involve
a journal of mathematics

Expected conflicts in pairs of rooted binary trees

Timothy Chu and Sean Cleary

msp

2013 vol. 6, no. 3

INVOLVE 6:3 (2013)
dx.doi.org/10.2140/involve.2013.6.323

Expected conflicts in pairs of rooted binary trees

Timothy Chu and Sean Cleary

(Communicated by Robert W. Robinson)

Rotation distance between rooted binary trees measures the extent of similarity of two trees with ordered leaves. There are no known polynomial-time algorithms for computing rotation distance. If there are common edges or immediately changeable edges between a pair of trees, the rotation distance problem breaks into smaller subproblems. The number of crossings or conflicts of a tree pair also gives some measure of the extent of similarity of two trees. Here we describe the distribution of common edges and immediately changeable edges between randomly selected pairs of trees via computer experiments, and examine the distribution of the amount of conflicts between such pairs.

1. Introduction

Binary trees are used in a broad spectrum of computational and mathematical applications. Binary search trees, for example, are widely used in databases and can be used to ensure efficient searches. The shape of a binary search tree is important in guaranteeing this efficiency — a balanced binary tree guarantees worst-case search time on the order of $\log(n)$, whereas a tree with a stringy shape will have worst-case search time on the order of n, where n is the number of nodes in the tree, or equivalently, items to be stored. Because of such applications, there has been a great deal of interest in operations which preserve the left-to-right order of the leaves of a tree while adjusting the shape of the tree. See [Knuth 1973] for background and numerous algorithms related to tree shape and balance. One widely studied approach to adjust tree shape uses rotations in binary trees where there is a left-to-right order on the leaves. A rotation is a single move at a particular node which promotes one of the grandchild nodes to a child node, switches another grandchild to have a different parent, and demotes a child node to a grandchild node, while preserving the order. Such an operation is pictured in Figure 1.

MSC2010: 05C05, 68P05.
Keywords: random binary tree pairs.
The authors gratefully acknowledge support from the National Science Foundation under grant 0811002.

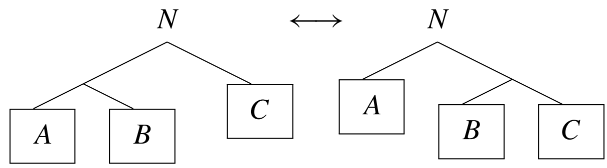

Figure 1. Rotation at a node N. Right rotation at N transforms the left tree to the right one, and left rotation at N is the inverse operation which transforms the right tree to the left one. A, B, and C represent leaves or subtrees, and the node N could be at the root or any other position in the tree.

Any shape tree of size n nodes can be converted to any other tree of the same size via a sequence of rotations, as described in [Culik and Wood 1982]. The minimum length of possible sequences of rotations converting a tree S with n nodes to a tree T with n nodes is the *rotation distance* between S and T. Though there are some properties of rotation distance that are well understood, there is no known effective algorithm for computing rotation distance. Sleator, Tarjan and Thurston [STT 1988] showed that the distance is never more than $2n-6$, and furthermore that for very large n that bound is achieved.

Here, we investigate some measures of tree similarity which are related to rotation distance. When there are common edges, described below, this reduces rotation distance and allows breaking of the problem into smaller parts. When there are one-off edges, described below, there is an immediate essential possible first move which then results in a common edge, again allowing reduction into parts. Another measure of tree similarity is the count of the number of conflicting edge pairs, described below. For each of these quantities, we investigate with a large number of computational experiments how quickly these quantities grow with tree size. In

