Generalized Binary Search Trees and Clock Trees Revisited

Gowtham Bellala
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109
E-mail: gowtham@umich.edu

1 Introduction

In this paper, we study two diverse problems from a random matrix perspective. The first one is the problem of binary testing (or object/entity identification) that arises in applications such as active learning, fault diagnosis and computer vision, and the second is the problem of zero or bounded skew clock tree construction which arises in applications such as VLSI circuit design and network multicasting. Though both these problems involve construction of binary trees, the objectives and the greedy algorithms used for binary tree construction are very different.

In the problem of binary testing, the goal is to identify an unknown object while minimizing the number of binary questions posed about that object. A binary decision tree is a solution to this problem, where often the goal is to minimize the average depth of the binary tree. *Generalized binary search* (GBS) is a greedy algorithm that is popularly used in the literature to construct near optimal binary decision trees. Here, we study the depth distribution of trees constructed using GBS and show that it converges to a strange distribution known as the Airy distribution under certain random matrix models. Refer Section 3 for more details.

Next, we study the zero or bounded skew clock tree problem. The skew of an edge-weighted rooted tree is defined to be the maximum difference between any two root-to-leaf path weights. Zero or bounded-skew trees are needed for achieving synchronization in applications such as network multicasting and VLSI clock routing, where the edge weights correspond to propagation delays. In these applications, the signal generated at the root should be received by multiple recipients located at the leaves (almost) simultaneously. The goal in these problems is to find a zero or bounded-skew tree of minimum total weight, since the weight of the tree corresponds to the amount of resources that must be allocated. Here, we study the skew distribution in clock trees and show that once again, this distribution converges to an Airy distribution as the size of the clock tree increases (refer Section 4).

These observations are both surprising and unexpected. Further, they raise several interesting questions regarding the connection of these problems to that of Catalan trees studied in the literature. Also, Airy distribution has been observed to arise as a limit in several other problems involving binary trees in the literature. Hence, these findings pave way for future investigations into these problems by exploiting their relation to previously studied problems.

2 Background

We begin by providing a brief description of Catalan trees along with Airy distribution and the relation between them.

2.1 Catalan trees

Before we describe Catalan trees, we need to briefly review *Catalan* numbers and *full* binary trees. Catalan numbers are a sequence of natural numbers that occur in various counting problems [1] and can be described

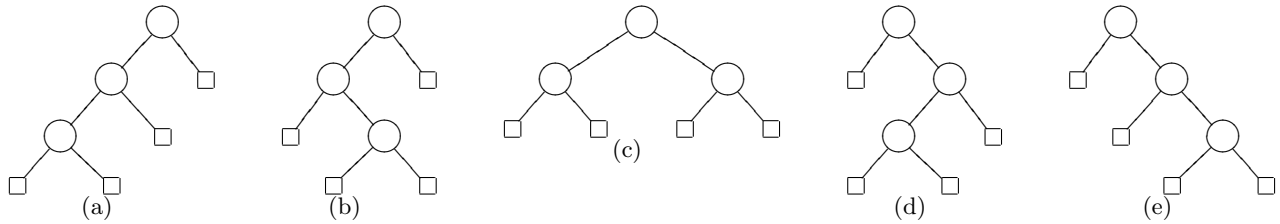


Figure 1: The five different full binary trees with 3 internal nodes

by the recurrence relation

$$C_0 = 1, \quad \text{and} \quad C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i} \text{ for } n \geq 0.$$

Then, the nth Catalan number can be expressed directly in terms of binomial coefficients as

$$C_n = \frac{1}{n+1}\binom{2n}{n}, \text{ for } n \geq 0.$$

A *full* binary tree is a rooted binary tree where every vertex has either two children or no children. The vertices with no children are often referred to as leaves and those with two children are referred to as internal nodes. It turns out that the number of full binary trees with n internal nodes (equivalently, $n+1$ leaves) is given by the nth Catalan number. Figure 1 shows the various full binary trees for $n = 3$.

