Near-Optimal Bounded-Degree Spanning Trees

Jennie C. Hansen
Department of Actuarial Mathematics & Statistics
Heriot-Watt University
Edinburgh, Scotland, UK
J.Hansen@ma.hw.ac.uk
Fax 0131-451-3249

Eric Schmutz
Department of Mathematics and Computer Science
Drexel University
Philadelphia, Pennsylvania, 19104
eschmutz@mcs.drexel.edu

September 14, 2000

Abstract

Random costs $C(i,j)$ are assigned to the arcs of a complete directed graph on n labelled vertices. Given the cost matrix $C_n = (C(i,j))$, let $T_k^* = T_k^*(C_n)$ be the spanning tree that has minimum cost among spanning trees with in-degree less than or equal to k. Since it is NP-hard to find T_k^*, we instead consider an efficient algorithm that finds a near-optimal spanning tree T_k^a. If the edge costs are independent, with a common exponential(1) distribution, then as $n \to \infty$

$$E(Cost(T_k^a)) = E(Cost(T_k^*)) + o(1).$$

Upper and lower bounds for $E(Cost(T_k^*))$ are also obtained for $k \geq 2$.

§1 Introduction

In this paper random costs are assigned to the edges of the complete directed graph, and the expected cost of the cheapest k-tree is estimated. (By "k-tree", we mean an in-directed spanning tree whose maximum in-degree is at most k.) Let $C_n = (C(i,j))$ denote an $n \times n$ matrix whose entries are independent exp(1) random variables. The variable $C(i,j)$ is understood to be the cost of the directed edge, (i,j), from vertex i to vertex j in the complete directed graph, D_n, with vertices labelled $1, 2, ..., n$. Note that loops (i,i) are included in D_n (so D_n has n^2 directed edges.)

It is an NP-hard problem to find an optimal k-tree given a cost matrix C_n as input (Lemma A.1 in the appendix). Since it is hard to find an optimal k-tree, we propose a heuristic algorithm. A near-optimal bounded-degree spanning tree can, with high probability, be obtained by easy modifications of the cheapest k-map. (A k-map is subgraph of D_n such that each vertex in the subgraph has out-degree 1 and has in-degree less than or equal to k.) This is significant because it is computationally easy to find the cheapest k-map. To analyze the algorithm, we prove that the average cost of the optimal k-tree is asymptotically close to the average cost of the cheapest k-map. We also provide upper and lower bounds for the expected cost of the optimal k-map (and hence, bounds for the expected cost of the optimal k-tree) for all $k \geq 2$.

We note that, for $k = 1$, a k-map is in fact a permutation. So our problem is related to the assignment problem. Our methods for obtaining a lower bound for the expected cost of the optimal k-map are different from those of Goemans and Kodialam [1] and Olin [2] who compute the expected value of a feasible solution to the dual linear program for the assignment problem. We believe their methods yield better bounds than ours in the case $k = 1$, but that our methods give better results for $k \geq 2$. Coppersmith and Sorkin[3] have recently made significant progress on the upper bound for the assignment problem. Their methods are appealing, and may be relevant here, but we have not been able to exploit them directly. The work of Frieze et.al. ([4],[5],[6]) deals with analagous problems for undirected graphs. There is also a literature dealing with bounded degree spanning trees in the plane and other "geometric"analogues (e.g. Khuller, Raghavachari, and Young[7]). These problems are only superficially similar to our problem.

To obtain an upper bound on the expected cost of the optimal k-tree, we bound the expected cost of a k-map that is constructed using a greedy heuristic algorithm. Although this greedy heuristic is not optimal, it yields a map whose expected cost is bounded and surprisingly close to the optimum. It is interesting to compare this with greedy heuristics for the assignment problem which yield very poor assignments having $\Theta(\log n)$ expected cost. As a referee pointed out, this is because the case $k = 1$ is more constrained: when edges are added using a greedy heuristic, the number of potential edges decreases steadily so that the last few edges contribute significantly to the expected cost. This is less of a problem for $k \geq 2$ because there is more flexibility in selecting edges.

Finally, we remark that the case $k = \infty$ is the problem of finding the optimal spanning tree with no degree restrictions. In this case, greedy heuristics work well asymptotically. In particular, greedy methods have been used to obtain limit theorems for the expected cost of the optimal spanning tree, and for the distribution of the cost of the optimal tree as $n \to \infty$. (Hansen[8], McDiarmid[9]). Thus, when estimating the expected cost of the optimal k-tree, $k = 1$ is the difficult case, $k = \infty$ is the easy case, and $2 \leq k < \infty$ is intermediate in difficulty.

