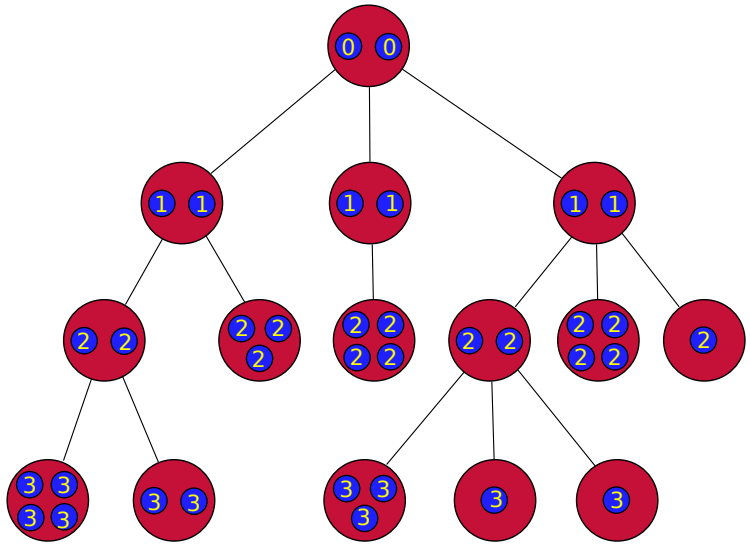
The Total Path Length of Split Trees

Cecilia Holmgren, Cambridge University
joint work with Nicolas Broutin, INRIA Rocquencourt

Bedlewo, Poland 17 June 2011



Aim of Study

To find the **asymptotic distribution of the total path length in random split trees**.

Split trees constitute a large class of random trees of logarithmic height defined by Devroye. Their study can be motivated by their use as models for tree data structures or sorting algorithms, e.g., the well-known Quicksort which can be depicted as the binary search tree. The total path length represents a natural cost measure or running time of these algorithms.

Examples of Split Trees

- The class of split trees includes many important random trees of logarithmic height, such as **binary search trees**, **m-ary search trees**, **quadtrees**, **median of $(2k+1)$-trees**, **simplex trees**, **tries** and **digital search trees**.

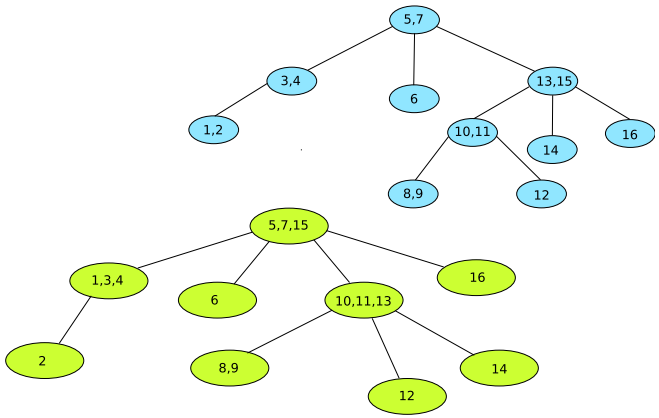


Figure: *A 3-ary and a 4-ary search tree.*

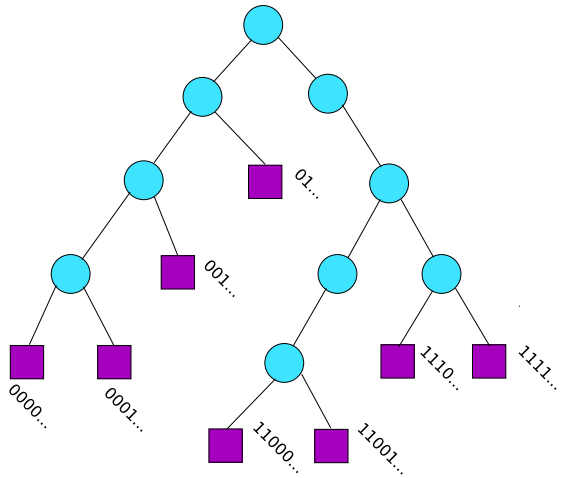
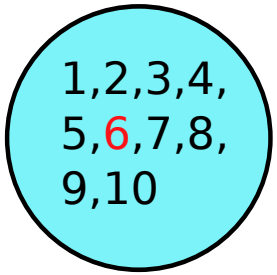


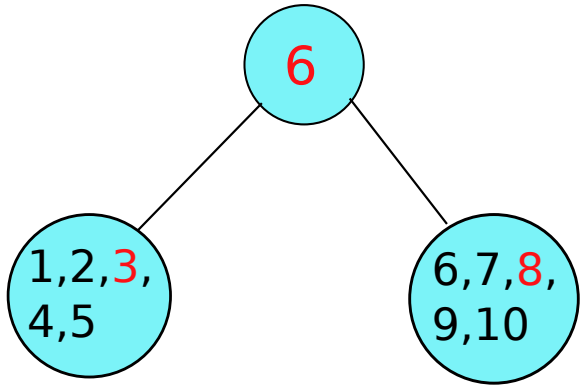
Figure: *A trie built from binary strings.*

The Binary Search Tree is an Example of a Split Tree: A Search Algorithm on n Items



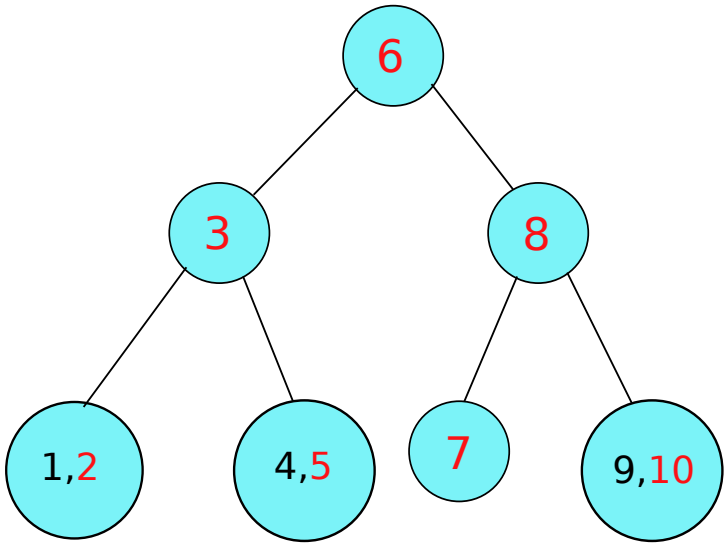
Start with a set of n ordered numbers/keys, e.g., the set $\{1, 2 \ldots, 10\}$. First all keys are associated to the root. Randomly draw one of the keys and store it at the root.

The Binary Search Tree is an Example of a Split Tree



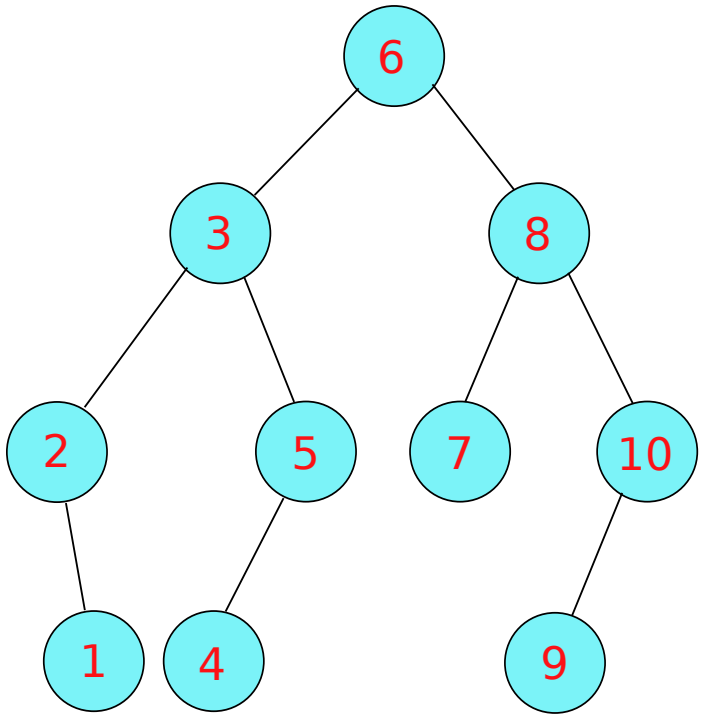
The remaining keys are then divided into two subgroups, depending on whether they are larger or smaller than the root's key. Randomly draw a new key in each of the two subgroups and store it in the left and right child respectively.

