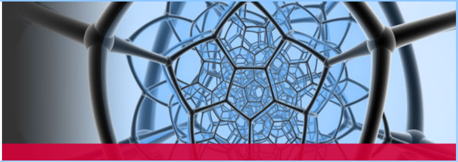
PHASE CHANGES IN RANDOM STRUCTURES AND ALGORITHMS

Hsien-Kuei Hwang

Summer School in Applied Probability

May 20, 2009

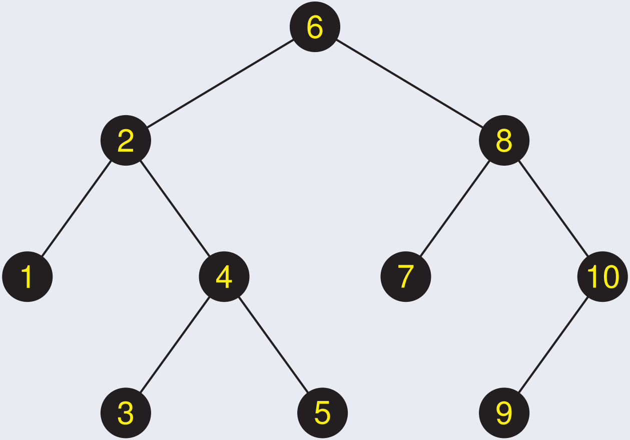
Carleton UNIVERSITY
Canada's Capital University



OUTLINE OF THE LECTURES

1. **Binary search trees, Quicksorts, and phase changes**
2. **Method of moments and its refinements**
3. **Differential equations with polynomial coefficients**
4. **Profiles of random log-trees**

BINARY SEARCH TREE CONSTRUCTED FROM {6,2,4,8,7,1,5,3,10,9}



WHO STUDIED BSTs FIRST? AND WHEN?

Knuth (1997, *Art Comput. Programming*), Vol. III, p. 453

- **Windley (1960) *Computer Journal***
- **Booth and Colin (1960) *Information and Control***
- **Hibbard (1962) *Journal ACM***
- **Hoare (1961) *Communications ACM* (Quicksort)**

The first published descriptions of tree insertion were by P. F. Windley [*Comp. J.* **3** (1960), 84–88], A. D. Booth and A. J. T. Colin [*Information and Control* **3** (1960), 327–334], and Thomas N. Hibbard [*JACM* **9** (1962), 13–28]. Each of these authors seems to have developed the method independently of the others, and each paper derived the average number of comparisons (6) in

WHY STUDY BSTs?

Computer algorithms

- ***Simple***, ***fundamental***, ***prototypical*** **data structure**
- **Many variants, generalizations: balanced BSTs, weighted BSTs, quadtrees, median BSTs, . . .**
- **Closely connected to quicksort (one of the most widely used sorting algorithms)**

Appeared in other fields

Statistical physics, probability, evolution, population genetics, chemistry, . . .

Mathematically intriguing

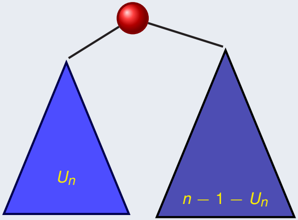
Many fascinating phenomena and challenging problems

RANDOM BSTs

The probability model

Assume that all $n!$ permutations of n elements are equally likely.

Construct the BST from a random permutation. Call it a *random BST*.



