Analysis of the total costs for variants of the Union-Find algorithm

Markus Kuba, Alois Panholzer

The Height of List-tries and TST

2007 Conference on Analysis of Algorithms, AofA 07 DMTCS proc. **AH**, 2007, 283–294

Analysis of the total costs for variants of the Union-Find algorithm

Markus Kuba[1] and Alois Panholzer[1†]

[1]*Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Wiedner Hauptstr. 8-10/104, A-1040 Wien, Austria.* {Markus.Kuba,Alois.Panholzer}@tuwien.ac.at

received 19 Feb 2007, revised 19th January 2008, accepted tomorrow.

We study the average behavior of variants of the UNION-FIND algorithm to maintain partitions of a finite set under the random spanning tree model. By applying the method of moments we can characterize the limiting distribution of the total costs of the algorithms "Quick Find Weighted" and "Quick Find Biased" extending the analysis of Knuth and Schönhage, Yao, and Chassaing and Marchand.

Keywords: Union-Find algorithm, average-case analysis, limiting distribution

Contents

1 Introduction

The so-called "Union-Find problem" (see (AHU74)) consists of maintaining a representation of equivalence classes or partitions of a finite set, such that the following two basic operations have to be supported, UNION: "merge two different equivalence classes s and t into a single equivalence class" and FIND: "find

†This work has been supported by the Austrian Science Foundation FWF, grant S9608-N13.

the equivalence class that contains a given element x". This problem arises naturally in several applications in computer science as, e.g., in minimum-cost spanning tree algorithms and algorithms for detecting the equivalence of finite automata.

Following (AHU74) the Union-Find problem for partitions $P(S)$ of a finite set S can be treated by introducing the following data structure:

> For every element $x \in S$ we store in $R[x]$ the name of the equivalence class containing x. Furthermore for every equivalence class $s \in P(S)$ we store in $N[s]$ the number of elements of s and in $L[s]$ we store the elements of s in a linked list.

(Yao76) has described two basic algorithms for implementing the operation UNION:

"Quick Find Weighted" (QFW): If we want to merge the different equivalence classes s and t then we update the class with less elements:

> if $N[s] \leq N[t]$ then set $R[x] := t$ for all x in $L[s]$, append $L[s]$ to $L[t]$, set $N[t] := N[t] + N[s]$ and call the new equivalence class t, otherwise set $R[x] := s$ for all x in $L[t]$, append $L[t]$ to $L[s]$, set $N[s] := N[s] + N[t]$ and call the new equivalence class s.

"Quick Find" (QF): If we want to merge the different equivalence classes s and t then we update one of the two classes *at random* according to the procedure described above, but we do not make use of the information stored in $N[s]$ and $N[t]$.

In (CM04) another variant of this algorithm is considered, which is of interest in some coalescence models:

"Quick Find Biased" (QFB): If we want to merge the different equivalence classes s and t then we update one of the two classes according to the procedure described above, where the probability that s is updated is given by $\frac{N[t]}{N[s]+N[t]}$ and the probability that t is updated is given by $\frac{N[s]}{N[s]+N[t]}$. So the smaller the size of the equivalence class the higher is the probability that it is updated.

The cost of the UNION-operation when merging the equivalence classes s and t can be measured by the number of updated elements, i.e., the number of allocations $R[x] := s$ (or $R[x] := t$). For QFW the cost of one merging step is thus given by the minimum of the class sizes $\min(N[s], N[t])$, whereas for QF the cost is given by $N[s]$ or $N[t]$ with equal probability $\frac{1}{2}$. For QFB the cost is given by $N[s]$ with probability $\frac{N[t]}{N[s]+N[t]}$ and by $N[t]$ with probability $\frac{N[s]}{N[s]+N[t]}$. When applying one of these algorithms the FIND-operation for an element x, i.e., finding the equivalence class where x is contained, simply consists in evaluating $R[x]$ and can thus be carried out in bounded time.

In order to measure the average behavior of the algorithms described above the following two models for sequences of UNION-operations have been introduced in (Yao76): the *random graph model* and the *random spanning tree model*. In both models we deal with a set S of size n, where at the beginning all elements $x \in S$ are forming an equivalence class $\{x\}$. These n equivalence classes will now be merged into larger and larger classes by carrying out the UNION-operations as described below.

