

involve

a journal of mathematics

Average reductions between random tree pairs

Sean Cleary, John Passaro and Yasser Toruno



Average reductions between random tree pairs

Sean Cleary, John Passaro and Yasser Toruno

(Communicated by Robert W. Robinson)

There are a number of measures of degrees of similarity between rooted binary trees. Many of these ignore sections of the trees which are in complete agreement. We use computational experiments to investigate the statistical characteristics of such a measure of tree similarity for ordered, rooted, binary trees. We generate the trees used in the experiments iteratively, using the Yule process modeled upon speciation.

1. Introduction

Rooted binary trees arise in a wide range of settings, from biological evolutionary trees to efficient structures for searching datasets. There are a number of measures of tree similarity which arise in these settings. Here we investigate a measure which is relevant for ordered, rooted, binary trees of the same size. Examples of trees satisfying such conditions include some binary search trees. Our approach is to consider pairs of such trees of increasing size n , selected via a random process, and investigate the degree of commonality given by a natural measure of the degree to which they agree completely on peripheral subtrees. Using experimental evidence, we find that the degree of commonality appears to grow linearly with tree size, and we estimate the average behavior.

There are a number of processes for selecting trees randomly. One method that is commonly studied is the uniform distribution on trees, where each tree is equally likely to be selected. Some properties of the reduction behavior of trees selected uniformly at random have been investigated by Cleary, Elder, Rechnitzer and Taback [Cleary et al. 2010] while studying statistical properties of Thompson's group F , showing that a tree pair selected from the uniform distribution on tree pairs is almost surely unreduced in the sense described below. The common subtrees investigated here via reduction are a particular case of common edges, where in the

MSC2010: 05C05, 68P05.

Keywords: random binary tree pairs.

Partial funding provided by NSF grants 0811002 and 1417820. Sean Cleary was partially supported by grant 234548 from the Simons Foundation.

common edge case the collections of common edges need not be peripheral. That is, in the more general case they need not include the complete subtree, extending to the leaves. For common edges of all types, the average number of common edges with respect to the uniform selection of trees at random case has been examined experimentally by Chu and Cleary [2013] and asymptotically by Cleary, Rechnitzer and Wong [Cleary et al. 2013]. Asymptotically, the expected number of reductions of a tree pair selected uniformly at random is

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

for reductions of a more general type, which is about

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

The experimental results in [Chu and Cleary 2013] show quick convergence to the dominant linear term of $0.092958n$. For the particular subtree peripheral reductions (that is, subtree reductions) considered here, a similar generating function analysis gives the asymptotic number of trees as $(7 - 4\sqrt{3})n$, which is about $0.0717968n$ when tree pairs are selected uniformly at random. So on average more than three quarters of the expected common edges lie in expected common peripheral subtrees.

Here, instead of considering trees selected uniformly at random, we study a process for generating trees at random motivated by biological questions, called the Yule process [Yule 1925; Harding 1971], also known as uniform speciation. A tree is grown iteratively from the root. At each step, a leaf is selected uniformly at random from the leaves present at that stage, and a new sibling pair is attached at that leaf, and then the process is iterated until we have a tree with the appropriate number of leaves. Such a distribution of trees also can arise from a variety of insertion scenarios in tree-structured data.

The distribution of the number of sibling pairs (“cherries”) of unordered trees was investigated by McKenzie and Steel [2000] for both the uniform and Yule tree distributions — asymptotically, there are $n/3$ expected sibling pairs for the Yule distribution and $n/4$ for the uniform distribution. Here we find experimentally that the expected number of subtree reductions is also larger for the Yule distribution than the uniform distribution, with almost 13% expected subtree reduction compared to the expected reduction of about 7% in the uniform case.

2. Background and definitions

We consider rooted binary trees on n leaves with a natural left-to-right order on leaves, numbered from 1 to n . The internal nodes of the trees we refer to as nodes and the external nodes we refer to as leaves. Two children of the same node which

are leaves form a sibling pair and their leaf numbers are necessarily of the form i and $i + 1$ for some i .