The Binary Search Tree is an Example of a Split Tree



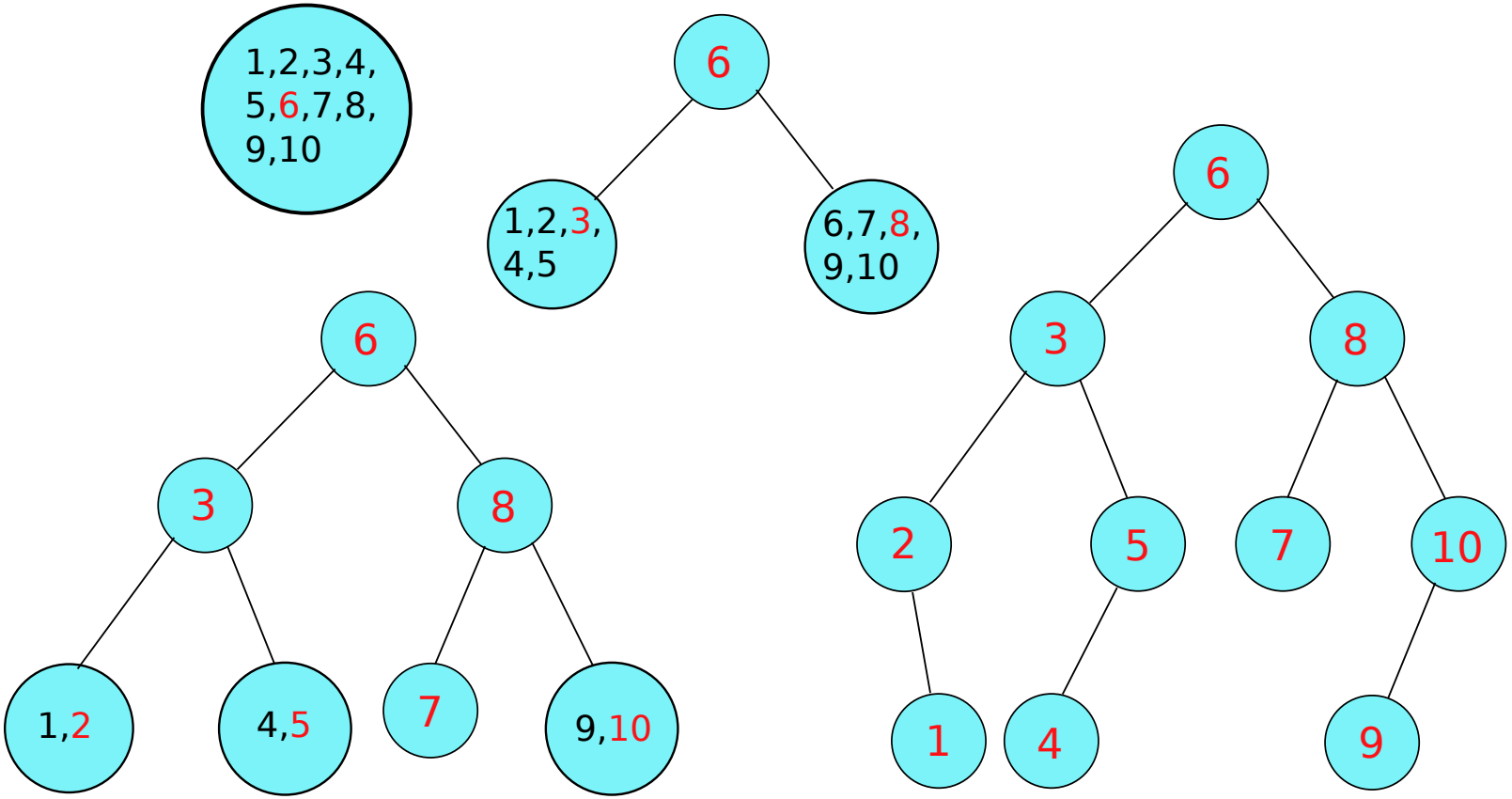
Compare the other keys in each subtree and again divide the keys into two groups in each subtree.

The Binary Search Tree is an Example of a Split Tree



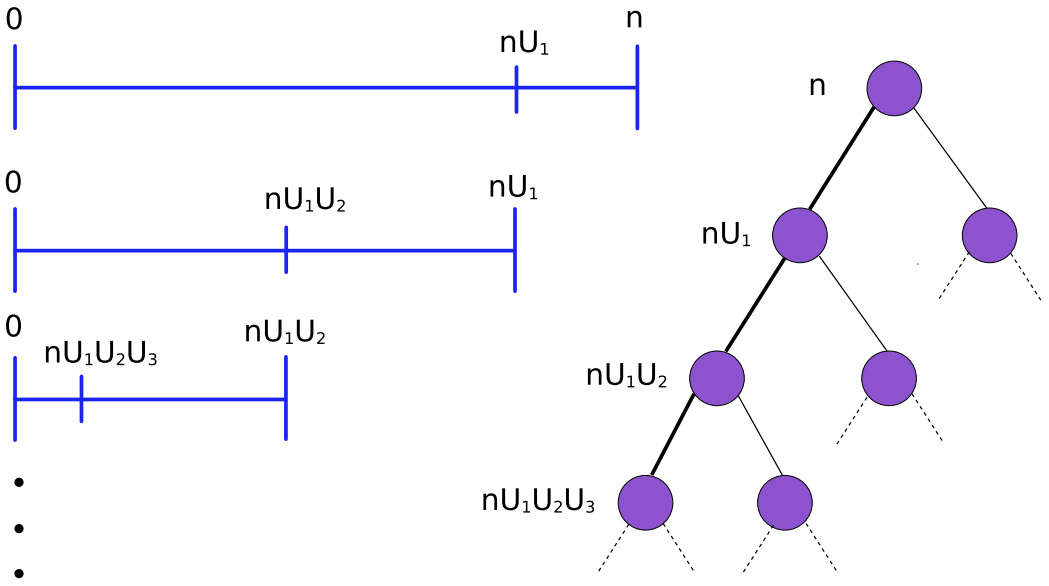
Proceed recursively in each subtree until the subtree holds exactly one key.

The Binary Search Tree is an Example of a Split Tree



In the final tree each node holds exactly one key.

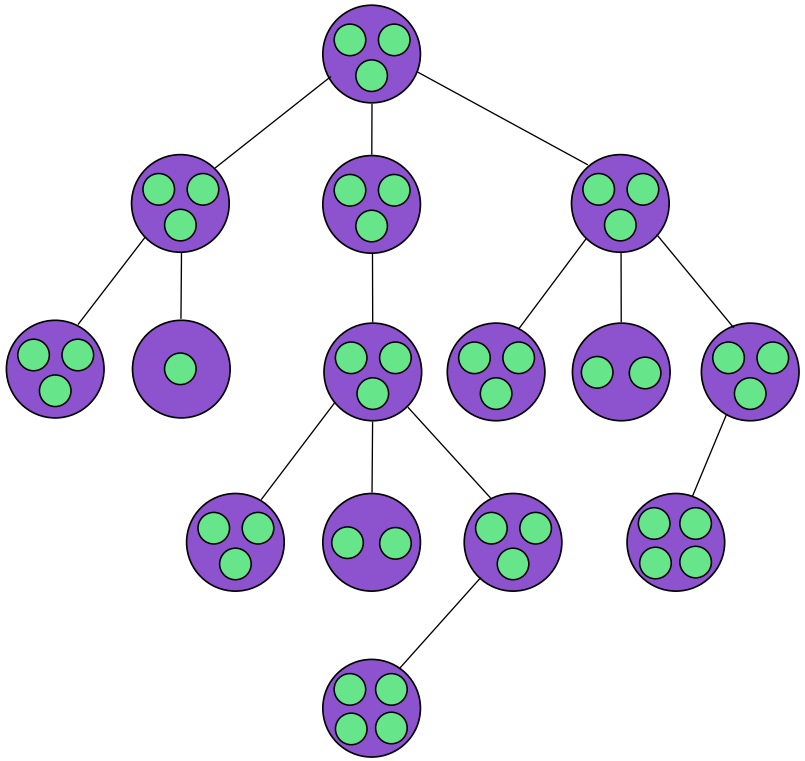
The Binary Search Tree (continued)



- Since the rank of the root's key is equally likely to be $\{1, 2, \ldots, n\}$, the size of its left subtree is distributed as $\lfloor nU \rfloor$, where U is a uniform $U(0, 1)$ random variable.