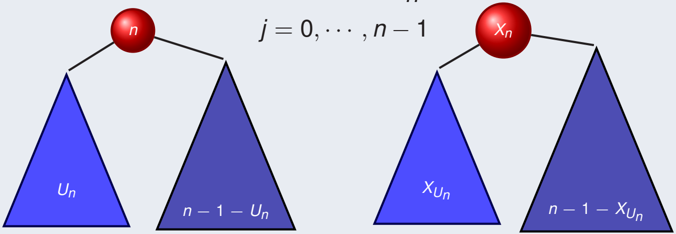
$$\mathbb{P}(U_n = j) = \frac{1}{n}$$

$$j = 0, \cdots, n-1$$

RANDOM BSTs

$$\mathbb{P}(U_n = j) = \frac{1}{n}$$

$$j = 0, \cdots, n-1$$



1. **probabilistic:** $X_n \stackrel{d}{=} X_{U_n} + X_{n-1-U_n} + T_n$
2. **recurrence:** $a_n = \dfrac{2}{n} \displaystyle\sum_{0 \le j < n} a_j + b_n$
3. **differential equation:** $f'(z) = \frac{2}{1-z} f(z) + g(z)$
4. **bivariate parameter:** $a_{n,k} = \dfrac{2}{n} \displaystyle\sum_{0 \le j < n} a_{j,k} + b_{n,k}$

CLOSELY CONNECTED STRUCTURES

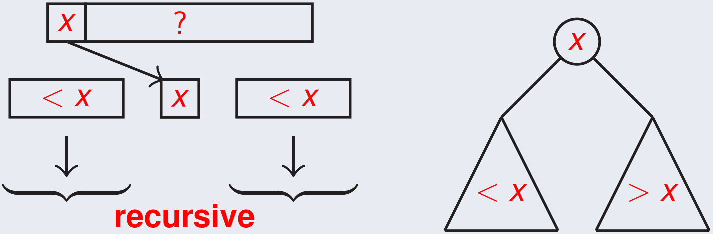
A short list

- **Quicksort algorithms**
- **Discrete probability: random pairwise selections (or an unfriendly seating arrangement problem)**
- **Statistical physics: Eden model, diffusion-limited aggregates, . . .**
- **Combinatorial structures: binary increasing trees**
- **Chemistry: random sequential adsorption (or random dimer filling, . . .)**
- **Evolutionary trees: Yule (or Markov) model**
- **Population genetics: Kingman's coalescent**
- **Parking problems: discrete and continuous**
- **Random fragmentation process**
- **Branching Markov processes**

CLOSELY CONNECTED STRUCTURES

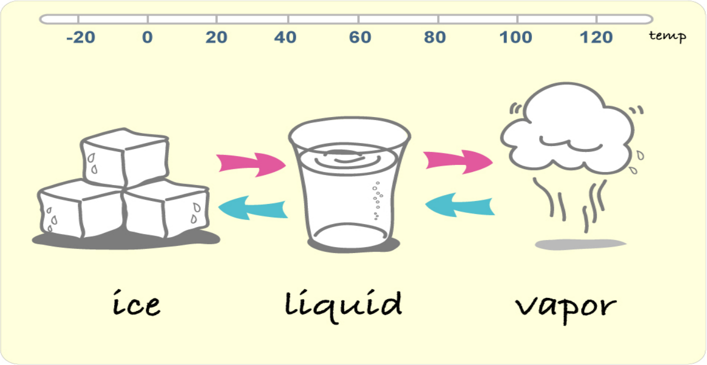
Quicksort (Hoare, 1961, *Comm. ACM*)

Selected to be among the *top 10 algorithms* in the 20th century with "*the greatest influence on the development and practice of science and engineering in the 20th century*" (appeared in the January/February issue of *Computing in Science & Engineering*).



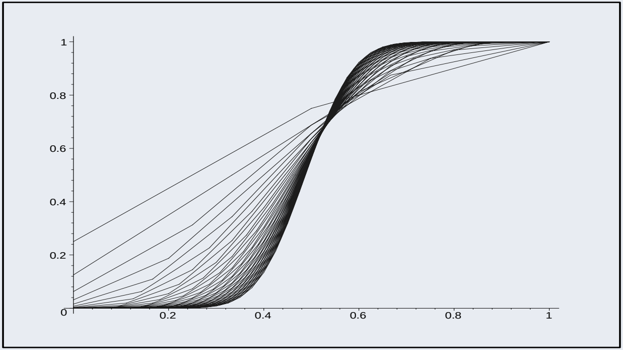
Widely used; many variants.

Random search trees and related models



PHASE CHANGE

$f(n;m)$ change behaviors as $n \to \infty$ and $m \geq m_0$



CLASSICAL CENTRAL LIMIT THEOREM

$X_1, \ldots, X_n$ iid, continuous, zero mean, finite variance $\sigma^2 > 0$

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{\sigma\sqrt{n}} < x\right) \to \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt;$$

so $\Phi(x)$ is used to bridge the transition between "*events unlikely to happen*" and "*events happening almost always*".