A tree pair (S, T) is *reduced* if there are no sibling pairs with leaves numbered i and $i + 1$ in S which have a corresponding sibling pair i and $i + 1$ as leaves in T . An *elementary reduction* for a tree pair (S, T) with n leaves with a common sibling pair $(i, i + 1)$ is a tree pair (S', T') with $n - 1$ leaves, where the common sibling pair has been removed in both S and T and the leaves have been appropriately renumbered. A *reduction* of a tree pair diagram is a sequence of elementary reductions. There may be many possible elementary reductions for an unreduced tree pair and thus many possible reductions, but for a given tree pair (S, T) , there is a unique reduced tree pair (S'', T'') which is itself a reduction of (S, T) and which has the property that any possible sequence of reductions from (S, T) will terminate in that reduced tree pair. An example of tree pair reduction is given in Figure 1.

The subtrees that are eliminated during the reduction process for a tree pair (S, T) are portions of the tree in which S and T agree completely. There are a number of metrics on spaces of trees of interest, coming from biological questions, database efficiency questions and more abstract approaches. For all of the standard metrics on spaces of trees with an order on the leaves, the parts of the trees which are in complete agreement do not contribute to the distance. That is, if a tree pair (S, T) reduces to a tree pair (S', T') , the distance of interest between S and T is the same as the distance between S' and T' . The fact that the trees S' and T' may be considerably smaller is of good use, particularly for distances which are

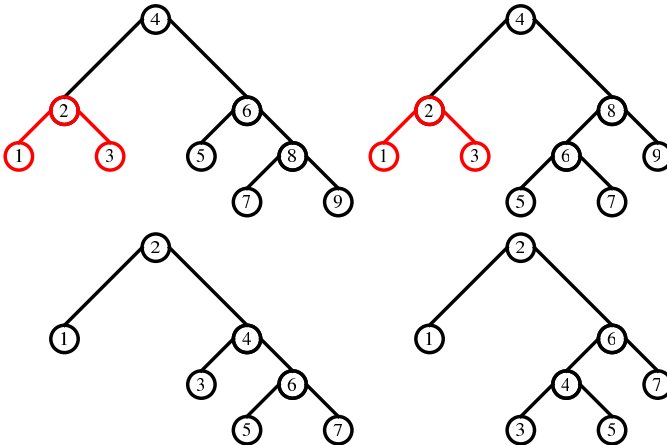


Figure 1. An unreduced tree pair and its reduction to a reduced tree pair. The top unreduced tree pair has a common subtree containing the sibling pair of nodes 1 and 3 in both trees, shown in red, which is then removed and the nodes renumbered, resulting in the lower tree pair which is reduced.

difficult to compute. Given that the best known algorithm for rotation distance is of exponential running time, and that many tree metrics of biological interest are proven to be of class NP, even a marginal reduction in the sizes of trees under consideration is worthwhile. This analysis is an effort to understand the degree to which such reductions typically reduce the size of tree pairs.

We generate trees using the Yule or *speciation* method as follows. We begin with a single node with two leaves, and then randomly select from the leaves and replace that leaf with a node with its own two leaves, renumbering the leaves as needed. We then choose randomly from the three current leaves, replacing that chosen leaf with a node and two leaves, and continue enlarging the tree in this process until it is the desired size.

As shown in Figure 2, there may be more than one way to generate a given tree using the Yule process. The process is generally more likely to generate balanced trees than stringy ones, so the distribution on trees is different than that for the uniform random selection of trees, as described in [Harding 1971]. This is also related to the difference in expected number of sibling pairs described in [McKenzie and Steel 2000].