In the random spanning tree model a spanning tree of the complete graph with vertex set S is chosen at random and then the edges of this spanning tree are randomly ordered, i.e., enumerated from 1 to $n-1$. Let us assume this leads to a sequence of edges $e_1 = (x_1, y_1)$, $e_2 = (x_2, y_2)$, $\ldots$, $e_{n-1} = (x_{n-1}, y_{n-1})$,

with $x_i, y_i \in S$. This gives then the following sequence of UNION-operations: UNION$(R[x_1], R[y_1])$, UNION$(R[x_2], R[y_2])$, ..., UNION$(R[x_{n-1}], R[y_{n-1}])$. Thus in this model all $n^{n-2}(n-1)!$ possible sequence of UNION-operations of that kind are equally likely.

Although we will not analyze the random graph model we will also give a brief description here for completeness. The random graph model can be described as follows. Let us assume we have already carried out a sequence of $i-1$ UNION-operations leading to some partition $P_i(S)$. We consider now the complete graph with vertex set S and consider further the set of all edges between nodes lying in different equivalence classes: $E_i = \{e = (x,y) : R[x] \neq R[y]\}$ (the edges connecting nodes, which are already in the same equivalence class are no more considered). Then in this model we chose for the i-th UNION-operation one of the edges $e = (x,y)$ in E_i at random and carry out UNION$(R[x], R[y])$.

The basic parameter of interest describing the average performance of the algorithms QFW, QF and QFB is then the total cost, i.e., the sum of the cost of every merging step, of merging the elements of a set S of size n, where at the beginning all elements are lying in different equivalence classes, into one equivalence class (containing all elements of S) by carrying out a sequence of $n-1$ UNION-operations according to the rules given in the random spanning tree model or the random graph model. This parameter, which can be considered as a random variable depending only on the size n of the set S of elements, is denoted by $X_n^{[QFW]}$, $X_n^{[QF]}$ or $X_n^{[QFB]}$. The QFW algorithm under the random spanning tree model is illustrated by an example in Figure 1.

The algorithms QFW and QF have been analyzed first by (Yao76) and (KS78) for both models described above studying the expected value $\mathbb{E}(X_n^{[QFW]})$ and $\mathbb{E}(X_n^{[QF]})$. For the random spanning tree model it was shown in (Yao76) that $\mathbb{E}(X_n^{[QFW]}) = \Theta(n \log n)$ and $\mathbb{E}(X_n^{[QF]}) = \Theta(n^{\frac{3}{2}})$. In (KS78) the following refined results are obtained by developing and applying the so-called "repertoire approach":

$$\mathbb{E}(X_n^{[QFW]}) = \frac{1}{\pi} n \log n + \mathcal{O}(n), \quad \text{and} \quad \mathbb{E}(X_n^{[QF]}) = \sqrt{\frac{\pi}{8}} n^{\frac{3}{2}} + \frac{1}{4} n \log n + \mathcal{O}(n). \tag{1}$$

Substantial progress on $X_n^{[QF]}$ has been made in (CM04) by establishing relations with Hashing with Linear Probing and using results from (FPV98) leading to the following characterization of the limiting distribution: $n^{-\frac{3}{2}} X_n^{[QF]} \xrightarrow{(d)} \int_0^1 e(t)dt$, where $(e(t))_{0 \leq t \leq 1}$ denotes the normalized Brownian excursion. With $X_n \xrightarrow{(d)} X$ we denote in this paper always the convergence in distribution of a sequence of random variables X_n to a random variable X, whereas $X \stackrel{(d)}{=} Y$ denotes equality in distribution of r.v. X and Y. For the algorithm QFW it has been conjectured in (CM04) a concentration result, namely that $\frac{X_n^{[QFW]}}{n \log n}$ converges in the $\mathcal{L}_2$-metric to $\frac{1}{\pi}$. For the algorithm QFB introduced and analyzed in (CM04) the corresponding concentration result has been proven: $\frac{X_n^{[QFB]}}{n \log n}$ converges in the $\mathcal{L}_2$-metric to $\frac{1}{2}$.