Important Parameters for Split Trees

(Devroye 1998)



- Branch factor b
- Cardinality n
- Vertex capacity s>0
- Internal vertex capacity $s_0 \leq s$
- An independent copy of the random split vector $\mathcal{V} = (V_1, V_2, \ldots, V_b)$ is attached to each vertex.

Split Trees

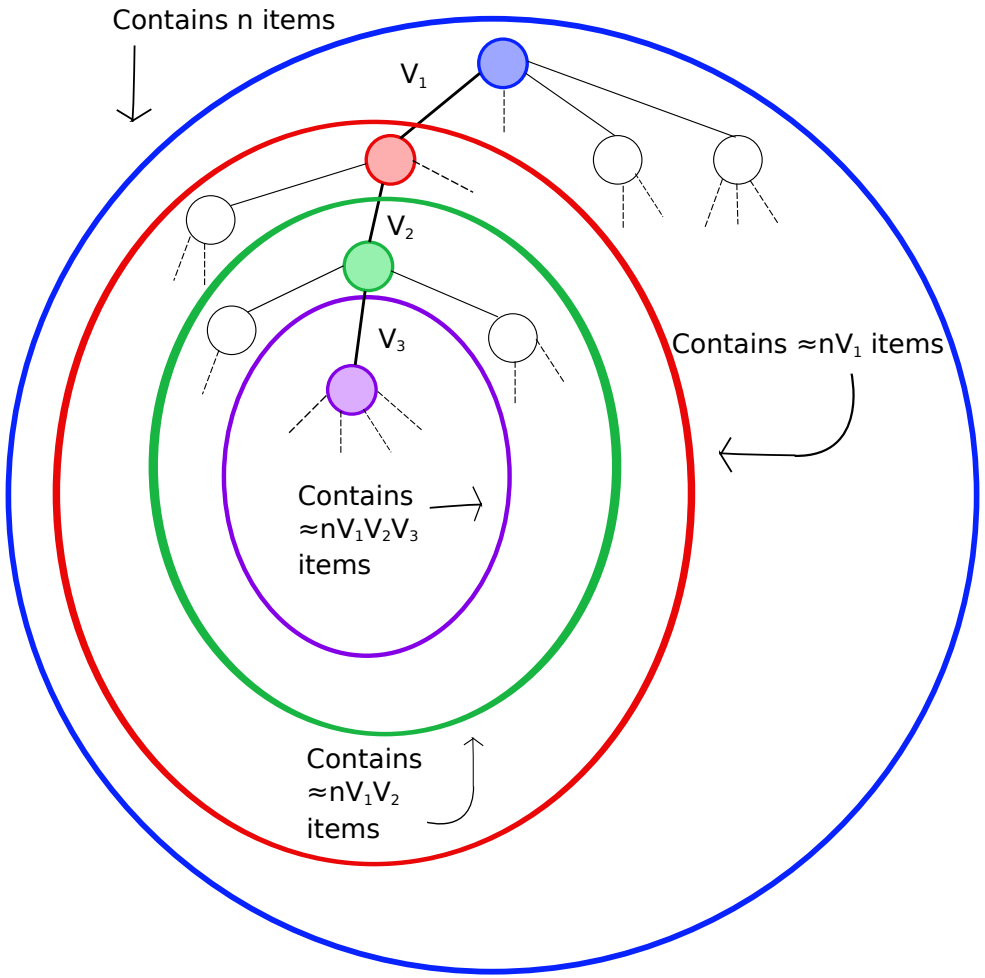


Figure: *Given all split vectors in the tree, n_v for v at depth d is close to $nL_v = n\prod_{j=1}^{d} V_j$, where the V_j's are i.i.d. random variables distributed as the components in the split vector.*

The Total Path Length/Running-Time of Sorting Algorithms

- Sorting algorithms sort a collection of data items (often called keys) by comparisons of the input data.

The Total Path Length/Running-Time of Sorting Algorithms

- Sorting algorithms sort a collection of data items (often called keys) by comparisons of the input data.
- The number of comparisons for a certain key is given by its depth in the tree.

The Total Path Length/Running-Time of Sorting Algorithms

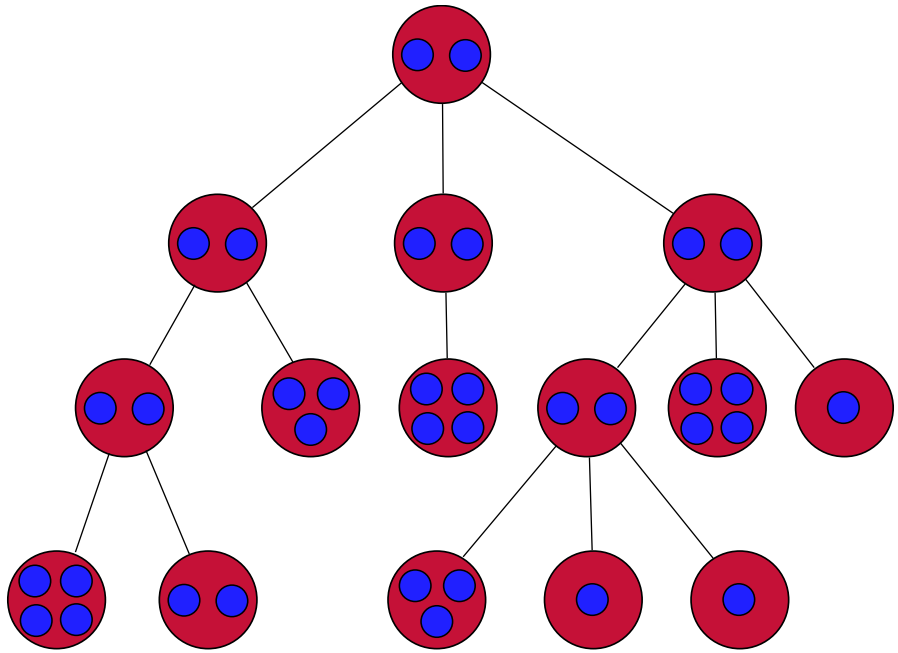
- Sorting algorithms sort a collection of data items (often called keys) by comparisons of the input data.
- The number of comparisons for a certain key is given by its depth in the tree.
- **The total number of comparisons is the total path length, which therefore represents a natural cost measure or running time of these algorithms.**

The Total Path Length/Running-Time of Sorting Algorithms

- Sorting algorithms sort a collection of data items (often called keys) by comparisons of the input data.
- The number of comparisons for a certain key is given by its depth in the tree.
- **The total number of comparisons is the total path length, which therefore represents a natural cost measure or running time of these algorithms.**
- Effective sorting algorithms are represented by $\log n$-trees with total path length, i.e., running time $\mathcal{O}(n \log n)$.

