}(n)$ denote the set of pairs $(A, B) \in \mathcal{B}(n)^2$ such that B can be obtained from A by performing a single TBR move.

It is well known that TBR moves are reversible [14]; that is, if B can be obtained from A via a TBR move, then A can be obtained from B via a TBR move. Thus,

FIGURE 1. The TBR operation

FIGURE 2. The SPR move

$$(A, B) \in \mathcal{TBR} \quad \text{if and only if} \quad (B, A) \in \mathcal{TBR}.$$

1.3. Subtree Prune and Regraft

An SPR move (Definition 1.5) is an operation on binary trees, whereby a subtree is removed from one part of the tree and regrafted to another part of the tree. SPR moves are studied to measure similarity between different trees [2, 4, 10].

Definition 1.5. A subtree-prune-and-regraft (SPR) move taking xy from ab to cd is an operation which can be performed on a tree $T \in \mathcal{B}(n)$, as long as xy, ax, xb, cd are distinct edges (vertices a, b, c, d need not be distinct, possibly $a = c, b = c, a = d$, or $b = d$) and the path from c to y contains x . An SPR move is a two-step process:

- (prune step) delete edges ax, xb and replace them with ab , then
- (regraft step) delete edge cd and replace it with cx and xd .

Let $\mathcal{SPR} = \mathcal{SPR}(n)$ denote the set of pairs (A, B) such that $A \in \mathcal{B}(n)$ and B is the result of an SPR move on A . If the path from 0 to y passes through x , then this is called a rooted-subtree-prune-and-regraft (rSPR) move. Let $\mathcal{rSPR} = \mathcal{rSPR}(n)$ denote the set of pairs $(A, B) \in \mathcal{SPR}(n)$ such that B is the result of an rSPR move on A .

See Fig. 2. By definition, an rSPR move is a special case of an SPR move. It is not difficult to check that an SPR move is a special case of a TBR move. Informally, an SPR move is a TBR move in which one of T_1 or T_2 is an isolated

FIGURE 3. Trees that differ by a single SPR, but at least $\frac{n-3}{2}$ rSPR moves

vertex or one of the end points of the reconnecting edge is the midpoint of one of the new edges. Therefore, any valid SPR move performed on a binary tree in $\mathcal{B}(n)$ will result in a tree in $\mathcal{B}(n)$, and moreover:

$$rSPR \subseteq SPR \subseteq TBR.$$

These inclusions have been noted in many publications [2, 4, 14]. If we think of the edges as being oriented away from the root, then an rSPR move is an SPR move that preserves these orientations. It is also known that SPR moves and rSPR moves are reversible [14]. Thus,

$$(A, B) \in SPR \iff (B, A) \in SPR \quad \text{and} \quad (A, B) \in rSPR \iff (B, A) \in rSPR.$$

Not all SPR moves can be achieved with a single rSPR move. Indeed, we may need an arbitrary number of rSPR moves to simulate one SPR move [4]. For example, the trees in Fig. 3 differ by a single SPR move by pruning off the single leaf labelled 0, and then regrafting it appropriately. In Example 2.3, we will show that at least $\frac{n-3}{2}$ rSPR moves are required to transform one into the other.

The following lemma appears as [2, Lemma 2.7 (2b)].

Lemma 1.6 [2]. *Any TBR move can be achieved using at most 2 SPRs; that is, if $(A, B) \in TBR$, then there exists some C such that $(A, C), (C, B) \in SPR$.*

1.4. Tree Metrics

The SPR distance (Definition 1.7) has been the subject of several publications, for example, [2, 8]. The problem of determining the rSPR distance between a pair of trees was shown to be *NP*-hard by Bordewich and Semple [4] in 2005. We shall see later, that all of the values in the following definition are finite.

Definition 1.7. Let $n \geq 2$ be a fixed integer. For $A, B \in \mathcal{B}(n)$ and any subtree transfer operation $\chi \in \{TBR, SPR, rSPR\}$, define the χ -distance between A and B (denoted $d_\chi(A, B)$), to be the minimum number of χ moves required to change A into B . Let $D_\chi(n)$ denote the χ -diameter of the class $\mathcal{B}(n)$,

$$D_\chi(n) := \max_{A, B \in \mathcal{B}(n)} d_\chi(A, B);$$

n	$R_{TBR}(n)$	$D_{TBR}(n)$	$R_{SPR}(n)$	$D_{SPR}(n)$	$R_{rSPR}(n)$	$D_{rSPR}(n)$
3	0	0	0	0	0	0
4	1	1	1	1	1	1
5	2	2	2	2	2	2
6	2	2	2	2	2	3
7	3	3	3	3	3	4
8	4	4	4	4	4	5
9	5	5	5	5	5	5

FIGURE 4. Small values of the radius and diameter

and moreover, let $R_\chi(n)$ denote the χ -radius of the class $\mathcal{B}(n)$,

$$R_\chi(n) := \min_{C \in \mathcal{B}(n)} \max_{A \in \mathcal{B}(n)} d_\chi(A, C).$$

The following is an immediate consequence of $rSPR \subseteq SPR \subseteq TBR$.

Proposition 1.8. *For any $\chi \in \{TBR, SPR, rSPR\}$, and any integer $n > 1$, we have $D_\chi(n) \geq R_\chi(n)$. Moreover,*

$$D_{rSPR}(n) \geq D_{SPR}(n) \geq D_{TBR}(n) \quad \text{and} \quad R_{rSPR}(n) \geq R_{SPR}(n) \geq R_{TBR}(n).$$

1.5. Main Theorems

We give both the worst-case and the average-case results on numbers of moves needed by each of the three subtree transfer operations, starting with the worst-case results. The values of the radius and diameter for $n \leq 9$ are given in Fig. 4. The asymptotic values of the radius and diameter are given in the following theorem.

Theorem 1.9. *For each $\chi \in \{TBR, SPR, rSPR\}$, both $D_\chi(n)$ and $R_\chi(n)$ are $n - \Theta(\sqrt{n})$.*

By Proposition 1.8, to prove Theorem 1.9, it suffices to give an upper bound for the rSPR diameter and a lower bound for the TBR radius, both of the form $n - \Theta(\sqrt{n})$. These bounds are given explicitly in the inequalities (2.2) and (4.3), respectively. One of the results contained within Theorem 1.9 is that $D_{TBR}(n) = n - \Theta(\sqrt{n})$. This result is not new: it was shown recently in [6], where the proof uses a similar approach to ours, based on agreement forests. As well as considering other types of move, concerning TBR moves we see here also that $R_{TBR}(n) = n - \Theta(\sqrt{n})$ and we generalise agreement forest results to arbitrary sets of trees.

Next, let us consider distances between random trees. The fundamental random distance to consider is that between one fixed tree and a second fixed tree with its leaves randomly relabelled. We shall see that the expected distance is $n - \Theta(n^{2/3})$, and the same estimate holds with high probability for random distances. Recall that T^π denotes the tree obtained from T by using π to permute the non-root leaf labels.

Theorem 1.10. *Let $\chi \in \{TBR, SPR, rSPR\}$ be fixed. There exist constants $c_1 > c_2 > 0$ and $c_3 > 0$ such that the following holds. Let $n \geq 3$ be an integer,*

let T_1 and T_2 be trees in $\mathcal{B}(n)$, and let the random permutation π be uniformly distributed over S_{n-1} . Then

$$n - c_1 n^{2/3} < \mathbb{E}[d_\chi(T_1, T_2^\pi)] < n - c_2 n^{2/3}, \quad (1.1)$$

and further

$$\mathbb{P}\left(n - c_1 n^{2/3} < d_\chi(T_1, T_2^\pi) < n - c_2 n^{2/3}\right) \geq 1 - e^{-c_3 n^{1/3}}. \quad (1.2)$$

Again by Proposition 1.8, we will prove the statement (1.1) by demonstrating that the expected rSPR distance is at most $n - \Omega(n^{2/3})$ and the expected TBR distance is at least $n - \mathcal{O}(n^{2/3})$; see the inequalities (3.3) and (5.1). The statement (1.2) will be proved from (1.1) and a lemma on concentration around the mean. In fact, since $0 \leq d_\chi(A, B) \leq n$, the second statement (1.2) easily implies the first (1.1) (with c_1 and c_2 replaced by $2c_1$ and $\frac{1}{2}c_2$, respectively).