A little notation is needed to proceed. For $1 \leq i \leq n$, let $c_{(1)}(i), c_{(2)}(i), ..., c_{(n)}(i)$ denote the order statistics of the variables $\{C(i,j) : 1 \leq j \leq n\}$. The joint distribution of the order statistics of i.i.d. exponential random variables is well understood. We will make use of the fact that $c_{(k)}(i) \sim R_n + R_{n-1} + ... + R_{n-k+1}$ where $\{R_m, 1 \leq m \leq n\}$ are independent random variables with $R_m \sim \exp(m)$. It is also a consequence of the 'memoryless' property of the exponential distribution that $c_{(1)}(i), c_{(2)}(i) - c_{(1)}(i), ..., c_{(n)}(i) - c_{(n-1)}(i)$ are independent with $c_{(k)}(i) - c_{(k-1)}(i) \sim \exp(n - k + 1)$ for $1 \leq k \leq n$. Finally, for $1 \leq i \leq n$ and any vertex v, define $X_{(i)}(v) = j$ if and only if $C(v,j) = c_{(i)}(v)$. For each vertex v, the vector $(X_{(1)}(v), X_{(2)}(v), ..., X_{(n)}(v))$ is a uniform random permutation of the vertices $1, 2, ..., n$. One can also verify that for each vertex v, the variables $\{X_{(i)}(v) : 1 \leq i \leq n\}$ and $\{c_{(i)}(v) : 1 \leq i \leq n\}$ are independent. It follows that the σ-algebras $\sigma\{X_{(i)}(v) : 1 \leq i \leq n, 1 \leq v \leq n\}$ and $\sigma\{c_{(i)}(v) : 1 \leq i \leq n, 1 \leq v \leq n\}$ are independent too.

Given C_n, let $T_k^* = T_k^*(C_n)$ be the cheapest k-tree, and let $M_k^* = M_k^*(C_n)$ be the cheapest k-map. It is helpful to think of M_k^* as a (non-uniform) random map; we write $M_k^*(v) = w$ iff $(v, w) \in M_k^*$. Hansen [8] observed that M_n^* is in some sense close to being a minimum spanning tree in D_n: by breaking a few cycles in M_n^* and redirecting some edges one can obtain a spanning tree whose expected cost is asymptotically optimal. A similar strategy is developed here for bounded-degree spanning trees. The idea is very simple. First create a forest by removing one edge from each cycle of M_k^*. Then patch together the components of the forest to form a tree. If r is the root of a tree in a forest, call v *available for* r if the in-degree of v is less than k and v is not in the same weak component as r. If we adjoin the edge from r to v, the result is a forest with one less component. This is the basis for

```
Algorithm 1
 1. Find $M_k^*$.
 2. Let $T_k^a$ be the forest obtained by deleting the most expensive edge from every
    cycle of $M_k^*$, and let $\kappa$ be the number of components that $T_k^a$ has.
 3. For $i = 1, \ldots, \kappa - 1$
        {
        Let $r_i$ be the root of the smallest component of $T_k^a$.
        Add to $T_k^a$ the cheapest edge in $D_n$ from $r_i$ to a vertex that is available
        to $r_i$.
        }
```

Algorithm 1 creates a k-tree T_k^a in polynomial time. To see this consider the following linear program $LP(C_n, k)$:

$$\texttt{Minimize } z = \sum_{i=1}^{n} \sum_{j=1}^{n} C(i,j) x_{i,j}$$

`Subject to:`

$$\sum_{i=1}^{n} x_{i,j} \le k \qquad (j = 1, 2, \ldots, n)$$

$$\sum_{j=1}^{n} x_{i,j} = 1 \qquad (i = 1, 2, \ldots, n)$$

$$x_{i,j} \ge 0 \qquad (1 \le i, j \le n)$$

Any 0-1 feasible solution to this LP corresponds to a k-map M. The correspondence is $x_{i,j} = 1$ if $(i, j) \in M$, and $x_{i,j} = 0$ otherwise. The first n constraints say that each vertex has in-degree less than or equal to k, and the second n constraints say that each vertex has out-degree one. It is a well known theorem in linear programming that the optimal solution to this kind of transportation problem in fact an integral solution, and so the optimal solution to the linear program $LP(C_n, k)$ is a 0-1 solution[10]. Since M_k^* corresponds to the optimal solution to the LP, the first step in Algorithm 1 can be solved in polynomial time and the remaining steps can also be carried out in polynomial time.

In this paper we prove that Algorithm 1 is asymptotically optimal: in section 2 we show that $E(Cost(T_k^a)) = E(Cost(T_k^*)) + o(1)$. In section 3 we obtain a lower bound for $E(Cost(T_k^*))$ and in section 4 we obtain an upper bound by analyzing a 'greedy' algorithm.

§2. Analysis of Algorithm 1

The main goal of this section is to prove

Theorem 2.1 *If $k \ge 2$, then* $E(Cost(T_k^a)) = E(Cost(T_k^*)) + o(1)$.