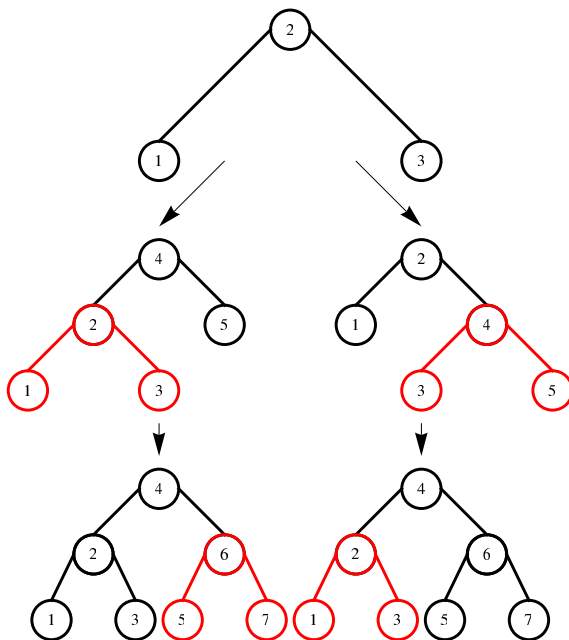


Figure 2. Some trees can be generated in several ways via the Yule process, such as this balanced tree with four leaves which can be generated in two ways. Every other tree with four leaves can be generated in just one way, resulting in a nonuniform distribution of random tree selection.

3. Experiments and conclusions

We constructed programs in C to create tree pairs of a specified size and count the reductions, iterating to obtain average values. Tree pairs with trees ranging from size 100 to 29,000 were generated and the total size of common subtrees was calculated and recorded for each pair generated, with the results summarized in Table 1. Generally, there were around 1000 tree pairs of each size generated and analyzed, sufficient to give small error bars in the analysis. The average reductions grew linearly, with about 12.8% average reduction in size, significantly more than the corresponding value of about 7.1% in the corresponding case for trees generated uniformly at random. As indicated in Figures 3 and 4, the relationship appears to be

| Tree size range | Average total subtree reduction | σ subtree reduction |
|-----------------|---------------------------------|----------------------------|
| 100– 2000 | 0.12846 | 0.013829 |
| 2001– 8000 | 0.12781 | 0.006034 |
| 8001–15000 | 0.12775 | 0.003462 |
| 15001–29000 | 0.12773 | 0.002402 |

Table 1. Average total size of common subtrees and corresponding sample standard deviations.

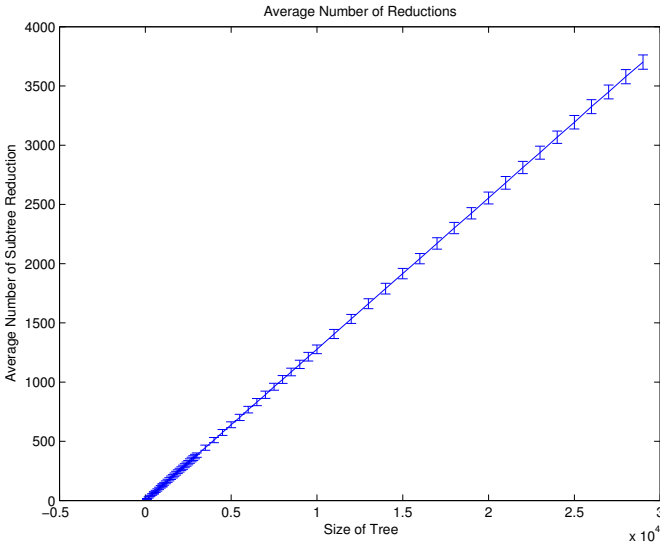


Figure 3. The average number of reductions grows linearly with tree size, with tight error bars from the sample sizes used over this range. The slope of the line of best fit is about 0.127. Error bars indicate 3 standard deviations from the sample averages.