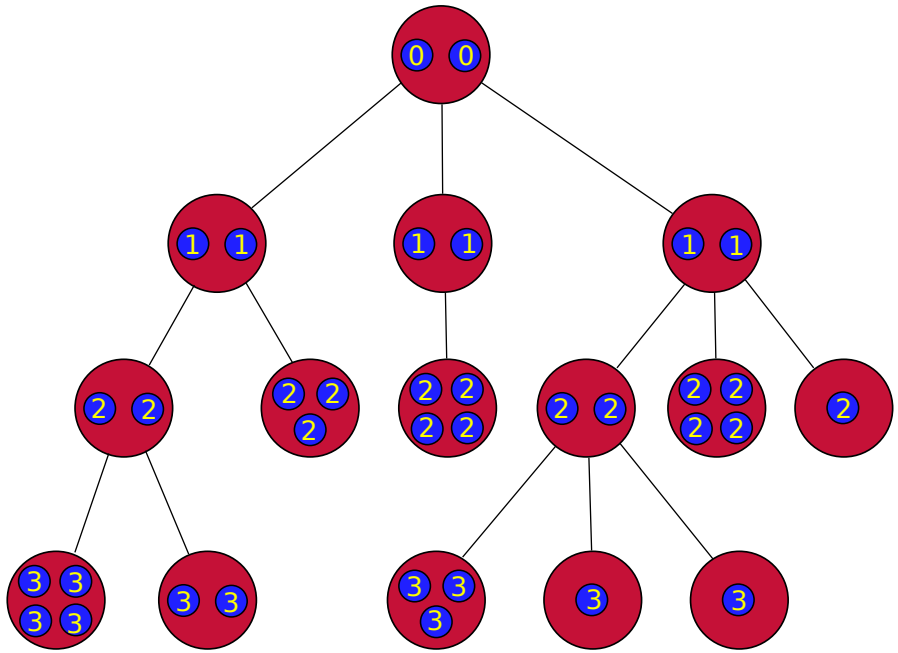
The Total Path Length of a Rooted Tree

- The total path length is the sum of the depths of all items (often represented by keys in tree data structures) in the tree.

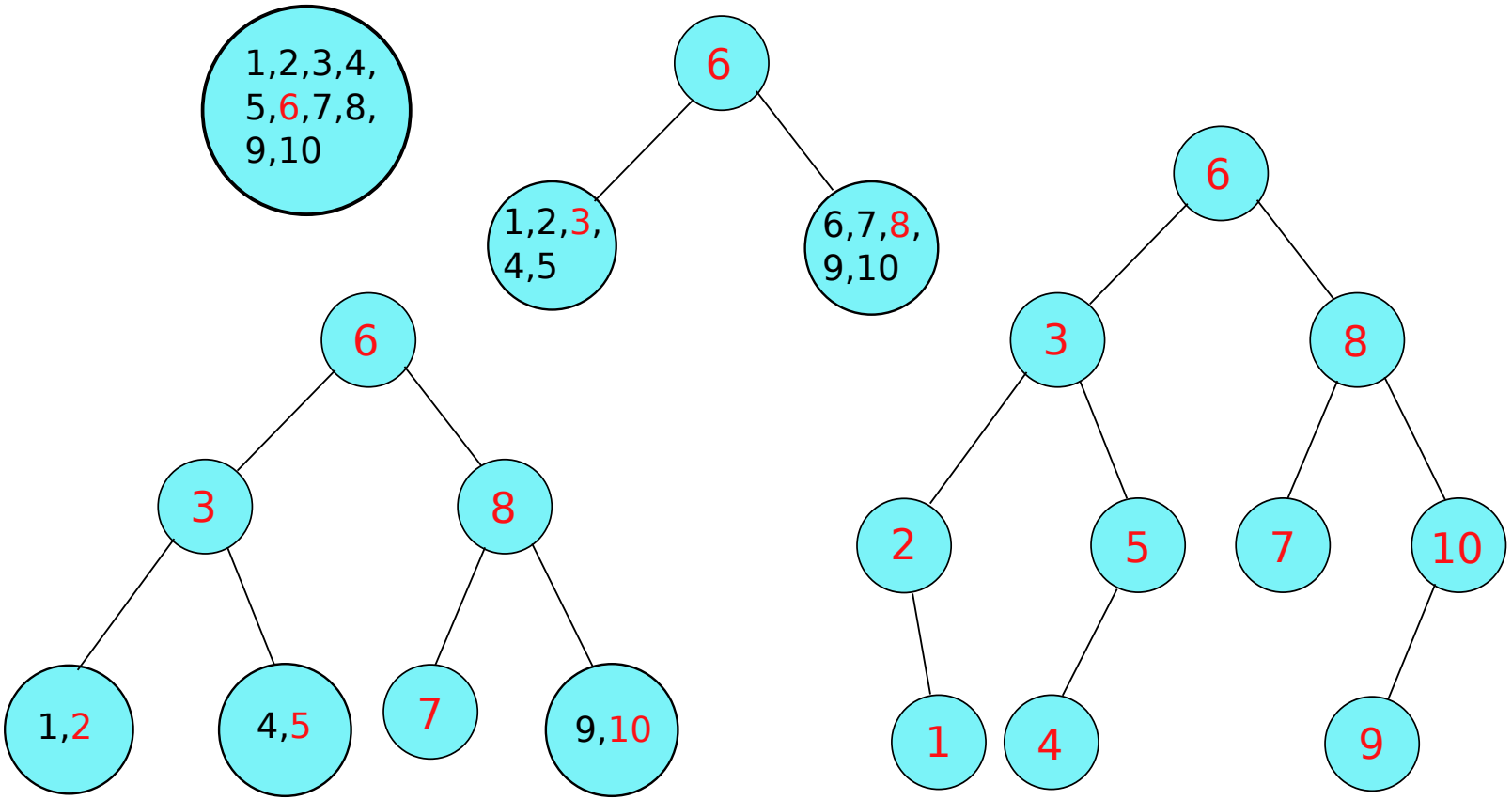


The Total Path Length of a Rooted Tree

- The total path length is the sum of the depths of all items in the tree.



The Running Time of Quicksort



Aim of Study

- To find the asymptotic distribution of the total path length in random split trees.

Background

- Devroye (1998) showed that the depth D_n of the last ball n in an arbitrary split tree obtains a normal limit law.

Background

- Devroye (1998) showed that the depth D_n of the last ball n in an arbitrary split tree obtains a normal limit law.
- The related total path length defined as the sum of the depths of the balls is more complicated since the depths are dependent.

Background

- Devroye (1998) showed that the depth D_n of the last ball n in an arbitrary split tree obtains a normal limit law.
- The related total path length defined as the sum of the depths of the balls is more complicated since the depths are dependent.
- Most studies of the total path length concern the model of binary search tree, or equivalently the cost of quicksort.

Background

- Devroye (1998) showed that the depth D_n of the last ball n in an arbitrary split tree obtains a normal limit law.
- The related total path length defined as the sum of the depths of the balls is more complicated since the depths are dependent.
- Most studies of the total path length concern the model of binary search tree, or equivalently the cost of quicksort.
- Properties of the asymptotic distribution and the rate of convergence of the total path length in quicksort have been studied e.g., by Régnier (1989), Rösler (1991), Tan and Hadjicostas (1995) and Janson and Fill (2001,2002).

Background cont.

- Rösler (1991) invented the *contraction method* which is the main technique in the study of the total path length for random trees.

Background cont.

- Rösler (1991) invented the *contraction method* which is the main technique in the study of the total path length for random trees.
- Neininger and Rüchendorf (1999,2004) further developed this method and used it e.g., to prove convergence in distribution for quadtrees.

Background cont.

- Our method relies on previous work by Neininger and Rüchendorf (1999) who gave a limit theorem for the path length of split trees under the assumption that the mean satisfies some precise asymptotic form, which we prove.