The aim of this paper is to describe the behavior of the total costs of the algorithms QFW and QFB for large n under the random spanning tree model by characterizing the limiting distribution of $X_n^{[QFW]}$ and $X_n^{[QFB]}$. As a consequence of our analysis the concentration result for QFW stated above also follows immediately. These results are given in Section 2. A brief sketch of the proof of the results for the QFW and the QFB algorithm are given in Section 3 and Section 4.

We only want to remark that the algorithms QFW and QF have a completely different behavior under the random graph model, which is a consequence of the analysis of (KS78) and (BS93) for the expected

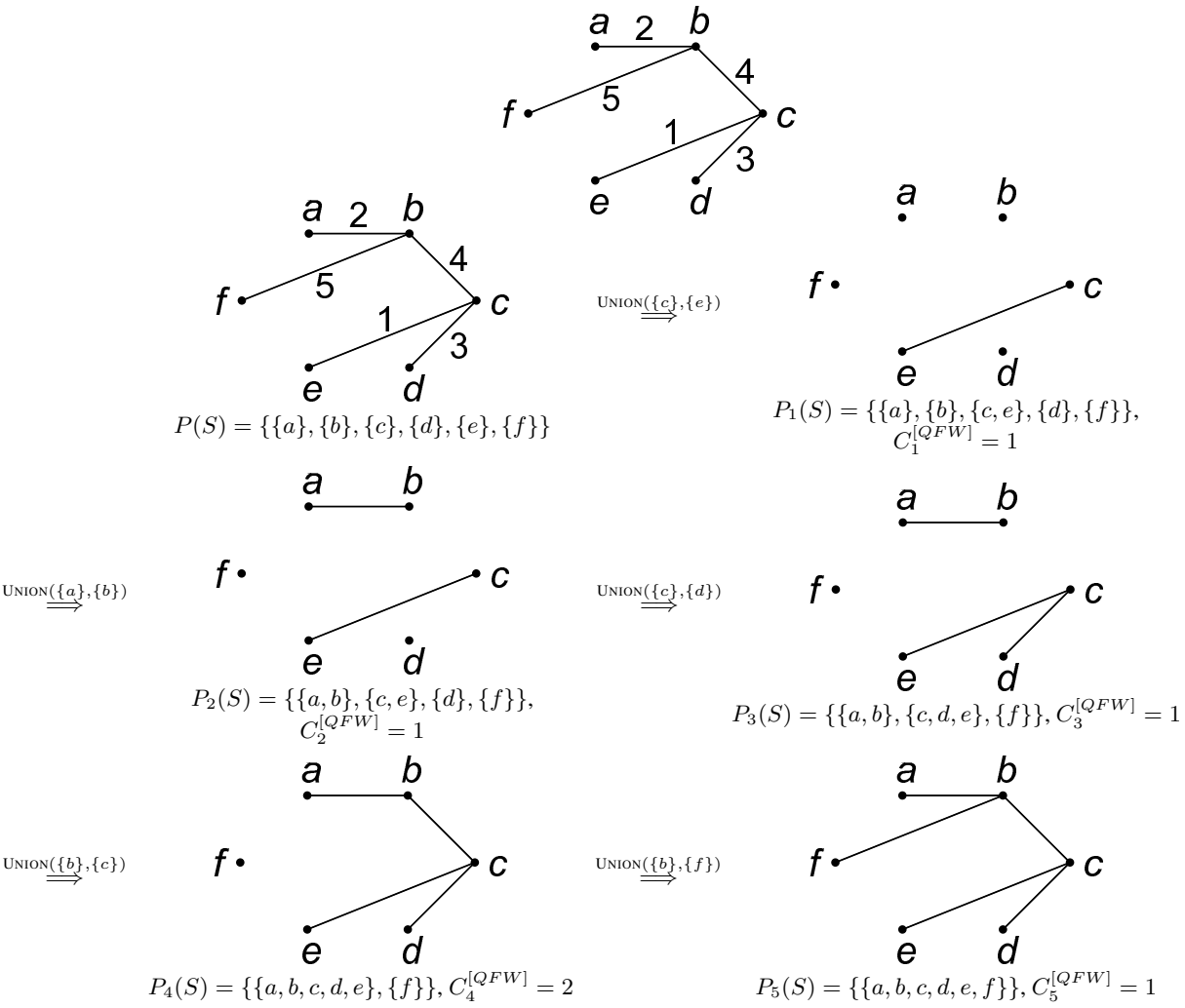

Fig. 1: Choosing the particular spanning tree given in the example the QFW algorithm has total cost $X^{[QFW]} = \sum_{i=1}^{5} C_i^{[QFW]} = 6$ to merge the elements $S = \{a, b, \ldots, f\}$ starting with the partition $P(S) = \{\{a\},\{b\},\ldots,\{f\}\}$. Here $C_i^{[QFW]}$ denotes the cost of the i-th merging step of the QFW algorithm.

values of the total costs: $\mathbb{E}(X_n^{[QF]}) \sim \frac{n^2}{8}$ and $\mathbb{E}(X_n^{[QFW]}) \sim cn$, where the constant c appearing can be described explicitly.

2 The results

Theorem 1 *Let $X_n^{[QFW]}$ denote the total cost of the algorithm "Quick Find Weighted" QFW to merge all elements of a finite set S of size n under the random spanning tree model.*

Then the expected value of $X_n^{[QFW]}$ has for $n \to \infty$ the following asymptotic expansion:

$$\mathbb{E}(X_n^{[QFW]}) = \frac{1}{\pi} n \log n + Cn + \mathcal{O}(n^{\frac{3}{4}}), \tag{2}$$