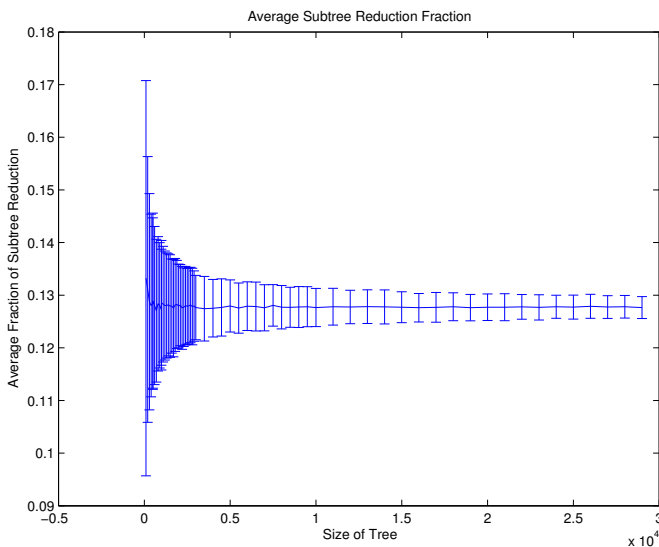


Figure 4. The average fraction of the tree pairs which are eliminated in the reduction process is close to 0.127, over the range shown. Error bars indicate 3 standard deviations from the sample averages.

linear, and a linear regression to the data gives an excellent fit with r^2 value of within one-millionth of 1. The line of best fit for the experimental data is $0.1277n + 0.268$.

What we find is that the fraction of the trees which reduce appears larger for the Yule distribution than for the uniform distribution.

References

- [Chu and Cleary 2013] T. Chu and S. Cleary, “Expected conflicts in pairs of rooted binary trees”, *Involve* **6**:3 (2013), 323–332. [MR 3101764](#) [Zbl 1274.05066](#)
- [Cleary et al. 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 et al. 2013] S. Cleary, A. Rechnitzer, and T. Wong, “Common edges in rooted trees and polygonal triangulations”, *Electron. J. Combin.* **20**:1 (2013), Paper 39, 22. [MR 3035049](#) [Zbl 1267.05249](#)
- [Harding 1971] E. F. Harding, “The probabilities of rooted tree-shapes generated by random bifurcation”, *Advances in Appl. Probability* **3** (1971), 44–77. [MR 43 #8162](#) [Zbl 0241.92012](#)
- [McKenzie and Steel 2000] A. McKenzie and M. Steel, “Distributions of cherries for two models of trees”, *Math. Biosci.* **164**:1 (2000), 81–92. [MR 2001e:92010](#) [Zbl 0947.92021](#)
- [Yule 1925] G. Yule, “A mathematical theory of evolution, based upon the conclusions of Dr. J. C. Willis, F.R.S.”, *Royal Society of London Philosophical Transactions, Series B* **213** (1925), 21–87.

cleary@sci.ccny.cuny.edu

*Department of Mathematics,
The City College of New York and the CUNY Graduate Center,
City University of New York, NAC R8133,
160 Convent Avenue, New York, NY 10031, United States*

john.a.passaro@gmail.com

*Department of Mathematics, The City College of New York,
City University of New York, New York, NY 10031,
United States*

yltoruno@gmail.com

*Department of Computer Science,
The City College of New York, City University of New York,
New York, NY 10031, United States*