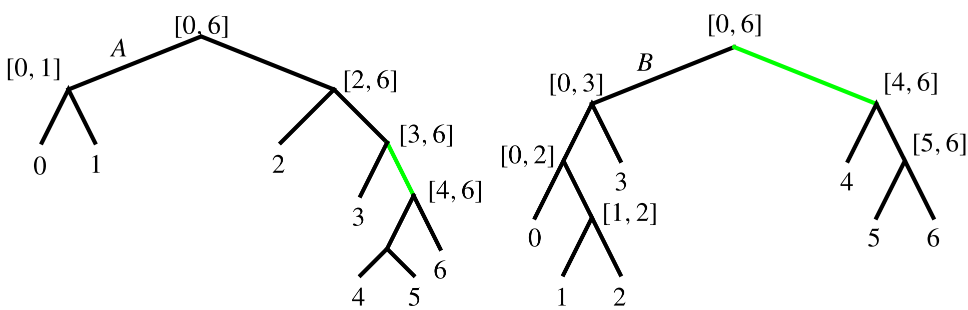

Figure 2. A common edge between a pair of trees. The green edge is common to both trees and separates leaves 4–6 from the other leaves in both trees. The root node interval is always $[0, n]$, and we do not consider that to represent a common edge.

the case of common edges and one-off edges, the growth is linear with tree size and in a manner consistent with the asymptotic behavior understood combinatorially. In the case of conflicting edges, we see growth which appears to lie between linear and quadratic.

By *binary tree of size n*, we mean a rooted binary tree with n leaves arranged in a left-to-right order, with leaves numbered from 0 to $n-1$. To each edge, we associate an interval $[i, j]$, where i is the leftmost leaf in the subtree attached to that edge's lower side, and similarly j is the rightmost such leaf. A pair of trees (S, T) has a *common edge* if an edge $[i, j]$ is present in both trees, as illustrated in Figure 2. An edge $[i, j]$ in S is a *one-off edge* with respect to T if it is itself not a common edge of S with respect to T, but there is a single rotation in S which changes $[i, j]$ to a new edge which is now in common with T.

An edge $[i, j]$ is *in conflict* with an edge $[l, m]$ if it is not possible for both edges to exist simultaneously in the same tree. We can readily detect edge conflicts by noting that each edge partitions the set of leaves into two sets, obtained by considering connected components of the forest obtained by deleting that edge. If the partitions of leaves are incompatible, the edges had a conflict. For example, in a tree with six leaves, an edge with interval label [2, 5] conflicts with an edge [0, 3]. The edge [2, 5] partitions the leaves into two sets: {0, 1} and {2, 3, 4, 5} and the

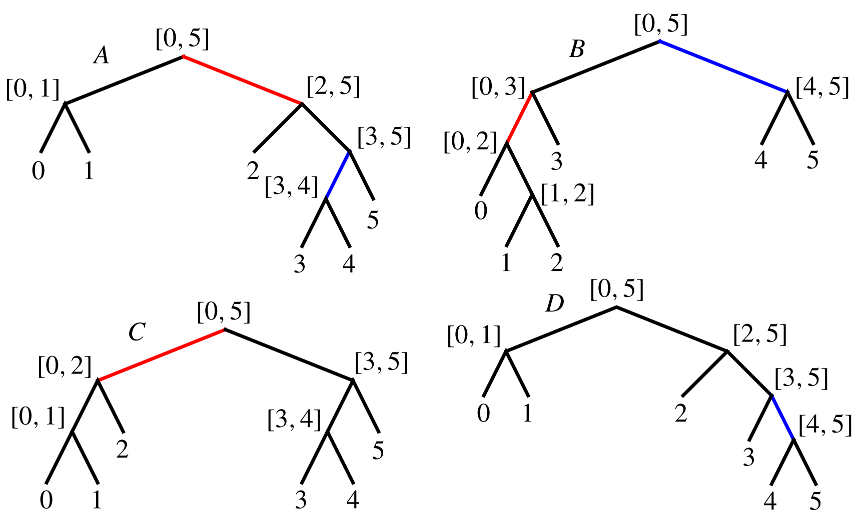

Figure 3. The tree pair (A, B) has two one-off edges in A, marked in red and blue. The tree pair (A, B) has no edges in common, but left rotation at the root of A gives tree C, which has the red edge [0, 2] in common with B. Similarly, right rotation at the node marked [3, 5] in A gives tree D which has the blue edge [4, 5] in common with B.

edge [0, 3] partitions the leaves into two sets: {0, 1, 2, 3} and {4, 5}. There is no overall partition compatible with both such partitions, so that edge pair is in conflict, as it would not be possible for those two edges to be present in the same tree simultaneously. An example illustrating this particular edge pair in conflict is given as trees A and B in Figure 3. Using the bijection between trees and triangulations of regular polygons, described in [STT 1988], each conflict can be counted as an intersection between edges of superimposed triangulations.