We establish Theorem 2.1 by showing that $E(Cost(M_k^*))$ is close to both $E(Cost(T_k^a))$ and $E(Cost(T_k^*))$. The argument is similar to Karp and Steele's [11] analysis of a patching algorithm for the asymmetric travelling salesman problem. The first step is to prove

Theorem 2.2 *If $k \ge 2$, then* $E(Cost(T_k^a)) = E(Cost(M_k^*)) + o(1)$.

Proof. Fix $k \ge 2$ and define the subgraph D'_n of D_n as follows: edge $(i, j) \in D'_n$ if and only if $C(i, j) < L(n)$ where $L(n) = \frac{50 \log^2 n}{n}$. Let M'_k denote the cheapest k-map in D'_n, provided such a map exists. (M'_k does not exist if, for example, there is a vertex i such that $(i, j) \notin D'_n$ for every $1 \le j \le n$). Let $\hat{M}_k = M'_k$ if M'_k exists; otherwise, let $\hat{M}_k = M_k^*$ and consider the following modification of Algorithm 1.

```
Algorithm 1′
1. Find $\hat{M}_k$.
2. Let $\hat{T}_k^a$ be the forest obtained by deleting the most expensive edge from every
   cycle of $\hat{M}_k$, and let $\hat{\kappa}$ be the number of components that $\hat{T}_k^a$ has.
3. For $i = 1, \ldots, \hat{\kappa} - 1$
       {
       Let $r_i$ be the root of the smallest component of $\hat{T}_k^a$ .
       Add to $\hat{T}_k^a$ the cheapest edge in $D_n$ from $r_i$ to a vertex that is available
       to $r_i$.
```

}

Now let $B_n = \{\hat{M}_k = M'_k, \hat{\kappa} < \log^2 n, F_i \leq \log^4 n, i = 1, 2, ..., n\}$ where $F_i = |\{j : C(i,j) \leq L(n)\}|$. *Given* B_n, the combined cost of the edges deleted in Step 2 of Algorithm $1'$ is at most $\hat{\kappa} \cdot L(n) \leq \frac{50 \log^4 n}{n}$. Hence,

$$E(Cost(\hat{M}_k)|B_n) - \frac{50 \log^4 n}{n} \leq E(Cost(\hat{T}_k^a)|B_n)$$

$$\leq E(Cost(\hat{M}_k)|B_n) + E(Cost(\text{added edges})|B_n).$$

We show below that $E(Cost(\text{added edges})|B_n) \leq \frac{100 \log^4 n}{n}$ by bounding the expected cost of each edge added by Algorithm $1'$.

For $i < \hat{\kappa} < \log^2 n$, consider the i'th iteration of step 3 in Algorithm $1'$. Let $\mathcal{A}_i = \mathcal{A}_i(r_i)$ denote the set of vertices that are available to r_i at the beginning of the i'th iteration of step 3. The edges out of r_i are examined in increasing order of cost until an edge that points to a vertex $v_i \in \mathcal{A}_i$ is found, then the edge (r_i, v_i) is added to $\hat{T}_k^a$ and the added cost is $C(r_i, v_i)$. Thus,

$$E(Cost(\text{added edges})|B_n) = E(\sum_{i=1}^{\hat{\kappa}-1} C(r_i, v_i)|B_n)$$

$$= \sum_{i=1}^{\log^2 n} E(C(r_i, v_i)|i < \hat{\kappa}, B_n) \Pr(i < \hat{\kappa}|B_n)$$

$$\leq \sum_{i=1}^{\log^2 n} E(C(r_i, v_i)|i < \hat{\kappa}, B_n).$$

To bound $E(C(r_i, v_i)|i < \hat{\kappa}, B_n)$ for each $i \leq \log^2 n$, we analyze a more expensive 'patching' operation which is described below.

Fix $i \leq \log^2 n$. Given $i < \hat{\kappa}$ and B_n, let $m_i = n - F_{r_i} = |\{j : C(r_i, j) > L(n)\}|$, and let $\mathcal{A}'_i = \{j \in \mathcal{A}_i : C(r_i, j) > L(n)\}$. Call any edge (r_i, j) with $C(r_i, j) > L(n)$ a *costly* edge. Now modify the i'th iteration of the patching operation in Step 3 of Algorithm $1'$ as follows: add to $\hat{T}_k^a$ the cheapest edge from r_i to a vertex in $\mathcal{A}'_i$. In other words, at the i'th iteration of the algorithm we examine only *costly* edges out of r_i in increasing order of cost until we encounter one that points to a vertex $w_i \in \mathcal{A}'_i$ and edge (r_i, w_i) is the new edge that is added to $\hat{T}_k^a$. Observe that $C(r_i, v_i) \leq C(r_i, w_i)$ always, where v_i is the cheapest vertex available to r_i in $\mathcal{A}_i$. Thus any upper bound for $E(C(r_i, w_i)|i < \hat{\kappa}, B_n)$ is also an upper bound for $E(C(r_i, v_i)|i < \hat{\kappa}, B_n)$.