EDITORS

MANAGING EDITOR

Kenneth S. Berenhaut, Wake Forest University, USA, berenhks@wfu.edu

BOARD OF EDITORS

| | | | |
|----------------------|---|------------------------|--|
| Colin Adams | Williams College, USA colin.c.adams@williams.edu | David Larson | Texas A&M University, USA larson@math.tamu.edu |
| John V. Baxley | Wake Forest University, NC, USA baxley@wfu.edu | Suzanne Lenhart | University of Tennessee, USA lenhart@math.utk.edu |
| Arthur T. Benjamin | Harvey Mudd College, USA benjamin@hmc.edu | Chi-Kwong Li | College of William and Mary, USA ckli@math.wm.edu |
| Martin Bohner | Missouri U of Science and Technology, USA bohner@mst.edu | Robert B. Lund | Clemson University, USA lund@clemson.edu |
| Nigel Boston | University of Wisconsin, USA boston@math.wisc.edu | Gaven J. Martin | Massey University, New Zealand g.j.martin@massey.ac.nz |
| Amarjit S. Budhiraja | U of North Carolina, Chapel Hill, USA budhiraj@email.unc.edu | Mary Meyer | Colorado State University, USA meyer@stat.colostate.edu |
| Pietro Cerone | La Trobe University, Australia P.Cerone@latrobe.edu.au | Emil Minchev | Ruse, Bulgaria eminchev@hotmail.com |
| Scott Chapman | Sam Houston State University, USA scott.chapman@shsu.edu | Frank Morgan | Williams College, USA frank.morgan@williams.edu |
| Joshua N. Cooper | University of South Carolina, USA cooper@math.sc.edu | Mohammad Sal Moslehian | Ferdowsi University of Mashhad, Iran moslehian@ferdowsi.um.ac.ir |
| Jem N. Corcoran | University of Colorado, USA corcoran@colorado.edu | Zuhair Nashed | University of Central Florida, USA znashed@mail.ucf.edu |
| Toka Diagana | Howard University, USA tdiagana@howard.edu | Ken Ono | Emory University, USA ono@mathcs.emory.edu |
| Michael Dorff | Brigham Young University, USA mdorff@math.byu.edu | Timothy E. O'Brien | Loyola University Chicago, USA tbriell@luc.edu |
| Sever S. Dragomir | Victoria University, Australia sever@matilda.vu.edu.au | Joseph O'Rourke | Smith College, USA orourke@cs.smith.edu |
| Behrouz Emamizadeh | The Petroleum Institute, UAE bemamizadeh@pi.ac.ae | Yuval Peres | Microsoft Research, USA peres@microsoft.com |
| Joel Foisy | SUNY Potsdam foisyjs@potsdam.edu | Y.-F. S. Pétermann | Université de Genève, Switzerland petermann@math.unige.ch |
| Errin W. Fulp | Wake Forest University, USA fulp@wfu.edu | Robert J. Plemmons | Wake Forest University, USA rplemmons@wfu.edu |
| Joseph Gallian | University of Minnesota Duluth, USA kgallian@d.umn.edu | Carl B. Pomerance | Dartmouth College, USA carl.pomerance@dartmouth.edu |
| Stephan R. Garcia | Pomona College, USA stephan.garcia@pomona.edu | Vadim Ponomarenko | San Diego State University, USA vadim@sciences.sdsu.edu |
| Anant Godbole | East Tennessee State University, USA godbole@etsu.edu | Bjorn Poonen | UC Berkeley, USA poonen@math.berkeley.edu |
| Ron Gould | Emory University, USA rg@mathcs.emory.edu | James Propp | U Mass Lowell, USA jpropp@cs.uml.edu |
| Andrew Granville | Université Montréal, Canada andrew.andrew@dms.umontreal.ca | József H. Przytycki | George Washington University, USA przytyck@gwu.edu |
| Jerrold Griggs | University of South Carolina, USA griggs@math.sc.edu | Richard Rebarber | University of Nebraska, USA rrebarbe@math.unl.edu |
| Sat Gupta | U of North Carolina, Greensboro, USA sgupta@uncg.edu | Robert W. Robinson | University of Georgia, USA rwr@cs.uga.edu |
| Jim Haglund | University of Pennsylvania, USA jhaglund@math.upenn.edu | Filip Saidak | U of North Carolina, Greensboro, USA f_saidak@uncg.edu |
| Johnny Henderson | Baylor University, USA johnny_henderson@baylor.edu | James A. Sellers | Penn State University, USA sellersj@math.psu.edu |
| Jim Hoste | Pitzer College jhoste@pitzer.edu | Andrew J. Sterge | Honorary Editor andy@ajsterge.com |
| Natalia Hritonenko | Prairie View A&M University, USA nahritonenko@pvamu.edu | Ann Trenk | Wellesley College, USA atrenk@wellesley.edu |
| Glenn H. Hurlbert | Arizona State University, USA hurlbert@asu.edu | Ravi Vakil | Stanford University, USA vakil@math.stanford.edu |
| Charles R. Johnson | College of William and Mary, USA crjohnso@math.wm.edu | Antonia Vecchio | Consiglio Nazionale delle Ricerche, Italy antonia.vecchio@cnr.it |
| K. B. Kulasekera | Clemson University, USA kk@ces.clemson.edu | Ram U. Verma | University of Toledo, USA verma99@msn.com |
| Gerry Ladas | University of Rhode Island, USA gladas@math.uri.edu | John C. Wierman | Johns Hopkins University, USA wierman@jhu.edu |
| | | Michael E. Zieve | University of Michigan, USA zieve@umich.edu |