Define an equivalence relation $\sim$ on $\mathcal{B}(n)$ by setting $T_1 \sim T_2$ when $T_1 = T_2^\sigma$ for some permutation $\sigma \in S_{n-1}$. A random tree T in $\mathcal{B}(n)$ has an *exchangeable* distribution [1] if $\mathbb{P}(T = T_1) = \mathbb{P}(T = T_2)$ whenever $T_1 \sim T_2$. Exchangeability is also known as leaf-invariance [13]. The uniform distribution and the Yule–Harding distribution [15] are both examples of exchangeable distributions. We shall deduce the following corollary quickly from Theorem 1.10. It shows in particular that

$$\mathbb{E}[d_\chi(A, B)] = n - \Theta(n^{2/3}).$$

Corollary 1.11. *Let $\chi \in \{TBR, SPR, rSPR\}$ be fixed. There exist constants $c_1 > c_2 > 0$ and $c_3 > 0$ such that the following holds. Let $n \geq 3$ be an integer, and let the random trees A and B in $\mathcal{B}(n)$ be sampled independently, where B has an exchangeable distribution (and A may have any distribution). Then,*

$$n - c_1 n^{2/3} < \mathbb{E}[d_\chi(A, B)] < n - c_2 n^{2/3},$$

and further

$$\mathbb{P}\left(n - c_1 n^{2/3} < d_\chi(A, B) < n - c_2 n^{2/3}\right) \geq 1 - e^{-c_3 n^{1/3}}.$$

The rest of this paper is devoted to proving the above results, or amplified versions of them.

2. Upper Bound for the rSPR Diameter

The upper bound in Theorem 1.9 is proved in this section using rooted agreement forests. The definition of a rooted agreement forest given here is equivalent to that used by Bordewich and Semple [4].

Definition 2.1. Let $\mathcal{A} \subseteq \mathcal{B}(n)$ be a set of $k \geq 2$ trees. A rooted agreement forest of $\mathcal{A}$ is a partition $L_1, L_2, \dots, L_m$ of the leaf-label set $[n]_0 = \{0, 1, 2, \dots, n-1\}$, such that for all A, B in $\mathcal{A}$:

FIGURE 5. Three minimal rooted agreement forests

- the subtrees $A|L_1, A|L_2, \dots, A|L_m$ are pairwise disjoint, and
- $A/(L_j \cup \{0\}) \equiv B/(L_j \cup \{0\})$ for $j = 1, 2, \dots, m$.

Let $M_r(\mathcal{A})$ be the minimal possible value of m , the minimal number of parts in a rooted agreement forest of $\mathcal{A}$. When $k = 2$, we write $M_r(A, B)$ for $M_r(\{A, B\})$.

If $|L_j \setminus \{0\}| \leq 2$ for each j , then the ‘isomorphism’ condition above must be satisfied, though not necessarily the ‘disjointness’ condition. We can always construct a trivial rooted agreement forest with $n - 2$ parts (when $n \geq 3$) by setting $L_1 = \{0, 1, 2\}$ and $L_{i-1} = \{i\}$ for $i = 3, 4, 5, \dots, n - 1$: thus, we always have $M_r(\mathcal{A}) \leq n - 2$. Three minimal agreement forests are depicted in Fig. 5.

The following result is proved by Bordewich and Sempel [4], built on the work of Hein et al. [9].

Lemma 2.2 (Rooted Agreement Forest Lemma). *If A and B are any two trees in $\mathcal{B}(n)$, then*

$$M_r(A, B) = d_{rSPR}(A, B) + 1.$$

The rSPR distance between a pair of binary trees is often used as a measure of the discrepancy between them. So $M_r(\mathcal{A})$ can be seen as generalisation of this notion from pairs of trees to arbitrary sets of trees.

Example 2.3. Let A and B be the two caterpillar trees in Fig. 3 (with A on the right), which differ only by a single SPR move changing the location of the root. The other leaves in A are ordered in the usual $(1, 2, 3, \dots, n - 1)$ order from the root, while in B they are in the reverse order. Notice that $A/\{0, a, b, c\}$ is never isomorphic to $B/\{0, a, b, c\}$ because the leaf nearest to 0 in $A/\{0, a, b, c\}$ is the smallest of $\{a, b, c\}$, while the leaf nearest to 0 in $B/\{0, a, b, c\}$ is the largest. Therefore if $\{L_1, L_2, \dots, L_m\}$ is a rooted agreement forest for $\{A, B\}$, with $0 \in L_1$, then $|L_1| \leq 3$ and $|L_i| \leq 2$ for all $i > 1$. So, $n \leq 3 + 2(m - 1)$ and therefore $m \geq \frac{n-1}{2}$. By the rooted agreement

forest lemma (Lemma 2.2), this means $d_{\text{rSPR}}(A, B) \geq \frac{n-3}{2}$ and so

$$D_{\text{rSPR}}(n) \geq \frac{n-3}{2}.$$

The next two lemmas (Lemmas 2.4, 2.5) are used in the proof of Theorem 1.9 and the proof of (1.1) in Theorem 1.10. Essentially, the lemmas demonstrate that any binary tree can be divided into disjoint subtrees, each having roughly the same number of leaves.

Lemma 2.4. *For an integer $n \geq 2$, let $|S| = n$, let $T \in \mathcal{B}(S)$ be a binary tree with root $0 \in S$, and let a be a real number with $1 < a \leq 2(n-1)$. It is possible to remove a single edge from T , so that the number of leaves in the component not containing the root is in the interval $[\frac{a}{2}, a)$.*

Proof. For each vertex $x \in T$, let $g(x)$ be the set of leaves y such that x is on the path between the root and y . For every non-leaf vertex p , there is a natural partition of $g(p)$, namely,

$$g(p) = g(c_1) \sqcup g(c_2),$$

where $c_1, c_2 \in \Gamma(p)$ are the neighbours of p which are not on the path from p to the root. Moreover, if v is a leaf then $|g(v)| = 1$ (unless v is the root, in which case $g(v) = S$). Therefore, for each tree node v_i , there exists a neighbour v_{i+1} such that

$$|g(v_i)| > |g(v_{i+1})| \geq \frac{1}{2}|g(v_i)|. \quad (2.1)$$

Now, construct a path $(v_0, v_1, \dots, v_m)$, such that v_0 is the root, v_1 is the neighbour of the root and Eq. (2.1) holds for each $i = 1, 2, \dots, m-1$. This terminates when v_m is a leaf. So the sequence $|g(v_i)|$ decreases from

$$|g(v_1)| = n-1 \geq \frac{a}{2} \quad \text{to} \quad |g(v_m)| = 1 < a,$$

and never decreases by a factor less than $\frac{1}{2}$ in a single step. Hence, there must be some $1 \leq j \leq m$ such that

$$\frac{a}{2} \leq |g(v_j)| < a.$$

If edge $v_{j-1}v_j$ is deleted, then $g(v_j)$ is precisely the set of leaves in the component containing v_j . $\square$

Lemma 2.5 is a direct consequence of Lemma 2.4.

Lemma 2.5. *For an integer $n \geq 2$, let $T \in \mathcal{B}(n)$ and let $a > 1$. It is possible to find disjoint subtrees of T , such that: each leaf lies in a subtree, each subtree contains less than a leaves and every subtree that does not contain the root contains at least $\frac{a}{2}$ leaves.*

Proof. The case $a > n$ is trivial. If $n \geq a$, then we can use Lemma 2.4 iteratively to prune off subtrees with less than a (but at least $\frac{a}{2}$) leaves until the remaining tree has less than a leaves left (this remainder might have less than $\frac{a}{2}$ leaves). This final remaining tree contains the root. $\square$

The following lemma is the crux of this section. In the proof, for any set of k binary trees, we find a large set of pairs of labels such that the paths between each pair are disjoint in each of the k trees. The rooted agreement forest which we construct in the proof contains only parts of size 1 or 2.