To bound $E(C(r_i, w_i)|i < \hat{\kappa}, B_n)$, observe that *given* that the edge $C(r_i, w) > L(n)$, we have $C(r_i, w) \sim L(n) + X$, where $X \sim \exp(1)$. (We use the fact that if M'_k exists, then a 'cheap' k-map has been constructed without examining any of the costly edges, so we have no *extra* information about the *costly* edges). It follows from standard results for order statistics of exponential random variables that if there are m_i *costly* edges out of vertex r_i and if the d'th cheapest *costly* edge out of r_i is added to $\hat{T}_k^a$, then

$$E\Big(C(r_i,w_i)\Big|d,m_i,i<\hat{\kappa},B_n\Big)=L(n)+\frac{1}{m_i}+\frac{1}{m_i-1}+...+\frac{1}{m_i-d+1}.$$

The random variable d has the same distribution as the number of draws, without replacement, until a black ball is drawn from an urn with $|\mathcal{A}_i'|$ black balls and $m_i-|\mathcal{A}_i'|$ white balls. In Lemma 2.9 we prove that $|\mathcal{A}_i|>\frac{n}{4}$, so $|\mathcal{A}_i'|\geq|\mathcal{A}_i|-F_{r_i}\geq\frac{n}{4}-\log^4 n\geq\frac{n}{5}$ for all large n. Hence

$$\Pr\left(d>5\log n\Big|m_i,i<\hat{\kappa},B_n\right)\leq(1-\frac{|\mathcal{A}_i'|}{n})^{5\log n}\leq\left(\frac{4}{5}\right)^{5\log n}\leq\frac{1}{n}$$

for all large n. Also, given B_n, we have $m_i\geq n-\log^4 n$ and thus

$$\begin{aligned}&E\Big(C(r_i,v_i)\Big|i<\hat{\kappa},B_n\Big)\leq E\Big(C(r_i,w_i)\Big|i<\hat{\kappa},B_n\Big)\\&\leq\leq L(n)+\frac{5\log n}{n-\log^4 n-5\log n}+\left(\sum_{k=1}^{n}\frac{1}{k}\right)\Pr\left(d>5\log n\Big|i<\hat{\kappa},B_n\right)\leq 2L(n)\end{aligned}$$

for all sufficiently large n. Hence

$$\begin{aligned}E(Cost(\text{added edges})|B_n)&\leq\sum_{i=1}^{\log^2 n}E(C(r_i,v_i)|i<\hat{\kappa},B_n)\\&\leq 2L(n)\log^2 n=\frac{100\log^4 n}{n}.\end{aligned}$$

In Lemma 2.8 we prove that $\Pr(B_n^c)=O(\frac{1}{n^5})$, so it follows that for all large n,

$$\Big|E(Cost(\hat{M}_k))-E(Cost(\hat{T}_k^a))\Big|\leq$$

$$\frac{150\log^4 n}{n}+\Big|E\Big((Cost(\hat{M}_k)-Cost(\hat{T}_k^a))\cdot 1\{B_n^c\}\Big)\Big|$$

$$\leq\frac{150\log^4 n}{n}+\left(E(Cost(\hat{M}_k)^2)^{1/2}+E(Cost(\hat{T}_k^a)^2)^{1/2}\right)(\Pr(B_n^c))^{1/2}.$$

$Cost(\hat{M}_k)\leq\sum_{i=1}^n c_{(n)}(i)$ always, so we have

$$E(Cost(\hat{M}_k)^2)\leq E(\sum_{i=1}^n c_{(n)}(i))^2=Var(\sum_{i=1}^n c_{(n)}(i))+(E(\sum_{i=1}^n c_{(n)}(i)))^2$$

$$= nVar(c_{(n)}(1)) + n^2(E(c_{(n)}(1)))^2 = n\sum_{k=1}^{n}\frac{1}{k^2} + n^2(\sum_{k=1}^{n}\frac{1}{k})^2 \leq 2n^2\log^2 n$$

since $c_{(n)}(1) \sim R_1 + R_2 + ... + R_n$. Similarly, $E(Cost(\hat{T}_k^a)^2) \leq 2n^2\log^2 n$. Thus

$$\Big|E(Cost(\hat{M}_k)) - E(Cost(\hat{T}_k^a))\Big| \leq \frac{250\log^4 n}{n}.$$