Binomial from Normal to Poisson

If $X_n \sim$ Binomial$(n; p)$, then

- $X_n \sim \mathcal{N}(pn, p(1-p)n)$ **if** $pn \to \infty$;
- $X_n \sim$ **Poisson**(λ) **if** $pn \to \lambda < \infty$.

PHASE TRANSITIONS (CHANGES): FEATURES

- ***Analytically*, singularity changes nature (or regular $\to$ singular)**

$$\mathbb{P}\left(X_1 + \cdots + X_n > x\right) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} t^{-1} e^{-ixt}\left(\mathbb{E}(e^{itX_1})\right)^n dt.$$

coalescence of pole and the saddle-point.

- ***Asymptotically*, the consequence of increasing errors: $\mathbb{E}(Y_n) = cn + O(n^\alpha)$,**

$$\mathbb{V}(Y_n) \asymp \begin{cases} n, & \textbf{if } \alpha < 1/2; \\ n\log n, & \textbf{if } \alpha = 1/2; \\ n^{2\alpha}, & \textbf{if } \alpha > 1/2. \end{cases}$$

- ***Algorithmically*, hardest instances often (but not always) occur in or near the phase transition range.**

DESIGN AND ANALYSIS OF ALGORITHMS: A NEW TREND

Massive data or data streams everywhere

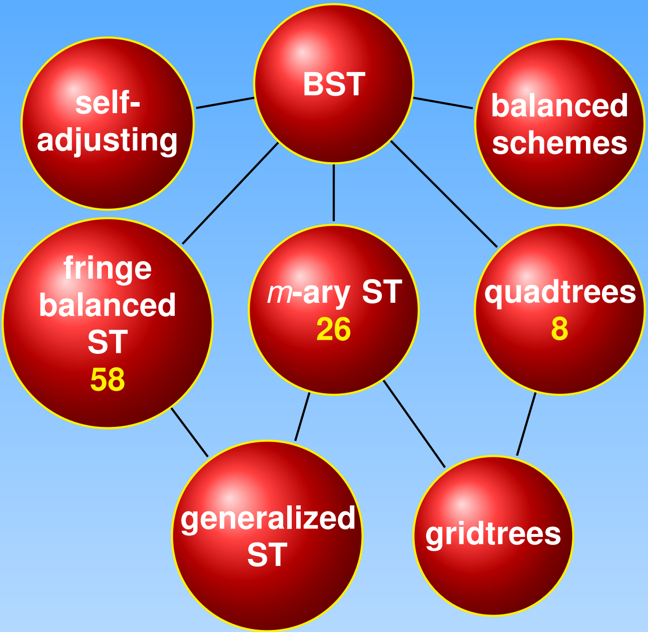
Many algorithms need to be redesigned and asymptotic analysis is gaining its increasing importance.

The current mega-giga-era will soon be replaced by the tera-peta-era.