Lemma 2.6. *For $k \geq 2$, and any set $\mathcal{A} = \{A_1, A_2, \dots, A_k\}$ of trees in $\mathcal{B}(n)$ where $n \geq 3$,*

$$M_r(\mathcal{A}) < n - \frac{1}{2(k+1)} \left(\frac{n}{k+1} \right)^{1/k}.$$

Proof. Since $M_r(\mathcal{A}) \leq n - 2$ for $n \geq 3$, we may assume, without loss of generality, that $n \geq k + 1$. With foresight, we define $t = \lfloor \left(\frac{n}{k+1} \right)^{1/k} \rfloor$. Divide each tree A_i into t (possibly empty) disjoint connected parts,

$$A_i = A_{i,1} \sqcup A_{i,2} \sqcup \dots \sqcup A_{i,t},$$

such that each part contains less than $\frac{2n}{t}$ leaves. This is possible by letting $a = \frac{2n}{t}$ in Lemma 2.5. Now, let us define a *token* to be a map $\tau: [k] \rightarrow [t]$. Thus, a token specifies a part for each tree A_i . For any token τ , define

$$I(\tau) := \bigcap_{i=1}^k L(A_{i,\tau(i)}) \quad \text{and} \quad U(\tau) := \bigcup_{i=1}^k L(A_{i,\tau(i)}).$$

For each label x , let τ_x be the token where, for each $i \in [k]$, we set $\tau_x(i) = j$ when $x \in L(A_{i,j})$. Observe that τ_x is the unique token τ such that $x \in I(\tau)$. Thus, the non-empty sets $I(\tau)$ partition the label set $[n]_0$. We say that the token τ is *nice* if $|I(\tau)| \geq 2$, and we say a pair τ, τ' are *friendly* if $\tau(i) \neq \tau'(i)$ for all i . Now, if F is a set of pairwise friendly, nice tokens, then we can construct an agreement forest of $n - |F|$ parts in the following manner:

- for each $\tau \in F$, construct a part of size 2 using labels in $I(\tau)$,
- all other labels are put in their own part of size 1.

In each tree in $\mathcal{A}$, the paths between labels in each part of this partition must be disjoint because the tokens are friendly. Hence, this is a valid agreement forest. So for any such set F , we have $M_r(\mathcal{A}) \leq n - |F|$. Now, let F be any maximal set of friendly nice tokens (i.e. every nice token that is not in F is unfriendly with a token in F). If the label x is not in $U(\tau)$ for any $\tau \in F$, then the token τ_x is not friendly with any $\tau \in F$; and so by the maximality of F , τ_x must not be nice (so $I(\tau_x) = \{x\}$). Thus for each label $x \in [n]_0$ either $x \in I(\sigma)$ for some not-nice token σ , or $x \in U(\tau)$ for some $\tau \in F$. So, let us count: there are

- n labels in total,
- at most t^k not-nice tokens σ , each with at most 1 label in $I(\sigma)$, and
- less than ka labels in $U(\tau)$ for each token $\tau \in F$.

Now, $F \neq \emptyset$ since $t^k < n$, so $t^k + ka|F| > n$. Substituting the value of t , we obtain

$$|F| > \frac{n - t^k}{ka} \geq \frac{n - \left(\frac{n}{k+1} \right)}{ka} = \frac{t}{2(k+1)}.$$

Since the above inequality is strict and t, k are both integers, we can tighten the bound to:

$$|F| \geq \frac{t+1}{2(k+1)} > \frac{1}{2(k+1)} \left(\frac{n}{k+1} \right)^{1/k}.$$

Finally, since $M_r(\mathcal{A}) \leq n - |F|$, we obtain the required inequality. $\square$

Set $k = 2$ in Lemma 2.6 and then apply the rooted agreement forest lemma (Lemma 2.2). We obtain

$$D_{\text{rSPR}}(n) < n - \frac{1}{6\sqrt{3}}\sqrt{n} - 1 \quad (2.2)$$

for each $n \geq 2$, which is a suitable upper bound in Theorem 1.9.

3. Upper Bound for the Expectation

In this section, we prove the upper bound in statement (1.1) of Theorem 1.10. The main step is provided by the following lemma.

Lemma 3.1. *Let $k \geq 2$ be a fixed integer, and let $\mathcal{T}$ be an arbitrary set of k trees in $\mathcal{B}(n)$ (where $n \geq 3$). Let $\mathcal{A}$ be obtained by performing a random permutation of the non-root leaf labels of each $T \in \mathcal{T}$, where the k permutations are chosen independently and uniformly from S_{n-1} . Then,*

$$\mathbb{E}[M_r(\mathcal{A})] \leq n - \Omega(n^{2/(k+1)})$$

as $n \rightarrow \infty$.

Proof. Since $M_r(\mathcal{A}) \leq n - 2$ when $n \geq 3$, we may assume, without loss of generality, that n is sufficiently large where needed. Set $b = \lfloor (n^{k-1})^{1/(k+1)} \rfloor$ and note that $b \geq 1$; and set $s = \lfloor \frac{n}{2b} \rfloor$. For each $T \in \mathcal{T}$, let $T_1, T_2, \dots, T_s$ be disjoint subtrees of T such that the number of leaves in each T_i is exactly b and none of them contain the root. This is possible by Lemma 2.5 (with $a = 2b$) and taking subtrees as necessary. Let $A \in \mathcal{A}$ be the tree obtained by the random labelling of T , and let $A_1, A_2, \dots, A_s$ be the subtrees of A corresponding to $T_1, T_2, \dots, T_s$, respectively.

Let $x_1, \dots, x_j$ be distinct non-root labels (i.e. distinct elements of $\{1, \dots, n-1\}$). Call the set $\{x_1, \dots, x_j\}$ *good* if for each $A \in \mathcal{A}$, the leaves labelled $x_1, \dots, x_j$ all lie in the same subtree A_i of A . We are most interested in *good pairs*, when $j = 2$. Observe that if $\{x_1, x_2\}$ and $\{x_1, x_3\}$ are good pairs, then the set $\{x_1, x_2, x_3\}$ is good. For each $A \in \mathcal{A}$, we say that a good pair $\{x_1, x_2\}$ is *spoiled in A* if there is another good pair in the same subtree A_i of A as x_1 and x_2 . We say that a good pair $\{x_1, x_2\}$ is *spoiled* if it is spoiled in A for some $A \in \mathcal{A}$. Note that a good pair $\{x_1, x_2\}$ might be spoiled by a good pair $\{x_1, x_3\}$ (with $x_2 \neq x_3$) or by a disjoint good pair $\{x_3, x_4\}$. In the first case, we say that $\{x_1, x_2\}$ is *triply spoiled*, and in the second case we say that it is *disjointly spoiled*. We call a pair *super* if it is good and not spoiled.

For distinct non-root labels x_1 and x_2 , we have

$$\mathbb{P}(\{x_1, x_2\} \text{ is good}) = \left(\frac{sb}{n-1} \times \frac{b-1}{n-2} \right)^k, \quad (3.1)$$

because for any $A \in \mathcal{A}$ the probability that $x_1 \in \bigcup_{i=1}^s L(A_i)$ is $\frac{sb}{n-1}$, and given that $x_1 \in L(A_i)$ the probability that $x_2 \in L(A_i)$ is $\frac{b-1}{n-2}$. Similarly, if x_1, x_2 , and x_3 are distinct non-root labels, we have

$$\mathbb{P}(\{x_1, x_2, x_3\} \text{ is good}) = \left(\frac{sb}{n-1} \times \frac{b-1}{n-2} \times \frac{b-2}{n-3} \right)^k,$$

because the probability that $\{x_1, x_2\} \subseteq L(A_i)$ for some i is $\frac{sb}{n-1} \times \frac{b-1}{n-2}$, and given that $\{x_1, x_2\} \subseteq L(A_i)$ the probability that $x_3 \in L(A_i)$ is $\frac{b-2}{n-3}$. Observe that $\frac{b}{n} \leq \frac{2}{3}$ for n sufficiently large, and then $\frac{b-2}{n-3} \leq \frac{b}{n}$.

Let x_1 and x_2 be distinct non-root labels. Let $\mathcal{B}$ be the collection of subsets of $\{1, \dots, n-1\}$ of size b which contain $\{x_1, x_2\}$. For each $A \in \mathcal{A}$ and $B \in \mathcal{B}$, let $E_{A,B}$ be the event that one of the subtrees A_i of A has $L(A_i) = B$. We shall need to condition on these events to control dependence.