To finish the proof, we note that whenever $\hat{M}_k = M_k^*$ we must have $\hat{T}_k^a = T_k^a$ too. Now it follows from Lemma 2.4 below that $\Pr(\hat{M}_k = M_k^*) \geq \Pr(\hat{M}_k = M_k' = M_k^*) = 1 - O(\frac{1}{n^5})$, and so by arguments similar to those given above, we have

$$|E(Cost(M_k^*) - Cost(T_k^a)| \leq \frac{250\log^4 n}{n} + \Big|E\Big((Cost(M_k^*) - Cost(T_k^a)) \cdot 1\{\hat{M}_k \neq M_k^*\}\Big)\Big|$$

$$\leq \frac{300\log^4 n}{n}$$

and except for the unproved lemmas that were cited, we have now completed the proof of the theorem. ■

The proof of Theorem 2.2 used several lemmas that must now be proved. In particular, a key step in the proof of Theorem 2.2 is the observation that M_k' exists and equals M_k^* with high probability. To establish this we modify an argument from Karp and Steele [11]. We begin by defining the directed subgraph $G(C_n)$ of D_n in which

$$(i,j) \text{ is an edge of } G(C_n) \;\Leftrightarrow C(i,j) < p(n) = \frac{16\log n}{n}.$$

For any subset $S \subseteq [n]$, define $\Gamma(S) = \big\{j : (i,j) \in G(C_n) \text{ for some } i \in S\big\}$ and $\Gamma^{-1}(S) = \big\{i : (i,j) \in G(C_n) \text{ for some } j \in S\big\}$. The directed graph $G(C_n)$ is called *expanding* if for any subset of vertices $S \subseteq [n]$, the following inequalities both hold:

$$|\Gamma(S)| \geq \min\big\{2|S| + 1, \frac{n+1}{2}\big\} \quad \text{and} \quad |\Gamma^{-1}(S)| \geq \min\big\{2|S| + 1, \frac{n+1}{2}\big\}.$$

Then we have

Lemma 2.3 $\Pr\big(G(C_n) \text{ is not expanding }\big) = O(\frac{1}{n^5})$.

Proof. This lemma is essentially Lemma 7 of Karp and Steele [11]. The only difference is that Karp and Steele use uniformly distributed cost variables and set $p(n) = 10\log n/n$ to obtain a probability bound which is $O(1/n^2)$. Their proof goes through with trivial modifications when the cost variables are exponential(1) and $p(n) = 16\log n/n$, so we do not repeat the argument here. $\blacksquare$

Lemma 2.4 *With probability* $1 - O(\frac{1}{n^5})$, *every edge of* M_k^* *has cost less than* $L(n) = \frac{50\log^2 n}{n}$.

Proof. Observe that by Lemma 2.3, it is enough to prove that if $G(C_n)$ is expanding, then every edge of M_k^* has cost less than $\frac{50\log^2 n}{n}$. So suppose $G(C_n)$ is expanding but $C(i', M_k^*(i')) \geq \frac{50\log^2 n}{n}$ for some $1 \leq i' \leq n$. We show that the mapping M_k^* can be modified to obtain a feasible soluition M_k which is cheaper than M_k^*.

The first step is to define a sequence of subsets of vertices as follows. Let $\Gamma(1) = \Gamma(\{i'\})$ and for $l \geq 2$, let $\Gamma(l) = \Gamma((M_k^*)^{-1}(\Gamma(l-1)))$. Let $\mathcal{A}(M_k^*)$ denote the set of vertices in M_k^* with in-degree less than k. Since each vertex has in-degree at most k under M_k^*, we must have $|\mathcal{A}(M_k^*)| \geq \lceil n/2 \rceil$. We claim that $\{l : \Gamma(l) \cap \mathcal{A}(M_k^*) \neq \emptyset\} \neq \emptyset$. To see this, note that if $|\Gamma(l)| \geq \frac{n+1}{2}$ for some l, then $\Gamma(l) \cap \mathcal{A}(M_k^*) \neq \emptyset$ (since $|\mathcal{A}(M_k^*)| \geq \lceil n/2 \rceil$). So if $\{l : \Gamma(l) \cap \mathcal{A}(M_k^*) \neq \emptyset\} = \emptyset$, then $|\Gamma(l)| < \frac{n+1}{2}$ for all $l \geq 1$. On the other hand, since $G(C_n)$ is expanding, $|\Gamma(1)| \geq \min(2, \frac{n+1}{2})$, and since $\Gamma(1) \cap \mathcal{A}(M_k^*) = \emptyset$, every vertex in $\Gamma(1)$ must have in-degree k under M_k^*. It follows that $|(M_k^*)^{-1}(\Gamma(1))| \geq 2k$ and $|\Gamma(2)| = |\Gamma((M_k^*)^{-1}(\Gamma(1)))| \geq \min(2^2 k, \frac{n+1}{2})$. Now induction shows that $|\Gamma(l)| \geq \min(2^l k^{l-1}, \frac{n+1}{2})$ for all $l \geq 1$, and so $|\Gamma(l)| \geq \frac{n+1}{2}$ for $l \geq 2\log n$. Thus we can't have $|\Gamma(l)| < \frac{n+1}{2}$ for all $l \geq 1$.