Main Theorem

Let $\Psi(T^n)$ be the total path length in a split tree with split vector $\mathcal{V} = (V_1, \ldots, V_b)$. Suppose that $\mathbf{P}(\exists i : V_i = 1) < 1$. Let

$$X_n := \frac{\Psi(T^n) - \mathbf{E}[\Psi(T^n)]}{n} \qquad \text{and} \qquad C(\mathcal{V}) = 1 + \frac{1}{\mu}\sum_{i=1}^{b} V_i \ln V_i,$$

where $\mu := b\mathbf{E}(-V \ln V)$. Then $X_n \to X$ in distribution, where X is the unique solution of the fixed point equation

$$X \stackrel{d}{=} \sum_{k=1}^{b} V_k X^{(k)} + C(\mathcal{V}), \tag{1}$$

where $X^{(k)}$ are independent and identically distributed copies of X, satisfying $\mathbf{E}[X] = 0$ and $\mathbf{Var}(X) < \infty$. Furthermore, exponential moments of X_n exist and converge $\mathbf{E}[e^{\lambda X_n}] \to \mathbf{E}[e^{\lambda X}]$ for any $\lambda \in \mathbb{R}$.

Precise asymptotics for the average path length implies the Main Theorem

The limit theorem for $\Psi(T^n)$ of a general split tree requires precise asymptotics for $\mathbf{E}[\Psi(T^n)]$. Let V be a random variable distributed as the components in the split vector $\mathcal{V}$.

Theorem 2

The expected value of the total path length $\Psi(T^n)$ exhibits the following asymptotics, as $n \to \infty$,

$$\mathbf{E}[\Psi(T^n)] = \mu^{-1} n \ln n + n\varpi(\ln n) + o(n). \tag{2}$$

where $\mu := b\mathbf{E}(-V \ln V)$ and ϖ is a continuous periodic function of period d. In particular, if $\ln V$ is not lattice, i.e., there is no $a \in \mathbb{R}$ such that $-\ln V \in a\mathbb{Z}$ almost surely, then $d = 0$ and ϖ is constant, i.e.,

$$\mathbf{E}[\Psi(T^n)] = \mu^{-1} n \ln n + cn + o(n). \tag{3}$$

Split Trees: Most Nodes Close to Depth $\mu^{-1} \ln n$.

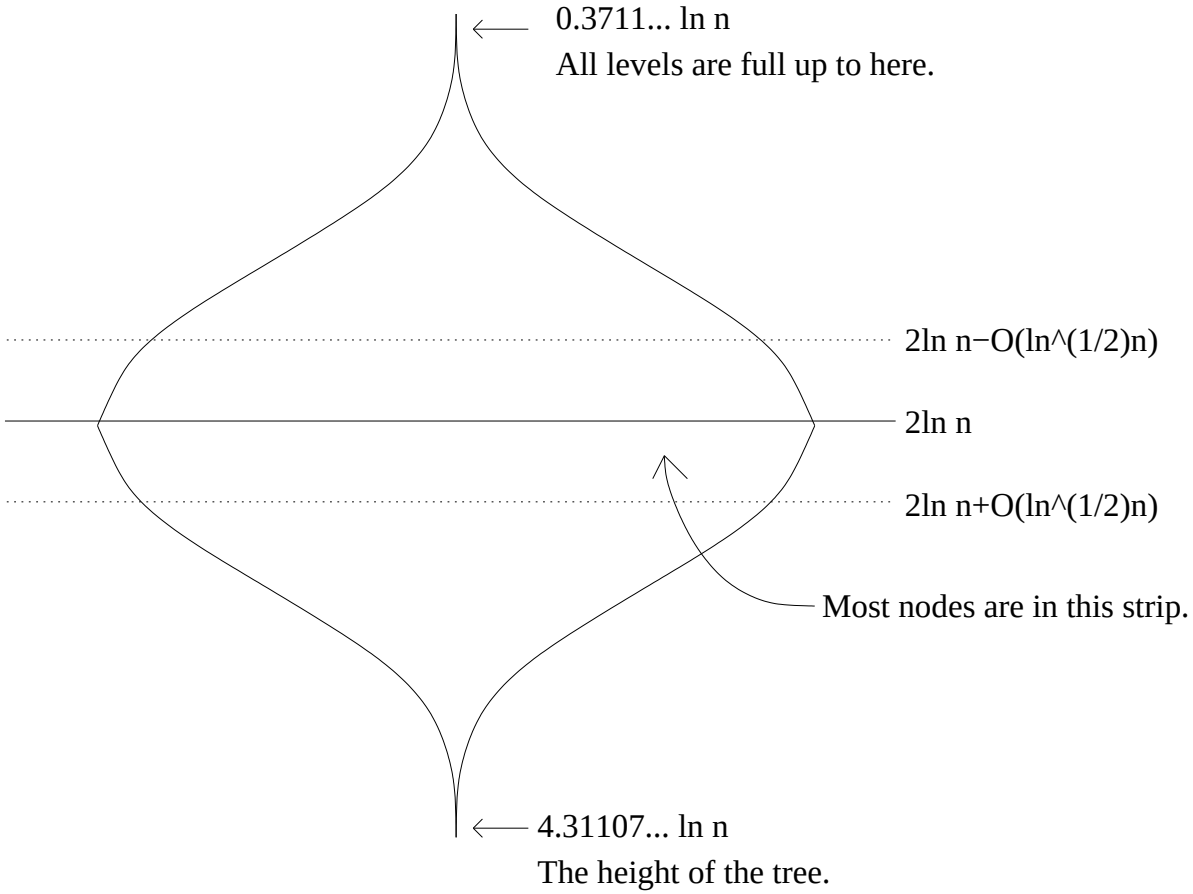


Figure: *The horizontal width represents the number of nodes at each level in a binary search tree. Most nodes are in a strip of width* $\mathcal{O}(\sqrt{\ln n})$ *around* $2 \ln n$.

Proving Theorem 2: Precise asymptotics for the average path length

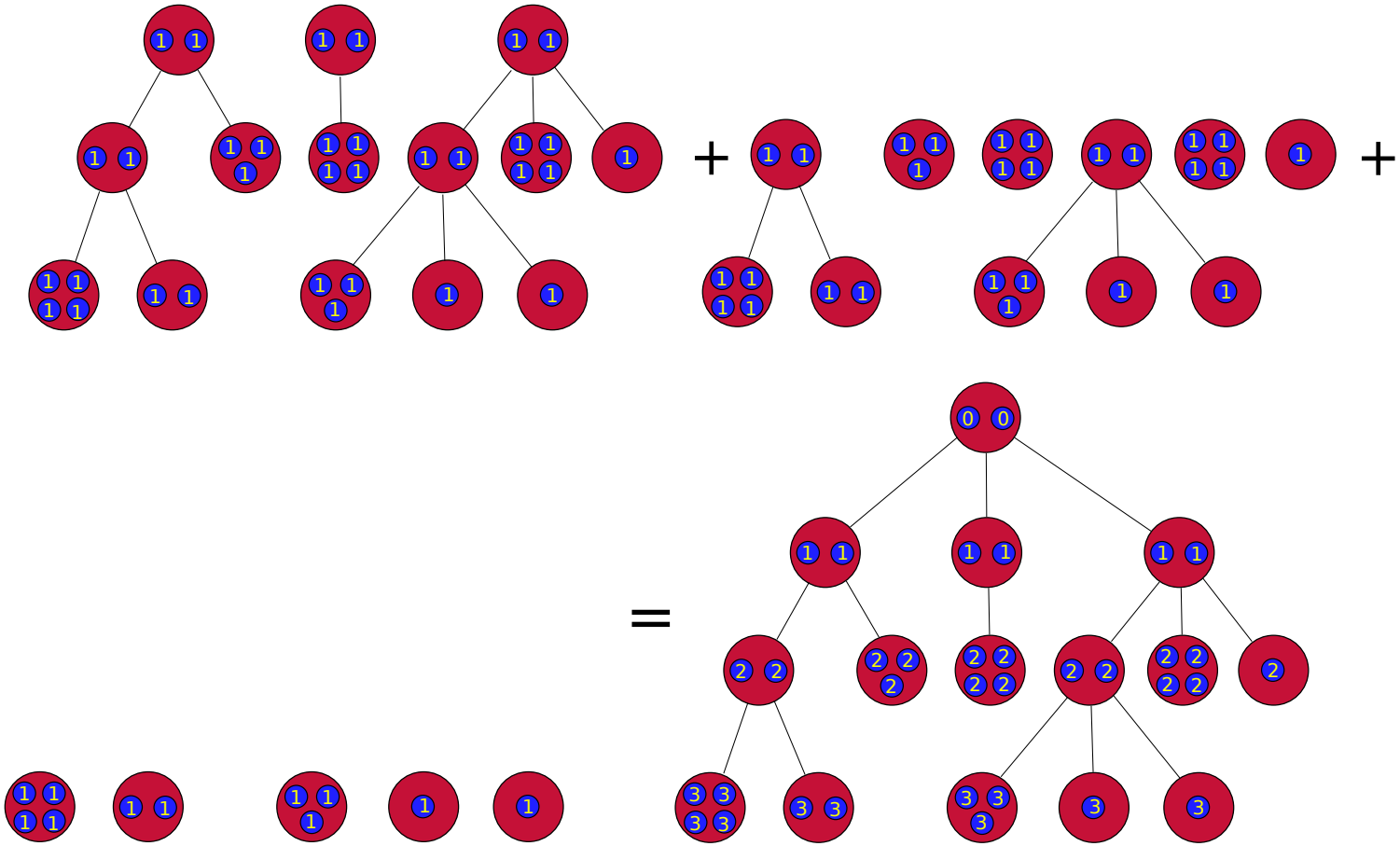
Basic idea: If $-\ln V$ has a non-lattice distribution: Consider two arbitrary values for the cardinality, say n and $\widehat{n}$ where $\widehat{n} \geq n$, and show that as $n \to \infty$