Since the number of trees of size n is the n-th Catalan number C_n, and C_n grows exponentially on the order of $4^n n^{-3/2}$, the number of pairs of trees of size n grows on the order of $16^n n^{-3}$. Thus, computing these quantities exhaustively is not possible except for very small n. Instead, we use sampling techniques, experimenting computationally by repeatedly choosing pairs of trees of size n uniformly at random and computing and tabulating the results.

2. Conflicts and one-off edges

As described in [Cleary and St. John 2010], common edges permit the subdivision of the rotation distance problem into smaller pieces. From [STT 1988], one-off edges can be moved immediately to find a geodesic, and the resulting common edge will then subdivide the problem as well.

The existence of common edges of a particular peripheral type was investigated by Cleary, Elder, Rechnitzer and Taback, who showed in [CERT 2010], in connection with using tree-pair diagrams to represent elements of Thompson's group F, that a randomly selected tree pair has at least one common peripheral edge. The number of such common edges with respect to trees generated randomly by the Yule process was investigated experimentally by Cleary, Passaro and Toruno [CPT 2013].

To understand the typical behavior of common edges, one-off edges, and conflicts between tree pairs, we performed a range of computational experiments to investigate. In each case, we generated tree pairs of a particular size randomly using Rémy's bijection [Rémy 1985], which allows efficient generation of trees of size n uniformly at random through an iterative process. After generating two trees randomly, we collected the relevant information about common edges, one-off edges, and conflicts and then iterated to collect large-sample data. As anticipated, the various measures of complexity grew with tree size. We present summaries of those experiments below.

We considered approximately 10 million tree pairs total of sizes ranging from 10 to 12,000. The bulk of the computational effort lay for trees of size 20 to 800. These results are presented in Tables 1–3.

The number of common edges grows linearly with size as shown in Figure 4, and the line of best fit for the data is an excellent match with the asymptotic exact

Tree size range	Average common edge fraction	σ edge fraction
≤ 40	0.1361	0.09462
41–80	0.1052	0.04494
81–120	0.1004	0.03422
121–200	0.09720	0.02613
201–400	0.09534	0.01945
401–1000	0.09409	0.01358
1001–12000	0.09310	0.004324

Table 1. Fractions of tree common edges and their standard deviations. The asymptotic fraction is known to be about 0.092958.

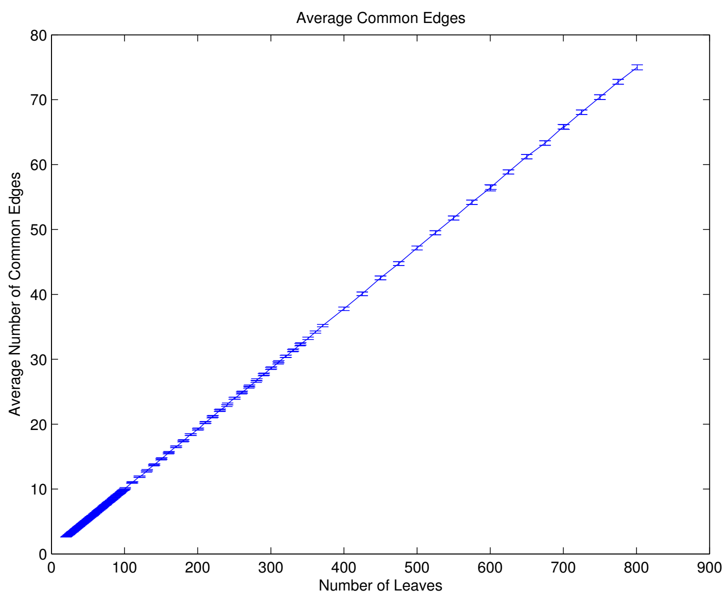

Figure 4. The average number of common edges grows linearly with tree size, with tight error bars from the large sample sizes used over this range. The slope of this line is very close to the expected 0.093 from the asymptotic analysis.

growth proven by Cleary, Rechnitzer and Wong in [CRW 2013].