A Catalan tree with n internal nodes is nothing but a full binary tree chosen uniformly at random from the space of C_n full binary trees. Since the probability of choosing a tree from this space of full binary trees is given by $1/C_n = (n+1)/\binom{2n}{n}$, it is commonly referred to as Catalan trees [2].

2.2 Airy Distribution

The *Airy Distribution* function describes the probability distribution of the *area* under a Brownian excursion over a unit interval. However for combinatorialists and theoretical computer scientists, this Airy distribution (of the "area type") arises in a surprising diversity of contexts like parking allocations, hashing tables, trees, discrete random walks, merge-sorting, etc. The most straightforward description of the Airy distribution is by its moments themselves defined by a simple nonlinear recurrence.

Definition 1. *The Airy distribution (of the "area type") is the probability distribution of a random variable X whose moments are given by*

$$\mathbb{E}(X^r) = \frac{-\Gamma(-\frac{1}{2})}{\Gamma(\frac{3r-1}{2})}\Omega_r, \quad r \geq 1,$$

where the "Airy constants" Ω_r are determined by the quadratic recurrence

$$\Omega_0 = -1, \quad 2\Omega_r = (3r-4)r\Omega_{r-1} + \sum_{j=1}^{r-1}\binom{r}{j}\Omega_j\Omega_{r-j} \quad (r \geq 1).$$

The normalized random variable $Y = X/\sqrt{8}$ is called the "Brownian excursion area" (BEA).

Figure 2 shows the first few values of Ω_r and of the moments $\mathbb{E}[X^r]$, while Figure 3 shows a standard Airy distribution. The right tail of an Airy distribution decays like a Gaussian at a rate $\sim e^{-x^2}$ where as the left tail decays at a much faster rate $\sim x^{-5}e^{-x^2}$. Some of the contexts in which an Airy distribution arises are discussed below.

r	0	1	2	3	4
Ω_r	-1	$\frac{1}{2}$	$\frac{5}{4}$	$\frac{45}{4}$	$\frac{3315}{16}$
$\mathbb{E}[X^r]$	1	$\sqrt{\pi}$	$\frac{10}{3}$	$\frac{15}{4}\sqrt{\pi}$	$\frac{884}{63}$

Figure 2: A Table of the Airy constants Ω_r and of the Airy moments $\mathbb{E}[X^r]$

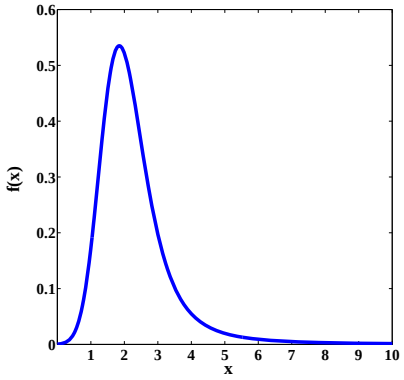


Figure 3: Plot of the Airy distribution function

1. Path length in trees (that is, the sum of distances from the root to all nodes) is asymptotically Airy distributed in Catalan trees as well as in other combinatorial families of trees [3, 4, 5, 6].

2. The total displacement of a random parking sequence or equivalently the construction cost of a hashing table under the linear probing strategy is Airy distributed in the limit [7, 8].

3. Also, the breadth first search traversal of random trees has a cumulated cost that is asymptotically Airy distributed [9].

For a more detailed discussion on the Airy distribution, refer [10].

3 The Binary Testing Problem

Binary Testing (also known as object/entity identification) is the problem of identifying an unknown object while minimizing the number of binary questions posed about that object. It often arises in applications such as active learning [11, 12], disease/fault diagnosis [13], toxic chemical identification [14], computer vision [15] or the adaptive traveling salesperson problem [16].