Now, let $A \in \mathcal{A}$ and $B \in \mathcal{B}$. Arguing as for (3.1) we may see that

$$\mathbb{P}(\{x_1, x_2\} \text{ good} \mid E_{A,B}) = \left(\frac{sb}{n-1} \times \frac{b-1}{n-2} \right)^{k-1}.$$

Let x_3 and x_4 be distinct elements of $B \setminus \{x_1, x_2\}$. Condition on $E_{A,B}$ and on $\{x_1, x_2\}$ being good. Then, for each $A' \neq A$ in $\mathcal{A}$, the probability that $x_3 \in \bigcup_{j=1}^s L(A'_j)$ is $\frac{sb-2}{n-3}$ and given that $x_3 \in L(A'_j)$, the probability that $x_4 \in L(A'_j)$ is at most $\frac{b-1}{n-4}$. Therefore,

$$\mathbb{P}(\{x_3, x_4\} \text{ good} \mid \{x_1, x_2\} \text{ good}, E_{A,B}) \leq \left(\frac{sb-2}{n-3} \times \frac{b-1}{n-4} \right)^{k-1} \leq \left(\frac{b}{n} \right)^{k-1}$$

(for sufficiently large n). Hence, using a union bound,

$$\begin{aligned} & \mathbb{P}(\{x_1, x_2\} \text{ good and disjointly spoiled in } A) \\ &= \sum_{B \in \mathcal{B}} \mathbb{P}(\{x_1, x_2\} \text{ good and disjointly spoiled in } A \mid E_{A,B}) \mathbb{P}(E_{A,B}) \\ &\leq \sum_{B \in \mathcal{B}} \left(\mathbb{P}(\{x_1, x_2\} \text{ good} \mid E_{A,B}) \right. \\ &\quad \times \sum_{x_3 \neq x_4 \in B \setminus \{x_1, x_2\}} \mathbb{P}(\{x_3, x_4\} \text{ good} \mid \{x_1, x_2\} \text{ good}, E_{A,B}) \mathbb{P}(E_{A,B}) \Big) \\ &\leq \sum_{B \in \mathcal{B}} \left(\frac{sb}{n-1} \times \frac{b-1}{n-2} \right)^{k-1} \binom{b-2}{2} \left(\frac{b}{n} \right)^{k-1} \mathbb{P}(E_{A,B}) \\ &= \mathbb{P}(\{x_1, x_2\} \text{ good}) \binom{b-2}{2} \left(\frac{b}{n} \right)^{k-1}, \end{aligned}$$

by (3.1) and using $\sum_{B \in \mathcal{B}} \mathbb{P}(E_{A,B}) = \frac{sb}{n-1} \times \frac{b-1}{n-2}$. The sum above involving x_3, x_4 is over unordered pairs (Figs. 6, 7).

Similarly, we may see that for each $x_3 \in B \setminus \{x_1, x_2\}$,

$$\mathbb{P}(\{x_1, x_2, x_3\} \text{ good} \mid \{x_1, x_2\} \text{ good}, E_{A,B}) \leq \left(\frac{b}{n} \right)^{k-1},$$

FIGURE 6. The construction in Lemma 2.4 for $n = 8$

FIGURE 7. A subdivision of a tree into 3 disjoint subtrees, each with 3 leaves

and

$$\begin{aligned}
 & \mathbb{P}(\{x_1, x_2\} \text{ good and triply spoiled in } A) \\
 &= \sum_{B \in \mathcal{B}} \mathbb{P}(\{x_1, x_2\} \text{ good and triply spoiled in } A \mid E_{A,B}) \mathbb{P}(E_{A,B}) \\
 &\leq \sum_{B \in \mathcal{B}} \left(\mathbb{P}(\{x_1, x_2\} \text{ good} \mid E_{A,B}) \right. \\
 &\quad \times \sum_{x_3 \in B \setminus \{x_1, x_2\}} \mathbb{P}(\{x_1, x_2, x_3\} \text{ good} \mid \{x_1, x_2\} \text{ good}, E_{A,B}) \mathbb{P}(E_{A,B}) \Big) \\
 &\leq \sum_{B \in \mathcal{B}} \left(\frac{sb}{n-1} \times \frac{b-1}{n-2} \right)^{k-1} (b-2) \left(\frac{b}{n} \right)^{k-1} \mathbb{P}(E_{A,B}) \\
 &= \mathbb{P}(\{x_1, x_2\} \text{ good}) (b-2) \left(\frac{b}{n} \right)^{k-1}.
 \end{aligned}$$

FIGURE 8. The paths between the two leaves in a super pair are disjoint because each super pair is contained within a single subtree A_i and no two super pairs share the same A_i

Since $\binom{b-2}{2} + (b-2) = \binom{b-1}{2}$, using a union bound, we see that

$$\mathbb{P}(\{x_1, x_2\} \text{ good and spoiled}) \leq \mathbb{P}(\{x_1, x_2\} \text{ good}) k \binom{b-1}{2} \left(\frac{b}{n}\right)^{k-1}.$$

But by our choice of b ,

$$k \binom{b-1}{2} \left(\frac{b}{n}\right)^{k-1} \leq \frac{1}{2} k \frac{b^{k+1}}{n^{k-1}} \leq \frac{1}{2},$$

and so

$$\mathbb{P}(\{x_1, x_2\} \text{ super}) \geq \frac{1}{2} \mathbb{P}(\{x_1, x_2\} \text{ good}) = \Omega(n^{-2k/(k+1)}).$$

Therefore, if S is the random number of super pairs, then

$$\mathbb{E}[S] = \binom{n-1}{2} \mathbb{P}(\{x_1, x_2\} \text{ super}) = \Omega(n^{2/(k+1)}).$$

Now, let us partition the labels such that each super pair forms a part of size two (super pairs are disjoint by definition, see Fig. 8) and all other labels are in individual parts. To prove that this partition is a valid agreement forest, it suffices to show that for any distinct super pairs $\mathbf{x}$ and $\mathbf{y}$, and each $A \in \mathcal{A}$, the paths $A|\mathbf{x}$ and $A|\mathbf{y}$ are disjoint. However, $A|\mathbf{x}$ is contained within subtree A_i and $A|\mathbf{y}$ is contained within subtree A_j for some $i \neq j$; and since A_i and A_j are disjoint, the partition must be a valid rooted agreement forest. It has $n - S$ components, so $M_r(\mathcal{A}) \leq n - S$ and hence

$$\mathbb{E}[M_r(\mathcal{A})] \leq n - \mathbb{E}[S] = n - \Omega(n^{2/(k+1)}),$$

as required. $\square$

FIGURE 9. An example of an unrooted agreement forest

FIGURE 10. The construction in Lemma 4.2

We shall deduce the required upper bound in statement (1.1) in Theorem 1.10 from Lemma 3.1. Initially, we phrase the argument generally, so we can reuse it later in the proof of the lower bound (Figs. 9, 10).

Fix $\chi \in \{rSPR, SPR, TBR\}$. Let $T_1, T_2 \in \mathcal{B}(n)$, and let π_1, π_2 be independently and uniformly distributed over S_{n-1} . If we apply the same permutation σ to the non-root leaf labels of two trees, then the χ -distance remains unchanged. Taking σ as π_1^{-1} , we see that

$$d_\chi(T_1^{\pi_1}, T_2^{\pi_2}) = d_\chi(T_1, (T_2^{\pi_2})^{\pi_1^{-1}}) = d_\chi(T_1, T_2^\pi), \quad (3.2)$$

where $\pi = \pi_2 \circ \pi_1^{-1}$ (with π_2 acting first). Further, π is uniformly distributed over S_{n-1} . To see this, let $\tau \in S_{n-1}$ be fixed. Then by conditioning on π_1 , using the independence of π_1 and π_2 , and then the uniformity of π_2 , we find that

$$\begin{aligned} \mathbb{P}(\pi = \tau) &= \mathbb{P}(\pi_2 = \tau \circ \pi_1) \\ &= \sum_{\tau'} \mathbb{P}(\pi_2 = \tau \circ \tau' | \pi_1 = \tau') \mathbb{P}(\pi_1 = \tau') \\ &= \sum_{\tau'} \mathbb{P}(\pi_2 = \tau \circ \tau') \mathbb{P}(\pi_1 = \tau') \\ &= \frac{1}{(n-1)!}, \end{aligned}$$

so indeed π is uniformly distributed over S_{n-1} .

By Lemma 3.1 with $k = 2$ and the rooted agreement forest lemma (Lemma 2.2), we obtain

$$\mathbb{E}[d_{\text{rSPR}}(T_1^{\pi_1}, T_2^{\pi_2})] \leq n - \Omega(n^{2/3}).$$

But by (3.2) above,

$$\mathbb{E}[d_{\text{rSPR}}(T_1^{\pi_1}, T_2^{\pi_2})] = \mathbb{E}[d_{\text{rSPR}}(T_1, T_2^{\pi})],$$

where π is uniformly distributed over S_{n-1} ; and so we have

$$\mathbb{E}[d_{\text{rSPR}}(T_1, T_2^{\pi})] \leq n - \Omega(n^{2/3}), \quad (3.3)$$

which proves the upper bound in statement (1.1) of Theorem 1.10.