The number of one-off edges grows linearly with size as shown in Figure 5, and the line of best fit for the data shows very close agreement with the number of common edges. This experimental, numerical observation led to renewed efforts using asymptotic combinatorial methods, and that equivalence is now proven asymptotically in [CRW 2013] by a delicate analysis of some of the relevant generating functions. We note that though the means are asymptotically the same,

Tree size range	Average one-off fraction	σ one-off fraction
≤ 40	0.1362	0.04890
41–80	0.1053	0.02591
81–120	0.1004	0.01986
121–200	0.09725	0.01516
201–400	0.09536	0.01129
401–1000	0.09408	0.007932
1001–12000	0.09307	0.002619

Table 2. Fractions of tree one-off edges and their standard deviations. The asymptotic fraction is known to be about 0.092958.

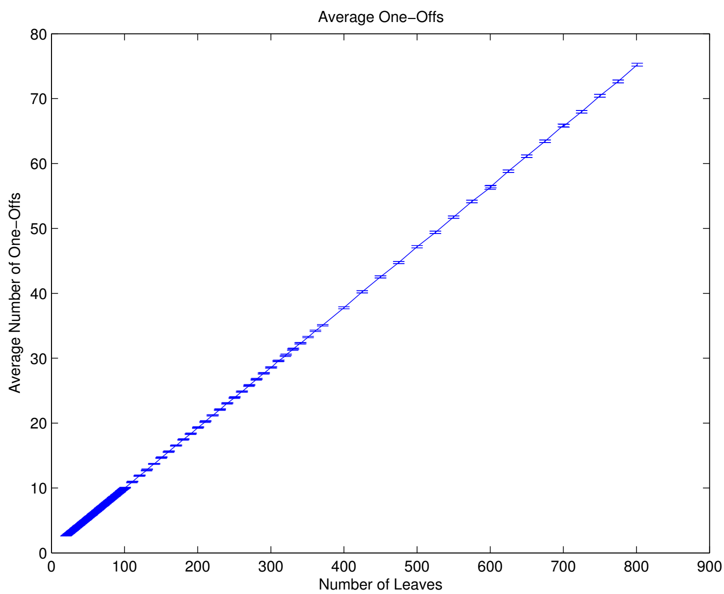

Figure 5. The average number of one-off edges grows linearly with tree size, again with tight error bars from the large sample sizes used over this range and this quantity is very close to the average number of common edges of similarly sized tree pairs. The slope of this line is very close to the expected 0.093 from the asymptotic analysis.

the distributions appear to be significantly different. The standard deviation for the number of one-offs generally has a standard deviation of a little more than half that of the number of common edges. The combinatorial analysis of [CRW 2013] only applies to the means — it appears that the distributions are genuinely different.

Tree size range	Average conflicts per edge	σ conflicts per edge
≤ 40	3.2418	0.9985
41–80	5.837	1.320
*81–120	6.968	1.409
121–200	8.321	1.533
201–400	9.768	1.628
*401–1000	11.71	1.792
*1001–12000	17.45	1.887

Table 3. Average number of conflicts per edge and their standard deviations.

The asymptotic analysis of [CRW 2013] shows that for large n, the expected number of common edges is

$$\frac{16-5\pi}{\pi}n+\frac{7\pi-20}{\pi}+O\left(\frac{\log n}{n}\right),$$

which is approximately

$$0.092958n+0.633802+O\left(\frac{\log n}{n}\right).$$

For the average number of common edges in a tree of size n, this experimental data yields a best linear fit of $0.092950n+0.643$, and similarly for the average number of one-off edges, the experimental data yields a best linear fit of $0.092867n+0.711$.