Following the terminology in [17], these problems can be formulated as containing a set $\Theta = \{\theta_1, \cdots, \theta_M\}$ of M different objects and a set $Q = \{q_1, \cdots, q_N\}$ of N distinct subsets of Θ known as queries. An unknown object θ is generated from this set Θ with a certain *prior* probability distribution $\Pi = (\pi_1, \cdots, \pi_M)$, i.e., $\pi_i = \Pr(\theta = \theta_i)$, and the goal is to uniquely identify this unknown object through as few queries from Q as possible, where a query $q \in Q$ returns a value 1 if $\theta \in q$, and 0 otherwise. An object identification problem can also be denoted using the pair $(\mathbf{B}, \Pi)$ where $\mathbf{B}$ is an $M \times N$ binary matrix with b_{ij} equal to 1 if $\theta_i \in q_j$, and 0 otherwise. Figure 4(b) shows a binary matrix representation of the toy dataset in Figure 4(a).

Given $(\mathbf{B}, \Pi)$, the goal of object identification is to construct an optimal binary decision tree, where each internal node in the tree is associated with a query from Q, and each leaf node corresponds to an object from Θ. Optimality is often with respect to the expected depth of the leaf node corresponding to the unknown object θ. In general the determination of an optimal tree is NP-complete [18]. Hence, various greedy algorithms [19, 20, 21] have been proposed to obtain a suboptimal binary decision tree. A well studied algorithm for this problem is known as the *splitting algorithm* [19] or *generalized binary search* (GBS) [11, 12]. This is the greedy algorithm which selects a query that most evenly divides the probability mass of the remaining objects [11, 12, 19].

3.1 Limiting Depth Distribution of Generalized Binary Search Trees

Let K be a random variable that denotes the number of queries required to identify an unknown object using a tree T. If the unknown object is θ_i, then K corresponds to the number of queries made along the

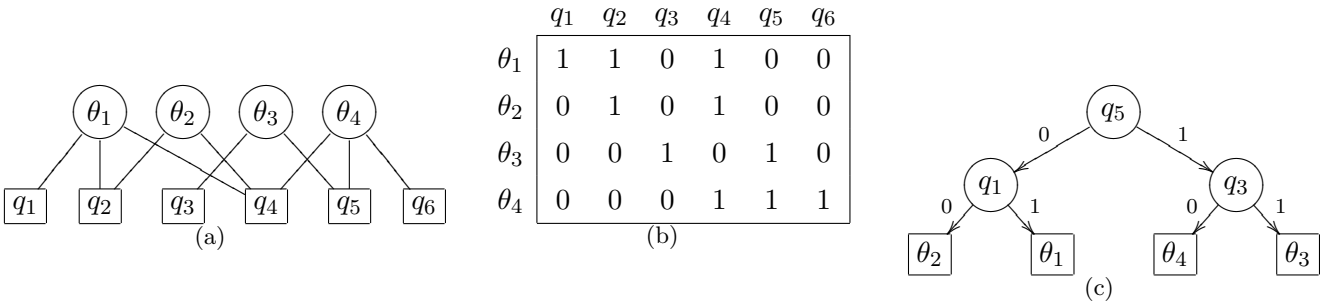


Figure 4: a) A toy object identification problem b) Corresponding binary matrix representation c) A GBS tree constructed on the toy problem under a uniform prior distribution on the objects.

path from the root node to the leaf node terminating in object θ_i, i.e., depth of the leaf node corresponding to object θ_i.

The trees generated using GBS have been extensively studied in terms of the expected depth of their leaf nodes [19, 11], i.e., the expected value of the random variable K. Here, we study the entire depth distribution of the leaf nodes. In particular, we study the limiting distribution of the random variable K in trees generated using GBS as M (the object set size) and N (the query set size) increases. This is the first study involving the entire depth distribution to the best of our knowledge.

We consider datasets ($M \times N$ binary matrices) generated using the standard Erdös-Rényi (ER) random graph model [22, 23]. We note that as M and N increases, the distribution of the random variable K in trees constructed using GBS converges to an Airy distribution. We observed the same phenomenon for different values of p (the probability of an edge) in the ER random graph model as shown in Figure 5.