Let $m = \min\{\ell : \Gamma(\ell) \cap \mathcal{A}(M_k^*) \neq \emptyset\}$ and note that it follows from the argument above that $m \leq 2\log n$. The next step is to define two sequences of vertices $i_1, i_2, ..., i_m$ and $j_1, j_2, ..., j_m$. Let $i_1 = i'$ and let j_m be a vertex in $\Gamma(m) \cap \mathcal{A}(M_k^*) \neq \emptyset$. Since $j_m \in \Gamma(m) \cap \mathcal{A}(M_k^*)$ there is a vertex $i_m \in (M_k^*)^{-1}(\Gamma(m-1))$ such that $C(i_m, j_m) < p(n) = \frac{16\log n}{n}$. The remaining vertices in the sequence are defined recursively as follows. For $1 \leq l \leq m-1$, let $j_l = M_k^*(i_{l+1}) \in \Gamma(l) = \Gamma((M_k^*)^{-1}(\Gamma(l-1)))$. For $2 \leq l \leq m-1$, choose $i_l \in (M_k^*)^{-1}(\Gamma(l-1))$ such that $C(i_l, j_l) < p(n)$. Observe that $j_m \neq M_k^*(i_{l+1}) = j_l$ for $1 \leq l \leq m-1$ since $M_k^*(i_{l+1}) \in \Gamma(l)$ and $\Gamma(l) \cap \mathcal{A}(M_k^*) = \emptyset$ for $1 \leq l \leq m-1$.

Given the two sequences $i_1, i_2, ..., i_m$ and $j_1, j_2, ..., j_m$, define a new mapping M_k by setting $M_k(i) = M_k^*(i)$ if $i \notin \{i_1, i_2, ..., i_m\}$ and setting $M_k(i_l) = j_l$ for $1 \leq l \leq n$. In other words, M_k is constructed from the optimal mapping M_k^* by deleting the edges $(i_l, M_k^*(i_l))$ and adding the edges (i_l, j_l) for $1 \leq l \leq n$. To see that M_k is a feasible solution, we note that for each $2 \leq l \leq m$, the deletion of edge $(i_l, M_k^*(i_l))$ makes vertex $M_k^*(i_l)$ 'available' and so the addtion of the edge $(i_{l-1}, j_{l-1}) = (i_{l-1}, M_k^*(i_l))$ does not violate the degree constraint at vertex $M_k^*(i_l)$. Also, since $j_m \neq j_l$ for $1 \leq l \leq m-1$, the vertex j_m is still 'available' after the addition of the edges $\{(i_l, j_l) : 1 \leq l \leq m-1\}$, so the addition of edge (i_m, j_m) does not violate the degree constraint at vertex j_m. So the mapping M_k is a feasible solution.

Finally, observe that

$$
\begin{aligned}
Cost(M_k) - Cost(M_k^*) &= \sum_{t=1}^{m} C(i_t, j_t) - \sum_{t=1}^{m} C(i_t, M_k^*(i_t)) \\
&\leq \sum_{t=1}^{m} C(i_t, j_t) - C(i_1, M_k^*(i_1)) \\
&\leq m \cdot p(n) - \frac{50 \log^2 n}{n} \\
&\leq 2 \log n \cdot p(n) - \frac{50 \log^2 n}{n} = \frac{-18 \log^2 n}{n}
\end{aligned}
$$

which contradicts the optimality of M_k^*. ■

Next we establish that, with high probabilty, M_k^* has less than $\log^2 n$ components. It is well known that, for *uniform* random maps, the number of components is $O(\log n)$ with high probabilty, and it would not be difficult to prove the same fact for uniform random k-maps. However M_k^* is a *non-uniform* random k-map. Nevertheless, the corresponding statement is a consequence of the following

Lemma 2.5 *Let f and g be two k-maps that differ only by a transposition of the values they assign to two vertices, i.e. there exist i_1, i_2 such that $f(i_1) = g(i_2)$, $f(i_2) = g(i_1)$, and for $v \neq i_1, i_2$, $f(v) = g(v)$. Then*

$$\Pr(f \text{ is optimal}) = \Pr(g \text{ is optimal})$$

Proof. Let $\mathcal{C}$ be the set of cost matrices, and for any k-map M, let $O_M \subseteq \mathcal{C}$ be the set of cost matrices for which M is optimal. We want to prove that $\Pr(O_f) = \Pr(O_g)$. Define $H : \mathcal{C} \to \mathcal{C}$ by $H(C) = C'$, where

$$C'(i,j) = \begin{cases} C(i_2, j) & \text{if } i = i_1 \\ C(i_1, j) & \text{if } i = i_2 \\ C(i,j) & \text{else} \end{cases}$$

We know $\Pr(O_f) = \Pr(H(O_f))$ because costs are assumed i.i.d. It therefore suffices to prove that $H(O_f) = O_g$.