with a certain constant $C \approx 0.6315$, which is given as follows:

$$C = \frac{\gamma + 2\log 2}{\pi} + \sum_{n\geq 0} \frac{1}{n+1}\Big[e^{-(n+1)}\Big(R_{n+2} - R_{n+1} - \sum_{k=0}^{n} \frac{(k+1)^{k+1}}{(k+2)!} R_{n-k}\Big) - \frac{1}{\pi}\Big],$$

with

$$R_n = \sum_{k=1}^{n-1} \frac{k^k (n-k)^{n-k-1}}{k!(n-k)!} \min(k, n-k).$$

The suitably shifted and scaled r.v. $X_n^{[QFW]}$ converges in distribution to a r.v. X, which can be characterized by its r-th integer moments:

$$\frac{X_n^{[QFW]} - \frac{1}{\pi} n \log n - Cn}{n} \xrightarrow{(d)} X, \quad \textit{with} \quad \mathbb{E}(X^r) = m_r,$$

where m_r is given recursively as follows:

$$m_r = \frac{\Gamma(r-1)}{2\sqrt{\pi}\Gamma(r-\frac{1}{2})} \sum_{\substack{r_1+r_2+r_3=r,\\ r_2,r_3<r}} \binom{r}{r_1, r_2, r_3} m_{r_2} m_{r_3} I_{r_1,r_2,r_3}, \quad \textit{for } r \geq 2,$$

with initial values $m_0 = 1$ and $m_1 = 0$ and

$$I_{r_1,r_2,r_3} = \int_0^1 \Big(\frac{1}{\pi}\big(x\log x + (1-x)\log(1-x)\big) + \min(x, 1-x)\Big)^{r_1} x^{r_2-\frac{1}{2}} (1-x)^{r_3-\frac{3}{2}} dx.$$

Theorem 2 *Let $X_n^{[QFB]}$ denote the total cost of the algorithm "Quick Find Biased" QFB to merge all elements of a finite set S of size n under the random spanning tree model.*

Then the expected value and the variance of $X_n^{[QFB]}$ are asymptotically for $n \to \infty$ given as follows:

$$\mathbb{E}(X_n^{[QFB]}) = \frac{1}{2} n \log n + \frac{\gamma + \log 2}{2} n + \mathcal{O}(\sqrt{n}), \quad \mathbb{V}(X_n^{[QFB]}) = (\frac{3}{2} - \frac{\pi^2}{8}) n^2 + \mathcal{O}(n^{\frac{3}{2}} \log^2 n). \tag{3}$$