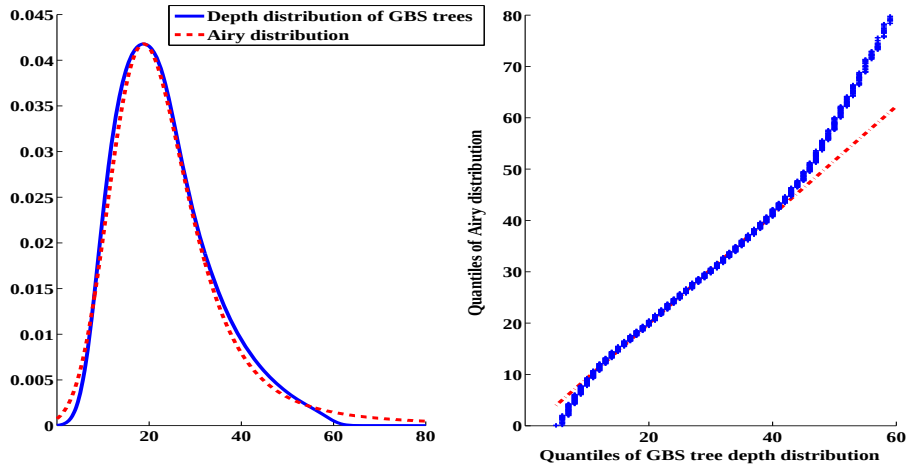
Figure 5 presents two plots for three different values of p in the ER random graph model. In each case, the first plot compares an Airy distribution to that of the leaf depth distribution (distribution of K) in GBS trees constructed on 1000 randomly generated datasets using the ER model. The second plot compares the quantiles of the two distributions. Note that this plot diverges from the straight line on the right end since the leaf depth distribution of GBS trees have a finite support as against the Airy distribution that has an infinite support on the right end.

4 The Zero and Bounded Skew Clock tree Problem

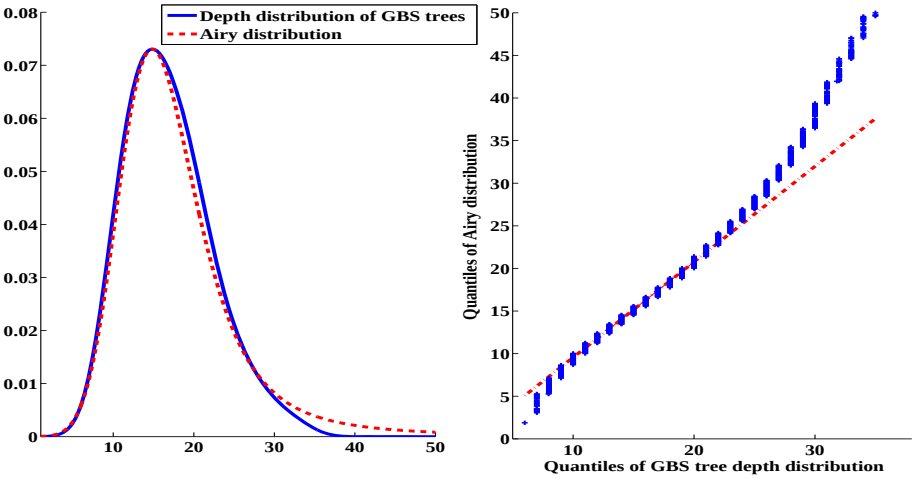
A fundamental problem in VLSI design is *clock routing*, i.e., distributing a clock signal to synchronous elements in a VLSI circuit so that the signal arrives at all elements simultaneously [24, 25]. The signal is distributed by means of a *clock* routing tree (also referred to as a H-tree) rooted at a global clock source. The difference in length between the longest and shortest root-leaf path is called the *skew* of the tree, where the edge lengths translate to propagation delays. To achieve synchronization, the skew should be zero, which is a desired property even in applications such as network multicasting [26] besides VLSI clock routing. Figure 7 shows a typical layout of a clock tree where the weights along the edges may correspond to propagation delays.