$$\left| \left(\frac{\mathbf{E}[\Psi(T^n)]}{n} - \frac{\ln n}{\mu} \right) - \left(\frac{\mathbf{E}[\Psi(T^{\widehat{n}})]}{\widehat{n}} - \frac{\ln \widehat{n}}{\mu} \right) \right| = o(1). \qquad (4)$$

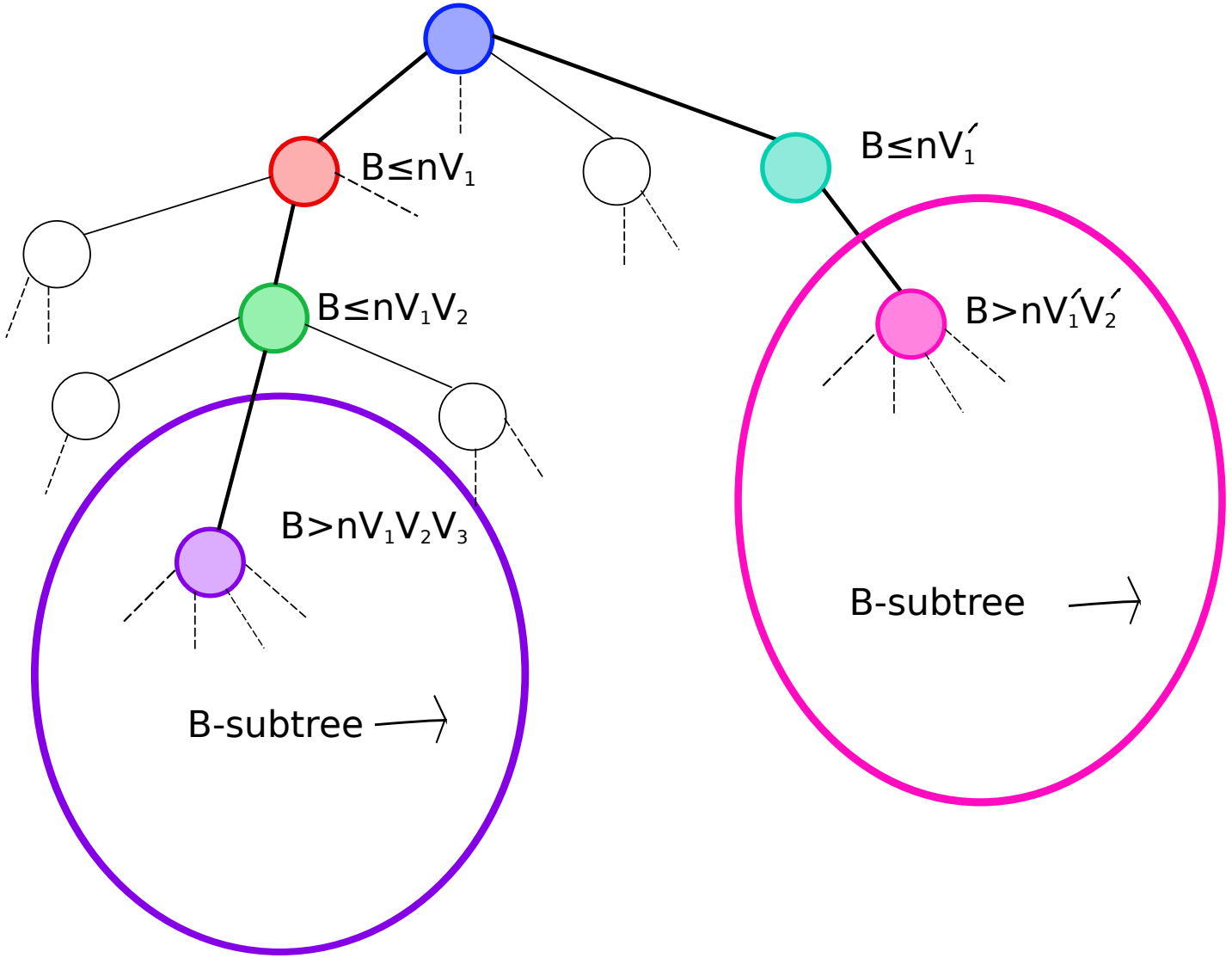
Thus, the sequence $(n^{-1}\mathbf{E}[\Psi(T^n)] - \mu^{-1} \ln n, n \geq 0)$ is Cauchy, implying $\mathbf{E}(\Psi(T^n)) = \mu^{-1} n \ln n + cn + o(n)$, for some constant c. If $-\ln V$ has a lattice distribution instead we can consider certain sub sequences and show that (4) holds for them. **The main tool for showing (4) is Renewal Theory.**

An Equivalent Definition of the Total Path Length

- The total path length can also be defined as the sum of all proper subtree sizes, i.e., $\Psi(T^n) = \sum_{v \neq \sigma} n_v$.



B-Subtrees



Decomposition of $\Psi(T^n)$ Using B-Subtrees

- The B-subtrees formally defined as $T_r,\ r \in R$ give the following decomposition of the total path length

$$\mathbf{E}[\Psi(T^n)] = \mathbf{E}\left[\sum_{v \neq \sigma} n_v \mathbf{1}_{\{nL_v \geq B\}}\right] + \mathbf{E}\left[\sum_{r \in R} \Psi(T^{n_r})\right]. \quad (5)$$

since the distribution of the total path length of T_r only depends on the number of items n_r.

Decomposition of $\Psi(T^n)$ Using B-Subtrees

- The B-subtrees formally defined as $T_r,\ r \in R$ give the following decomposition of the total path length

$$\mathbf{E}[\Psi(T^n)] = \mathbf{E}\left[\sum_{v \neq \sigma} n_v \mathbf{1}_{\{nL_v \geq B\}}\right] + \mathbf{E}\left[\sum_{r \in R} \Psi(T^{n_r})\right]. \qquad (5)$$

since the distribution of the total path length of T_r only depends on the number of items n_r.

- **Both sums are treated by using arguments from renewal theory.**

Decomposition of $\Psi(T^n)$ Using B-Subtrees

- The B-subtrees formally defined as $T_r,\ r \in R$ give the following decomposition of the total path length

$$\mathbf{E}[\Psi(T^n)] = \mathbf{E}\left[\sum_{v \neq \sigma} n_v \mathbf{1}_{\{nL_v \geq B\}}\right] + \mathbf{E}\left[\sum_{r \in R} \Psi(T^{n_r})\right]. \qquad (5)$$

  since the distribution of the total path length of T_r only depends on the number of items n_r.
- **Both sums are treated by using arguments from renewal theory.**
- The first sum can be approximated by a sum of $i.i.d.$ random variables due to the fact that n_v is well approximated by $nL_v = n\prod_{j=1}^{d} V_j$.