The suitably shifted and scaled r.v. $X_n^{[QFB]}$ converges in distribution to a r.v. X, which can be characterized by its r-th integer moments:

$$\frac{X_n^{[QFB]} - \frac{1}{2}n\log n - \frac{\gamma+\log 2}{2}n}{n} \xrightarrow{(d)} X, \quad \textit{with} \quad \mathbb{E}(X^r) = m_r,$$

where m_r is given recursively as follows:

$$m_r = \frac{\Gamma(r-1)}{2\sqrt{\pi}\Gamma(r-\frac{1}{2})} \sum_{\substack{r_1+r_2+r_3=r,\\ r_2,r_3<r}} \binom{r}{r_1, r_2, r_3} m_{r_2} m_{r_3} I_{r_1,r_2,r_3}, \quad \textit{for } r \geq 2,$$

with inital values $m_0 = 1$ and $m_1 = 0$ and

$$I_{r_1,r_2,r_3} = \int_0^1 \Big((1-x)\Big[\frac{1}{2}\big(x\log x + (1-x)\log(1-x)\big) + x\Big]^{r_1} + x\Big[\frac{1}{2}\big(x\log x + (1-x)\log(1-x)\big) + 1 - x\Big]^{r_1}\Big) x^{r_2-\frac{1}{2}}(1-x)^{r_3-\frac{3}{2}}\,dx.$$

3 Sketch of the proof of the results for QFW

3.1 General remarks

In order to analyze the total cost in the merging algorithms described above under the random spanning tree model two main approaches have been used. The first one uses a description via a random coagulation model for particles: it has been pointed out in (CM04) that the random spanning tree model corresponds to the additive Marcus-Lushnikov process. The second one considers the "inverse process": instead of merging equivalence classes by carrying out UNION-operations and thus adding successively edges until we obtain a spanning tree, we start with a random spanning tree and remove successively edges until all nodes are isolated. The basis of the approach is the following simple fact. Let us assume we start with a random unrooted labeled tree of size n (this corresponds to the random spanning tree of the complete graph of a set S of size n) and remove one edge at random (this corresponds to the edge, which has been added in the final, i.e., the $n-1$-th, merging step). Then it holds that both resulting subtrees, let us assume they are of sizes k and $n-k$, with $1 \leq k \leq n-1$, are itself *random* unrooted labeled trees of smaller sizes k and $n-k$, respectively. In the QFW algorithm the cost of this edge-removal step is given by $\min(k, n-k)$, in the QF algorithm the cost is k or $n-k$ each with probability $\frac{1}{2}$. In the QFB algorithm the cost is k with probability $\frac{n-k}{k}$ and $n-k$ with probability $\frac{k}{n}$. This decomposition of the problem gives rise to a recursive approach, which has been introduced in (KS78).

An analysis of the algorithms QFW and QFB (and also QF) can be carried out by studying the distributional recurrence (4) for some deterministic (QFW) or non-deterministic (QFB and QF) toll function $t_{n,k}$. Strictly speaking the recurrence (4) describes the edge-removal procedure for *rooted* labeled trees, but since the costs studied here are independent of the actual root of the tree, this makes absolutely no difference. A treatment of (4) has been given in (FKP06) for deterministic toll functions $t_n = t_{n,k} = n^\alpha$, with $\alpha > 0$, by applying extensions of singularity analysis to the Hadamard-product of generating functions, see (FFK05), and using the method of moments. However for an analysis of the total cost of QFW and QFB two difficulties are appearing: (i) in the case of QFW the toll function is of a kind, such that

singularity analysis (with its extensions) is not directly applicable (the analytic behavior of the generating functions including $\min(k, n-k)$-terms is not obvious) and (ii) for QFW and QFB the toll function $t_{n,k}$ is dependent on two parameters, namely the size n of the tree and the size k of the subtree after the edge-removal, for QFB the toll function is not even deterministic but a random variable.

These difficulties lead us to a different approach, where generating functions are only used to obtain an exact solution of the recurrence for the moments of $X_n^{[QFW]}$ and $X_n^{[QFB]}$. The asymptotic behavior of these explicit solutions is then extracted by "classical real analysis", i.e., dissecting summation intervals, approximating sums by integrals, etc., see also (Pan04), where similar methods have been applied. The analytic considerations required to show our results are not included in this extended abstract, but it is planned to give them, together with a refined analysis of related algorithms studied in (Yao76) and (KS78): "Quick Merge" and "Quick Merge Weighted", in the full version.