Though it is easy to produce zero skew clock routing trees [27], naive algorithms may lead to trees that are expensive in terms of total *wirelength* (i.e., the sum of the edge lengths in the tree). This total wirelength corresponds to circuit area or power for clock routing in VLSI, and bandwidth or buffers for network multicasting. Thus, an ideal clock tree routing algorithm would produce a zero skew clock tree with minimal total wirelength.

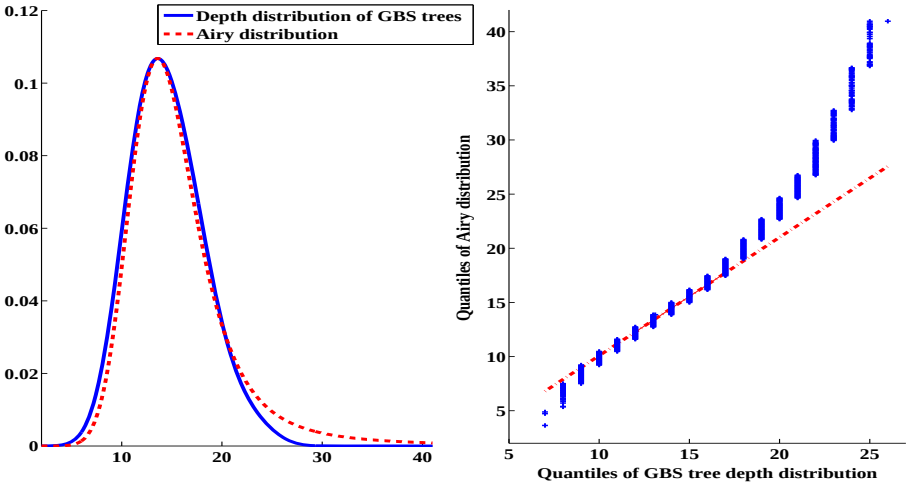
Let M be some metric space and let d denote a distance function defined on this metric space. Let S be a set of points in M that are designated as sinks or terminals. These sinks correspond to leaves of a clock tree. Then, the *cost* of a clock tree T is defined as the sum of the lengths of all edges of T. A clock tree with skew$= 0$, is referred to as a *zero skew clock tree* and one that has a skew at most b is referred to



(a) $p = 0.05$



(b) $p = 0.1$



(c) $p = 0.15$

Figure 5: Limiting distribution of leaf depth in GBS trees under different parameter values in the Erdös-Rényi random graph model.

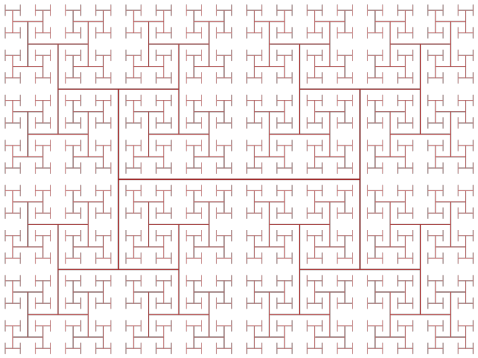


Figure 6: Typical layout of a clock tree.

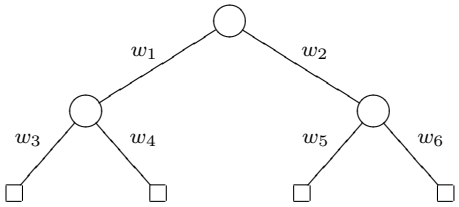


Figure 7: Tree view of the layout where each edge weight corresponds to propagation delay

as a *b-bounded skew clock tree*. Two problems that are often studied in this context are the,

Zero-Skew Tree (ZST) Problem: Given a set of sinks/terminals S in a metric space (M, d), find a minimum cost zero-skew tree for S.

Bounded-Skew Tree (BST) Problem: Given a set of sinks S in a metric space (M, d) and a bound $b > 0$, find a minimum cost b-bounded-skew tree for S.