Let $C \in O_f$. To prove that $H(O_f) \subseteq O_g$, it suffices to show that g is optimal for the instance C'. First note that $Cost(f, C) = Cost(g, C')$:

$$Cost(f, C) = C(i_1, f(i_1)) + C(i_2, f(i_2)) + \sum_{i \neq i_1, i_2} C(i, f(i))$$

$$
\begin{aligned}
&= C(i_1, g(i_2)) + C(i_2, g(i_1)) + \sum_{i \neq i_1, i_2} C(i, g(i)) \\
&= C'(i_2, g(i_2)) + C'(i_1, g(i_1)) + \sum_{i \neq i_1, i_2} C'(i, g(i)) \\
&= Cost(g, C').
\end{aligned}
$$

Now we prove by contradiction that g is optimal for C'. Suppose on the contrary, that h is a k-map and $Cost(h, C') < Cost(g, C')$. Define h' by $h'(i_1) = h(i_2)$, $h'(i_2) = h(i_1)$, and for $i \neq i_1, i_2$, $h'(i) = h(i)$. Then $Cost(h', C) = Cost(h, C') < Cost(g, C') = Cost(f, C)$ This contradicts the optimality of f, and completes the proof that

$$H(O_f) \subseteq O_g. \tag{2.1}$$

By the same argument,

$$H(O_g) \subseteq O_f. \tag{2.2}$$

Observe that H^2 is the identity. Hence, by applying H to equation (2.2) we get

$$O_g \subseteq H(O_f). \tag{2.3}$$

Combining (2.1) and (2.3) we get $H(O_f) = O_g$. ∎

For any map M, let $Z = Z(M) = \{i : M^t(i) = i \text{ for some } t \geq 1\}$ be the set of cyclic vertices of M. Then $M|_Z$ is a permutation on Z. The following corollary asserts that, given the set Z of cyclic vertices, all permutations are equally likely to occur as $M_k^*|_Z$.

Corollary 2.6 *For any set $\mathcal{Z} \subseteq \{1, 2, .., n\}$ and any permutation σ on $\mathcal{Z}$,*

$$\Pr\Big(M_k^*|_{\mathcal{Z}} = \sigma \Big| Z(M_k^*) = \mathcal{Z}\Big) = \frac{1}{|\mathcal{Z}|!}$$

Proof. For any two permutations π, σ of $\mathcal{Z}$, there is a sequence $\tau_1, \tau_2, \ldots, \tau_m$ of transpositions such that $\pi = \sigma \circ \prod_{i=1}^{m} \tau_i$. By Lemma 2.5, we have

$$\Pr\Big(M_k^*|_{\mathcal{Z}} = \sigma \circ \prod_{i=1}^{\ell-1} \tau_i \ \Big| Z(M_k^*) = \mathcal{Z}\Big) = \Pr\Big(M_k^*|_{\mathcal{Z}} = \sigma \circ \prod_{i=1}^{\ell} \tau_i \ \Big| Z(M_k^*) = \mathcal{Z}\Big)$$

for $1 \leq \ell \leq m$. ∎

Lemma 2.7 *With probabilty $1 + o(\frac{1}{n^5})$, M_k^* has less than $\log^2 n$ components.*

Proof. Let $m \leq n$ be the number of cyclic vertices, and κ the number of cycles. By the Corollary 2.6, we need only estimate the probability that a uniform random permutation on m letters has more than $\log^2 n$ cycles. It is well known (e.g. Flajolet and Soria[12]) that this probability is negligible: there is a constant $C > 0$ and a positive constant $\alpha < 1$ such that for all $t > 0$

$$\Pr\left(\frac{\kappa - \log m}{\sqrt{\log m}}) > t \Big| |Z(M_k^*)| = m\right) < C\alpha^t.$$

Take $t = \log^{4/3} n$ to obtain the result. ■

Lemma 2.8 $\Pr(B_n) = 1 - O(\frac{1}{n^5})$.

Proof. Recall that $B_n = \{\hat{M}_k = M_k', \hat{\kappa} < \log^2 n, F_i \leq \log^4 n, i = 1, 2, ..., n\}$ where $F_i = |\{j : C(i,j) \leq L(n)\}|$. Therefore

$$\Pr(B_n^c) \leq \Pr(\hat{M}_k \neq M_k') + \Pr(\hat{\kappa} \geq \log^2 n) + \sum_{i=1}^{n} \Pr(F_i > \log^4 n).$$

Now if every edge of M_k^* has cost less than $L(n) = \frac{50 \log^2 n}{n}$ then M_k' must exist and $M_k' = M_k^*$. So it follows from Lemma 2.4 that

$$P(\hat{M}_k = M_k') \geq \Pr(\hat{M}_k = M_k' = M_k^*) \geq 1 - O(\frac{1}{n^5}).$$