3.2 The recurrence

The total cost $X_n := X_n^{[QFW]}$ of the algorithm QFW under the random spanning tree model, when merging the elements of a set of size n, satisfy, for $n \geq 2$, the following distributional recurrence (with $X_1 = 0$):

$$X_n \stackrel{(d)}{=} X_{S_n} + X^*_{n-S_n} + t_{n,S_n}, \quad \text{for } n \geq 2, \tag{4}$$

where S_n, which describes the distribution of the size of the subtree containing the root node after removing a random edge of a randomly chosen labeled rooted tree of size n, is independent of $(X_j)_{j\geq 1}$ and $(X_j^*)_{j\geq 1}$, which are independent copies of each other. The toll function $t_{n,k}$ is for QFW given by

$$t_{n,k} := \min(k, n-k).$$

Furthermore S_n is distributed as follows:

$$\mathbb{P}\{S_n = k\} = \frac{kT_kT_{n-k}}{(n-1)T_n}, \quad \text{for } 1 \leq k \leq n-1, \tag{5}$$

where we use throughout this paper the notation $T_n := \frac{n^{n-1}}{n!}$, for $n \geq 1$, and $T_0 = 0$.

3.3 The expectation

It is crucial for our approach to obtain a detailed asymptotic expansion of the expectation $\mu_n := \mathbb{E}(X_n)$ refining the corresponding one in (1).

Due to the distributional recurrence (4) the expectation μ_n satisfies the following recurrence:

$$(n-1)T_n\mu_n = \sum_{k=1}^{n-1} kT_kT_{n-k}(\mu_k + \mu_{n-k}) + R_n, \quad \text{for } n \geq 2, \tag{6}$$

with $\mu_1 = 0$ and $R_n = \sum_{k=1}^{n-1} kT_kT_{n-k}\min(k, n-k)$.

To treat recurrence (6) we use a generating functions approach leading to an explicitly solvable differential equation. Introducing the generating functions

$$C(z) := \sum_{n\geq 1} T_n\mu_n z^n, \quad R(z) := \sum_{n\geq 1} R_n z^n, \quad T(z) := \sum_{n\geq 1} T_n z^n,$$

one obtains the differential equation

$$z(1-T(z))C'(z) - (1+zT'(z))C(z) = R(z), \tag{7}$$

with initial condition $C'(0) = 0$. Using the functional equation $T(z) = ze^{T(z)}$ satisfied by the tree function $T(z)$ one can describe the solution of equation (7) in the following compact form:

$$C(z) = \frac{T(z)}{1-T(z)} \int_0^z \frac{R(t)}{tT(t)} dt. \tag{8}$$

The asymptotic behavior of the coefficients of $C(z)$ and thus of μ_n can be extracted from (8) "at the level of the coefficients" starting with the expansion $R_n = [z^n]R(z) = \frac{e^n}{\pi}\big(1 + \mathcal{O}(n^{-\frac{1}{2}})\big)$. A careful analysis leads then to the expansion given in (2), where we do not expect that the remainder bound given there is tight.

3.4 The higher moments

For establishing a limiting distribution result we consider for $n \geq 1$ the shifted random variable

$$\tilde{X}_n = X_n - \frac{1}{\pi} n \log n - Cn,$$

where the constant C is specified in (2). $\tilde{X}_n$ satisfies then the distributional recurrence:

$$\tilde{X}_n \stackrel{(d)}{=} \tilde{X}_{S_n} + \tilde{X}^*_{n-S_n} + \tilde{t}_{n,S_n}, \quad \text{for } n \geq 2, \tag{9}$$

with $X_1 = -C$ and where $\tilde{t}_{n,k}$ is given as follows:

$$\tilde{t}_{n,k} = \frac{1}{\pi}\big(k \log k + (n-k)\log(n-k) - n \log n\big) + \min(k, n-k).$$

S_n and the independence conditions are given as in the definition of X_n in Subsection 3.2.