For general metric spaces, it is known that the ZST and BST problems are NP-complete [28]. Hence several approximation algorithms have been proposed in literature [28, 29]. These greedy algorithms achieve a zero-skew tree or a b-bounded-skew tree whose total wirelength is a constant factor to the total wirelength of their optimal counterparts.

4.1 Limiting Skew Distribution of Clock Trees

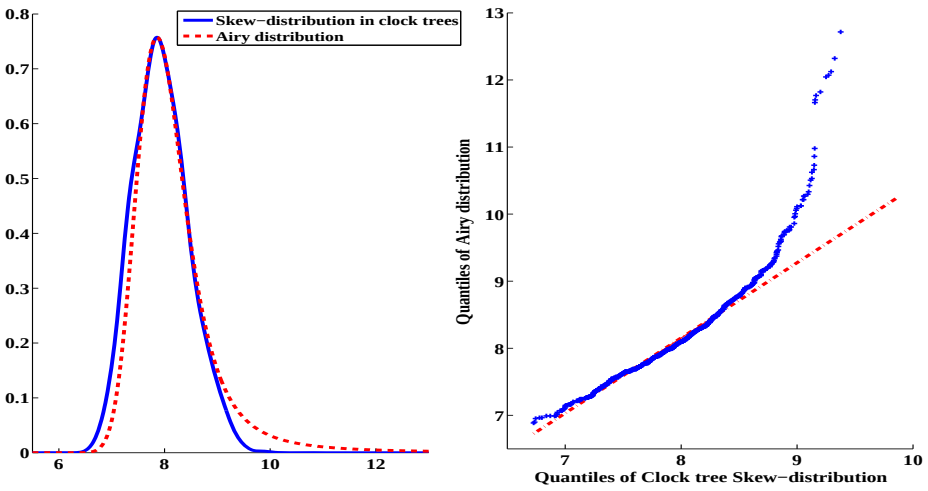
Here, we study the skew distribution of clock trees under different models for the propagation delay. In particular, we are interested in addressing the following questions - How does this skew distribution vary with increase in the size of the clock trees? Is there a limiting distribution? If so, is it a universal limiting distribution, in that it does not depend on the underlying distribution of propagation delays? Addressing these questions could lead to insights for the design of high density VLSI circuits.

We present the results of our empirical analysis that provides answers to these questions. We considered the following experimental setting. We generated 1000 clock trees whose edge weights (propagation delays) are generated randomly from a fixed distribution. For each clock tree generated, we compute its skew and observe the skew distribution as we increase the size of the clock trees. Figure 8 presents two plots for three different probability distribution on the edge weights. In each case, the first plot compares the skew distribution of clock trees with 2^{20} sinks to that of an Airy distribution, and the second plot compares their quantiles. Once again, the quantiles plot diverges from the straight line at the right end since the skew distribution has a finite support whereas the Airy distribution has an infinite support on the right end.

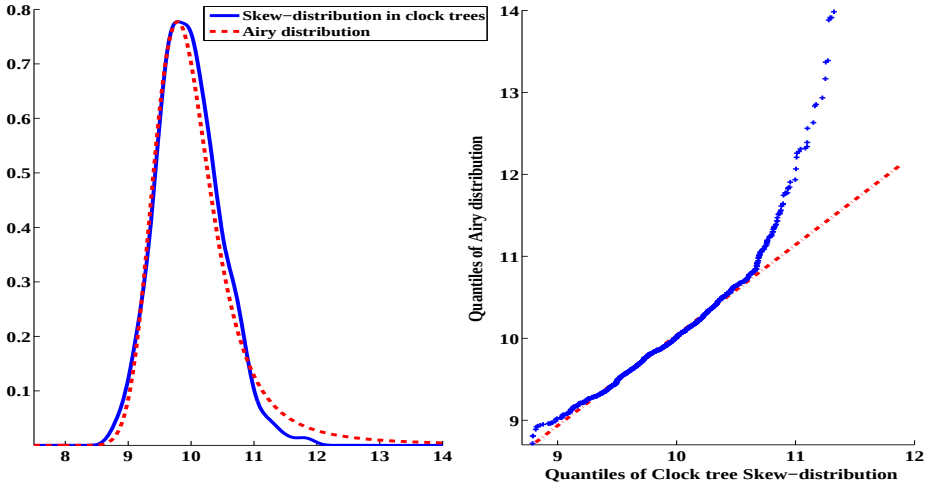
Note from these plots that the skew distribution tends to converge to an Airy distribution for large clock trees (i.e., large number of sinks). Moreover, this phenomenon seems to be invariant to the distribution of edge weights/propagation delays.