4. Lower Bound for the Radius

The lower bound in Theorem 1.9 is established in Lemma 4.5 at the end of this section. The proofs in this section use a notion similar to the rooted agreement forest from the previous section; Definition 4.1 and Lemma 4.2 below correspond, respectively, to Definition 2.1 and Lemma 2.2, for TBRs instead of rSPRs.

Definition 4.1. Let $\mathcal{A} \subset \mathcal{B}(n)$ be a set of $k \geq 2$ trees. An unrooted agreement forest of $\mathcal{A}$ is a partition $L_1, \dots, L_m$ of the label set $[n]_0 = \{0, 1, 2, \dots, n-1\}$, such that for all $A, B \in \mathcal{A}$:

- The subtrees $A|L_1, A|L_2, \dots, A|L_m$ are pairwise disjoint, and
- $A|L_j \equiv B|L_j$ for $j = 1, \dots, m$.

Let $M_u(\mathcal{A})$ be the minimal possible value for m , the minimal number of parts in an unrooted agreement forest of $\mathcal{A}$. When $k = 2$, we write $M_u(A, B)$ for $M_u(\{A, B\})$.

Any rooted agreement forest is also an unrooted agreement forest, so we always have $M_u(\mathcal{A}) \leq M_r(\mathcal{A})$. The next result (the Unrooted Agreement Forest Lemma) appears in [2].

Lemma 4.2 (Unrooted Agreement Forest Lemma). *If A and B are any two trees in $\mathcal{B}(n)$, then,*

$$M_u(A, B) = d_{\text{TBR}}(A, B) + 1.$$

We now present two preliminary lemmas, which we shall use in the proof of the main result of this section, Lemma 4.5. The first shows that by choosing k permutations, we may ensure that each pair of labels is spread well apart by at least one of the permutations.

Lemma 4.3. *Let n, b , and k be integers such that $b, k \geq 2$ and $n \geq b^k$. Then there exist k permutations $\{\phi_j: [n]_0 \rightarrow [n]_0\}_{j=1}^k$ such that for any distinct $x, y \in [n]_0$,*

$$\max_i |\phi_i(x) - \phi_i(y)| > b^{k-1} - 2b^{k-3}. \quad (4.1)$$

TABLE 1. For $n = 8$, $k = 3$, and $b = 2$, the table shows ϕ_1 , ϕ_2 , and ϕ_3 (listing the elements of $[8]_0$ in order for each ϕ_i)

ϕ_1	0	1	2	3	4	5	6	7
ϕ_2	2	3	6	7	0	1	4	5
ϕ_3	1	3	5	7	0	2	4	6

For each $x < y$, $\max_i |\phi_i(x) - \phi_i(y)| \geq 3 (=b^2-1)$, with equality for 0, 3 and 4, 7

FIGURE 11. For $n = 28$, $k = 2$, and $b = 5$, the permutations ϕ_1 (left) and ϕ_2 (right) as constructed in Lemma 4.3 are shown. The depicted paths trace from $\phi_i(0)$ to $\phi_i(n-1)$ in order. For example, $\phi_2(7) - \phi_2(14) = -d_2b + d_1 = 2 \cdot 5 - 1 = 9$; see (4.2)

Note that (4.1) gives $\max_i |\phi_i(x) - \phi_i(y)| \geq b$ when $k = 2$, and $\max_i |\phi_i(x) - \phi_i(y)| \geq b^2 - 1$ when $k = 3$.

Proof. We may represent each $x \in [n]_0$ uniquely by a k -tuple $(x_1, x_2, \dots, x_k)$ of non-negative integers such that

- $x = x_1b^{k-1} + x_2b^{k-2} + \dots + x_k$, and
- $0 \leq x_i < b$ for all $1 < i \leq k$.

If $x < b^k$, then this exactly represents x in base b . For larger x , we allow the coefficient of b^{k-1} to exceed $b-1$. Let ϕ_1 be the identity permutation, and for each $2 \leq i \leq k$ define ϕ_i so that

$$\phi_i(x) > \phi_i(y) \quad \text{if and only if} \quad \begin{cases} \text{either,} & x_i < y_i, \\ \text{or,} & x_i = y_i \text{ and } x > y. \end{cases}$$

See Table 1 and Fig. 11 for illustrations. Given $x, y \in [n]_0$, let $d_i = x_i - y_i$ for each i . Now, let $2 \leq i \leq k$ and suppose that $\phi_i(x) > \phi_i(y)$ (so $d_i \leq 0$). We claim that

$$\phi_i(x) - \phi_i(y) \geq -d_i b^{k-1} + \sum_{j=1}^{i-1} d_j b^{k-j-1} + \sum_{j=i+1}^k d_j b^{k-j}. \quad (4.2)$$

We may establish this claim as follows. Define the function f_i on $[n]_0$ by setting

$$f_i(z) = \sum_{j=1}^{i-1} z_j b^{k-j-1} + \sum_{j=i+1}^k z_j b^{k-j}.$$

Thus, the claim (4.2) says that

$$\phi_i(x) - \phi_i(y) \geq -d_i b^{k-1} + f_i(x - y) = -d_i b^{k-1} + f_i(x) - f_i(y).$$

Suppose first that $d_i = 0$, and say $x_i = y_i = t$. Let $S = \{z \in [n]_0 : z_i = t\}$. Restricted to S , f_i is a monotonic bijection onto the interval of integers $\{0, 1, \dots, |S| - 1\}$. Thus,

$$\phi_i(x) - \phi_i(y) = |\{z \in S : x > z \geq y\}| = f_i(x) - f_i(y).$$

Now, suppose that $x_i < y_i$. Let $A_t = \{z \in [n]_0 : \phi_i(x) > \phi_i(z) \geq \phi_i(y) \text{ and } z_i = t\}$ and let $a_t = |A_t|$. Thus, $\phi_i(x) - \phi_i(y) = \sum_{t=x_i}^{y_i} a_t$. For any integer t with $x_i + 1 \leq t \leq y_i - 1$ we have $a_t \geq b^{k-1}$ (with equality if $n = b^k$), so the total contribution from such t is at least $(y_i - x_i - 1)b^{k-1}$. Now consider $t = x_i$. Each $z \in A_{x_i}$ satisfies $\phi_i(z) > \phi_i(y)$, so $A_{x_i} = \{z \in [n]_0 : \phi_i(x) > \phi_i(z) \text{ and } z_i = x_i\}$; and thus $a_{x_i} = f_i(x)$. Finally, consider $t = y_i$. Each $z \in A_{y_i}$ satisfies $\phi_i(z) < \phi_i(x)$, so $A_{y_i} = \{z \in [n]_0 : \phi_i(z) \geq \phi_i(y) \text{ and } z_i = y_i\}$, and thus $a_{y_i} = f_i(n-1) - f_i(y) + 1$. Thus,

$$\phi_i(x) - \phi_i(y) = \sum_{t=x_i}^{y_i} a_t \geq (y_i - x_i - 1)b^{k-1} + f_i(n-1) + 1 + f_i(x_i) - f_i(y)$$

(with equality if $n = b^k$). Now suppose that $n = b^k$. Then, $f_i(n-1) + 1 = b^{k-1}$ and we have

$$\phi_i(x) - \phi_i(y) = (y_i - x_i)b^{k-1} + f_i(x) - f_i(y),$$

so (4.2) holds with equality. But f_i is an increasing function, so (4.2) holds for all $n \geq b^k$, as required to establish the claim.

Now, we can prove the inequality (4.1). Fix distinct $x, y \in [n]_0$. Suppose first that $|x_i - y_i| > 1$ for some $i \in \{1, \dots, k\}$. All $z \in [n]_0$ such that $z_i = \min\{x_i, y_i\} + 1$ must have $\phi_i(z)$ strictly between $\phi_i(x)$ and $\phi_i(y)$. But there are at least b^{k-1} such values z , so (4.1) holds.

Thus, we may assume from now on that $|x_i - y_i| \leq 1$ for each i . Also, we may assume, without loss of generality, that $x_1 \geq y_1$. We now treat three cases separately:

- (1) $x_1 = y_1$.
- (2) $x_1 = y_1 + 1$ and $x_i \geq y_i$ for all i .
- (3) $x_1 = y_1 + 1$ and $x_i < y_i$ for some $i \geq 2$.