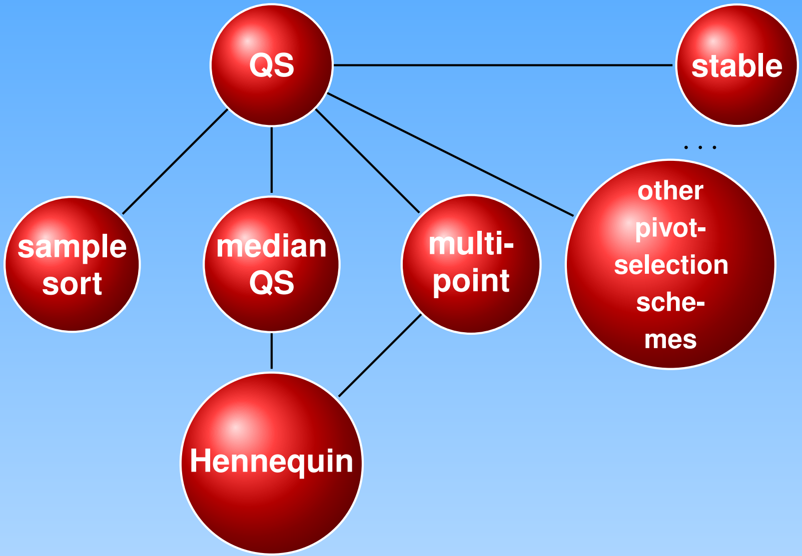
PHASE CHANGES ARE IMPORTANT

- *Quantitatively*, phase transitions more informative than static states
- *Structurally*, phase transitions useful in describing the structural stability
- *Theoretically*, many aspects of phase transitions like classification, universality $\Longrightarrow$ theory
- *Computationally*, identifying phase transitions helpful in improving algorithms
- *Methodologically*, more powerful tools always needed

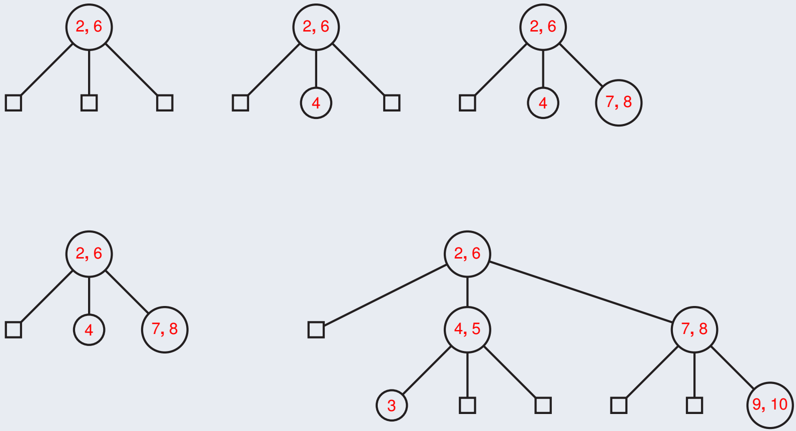
VARIANTS OF BSTs



VARIANTS OF QUICKSORT

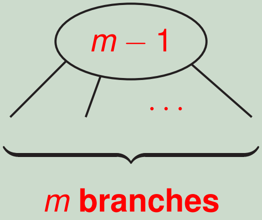


TERNARY SEARCH TREE BUILT FROM $\{6,2,4,8,7,1,5,3,10,9\}$



m-ARY SEARCH TREES

Introduced by Muntz and Uzgalis (1971)



An example: space requirements

Mahmhoud and Pittel (1989), Lew and Mahmoud (1994), Chern and H. (2001): The space requirement X_n exhibits the phase change: if $3 \le m \le 26$, then

$$\frac{X_n - \mu n}{\sigma \sqrt{n}} \longrightarrow N(0,1);$$

if $m \ge 27$, then the sequence of random variables $(X_n - \mathbb{E}(X_n))/\sqrt{\mathbb{V}(X_n)}$ does not converge (periodicity dominates).

For other results, see H. (2003), Janson (2005), Chauvin and Pouyanne (2005), Fill and Kapur (2005), Dean and Majumdar (2005).

Convergence rate to normal limit law

H. (2003)

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}(X_n)}} < x \right) - \Phi(x) \right|$$
$$= \begin{cases} O(n^{-1/2}), & \text{if } 3 \le m \le 19; \\ O(n^{-3(\frac{3}{2}-\alpha)}), & \text{if } 20 \le m \le 26, \end{cases}$$

where $\alpha \in (\frac{4}{3}, \frac{3}{2})$ denotes the *real part of the second largest zero* of the equation

$$z(z+1)\cdots(z+m-2) = m!.$$

Rate optimal, up to implied constants

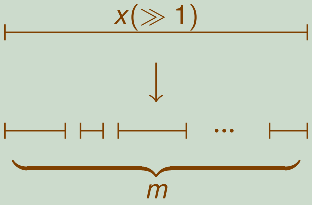
Approximate numerical values of α and $3(\frac{3}{2} - \alpha)$

m	α	$3(\frac{3}{2} - \alpha)$
20	1.34892881	0.45321354
21	1.38079786	0.35760639
22	1.40936978	0.27189065
23	1.43512896	0.19461309
24	1.45847025	0.12458925
25	1.47971848	0.06084455
26	1.49914326	0.00257020

RANDOM FRAGMENTATION PROCESS

Dean and Majumdar (2002)

a physical model for random *m*-ary search trees.