5 Conclusions and Future Work

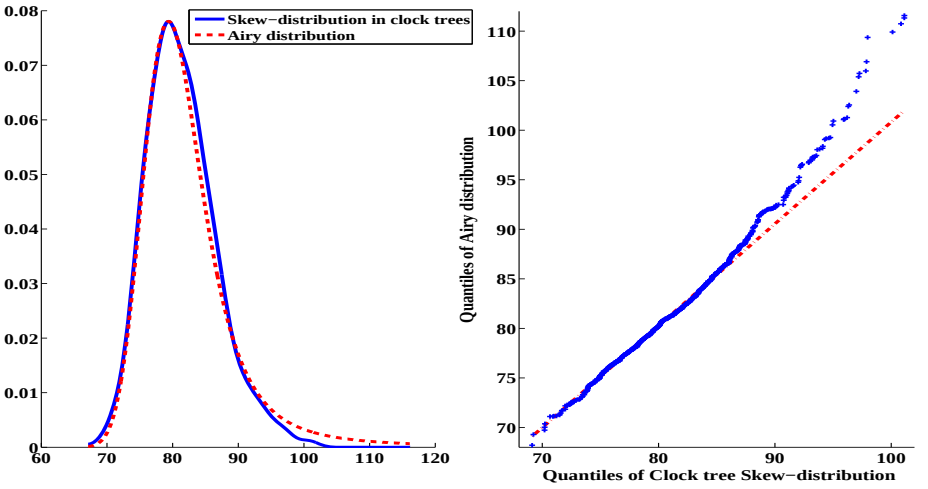
We present two interesting findings in this work. First, we show that the depth distribution of the leaf nodes in trees constructed using GBS converge to a strange distribution known as the Airy distribution.



(a)



(b)



(c)

Figure 8: Limiting skew distribution of clock trees under different distributions on edge weights. a) Uniform distribution over the interval $[0, 1]$ b) Normal distribution $\mathcal{N}(1, 0.25)$ c) Exponential distribution $\sim e^{-x/2}$.

Then, we show that the skew distribution of clock trees also converge to this Airy distribution as the number of sinks increases. These observations lead to interesting questions regarding the relation of these problems to that of Catalan trees and other problems involving binary trees as described in Section 2.2. Future work should investigate more in to these relations.

References

[1] R. P. Stanley, *Enumerative Combinatorics. Vol. 2.* Cambridge University Press, 1999.

[2] N. Kapur, "Additive functionals on random search trees," Ph.D. dissertation, The John Hopkins University, 2003.

[3] L. Takacs, "A Bernoulli excursion and its various applications," *Advances in Applied Probability*, vol. 23, pp. 557–585, 1991.

[4] ——, "Conditional limit theorems for branching processes," *Journal of Applied Mathematics and Stochastic Analysis*, vol. 4(4), pp. 262–292, 1991.

[5] ——, "On a probability problem connected with railway traffic," *Journal of Applied Mathematics and Stochastic Analysis*, vol. 4(1), pp. 1–27, 1991.

[6] ——, "On the total height of random rooted trees," *Journal of Applied Probability*, vol. 29(3), pp. 543–556, 1992.