Decomposition of $\Psi(T^n)$ Using B-Subtrees

- The B-subtrees formally defined as $T_r,\ r \in R$ give the following decomposition of the total path length

$$\mathbf{E}[\Psi(T^n)] = \mathbf{E}\left[\sum_{v \neq \sigma} n_v \mathbf{1}_{\{nL_v \geq B\}}\right] + \mathbf{E}\left[\sum_{r \in R} \Psi(T^{n_r})\right]. \qquad (5)$$

  since the distribution of the total path length of T_r only depends on the number of items n_r.
- **Both sums are treated by using arguments from renewal theory.**
- The first sum can be approximated by a sum of $i.i.d.$ random variables due to the fact that n_v is well approximated by $nL_v = n\prod_{j=1}^{d} V_j$.
- The number of items n_r in the subtrees T_r, $r \in R$, can be estimated using the overshoot in renewal theory.

Proving Theorem 2: Precise asymptotics for the average path length

- Recall that we aim to prove Theorem 2 by showing that for two arbitrary values n and $\widehat{n}$, where $\widehat{n} \geq n$ and $n \to \infty$,

$$\left| \left(\frac{\mathbf{E}[\Psi(T^n)]}{n} - \mu^{-1} \ln n \right) - \left(\frac{\mathbf{E}[\Psi(T^{\widehat{n}})]}{\widehat{n}} - \mu^{-1} \ln \widehat{n} \right) \right| = o(1). \tag{6}$$

Proving Theorem 2: Precise asymptotics for the average path length

- Recall that we aim to prove Theorem 2 by showing that for two arbitrary values n and $\widehat{n}$, where $\widehat{n} \geq n$ and $n \to \infty$,

$$\left| \left(\frac{\mathbf{E}[\Psi(T^n)]}{n} - \mu^{-1} \ln n \right) - \left(\frac{\mathbf{E}[\Psi(T^{\widehat{n}})]}{\widehat{n}} - \mu^{-1} \ln \widehat{n} \right) \right| = o(1). \tag{6}$$

- We will prove this by showing that for $v \neq \sigma$,

$$\left| \frac{\mathbf{E}\left[\sum_v n_v \mathbf{1}_{\{nL_v \geq B\}}\right]}{n} - \frac{\ln n}{\mu} - \frac{\mathbf{E}\left[\sum_v n_v \mathbf{1}_{\{nL_v \geq B\}}\right]}{\widehat{n}} + \frac{\ln \widehat{n}}{\mu} \right| = o(1). \tag{7}$$

and

$$\left| \frac{\mathbf{E}\left[\sum_{r \in R} \Psi(T^{n_r})\right]}{n} - \frac{\mathbf{E}\left[\sum_{r \in R} \Psi(T^{\widehat{n}_r})\right]}{\widehat{n}} \right| = o(1). \tag{8}$$

Renewal Theory

Study sums $S_k = \sum_i^k X_i$ of $i.i.d$ random variables.

- A light bulb has a random life time X_1 with some distribution and when it breaks it has to be replaced by a new light bulb with a life time X_2 of the same distribution.

Renewal Theory

Study sums $S_k = \sum_i^k X_i$ of $i.i.d$ random variables.

- A light bulb has a random life time X_1 with some distribution and when it breaks it has to be replaced by a new light bulb with a life time X_2 of the same distribution.
- How many light bulbs are needed say in 1 year time?

Renewal Theory

Study sums $S_k = \sum_i^k X_i$ of $i.i.d$ random variables.

- A light bulb has a random life time X_1 with some distribution and when it breaks it has to be replaced by a new light bulb with a life time X_2 of the same distribution.
- How many light bulbs are needed say in 1 year time?
- This number is a random variable called counting process $\mathcal{N}(t) := \max\{k : S_k \leq t\}$ in renewal theory.

Renewal Theory

Study sums $S_k = \sum_i^k X_i$ of $i.i.d$ random variables.

- A light bulb has a random life time X_1 with some distribution and when it breaks it has to be replaced by a new light bulb with a life time X_2 of the same distribution.
- How many light bulbs are needed say in 1 year time?
- This number is a random variable called counting process $\mathcal{N}(t) := \max\{k : S_k \leq t\}$ in renewal theory.
- Let $F(t) = \mathbf{P}(X_1 \leq t)$. The expected value of $\mathcal{N}(t)$ is given by the renewal function

$$V(t) = \mathbf{E}(\mathcal{N}(t)) := \sum_{k=1}^{\infty} \mathbf{P}(S_k \leq t)$$

$$F(t) + \int_0^t V(t-s)dF(s) = F(t) + (V * dF)(t).$$

Renewal Theory

Study sums $S_k = \sum_i^k X_i$ of $i.i.d$ random variables.

- A light bulb has a random life time X_1 with some distribution and when it breaks it has to be replaced by a new light bulb with a life time X_2 of the same distribution.
- How many light bulbs are needed say in 1 year time?
- This number is a random variable called counting process $\mathcal{N}(t) := \max\{k : S_k \leq t\}$ in renewal theory.
- Let $F(t) = \mathbf{P}(X_1 \leq t)$. The expected value of $\mathcal{N}(t)$ is given by the renewal function

$$V(t) = \mathbf{E}(\mathcal{N}(t)) := \sum_{k=1}^{\infty} \mathbf{P}(S_k \leq t)$$

$$F(t) + \int_0^t V(t-s)dF(s) = F(t) + (V * dF)(t).$$

- Law of large numbers suggests $V(t) = \frac{t}{\mathbf{E}(X)}$+o(t).

Applying Renewal Theory

- Recall that the subtree sizes n_v for v at depth d are approximated by $n \prod_{j=1}^{d} V_j$.

Applying Renewal Theory

- Recall that the subtree sizes n_v for v at depth d are approximated by $n \prod_{j=1}^{d} V_j$.
- Let $S_d := -\sum_{j=1}^{d} \ln V_j$. Note that $n \prod_{j=1}^{d} V_j = n e^{-S_d}$.

Applying Renewal Theory

- Recall that the subtree sizes n_v for v at depth d are approximated by $n\prod_{j=1}^{d} V_j$.
- Let $S_d := -\sum_{j=1}^{d} \ln V_j$. Note that $n\prod_{j=1}^{d} V_j = ne^{-S_d}$.
- Define the renewal function

$$U(t) := \sum_{k=1}^{\infty} b^k \mathbf{P}(S_k \leq t), \tag{9}$$

and let $F(t) := b\mathbf{P}(-\ln V \leq t)$.

Applying Renewal Theory

- Recall that the subtree sizes n_v for v at depth d are approximated by $n\prod_{j=1}^{d} V_j$.
- Let $S_d := -\sum_{j=1}^{d} \ln V_j$. Note that $n\prod_{j=1}^{d} V_j = ne^{-S_d}$.
- Define the renewal function

$$U(t) := \sum_{k=1}^{\infty} b^k \mathbf{P}(S_k \leq t), \tag{9}$$

  and let $F(t) := b\mathbf{P}(-\ln V \leq t)$.
- For $U(t)$ we obtain the following renewal equation

$$U(t) = F(t) + (U * dF)(t). \tag{10}$$

Applying Renewal Theory

- Recall that the subtree sizes n_v for v at depth d are approximated by $n\prod_{j=1}^{d} V_j$.
- Let $S_d := -\sum_{j=1}^{d} \ln V_j$. Note that $n\prod_{j=1}^{d} V_j = ne^{-S_d}$.
- Define the renewal function

$$U(t) := \sum_{k=1}^{\infty} b^k \mathbf{P}(S_k \leq t), \tag{9}$$

  and let $F(t) := b\mathbf{P}(-\ln V \leq t)$.
- For $U(t)$ we obtain the following renewal equation

$$U(t) = F(t) + (U * dF)(t). \tag{10}$$

- As $t \to \infty$, $U(t)$ satisfies

$$U(t) = (\mu^{-1} + o(1))e^t. \tag{11}$$

Contribution of the Top

- In the top the sizes n_v are well approximated by $n \prod_{i=1}^{k} V_i$.