We now turn to the number of conflicts between randomly selected pairs of trees, shown in Figure 6. It is apparent from this data that the typical number of conflicts per edge grows quite slowly. Even in trees of size multiple thousands, where in theory an edge could cross hundreds of other edges, typically the mean number of conflicts per edge is quite small, for example about 17. This illustrates that a tree of size n selected uniformly at random tends to be rather "stringy" rather than balanced (see [Knuth 1969]), and a pair of such stringy trees is not likely to have edges that conflict with large swaths of the other trees. Though it is possible to construct tree pairs of increasing size whose number of conflicts grows quadratically, these constructions do not represent typical behavior of randomly selected tree pairs.

This data shows that the average number of conflicts grows more than linearly with n, but subquadratically. The maximum possible number of conflicts is bounded above by a quadratic function — each edge could conflict with at most $n-1$ other edges, giving a crude upper bound of n^2-n. Since each unmatched edge gives at least one conflict, the average number of conflicts thus is growing somewhere between linearly and quadratically. It is difficult to ascertain the exact growth of the mean number of conflicts. Using the data set over the range from 5 to 12,000, a log-log analysis, as shown in Figure 7, suggests that the power law of best fit over

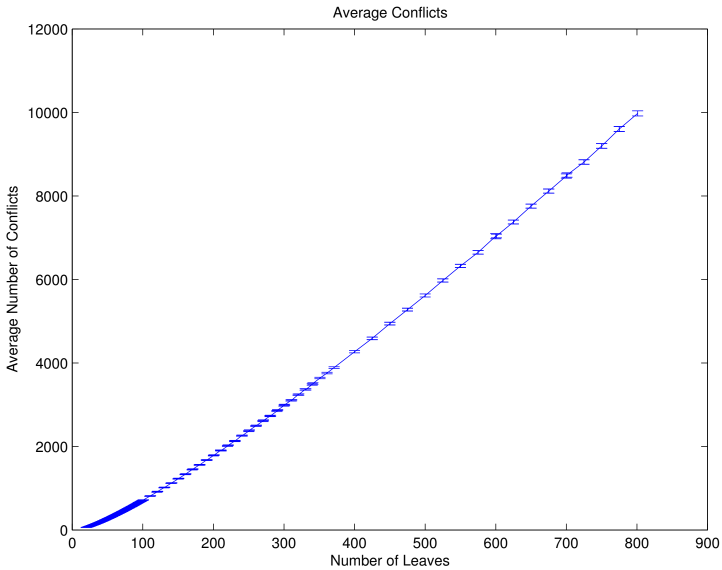

Figure 6. The number of conflicts grows with tree size in a manner which appears to be between linear and quadratic, with a slight upward concavity apparent over this range.

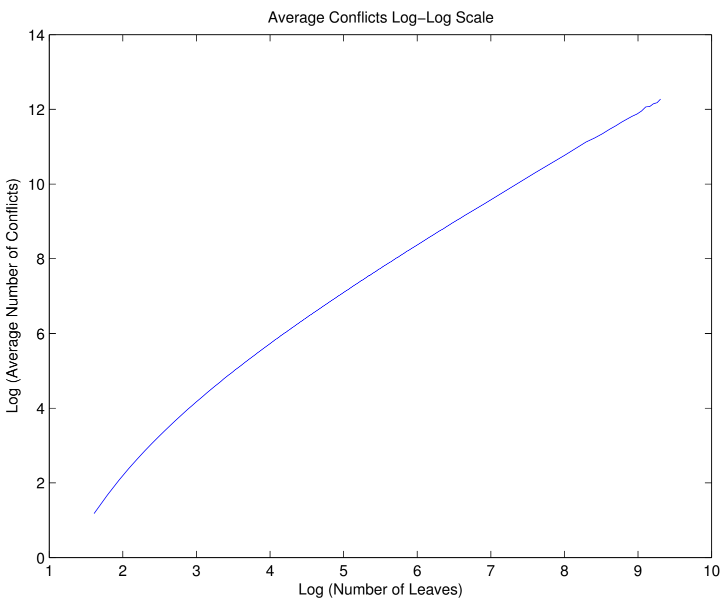

Figure 7. Log-log plot of the average number of conflicts against size. This relationship appears to be slightly concave downward, indicating that the asymptotic behavior is probably a lower-degree power law than the straight-line fit analysis would indicate.

the pictured range is 1.3173. There may also be logarithmic terms present in the growth which are difficult to detect experimentally.