[7] P. Flajolet, P. Poblete, and A. Viola, "On the analysis of linear probing hasing," *Algorithmica*, vol. 22(4), pp. 490–515, 1998.

[8] D. E. Knuth, *The Art of Computer Programming: Volume 2*, 3rd ed. Addison-Wesley, 1998.

[9] P. Chassaing and J. F. Marckert, "Parking functions, empirical processes and the width of rooted labeled trees," University of Nancy, Tech. Rep., 1999.

[10] P. Flajolet and G. Louchard, "Analytic Variations on the Airy Distribution," *Algorithmica*, vol. 31(3), pp. 361–377, 2001.

[11] S. Dasgupta, "Analysis of a greedy active learning strategy," *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[12] R. Nowak, "Generalized binary search," *Proceedings of the Allerton Conference*, 2008.

[13] F. Yu, F. Tu, H. Tu, and K. Pattipati, "Multiple disease (fault) diagnosis with applications to the QMR-DT problem," *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 1187–1192, October 2003.

[14] S. Bhavnani, A. Abraham, C. Demeniuk, M. Gebrekristos, A. Gong, S. Nainwal, G. Vallabha, and R. Richardson, "Network analysis of toxic chemicals and symptoms: Implications for designing first-responder systems," *Proceedings of American Medical Informatics Association*, 2007.

[15] D. Geman and B. Jedynak, "An active testing model for tracking roads in satellite images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 1–14, 1996.

[16] A. Gupta, R. Krishnaswamy, V. Nagarajan, and R. Ravi, "Approximation algorithms for optimal decision trees and adaptive TSP problems," 2010, available online at arXiv.org:1003.0722.

[17] G. Bellala, S. K. Bhavnani, and C. Scott, "Extensions of Generalized Binary Search to Group Identification and Exponential Costs," *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[18] L. Hyafil and R. Rivest, "Constructing optimal binary decision trees is NP-complete," *Information Processing Letters*, vol. 5(1), pp. 15–17, 1976.

[19] D. W. Loveland, "Performance bounds for binary testing with arbitrary weights," *Acta Informatica*, 1985.

[20] S. R. Kosaraju, T. M. Przytycka, and R. S. Borgstrom, "On an optimal split tree problem," *Proceedings of 6th International Workshop on Algorithms and Data Structures, WADS*, pp. 11–14, 1999.

[21] S. Roy, H. Wang, G. Das, U. Nambiar, and M. Mohania, "Minimum-effort driven dynamic faceted search in structured databases," *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 13–22, 2008.

[22] P. Erdös and A. Rényi, "On random graphs I," *Publications Mathematics*, pp. 290–297, 1959.

[23] J. L. Guillaume and M. Latapy, "Bipartite graphs as models of complex networks," *Proceedings of the 1st International Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN), Lecture Notes in Computer Sciences (LNCS)*, 2004.

[24] H. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI.* Addison-Wesley, 1990.

[25] A. Kahng and G. Robins, *On Optimal Interconnections for VLSI.* Kluwer Academic Publishers, 1995.

[26] G. Rouskas and I. Baldine, "Multicast routing with end-to-end delay and delay variation constraints," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 346–356, 1997.

[27] T. H. Chao, Y. C. Hsu, J. M. Ho, K. D. Boese, and A. B. Kahng, "Zero skew clock routing with minimum wirelength," *IEEE Transactions on Circuits and Systems Part 2: Analog and Digital Signal Processing*, vol. 39, pp. 799–814, 1992.

[28] M. Charikar, J. Kleinberg, R. Kumar, S. Rajagopalan, A. Sahai, and A. Tomkins, "Minimizing wirelength in zero and bounded skew clock trees," *SIAM Journal of Discrete Mathematics*, vol. 17, no. 4, pp. 582–595, 2004.

[29] Z. Zelikovsky and I. I. Măndoiu, "Practical approximation algorithms for zero and bounded skew trees," *SIAM Journal of Discrete Mathematics*, vol. 15, pp. 97–111, 2002.