Case 1. We may assume, without loss of generality that $x < y$. Then there is an $i \geq 2$ such that $d_i = -1$ and $d_j = 0$ for each $1 \leq j < i$.

Substituting this into Eq. (4.2) yields

$$\phi_i(x) - \phi_i(y) \geq b^{k-1} - \sum_{j=0}^{k-i-1} b^j \geq b^{k-1} - \sum_{j=0}^{k-3} b^j > b^{k-1} - 2b^{k-3}.$$

Case 2. Since $d_1 = 1$ and $d_j \geq 0$ for all j , we have $\phi_1(x) - \phi_1(y) = x - y \geq b^{k-1}$.

Case 3. We have $d_1 = 1$, and there is some $i \geq 2$ such that $d_i = -1$ and $d_j \geq 0$ for all $j < i$. Substituting this into Eq. (4.2) yields

$$\phi_i(x) - \phi_i(y) \geq b^{k-1} + b^{k-2} - \sum_{j=0}^{k-3} b^j > b^{k-1}.$$

Thus the inequality (4.1) holds in each case, which completes the proof of the lemma. $\square$

The second preliminary lemma gives a property of a set $\mathcal{A}$ of trees which ensures that $M_u(\mathcal{A})$ is large. It will be used in the proofs of both Lemmas 4.5 and 5.1.

Lemma 4.4. *Let n , k , and t be positive integers, and let $\mathcal{A} \subseteq \mathcal{B}(n)$ be a set of k trees. For each $A \in \mathcal{A}$, let $A = A_1 \sqcup A_2 \sqcup A_3 \sqcup \dots$ be a partition of the vertices of A into at most t connected parts. Let $z \geq 0$, and suppose that for all but at most z pairs of distinct labels $x, y \in [n]_0$ there is at least one $A \in \mathcal{A}$ such that the leaves labelled with x and y are in different parts of A . Then,*

$$M_u(\mathcal{A}) \geq n - k(t-1) - z.$$

Proof. Let $\{L_1, L_2, \dots, L_m\}$ be an unrooted agreement forest of $\mathcal{A}$. Let $\tilde{A}$ be a tree in $\mathcal{A}$, and let $F = \tilde{A}/L_1 \sqcup \tilde{A}/L_2 \sqcup \dots \sqcup \tilde{A}/L_m$ be the corresponding forest with leaves labelled by $[n]_0$ (by Definition 4.1 it will not matter which tree in $\mathcal{A}$ we choose). Each tree $A \in \mathcal{A}$ can be divided into the given forest $A = A_1 \sqcup A_2 \sqcup A_3 \sqcup \dots$ by deleting at most $t-1$ edges. Let these edges be called *hot edges*. There is a total of at most $k(t-1)$ hot edges (at most $t-1$ in each $A \in \mathcal{A}$). An edge in A might not correspond to any edge in F , moreover there may be several edges that correspond to the same edge in F . In any case, each edge corresponds to at most one edge in F ; and in particular this holds for hot edges. Now form the forest F' by deleting from F all the edges in F corresponding to hot edges. Then we have deleted at most $k(t-1)$ edges (Fig. 12).

Consider two distinct leaves, labelled x and y , which are in the same component of F' . The unique path in F' between these leaves must be contained in $\tilde{A}/L_j$ for some j . Since no edge corresponding to a hot edge was deleted from this path, the leaves labelled x and y must be in the same component of each tree $A \in \mathcal{A}$. Thus, the number p of pairs of distinct leaves which are in the same component of F' satisfies $p \leq z$. Suppose that F' has j components, and pick a leaf in each component. The remaining $n-j$ leaves each contribute at least 1 to p , so $p \geq n-j$ and $j \geq n-p \geq n-z$; that is, F' has at least $n-z$ components. Therefore, F must have had $m \geq n-z-k(t-1)$ components. $\square$

FIGURE 12. Edges e_1, e_2 in A correspond to the same edge in $A/\{4, 5, 6\}$

Lemma 4.5. *Let $k \geq 2$ and $n \geq (3k)^k$ be integers. If $\mathcal{T} = \{T_1, T_2, \dots, T_k\}$ is a set of binary trees with n leaves, then there exists $\mathcal{A} = \{A_1, A_2, \dots, A_k\} \subseteq \mathcal{B}(n)$ such that A_i is obtained by labelling the leaves of T_i (for each i), and*

$$M_u(\mathcal{A}) > n - 3k n^{1/k}.$$

Proof. First set

$$b = \lfloor n^{1/k} \rfloor, \quad t = \left\lceil \frac{2n}{b^{k-1} - 2b^{k-3}} \right\rceil, \quad \text{and} \quad a = \frac{2n}{t} \leq b^{k-1} - 2b^{k-3}.$$

Then, divide each tree T_i into at most t connected parts, $T_i = T_{i1} \sqcup T_{i2} \sqcup T_{i3} \sqcup \dots$ (using Lemma 2.5) so each part contains less than a leaves. We shall label the leaves of each T_i in such a way that, for each pair of distinct labels, there is some i such that the leaves with these labels are in different parts of T_i , and then apply Lemma 4.4 (with $z = 0$).

Since $n \geq b^k$, by Lemma 4.3, we can find k permutations $\phi_1, \dots, \phi_k$ of $[n]_0$ such that for any two distinct labels $x, y \in [n]_0$ there is some $1 \leq i \leq k$ such that $|\phi_i(x) - \phi_i(y)| \geq a$. For each tree T_i , assign labels to the leaves of T_i in the order given by ϕ_i , firstly to all the leaves of T_{i1} , then T_{i2} , then T_{i3} , and so on. For any distinct labels x and y , consider a permutation ϕ_i in which $|\phi_i(x) - \phi_i(y)| \geq a$. If in the tree T_i some T_{ij} contained both x and y , then it would also have to contain all the labels z with $\phi_i(z)$ between $\phi_i(x)$ and $\phi_i(y)$: but this is impossible since each T_{ij} has less than a leaves. Thus by Lemma 4.4,

$$M_u(\mathcal{A}) \geq n - k(t - 1).$$

For constant k , one can easily observe that $t \sim 2n^{1/k}$ as $n \rightarrow \infty$. For an explicit bound (using $n \geq (3k)^k$), we can show that $t < 3n^{1/k} + 1$, in the following manner. Observe that $b = \lfloor n^{1/k} \rfloor$ satisfies $b \geq 3k$ and

$$b > n^{1/k} (1 - n^{-1/k}) \geq n^{1/k} \left(1 - \frac{1}{3k}\right).$$

Thus,

$$\begin{aligned} b^{k-1} - 2b^{k-3} &= b^{k-1} \left(1 - \frac{2}{b^2}\right) \\ &> n^{(k-1)/k} \left(1 - \frac{1}{3k}\right)^{k-1} \left(1 - \frac{2}{9k^2}\right) > \frac{2}{3} n^{(k-1)/k}, \end{aligned}$$

where the last inequality holds since

$$\left(1 - \frac{1}{3k}\right)^{k-1} \left(1 - \frac{2}{9k^2}\right) > 1 - \frac{k-1}{3k} - \frac{2}{9k^2} > \frac{2}{3}.$$

Hence, by the definition of t ,

$$t \leq \left\lceil \frac{2n}{\frac{2}{3} n^{(k-1)/k}} \right\rceil < 3n^{1/k} + 1.$$

Finally, we have

$$M_u(\mathcal{A}) \geq n - k(t-1) > n - 3kn^{1/k},$$

as required. $\square$

If $\mathcal{A}' = \{A'_1, \dots, A'_k\}$ is obtained from $\mathcal{A}$ by a single permutation π of the labels (i.e. the leaf labelled x in A_i is labelled $\pi(x)$ in A'_i for all $x \in [n]_0$ and all $i \in [k]$), then $M_u(\mathcal{A}') = M_u(\mathcal{A})$. Therefore, in the above lemma, we could have constructed $\mathcal{A}$ so that one of the trees has a particular labelling, say A_1 . Now if we set $k = 2$ and apply the unrooted agreement forest lemma (Lemma 4.2), then we find

$$R_{TBR}(n) > n - 6\sqrt{n} - 1 \quad (4.3)$$

for all integers $n \geq 36$, and the result is trivially true for $n < 36$ since the right hand side is negative. This is a suitable lower bound for Theorem 1.9.

When $k \geq 2$ is fixed and $n \rightarrow \infty$, Lemmas 2.6 and 4.5 show that the maximum value of $M(\mathcal{A})$ over any set $\mathcal{A}$ of k trees in $\mathcal{B}(n)$ is $n - \Theta(n^{1/k})$, and the same holds for $M_F(\mathcal{A})$.