PRODUCTION


Silvio Levy, Scientific Editor

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2015 is US \$140/year for the electronic version, and \$190/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/>

© 2015 Mathematical Sciences Publishers

involve

2015

vol. 8

no. 1

| | |
|---|-----|
| Efficient realization of nonzero spectra by polynomial matrices | 1 |
| NATHAN MCNEW AND NICHOLAS ORMES | |
| The number of convex topologies on a finite totally ordered set | 25 |
| TYLER CLARK AND TOM RICHMOND | |
| Nonultrametric triangles in diametral additive metric spaces | 33 |
| TIMOTHY FAVER, KATELYNN KOCHALSKI, MATHAV KISHORE MURUGAN, HEIDI VERHEGGEN, ELIZABETH WESSON AND ANTHONY WESTON | |
| An elementary approach to characterizing Sheffer A-type 0 orthogonal polynomial sequences | 39 |
| DANIEL J. GALIFFA AND TANYA N. RISTON | |
| Average reductions between random tree pairs | 63 |
| SEAN CLEARY, JOHN PASSARO AND YASSER TORUNO | |
| Growth functions of finitely generated algebras | 71 |
| ERIC FREDETTE, DAN KUBALA, ERIC NELSON, KELSEY WELLS AND HAROLD W. ELLINGSEN, JR. | |
| A note on triangulations of sumsets | 75 |
| KÁROLY J. BÖRÖCZKY AND BENJAMIN HOFFMAN | |
| An exploration of ideal-divisor graphs | 87 |
| MICHAEL AXTELL, JOE STICKLES, LANE BLOOME, ROB DONOVAN, PAUL MILNER, HAILEE PECK, ABIGAIL RICHARD AND TRISTAN WILLIAMS | |
| The failed zero forcing number of a graph | 99 |
| KATHERINE FETCIE, BONNIE JACOB AND DANIEL SAAVEDRA | |
| An Erdős–Ko–Rado theorem for subset partitions | 119 |
| ADAM DYCK AND KAREN MEAGHER | |
| Nonreal zero decreasing operators related to orthogonal polynomials | 129 |
| ANDRE BUNTON, NICOLE JACOBS, SAMANTHA JENKINS, CHARLES MCKENRY JR., ANDRZEJ PIOTROWSKI AND LOUIS SCOTT | |
| Path cover number, maximum nullity, and zero forcing number of oriented graphs and other simple digraphs | 147 |
| ADAM BERLINER, CORA BROWN, JOSHUA CARLSON, NATHANAEL COX, LESLIE HOGBEN, JASON HU, KATRINA JACOBS, KATHRYN MANTERNACH, TRAVIS PETERS, NATHAN WARNBERG AND MICHAEL YOUNG | |
| Braid computations for the crossing number of Klein links | 169 |
| MICHAEL BUSH, DANIELLE SHEPHERD, JOSEPH SMITH, SARAH SMITH-POLDERMAN, JENNIFER BOWEN AND JOHN RAMSAY | |