Stop if length < 1

Number of nodes in the corresponding tree is an RV

RANDOM FRAGMENTATION TREE

Dean and Majumdar (2002)

They argue heuristically (called *scientifically modeling math* by Aldous, in contrast to *theorem-proof math*) that

$$\mathbb{V}(X_n) \asymp \begin{cases} n, & \textbf{if } 3 \le m \le 26; \\ n^{2\alpha-2}, & \textbf{if } m > 26. \end{cases}$$

They conclude

. . . we have shown that a fragmentation process with an atomic threshold can undergo a nontrivial phase transition in the fluctuations of the number of splittings at a critical value of the branching number m. . . . The mechanism of this transition is remarkably simple and therefore one expects it to be rather generic with broad applications . . . .

RANDOM FRAGMENTATION TREE

Dean and Majumdar (2002): a cuboid splitting tree

- **A random point splits** $[0, x]^d$ **into** 2^d **hyper-rectangles.**
- **Continue as long as the volume is** > 1**.**

Call the tree corresponds to the resulting configuration a *random fragmentation tree*.

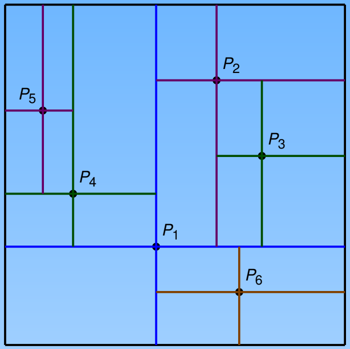
The phase change at $d = 8$

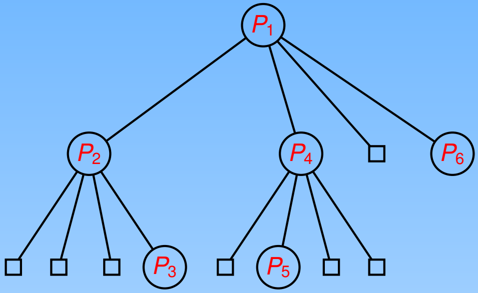
The number X_n **of nodes in the tree undergoes a phase change:**

$$\mathbb{V}(X_n) \asymp \begin{cases} n, & \textbf{if } 1 \le d \le 8; \\ n^{2\cos(2\pi/d)-1}, & \textbf{if } d > 8. \end{cases}$$

$$\{2\cos\tfrac{2\pi}{d} - 1\}_{d\ge 5} = \{-.38, 0, 0.24, 0.41, -.53, 0.61, \dots\}$$

A 2-DIMENSIONAL POINT QUADTREE

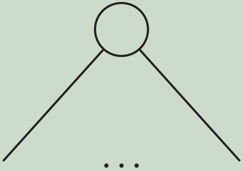




d-DIMENSIONAL QUADTREE

Introduced by Finkel and Bentley (1974)

$x \in \mathbb{R}^d$



$$\underbrace{\qquad\cdots\qquad}_{2^d \textbf{ subtrees}}$$

The model

If the n given points are iid from $[0,1]^d$, then the resulting tree is called a *random quadtree*.

The phase change

Chern, Fuchs, H. (2005): ***If $1 \leq d \leq 8$, then the number X_n of leaves is asymptotically normally distributed; if $d \geq 9$, then the random variables $(X_n - \mathbb{E}(X_n))/\sqrt{\mathbb{V}(X_n)}$ do not converge.***

Second phase change at $d = 7$.