We close this section by looking briefly at the general tool Lemma 4.3, again with $k \geq 2$ fixed and $n \rightarrow \infty$. The lemma shows that with k permutations, we can ensure a ‘spread’ of at least $(1 + o(1)) n^{(k-1)/k}$. (Here ‘spread’ refers to the value $\max_i |\phi_i(x) - \phi_i(y)|$.) Let us show that this bound cannot be much improved. Consider first $k = 2$. We may assume, without loss of generality, that ϕ_1 is the identity. Suppose that $\max_i |\phi_i(x) - \phi_i(y)| \geq a$ for each pair x, y of distinct labels. By a classical result of Erdős and Szekeres [7] in 1935, the n -sequence corresponding to ϕ_2 has a monotonic subsequence of length $t = \lceil \sqrt{n} \rceil$. By considering the $t-1$ gaps between successive terms in such a subsequence, we may see that $(t-1)/2 \cdot (a+1) \leq n-1$. This shows that $a \leq 2\sqrt{n} + O(1)$. Thus, we cannot improve the bound in (4.1) by a factor greater than 2.

For general $k \geq 2$, we cannot improve the bound by a factor $\alpha > 4k(k+1)^{(k+1)/k}$. For, if we could, then in the proof of Lemma 4.5 we could take t

about $(2/\alpha)n^{1/k}$, and that would yield

$$M_u(\mathcal{A}) \geq n - k(t-1) \geq n - (1 + o(1))\frac{2k}{\alpha}n^{1/k},$$

which would contradict Lemma 2.6 for large n .

5. Lower Bound for the Expectation

In Lemma 5.1 below, we give a lower bound on the expected size of an unrooted maximum agreement forest, which asymptotically matches the upper bound in Lemma 3.1 for the rooted case; see (5.3). This will yield the required lower bound on the expected TBR distance in statement (1.1) of Theorem 1.10, and the upper bound (5.2) on the expected size of a maximum agreement subtree.

Lemma 5.1. *Let $k \geq 2$ and $n \geq 2$ be integers, and let $\mathcal{T} = \{T_1, T_2, \dots, T_k\} \subseteq \mathcal{B}(n)$ be a set of k binary trees. For each $i = 1, 2, \dots, k$, let A_i be the tree $T_i^{\pi_i}$ obtained from T_i by a random relabelling π_i of the non-root labels, where the π_i are sampled independently and uniformly from S_{n-1} . If $\mathcal{A} = \{A_1, A_2, \dots, A_k\}$ then*

$$\mathbb{E}[M_u(\mathcal{A})] > n - (2k + \tfrac{1}{2})n^{2/(k+1)}.$$

Proof. Set $a = n^{\frac{k-1}{k+1}}$ and $t = \lceil \frac{2n}{a} \rceil$, so $t-1 < \frac{2n}{a} = 2n^{\frac{2}{k+1}}$. Now, using Lemma 2.5, divide each tree T_i into t (possibly empty) connected parts, so that each part has less than a leaves; and let $A_i = A_{i1} \sqcup \dots \sqcup A_{it}$ be the corresponding partition of A_i . For distinct $x, y \in [n]_0$, notice that

$$\mathbb{P}(y \in L(A_{ij}) \mid x \in L(A_{ij})) = \frac{|L(A_{ij})| - 1}{n-1} < \frac{a-1}{n-1} < \frac{a}{n} = n^{-\frac{2}{k+1}}.$$

Now, let Z be the set of pairs $\{x, y\}$ of distinct labels such that, for each i , x and y are in the same part A_{ij} . Since the relabellings π_i are sampled independently, $\mathbb{P}(\{x, y\} \in Z) < (n^{-\frac{2}{k+1}})^k$ for all distinct $x, y \in [n]_0$. By the linearity of expectation,

$$\mathbb{E}[|Z|] = \binom{n}{2} \mathbb{P}(\{x, y\} \in Z) < \binom{n}{2} n^{-\frac{2k}{k+1}} < \tfrac{1}{2} n^{\frac{2}{k+1}}.$$

Since $t-1 < 2n^{\frac{2}{k+1}}$, we can apply Lemma 4.4 to get

$$\begin{aligned} \mathbb{E}[M_u(\mathcal{A})] &\geq n - k(t-1) - \mathbb{E}[|Z|] \\ &> n - (2k + \tfrac{1}{2})n^{\frac{2}{k+1}}, \end{aligned}$$

as required. $\square$

Setting $k = 2$ in Lemma 5.1 gives $\mathbb{E}[M_u(T_1^{\pi_1}, T_2^{\pi_2})] > n - \frac{9}{2}n^{2/3} \geq n - 5n^{2/3} + 1$ for $n \geq 3$. Applying the unrooted agreement forest lemma (Lemma 4.2), we get

$$\mathbb{E}[d_{TBR}(T_1^{\pi_1}, T_2^{\pi_2})] > n - 5n^{2/3}$$

for all positive integers n . Now, we may complete the proof of the lower bound in statement (1.1) of Theorem 1.10, as in the proof of the upper bound.

By (3.2), $d_{TBR}(T_1^{\pi_1}, T_2^{\pi_2}) = d_{TBR}(T_1, T_2^\pi)$ where π is uniformly distributed over S_{n-1} , and we have

$$\mathbb{E}[d_{TBR}(T_1, T_2^\pi)] > n - 5n^{2/3}, \quad (5.1)$$

which is an explicit version of the required lower bound in (1.1).

Given a list $\mathcal{A}$ of k trees, let $\mathcal{MAST}(\mathcal{A})$ denote the maximum size of an (unrooted) agreement subtree, where we measure size by the number of leaves. Since $M_u(\mathcal{A}) \leq n + 1 - \mathcal{MAST}(\mathcal{A})$, for $\mathcal{A}$ as in Lemma 5.1, we have

$$\mathbb{E}[\mathcal{MAST}(\mathcal{A})] \leq n + 1 - \mathbb{E}[M_u(\mathcal{A})] = \mathcal{O}(n^{2/(k+1)}). \quad (5.2)$$

We already noted that, for any set $\mathcal{A}$ of $k \geq 2$ trees in $\mathcal{B}(n)$, we have $M_u(\mathcal{A}) \leq M_r(\mathcal{A})$. Putting together Lemmas 3.1 and 5.1, we see that, for $\mathcal{A}$ as in these lemmas, we have both

$$\mathbb{E}[M_u(\mathcal{A})] = n - \Theta(n^{2/(k+1)}) \quad \text{and} \quad \mathbb{E}[M_r(\mathcal{A})] = n - \Theta(n^{2/(k+1)}). \quad (5.3)$$

6. Concentration Results

In this section, we shall prove statement (1.2) of Theorem 1.10 and deduce Corollary 1.11. Fix $\chi \in \{TBR, SPR, rSPR\}$. Let $n \geq 2$ be an integer and let T_1 and T_2 be fixed trees in $\mathcal{B}(n)$. We investigate the distribution of the random variable $X = d_\chi(T_1, T_2^\pi)$, where π is uniformly distributed on S_{n-1} . Specifically, we shall show that X is highly concentrated about its mean, using the following result from [12] (Theorem 7.4 applied to Example 7.2).