3. Discussion

For common edges and one-off edges, these experiments worked well to establish the behavior for growing tree size. These computational experiments confirm close agreement to the asymptotic behavior of common edges, with relatively quick convergence to the asymptotic limit. Furthermore, if we ignore the smaller tree pairs of size 16 and less, the agreement is even stronger with the asymptotic behavior. These experiments also suggested that the average number of one-off edges was the same as common edges, which in other work has now been proven asymptotically to be the case. Again we saw close agreement with the asymptotic limit and relatively quick convergence.

For conflicts between tree pairs, these experiments gave some indication of the order of growth, with some insight coming from the relatively low overall average conflicts between randomly selected tree pairs. But there was not conclusive enough behavior to establish a likely asymptotic estimate for the growth of the number of conflicts as tree size increases. The possibility of logarithmic terms in the asymptotic terms suggests that it may be difficult to detect the asymptotic behavior more precisely with experimental methods. With fixed computational resources and computation time, investigating larger size trees gives a significant increase in the resulting error bars. Nevertheless, the power-law behavior is apparent from the analysis and we expect the possible logarithmic correction terms, for example, to not be dramatic.

References

[CERT 2010] S. Cleary, M. Elder, A. Rechnitzer, and J. Taback, "Random subgroups of Thompson's group *F*", *Groups Geom. Dyn.* **4**:1 (2010), 91–126. MR 2011e:20062 Zbl 1226.20034

[Cleary and St. John 2010] S. Cleary and K. St. John, "A linear-time approximation for rotation distance", *J. Graph Algorithms Appl.* **14**:2 (2010), 385–390. MR 2011m:68205 Zbl 1215.68271

[CPT 2013] S. Cleary, J. Passaro, and Y. Toruno, "Average reductions in Yule-generated binary trees", *Involve* **6**:4 (2013).

[CRW 2013] S. Cleary, A. Rechnitzer, and T. Wong, "Common edges in rooted trees and polygonal triangulations", *Electron. J. Combin.* **20**:1 (2013), Paper 39. MR 3035049

[Culik and Wood 1982] K. Culik, II and D. Wood, "A note on some tree similarity measures", *Inform. Process. Lett.* **15**:1 (1982), 39–42. MR 83k:68059 Zbl 0489.68058

[Knuth 1969] D. E. Knuth, *The art of computer programming, I: Fundamental algorithms*, Second printing, Addison-Wesley, Reading, Massachusetts, 1969. MR 44 #3530 Zbl 0895.68055

[Knuth 1973] D. E. Knuth, *The art of computer programming, III: Sorting and searching*, Addison-Wesley, Reading, Massachusetts, 1973. MR 56 #4281

[Rémy 1985] J.-L. Rémy, "Un procédé itératif de dénombrement d'arbres binaires et son application à leur génération aléatoire", *RAIRO Inform. Théor.* **19**:2 (1985), 179–195. MR 87h:68132 Zbl 0565.05037

[STT 1988] D. D. Sleator, R. E. Tarjan, and W. P. Thurston, "Rotation distance, triangulations, and hyperbolic geometry", *J. Amer. Math. Soc.* **1**:3 (1988), 647–681. MR 90h:68026 Zbl 0653.51017

Received: 2012-06-26 Revised: 2012-10-22 Accepted: 2012-10-23

timchu100@gmail.com *Department of Computer Science, The City College of New York, City University of New York, 160 Convent Avenue, New York, NY 10031, United States*

cleary@sci.ccny.cuny.edu *Department of Mathematics, The City College of New York, City University of New York, 160 Convent Avenue, New York, NY 10031, United States*
http://www.sci.ccny.cuny.edu/~cleary

mathematical sciences publishers msp

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2013 is US $105/year for the electronic version, and $145/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

involve

2013 vol. 6 no. 3

1944-4176(2013)6:3;1-8