The r-th moments $\tilde{\mu}_n^{[r]} := \mathbb{E}\big(\tilde{X}_n^r\big)$ of $\tilde{X}_n$ satisfy due to (9) for $r \geq 1$ the following recurrence:

$$(n-1)T_n \tilde{\mu}_n^{[r]} = \sum_{k=1}^{n-1} k T_k T_{n-k} \big(\tilde{\mu}_k^{[r]} + \tilde{\mu}_{n-k}^{[r]}\big) + \tilde{R}_n^{[r]}, \quad \text{for } n \geq 2, \tag{10}$$

with $\tilde{\mu}_1^{[r]} = (-C)^r$ and $\tilde{R}_n^{[r]} = \sum_{\substack{r_1+r_2+r_3=r,\\ r_2,r_3<r}} \binom{r}{r_1,r_2,r_3} \sum_{k=1}^{n-1} k T_k T_{n-k} \big(\tilde{t}_{n,k}\big)^{r_1} \tilde{\mu}_k^{[r_2]} \tilde{\mu}_{n-k}^{[r_3]}$. Of course, we have $\tilde{\mu}_n^{[0]} = 1$, for $n \geq 1$, and $\tilde{\mu}_n^{[1]} = \mathcal{O}(n^{\frac{3}{4}})$.

Since recurrence (10) is apart from the inhomogeneous part and the initial value the same as the recurrence (6) for the expectation μ_n, we can treat it again by introducing suitable generating functions:

$$\tilde{C}^{[r]}(z) := \sum_{n \geq 1} T_n \tilde{\mu}_n^{[r]} z^n, \quad \tilde{R}^{[r]}(z) := \sum_{n \geq 1} \tilde{R}_n^{[r]} z^n.$$

The solution of the differential equation appearing is then given as follows:

$$\tilde{C}^{[r]}(z) = \frac{T(z)}{1-T(z)} \int_0^z \frac{\tilde{R}^{[r]}(t)}{tT(t)} dt + \frac{\tilde{\mu}_1^{[r]} T(z)}{1-T(z)}. \tag{11}$$

Using the representation of $\tilde{C}^{[r]}(z)$ given in equation (11) one can show inductively the following asymptotic equivalents of the moments $\tilde{\mu}_n^{[r]}$:

$$\tilde{\mu}_n^{[r]} \sim m_r n^r,$$

where the coefficients m_r are defined in Theorem 1. In order to show that the sequence of moments $(m_r)_{r\geq 0}$ indeed characterizes the limiting distribution one has to show some growth estimates on m_r and applying Carleman's criterion. This can be done similar to (HN02) leading to estimates of the kind $m_r \leq r! K^r$, for all $r \geq 0$, with some constant K. This completes the proof of Theorem 1.

4 Sketch of the proof of the results for QFB

An analysis of the total cost $X_n := X_n^{[QFB]}$ of the algorithm QFB under the random spanning tree model, when merging the elements of a set of size n, can be carried out very similar to the corresponding one for QFW in Section 3.

One starts again with the distributional recurrence (with $X_1 = 0$):

$$X_n \stackrel{(d)}{=} X_{S_n} + X^*_{n-S_n} + t_{n,S_n}, \quad \text{for } n \geq 2, \tag{12}$$

where the non-deterministic toll function $t_{n,k}$ is now given as follows:

$$t_{n,k} \stackrel{(d)}{=} (1 - B_{n,k}) \cdot k + B_{n,k} \cdot (n-k).$$

Here $B_{n,k}$ denotes a Bernoulli r.v. with success-probability $\mathbb{P}\{B_{n,k} = 1\} = \frac{k}{n}$ and $\mathbb{P}\{B_{n,k} = 0\} = 1 - \frac{k}{n}$. Also S_n and $(B_{n,k})_{1\leq k\leq n-1}$ are independent of $(X_j)_{j\geq 1}$ and $(X^*_j)_{j\geq 1}$, which are independent copies of each other. Again S_n has the distribution given in (5).

The expectation μ_n satisfies now the recurrence:

$$(n-1)T_n\mu_n = \sum_{k=1}^{n-1} kT_kT_{n-k}(\mu_k + \mu_{n-k}) + R_n, \quad \text{for } n \geq 2, \tag{13}$$

with $\mu_1 = 0$ and $R_n = \frac{2}{n}\sum_{k=1}^{n-1} k^2(n-k)T_kT_{n-k}$.