Lemma 6.1 [12]. *Let S_m denote the group of all permutations of $\{1, 2, \dots, m\}$, and let $k > 0$ be a constant. Let $f: S_m \rightarrow \mathbb{R}$ be a function such that for any $\sigma \in S_m$ and any transposition τ , we have*

$$|f(\sigma \circ \tau) - f(\sigma)| \leq k.$$

Let π be chosen uniformly at random from S_m , let X be the random variable $X = f(\pi)$, and let μ denote $\mathbb{E}[X]$. Then for any $\alpha > 0$, we have

$$\mathbb{P}(|X - \mu| \geq \alpha) \leq 2 \exp\left(-\frac{2\alpha^2}{(m-1)k^2}\right).$$

For any tree $T \in \mathcal{B}(n)$, if T' is the result of a transposition of the labels of two non-root leaves of T , then $d_\chi(T, T') \leq 2$ because we can obtain T' from T by swapping the two leaves with 2 rSPR moves. Therefore, we can apply Lemma 6.1 to the random variable $X = d_\chi(T_1, T_2^\pi)$ with $k = 2$, and $m = n - 1$. By statement (1.1) of Theorem 1.10, there are constants $\lambda_1 > \lambda_2 > 0$ such that

$$n - \lambda_1 n^{2/3} < \mu < n - \lambda_2 n^{2/3}$$

for all integers $n \geq 2$. Let $\alpha = \frac{1}{2}\lambda_2 n^{2/3}$. Then by Lemma 6.1,

$$\begin{aligned}\mathbb{P}(|X - \mu| \geq \alpha) &\leq 2 \exp \left(-\frac{2 \left(\frac{1}{2} \lambda_2 n^{2/3} \right)^2}{4(n-2)} \right) \\ &\leq 2 \exp \left(-\frac{1}{8} \lambda_2^2 n^{1/3} \right).\end{aligned}$$

Choose constants $c_1 > c_2 > 0$ and $c_3 > 0$ by setting $c_1 = \lambda_1 + \frac{1}{2} \lambda_2$ and $c_2 = \frac{1}{2} \lambda_2$ and $c_3 = \frac{1}{8} \lambda_2^2$. Then by the above,

$$\mathbb{P} \left(n - c_1 n^{2/3} < X < n - c_2 n^{2/3} \right) \geq 1 - 2 \exp \left(-c_3 n^{1/3} \right).$$

Finally, recall that $0 \leq X \leq n - 2$ for each $n \geq 2$. Hence, by replacing c_3 by $\frac{1}{2} c_3$, and choosing c_1 larger and c_2 smaller if necessary, we may drop the factor 2 in the last term. This completes the proof of statement (1.2) of Theorem 1.10.

It remains to deduce Corollary 1.11. Observe first that, if the random tree T in $\mathcal{B}(n)$ has an exchangeable distribution, and T_0 is a fixed tree with $\mathbb{P}(T = T_0) > 0$, then conditioned on $T \sim T_0$, T is uniformly distributed on the equivalence class of T_0 under $\sim$, and so T has the same distribution as T_0^π where π is uniformly distributed on S_{n-1} .

Let A and B be sampled independently in $\mathcal{B}(n)$, where B has an exchangeable distribution. Let T_1 be any tree in $\mathcal{B}(n)$ with $\mathbb{P}(A = T_1) > 0$, and let T_2 be any tree in $\mathcal{B}(n)$ with $\mathbb{P}(B = T_2) > 0$. Conditioned on $A = T_1$ and $B \sim T_2$, $d_\chi(A, B)$ has the same distribution as $d_\chi(T_1, T_2^\pi)$ where π is uniformly distributed on S_{n-1} . Let I be the interval $(n - c_1 n^{2/3}, n - c_2 n^{2/3})$ in statement (1.1) of Theorem 1.10. Then by that result,

$$\mathbb{E}[d_\chi(A, B) \mid (A = T_1) \wedge (B \sim T_2)] = \mathbb{E}[d_\chi(T_1, T_2^\pi)] \in I.$$

But $\mathbb{E}[d_\chi(A, B)]$ is a weighted average of the terms on the left, so $\mathbb{E}[d_\chi(A, B)] \in I$. Similarly, let J be the interval $(n - c_1 n^{2/3}, n - c_2 n^{2/3})$ in statement (1.2) of Theorem 1.10. Then by that result,

$$\mathbb{P}(d_\chi(A, B) \in J \mid (A = T_1) \wedge (B \sim T_2)) = \mathbb{P}(d_\chi(T_1, T_2^\pi) \in J) \geq 1 - e^{-c_3 n^{1/3}}.$$

But $\mathbb{P}(d_\chi(A, B) \in J)$ is a weighted average of the terms on the far left, so

$$\mathbb{P}(d_\chi(A, B) \in J) \geq 1 - e^{-c_3 n^{1/3}}.$$

This completes the proof of Corollary 1.11.

7. Concluding Remarks

For each move $\chi \in \{TBR, SPR, rSPR\}$ we have estimated the diameter $D_\chi(n)$ and the radius $R_\chi(n)$ of the class $\mathcal{B}(n)$ quite precisely: we saw in Theorem 1.9 that both $D_\chi(n)$ and $R_\chi(n)$ are $n - \Theta(\sqrt{n})$. It would be interesting to learn about finer behaviour.

- Can we separate diameter and radius? Are there constants $0 < c_\chi < c'_\chi$ such that

$$D_\chi(n) > n - (c_\chi + o(1))\sqrt{n} \quad \text{and} \quad R_\chi(n) < n - (c'_\chi + o(1))\sqrt{n}?$$

- Can we separate, for example, rSPR diameter and (unrooted) SPR diameter? Are there constants $0 < c_r < c_u$ such that

$$D_{rSPR}(n) > n - (c_r + o(1))\sqrt{n} \quad \text{and} \quad D_{SPR}(n) < n - (c_u + o(1))\sqrt{n}?$$

- What about even finer behaviour? Are there constants $c_\chi > 0$ and $c'_\chi > 0$ such that

$$D_\chi(n) = n - (c_\chi + o(1))\sqrt{n} \quad \text{and} \quad R_\chi(n) = n - (c'_\chi + o(1))\sqrt{n}?$$

If so, echoing the first two questions, is the inequality $c_\chi \leq c'_\chi$ strict, and does c_χ or c'_χ depend on χ ?

There are similar natural questions arising from Theorem 1.10 and Corollary 1.11 on the behaviour of random distances.

Acknowledgements

We would like to thank Charles Semple and the two anonymous referees for several helpful comments, and Mike Steel for drawing Ref. [6] to our attention.

Open Access. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- [1] Aldous, D.: Probability distributions on cladograms. In: Aldous, D., Pemantle, R. (eds.) Random Discrete Structures, IMA Vol. Math. Appl., Vol. 76, 1–18. Springer, New York (1996)
- [2] Allen, B., Steel, M.: Subtree transfer operations and their induced metrics on evolutionary trees. *Ann. Combin.* 5(1), 1–15 (2001)
- [3] Bernstein, D.I., Ho, L.S.T., Long, C., Steel, M., St. John, K., Sullivant, S.: Bounds on the expected size of the maximum agreement subtree. *SIAM J. Discrete Math.* 29(4), 2065–2074 (2015)
- [4] Bordewich, M., Semple, C.: On the computational complexity of the rooted subtree prune and regraft distance. *Ann. Combin.* 8(4), 409–423 (2005)
- [5] Bryant, D., McKenzie, A., Steel, M.: The size of a maximum agreement subtree for random binary trees. In: Janowitz, M.F., Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (eds.) BioConsensus, Discrete Math. Theoret. Comput. Sci., Vol. 61, 55–65. American Mathematical Society, Providence, RI (2003)

- [6] Ding, Y., Grünewald, S., Humphries, P.J.: On agreement forests. *J. Combin. Theory Ser. A* 118(7), 2059–2065 (2011)
- [7] Erdős, P., Szekeres, G.: A combinatorial problem in geometry. *Compositio Math.* 2, 463–470 (1935)
- [8] Hein, J.: Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.* 98(2), 185–200 (1990)
- [9] Hein, J., Jiang, T., Wang, L., Zhang, K.: On the complexity of comparing evolutionary trees. *Discrete Appl. Math.* 71(1-3), 153–169 (1996)
- [10] Humphreys, P.J., Semple, C.: Note on the hybridization number and subtree distance in phylogenetics. *App. Math. Lett.* 22(4), 611–615 (2009)
- [11] Martin, D.M., Thatte, B.D.: The maximum agreement subtree problem. *Discrete Appl. Math.* 161(13-14), 1805–1817 (2013)
- [12] McDiarmid, C.: On the method of bounded differences. In: Siemons, J. (ed.) *Surveys in Combinatorics*, London Mathematical Society Lecture Note Series, Vol. 141, pp. 148–188. Cambridge University Press, Cambridge (1989)
- [13] Penny, D., Steel, M.: Distributions of tree comparison metrics—some new results. *Systematic Biol.* 42(2), 126–141 (1993)
- [14] Semple, C., Steel, M.: *Phylogenetics*. Oxford University Press, Oxford (2003)
- [15] Zhu, S., Steel, M.: Does random tree puzzle produce Yule-Harding trees in the many-taxon limit? *Math. Biosci.* 243(1), 109–116 (2013)

Ross Atkins and Colin McDiarmid
Department of Statistics
University of Oxford
24-29 St Giles'
Oxford OX1 3LB
UK
e-mail: cmcd@stats.ox.ac.uk

Ross Atkins
e-mail: ross.atkins@univ.ox.ac.uk

Received: 29 January 2015.