RANDOM d-DIMENSIONAL GRID-TREES

A combination of m-ary search tree and quadtree

Devroye (1998):

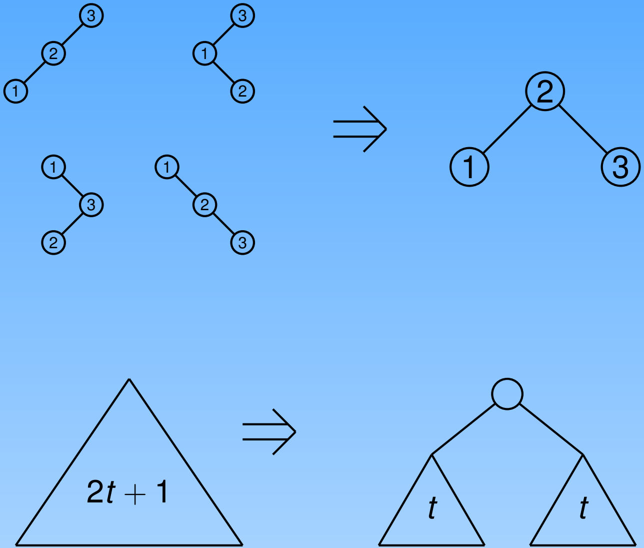
- $m-1$ **random points split** $[0, x]^d$ **into** m^d **grids.**
- **Repeat if volume** > 1**.**

Call the corresponding tree a *random grid-tree*.

Chern, Fuchs, H. (2005): all pairs (m, d) leading to asymptotic normality for the number of leaves

m	2	3	4	$5, \dots, 8$	$9, \dots, \underline{26}$
d	$1, \dots, \underline{8}$	$1, \dots, 4$	$1, \dots, 3$	$1, 2$	1

MEDIAN-OF-$(2t+1)$ BSTs (FRINGE-BALANCED)



MEDIAN-OF-$(2t+1)$ BSTs

How to construct from a sequence of numbers?

Bell (1965) and Walker and Wood (1976):

- **Find the median of $2t+1$ random elements**
- **Place this median at the root with the smaller t elements going to the left, larger to the right**
- **Keep inserting as usual (small $\rightarrow$ L, large $\rightarrow$ R)**
- **Split recursively if size $= 2t+1$**

Phase change at $t = 58$

Chern and H. (2001): ***The number of nodes with subtree sizes*** $\geq 2t+1$ ***is asymptotically normally distributed for*** $1 \leq t \leq 58$***, and does not converge for*** $t > 58$**.**

Combine *m*-ary search trees and median BSTs

All pairs (m, t) for which asymptotic normality holds.

m	2	3	4	5	6
t	$1, ..., 58$	$0, ..., 19$	$0, ..., 10$	$0, ..., 6$	$0, ..., 4$
m	7	$8, 9$	$10, ..., 13$	$14, ..., 26$	> 27
t	$0, ..., 3$	$0, 1, 2$	$0, 1$	0	$\emptyset$

A SIMPLE SCHEME FOR PHASE CHANGES

Chern, H. and Tsai (2002)

Phase changes are clarified for random variables defined recursively by

$$X_n \stackrel{d}{=} X_{I_n} + X^*_{n-1-I_n} + 1 \qquad (n \geq r),$$

where

$$\mathbb{P}(I_n = k) = \sum_{0 \leq j < r} p_j \frac{\binom{k}{j}\binom{n-1-k}{r-1-j}}{\binom{n}{r}}, \qquad \sum_{0 \leq j < r} p_j = 1.$$

More than a dozen of examples addressed there.

A DIFFERENT TYPE OF PHASE CHANGE

H. and Neininger (2002)

General cost measures on BSTs satisfy

$$X_n \stackrel{d}{=} X_{\textbf{uniform}_{[0,n-1]}} + X^*_{n-1-\textbf{uniform}_{[0,n-1]}} + Y_n \qquad (n \geq 2),$$

where Y_n is known.

The phase change

If $Y_n = O(n^{1/2})$, then the limit law of X_n is normal; if $Y_n \gg n^{1/2}$, then nonnormal.

A large number of applications.

CONCLUSION

Analysis of algorithms: a rich source of phase changes

- ***Many intriguing phenomena and challenging math***
- **More research needed to unveil new phase changes**
- **More collaboration needed (with statistical physicists, biologists, . . .)**
- ***Simple models are often ubiquitous***