Introducing suitable generating functions as in Subsection 3.3 we obtain the following generating functions solution:

$$C(z) = \frac{T(z)}{1-T(z)} \log\Big(\frac{1}{1-T(z)}\Big). \tag{14}$$

Thus $C(z)$ is amenable to singularity analysis, which immediately leads to the asymptotic expansion for the expectation μ_n given in (3). In a similar manner one can obtain a closed form expression for the generating function of the second moment of $X_n^{[QFB]}$, which leads to the asymptotic formula for the variance given in (3).

For establishing a limiting distribution result one considers for $n \geq 1$ the shifted random variable $\tilde{X}_n = X_n - \frac{1}{2}n\log n - \frac{\gamma+\log 2}{2}n$. The r-th moments $\tilde{\mu}_n^{[r]} := \mathbb{E}\big(\tilde{X}_n^r\big)$ of $\tilde{X}_n$ satisfy for $r \geq 1$ the following recurrence:

$$(n-1)T_n\tilde{\mu}_n^{[r]} = \sum_{k=1}^{n-1} kT_kT_{n-k}\big(\tilde{\mu}_k^{[r]} + \tilde{\mu}_{n-k}^{[r]}\big) + \tilde{R}_n^{[r]}, \quad \text{for } n \geq 2, \tag{15}$$

with $\tilde{\mu}_1^{[r]} = (-\frac{\gamma+\log 2}{2})^r$ and $\tilde{R}_n^{[r]} = \sum_{\substack{r_1+r_2+r_3=r,\\ r_2,r_3<r}} \binom{r}{r_1,r_2,r_3} \sum_{k=1}^{n-1} kT_kT_{n-k}\tilde{t}_{n,k}^{[r_1]}\tilde{\mu}_k^{[r_2]}\tilde{\mu}_{n-k}^{[r_3]}$, where

$$\tilde{t}_{n,k}^{[r]} = \big(1-\frac{k}{n}\big)\Big(\frac{1}{2}k\log k + \frac{1}{2}(n-k)\log(n-k) - \frac{1}{2}n\log n + k\Big)^r + \frac{k}{n}\Big(\frac{1}{2}k\log k + \frac{1}{2}(n-k)\log(n-k) - \frac{1}{2}n\log n + n - k\Big)^r.$$

Recurrence (15) can be treated analogous to (10) leading to Theorem 2.

References

[AHU74] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, 1974.

[BS93] B. Bollobás and I. Simon. Probabilistic analysis of disjoint set union algorithms. *SIAM Journal on Computing*, 22:1053–1074, 1993.

[CM04] P. Chassaing and R. Marchand. Merging costs for the additive Marcus-Lushnikov process, and Union-Find algorithms. manuscript, 2004.

[FFK05] J. A. Fill, P. Flajolet, and N. Kapur. Singularity analysis, Hadamard products, and tree recurrences. *Journal of Computational and Applied Mathematics*, 174:271–313, 2005.

[FKP06] J. A. Fill, N. Kapur, and A. Panholzer. Destruction of very simple trees. *Algorithmica*, 46:345–366, 2006.

[FPV98] P. Flajolet, P. Poblete, and A. Viola. On the analysis of linear probing hashing. *Algorithmica*, 22:490–515, 1998.

[HN02] H.-K. Hwang and R. Neininger. Phase change of limit laws in the quicksort recurrence under varying toll functions. *SIAM Journal on Computing*, 31:1687–1722, 2002.

[KS78] D. E. Knuth and A. Schönhage. The expected linearity of a simple equivalence algorithm. *Theoretical Computer Science*, 6:281–315, 1978.

[Pan04] A. Panholzer. Destruction of recursive trees. In *Mathematics and Computer Science. III*, Trends in Mathematics, pages 267–280, Basel, 2004. Birkhäuser.

[Yao76] A. C.-C. Yao. On the average behavior of set merging algorithms (extended abstract). In *Conference Record of the Eight Annual ACM Symposium on Theory of Computing*, pages 192–195, 1976.