Contribution of the Top

- In the top the sizes n_v are well approximated by $n\prod_{i=1}^{k} V_i$.
- Let V_i, $i \geq 1$ be i.i.d. copies of V. Let $L_k = \prod_{i=1}^{k} V_i$ and $S_k = -\ln L_k$. Then, renewal theory for $U(t)$ gives

$$\mathbf{E}\left[\sum_v n_v \mathbf{1}_{\{nL_v \geq B\}}\right] = n\int_0^{\ln(n/B)} e^{-t} dU(t) \tag{12}$$

$$= \frac{1}{\mu} n \ln\left(\frac{n}{B}\right) + n\frac{\sigma^2 - \mu^2}{2\mu^2} + o(n). \tag{13}$$

Contribution of the Top

- In the top the sizes n_v are well approximated by $n \prod_{i=1}^{k} V_i$.
- Let V_i, $i \geq 1$ be i.i.d. copies of V. Let $L_k = \prod_{i=1}^{k} V_i$ and $S_k = -\ln L_k$. Then, renewal theory for $U(t)$ gives

$$\mathbf{E}\left[\sum_v n_v \mathbf{1}_{\{nL_v \geq B\}}\right] = n \int_0^{\ln(n/B)} e^{-t} dU(t) \tag{12}$$

$$= \frac{1}{\mu} n \ln\left(\frac{n}{B}\right) + n\frac{\sigma^2 - \mu^2}{2\mu^2} + o(n). \tag{13}$$

- Thus,

$$\left| \frac{\mathbf{E}\left[\sum_v n_v \mathbf{1}_{\{nL_v \geq B\}}\right]}{n} - \frac{\ln n}{\mu} - \frac{\mathbf{E}\left[\sum_v n_v \mathbf{1}_{\{nL_v \geq B\}}\right]}{\widehat{n}} + \frac{\ln \widehat{n}}{\mu} \right| = o(1). \tag{14}$$

B-subtrees: Convergence of the Overshoot

- For the distribution function of n_r we shall again apply renewal theory.

B-subtrees: Convergence of the Overshoot

- For the distribution function of n_r we shall again apply renewal theory.
- We approximate n_r by the product $n \prod_{j=1}^{d(r)} V_j$.

B-subtrees: Convergence of the Overshoot

- For the distribution function of n_r we shall again apply renewal theory.
- We approximate n_r by the product $n \prod_{j=1}^{d(r)} V_j$.
- The idea is to show that the distribution is independent of which n we started with.

B-subtrees: Convergence of the Overshoot

- For the distribution function of n_r we shall again apply renewal theory.
- We approximate n_r by the product $n \prod_{j=1}^{d(r)} V_j$.
- The idea is to show that the distribution is independent of which n we started with.
- It is important that $B << n$ for the error terms to be ignored.

B-subtrees: Convergence of the Overshoot

- For the distribution function of n_r we shall again apply renewal theory.
- We approximate n_r by the product $n\prod_{j=1}^{d(r)} V_j$.
- The idea is to show that the distribution is independent of which n we started with.
- It is important that $B << n$ for the error terms to be ignored.
- Recall that $S_d := -\sum_{j=1}^{d} \ln V_j$. The renewal counting process $\mathcal{N}(t)$ is defined as

$$\mathcal{N}(t) := \max\{k : S_k \leq t\}. \tag{15}$$

The residual lifetime or overshoot $\mathcal{R}(t)$ is defined as

$$\mathcal{R}(t) := S_{\mathcal{N}(t)+1} - t. \tag{16}$$

A well-known result in renewal theory is that $\mathcal{R}(t)$ converges in distribution as $t \to \infty$.

Convergence of Overshoot (cont.)

- Let $q = \ln(\frac{n}{B})$ For $x \in \{\frac{1}{B}, \frac{2}{B}, \ldots, \frac{B}{B}\}$ we have

$$\mathbf{P}\Big(\frac{n\prod_{j=1}^{d(r)} V_j}{B} < x\Big) = 1 - \mathbf{P}\Big(\mathcal{R}(q) < -\ln x\Big), \qquad (17)$$

  which converges as $q \to \infty$.

Convergence of Overshoot (cont.)

- Let $q = \ln(\frac{n}{B})$ For $x \in \{\frac{1}{B}, \frac{2}{B}, \ldots, \frac{B}{B}\}$ we have

$$\mathbf{P}\left(\frac{n\prod_{j=1}^{d(r)} V_j}{B} < x\right) = 1 - \mathbf{P}\Big(\mathcal{R}(q) < -\ln x\Big), \qquad (17)$$

  which converges as $q \to \infty$.
- This means that the distribution of the sizes n_r is independent of which n we started with.

Distribution of B-Subtrees Independent of n

- Since almost all n items are in the $T_r,\ r \in R$ subtrees (the B-subtrees),

$$\mathbf{E}\left[\sum_{r\in R}\Psi(T^{n_r})\right] \approx \frac{n}{\mathbf{E}\left[n_r\right]}\mathbf{E}\left[\Psi(T^{n_r})\right]. \tag{18}$$

Distribution of B-Subtrees Independent of n

- Since almost all n items are in the $T_r,\ r \in R$ subtrees (the B-subtrees),

$$\mathbf{E}\left[\sum_{r\in R}\Psi(T^{n_r})\right] \approx \frac{n}{\mathbf{E}\left[n_r\right]}\mathbf{E}\left[\Psi(T^{n_r})\right]. \tag{18}$$

- Since the distribution of the sizes of the B-subtrees is not depending on n, up to small error terms,

$$\mathbf{E}\left[\sum_{r\in R}\Psi(T^{n_r})\right] = nf(B) + o(n). \tag{19}$$

Distribution of B-Subtrees Independent of n

- Since almost all n items are in the T_r, $r \in R$ subtrees (the B-subtrees),

$$\mathbf{E}\left[\sum_{r\in R}\Psi(T^{n_r})\right] \approx \frac{n}{\mathbf{E}[n_r]}\mathbf{E}\left[\Psi(T^{n_r})\right]. \tag{18}$$

- Since the distribution of the sizes of the B-subtrees is not depending on n, up to small error terms,

$$\mathbf{E}\left[\sum_{r\in R}\Psi(T^{n_r})\right] = nf(B) + o(n). \tag{19}$$

- Hence, for two arbitrary values n and $\widehat{n}$, where $\widehat{n} \geq n$,

$$\lim_{n\to\infty}\left|\frac{\mathbf{E}\left[\sum_{r\in R}\Psi(T^{n_r})\right]}{n} - \frac{\mathbf{E}\left[\sum_{r\in R}\Psi(T^{\widehat{n}_r})\right]}{\widehat{n}}\right| = o(1). \tag{20}$$

Conclusions

- **Renewal theory was used to prove precise asymptotics of the average total path length $\Psi(T^n)$ of random split trees (a large class of random trees of logarithmic height).** More precisely it was shown that

$$\mathbf{E}[\Psi(T^n)] = \mu^{-1} n \ln n + n\varpi(\ln n) + o(n). \qquad (21)$$

where μ is a constant and ϖ is a continuous periodic function depending on the split vector, which is a constant if $-\ln V$ has a non-lattice distribution.

Conclusions

- **Renewal theory was used to prove precise asymptotics of the average total path length $\Psi(T^n)$ of random split trees (a large class of random trees of logarithmic height).** More precisely it was shown that

$$\mathbf{E}[\Psi(T^n)] = \mu^{-1} n \ln n + n\varpi(\ln n) + o(n). \qquad (21)$$

  where μ is a constant and ϖ is a continuous periodic function depending on the split vector, which is a constant if $-\ln V$ has a non-lattice distribution.

- **This result in (21) enabled the use of the contraction method to prove that the total path length $\Psi(T^n)$ in a general split tree converges in distribution to a random variable characterized by some fixed point equation.**

Thank you for listening!