Thus

$$\begin{aligned}\Pr(B_n^c) &\leq \Pr(\hat{\kappa} \geq \log^2 n, \hat{M}_k = M_k' = M_k^*) + O(\frac{1}{n^5}) + \sum_{i=1}^{n} \Pr(F_i > \log^4 n) \\ &\leq \Pr(\kappa \geq \log^2 n) + n \Pr(F_1 > \log^4 n) + O(\frac{1}{n^5}).\end{aligned}$$

By Lemma 2.7, $\Pr(\kappa \geq \log^2 n) = o(\frac{1}{n^5})$. Finally, since $F_1 \sim Bin(n, 1 - \exp(-L(n)))$,

$$\Pr(F_1 > \log^4 n) = \Pr(e^{F_1} > n^{\log^3 n}) \leq \frac{E(e^{F_1})}{n^{log^3 n}} \leq \frac{e^{50(e-1)\log^2 n}}{n^{log^3 n}} = o(\frac{1}{n^6}).$$

■

Lastly, we prove

Lemma 2.9 *For* $1 \leq i < \hat{\kappa}$, $|\mathcal{A}_i| > \frac{n}{4}$

Proof. Let $\mathcal{F}_i$ be the forest that consists of all components other than that of r_i. Let $d_j(\mathcal{F}_i)$ denote the number of vertices in $\mathcal{F}_i$ having in-degree j. The number of available vertices is

$$|A_i| = \sum_{j=0}^{k-1} d_j(\mathcal{F}_i)$$

The number of vertices in $\mathcal{F}_i$ is

$$v(i) = \sum_{j=0}^{k} d_j(\mathcal{F}_i).$$

Since $\mathcal{F}_i$ has $v(i)$ vertices and $\hat{\kappa} - i$ components number of edges of $\mathcal{F}_i$ is

$$v(i) - \hat{\kappa} + i = \sum_{j=1}^{k} j d_j(\mathcal{F}_i).$$

It follows that

$$0 < \hat{\kappa} - i = v(i) - \sum_{j=1}^{k} j d_j(\mathcal{F}_i) < v(i) - d_1(\mathcal{F}_i) - 2(\sum_{j=2}^{k} d_j(\mathcal{F}_i)) = -v(i) + 2d_0(\mathcal{F}_i) + d_1(\mathcal{F}_i).$$

Hence

$$2(d_0(\mathcal{F}_i) + d_1(\mathcal{F}_i)) \geq v(i).$$

Provided $k \geq 2$, we have $|A_i| \geq (d_0(\mathcal{F}_i) + d_1(\mathcal{F}_i))$. We know $v(i) \geq \frac{n}{2}$ because r_i was the root of the *smallest* component. Therefore $|A_i| \geq \frac{n}{4}$.

■

Now that the lemmas are proved, the proof of Theorem 2.2 is complete. However our main aim is to prove

Theorem 2.1 *If $k \geq 2$, then* $E(Cost(T_k^a)) = E(Cost(T_k^*)) + o(1)$.

Proof. Let r be the root of T_k^*, and let (r, w) be the cheapest edge from r to a vertex having in-degree less than k in T_k^*. Then

$$Cost(M_k^*) \leq Cost(T_k^*) + Cost((r, w))$$

and consequently

$$E(Cost(T_k^*)) \geq E(Cost(M_k^*)) - E(Cost((r, w))$$

We claim that $E(Cost((r, w))) \leq \frac{2 \log^4 n}{n}$ for all large n. To prove this we show that $C(r, w)$ is bounded above by a random variable Y whose expected value is bounded by $\frac{2 \log^4 n}{n}$. To define Y we introduce some notation. For each $1 \leq i \leq n$, let $T_k(i)$ be the cheapest k-tree rooted at vertex i, and let (i, w_i) denote the cheapest edge from i to a vertex having in-degree less than k in $T_k(i)$. Note that the variables $Cost(1, w_1), Cost(2, w_2), ..., Cost(n, w_n)$ are identically distributed, though not independent. Now let $Y = \max_{1 \leq i \leq n} Cost(i, w_i)$.

Since $T_k^* = T_k(i)$ and $(r,w) = (i,w_i)$ for *some* $1 \le i \le n$, we have $Cost(r,w) \le Y$. Observe that

$$\begin{aligned} E(Y) &\le \frac{\log^4 n}{n} + E\left(Y \cdot 1\{Y \ge \frac{\log^4 n}{n}\}\right) \\ &\le \frac{\log^4 n}{n} + (E(Y^2))^{1/2}\left(\Pr\left(Y \ge \frac{\log^4 n}{n}\right)\right)^{1/2} \\ &\le \frac{\log^4 n}{n} + (E(Y^2))^{1/2}\left(\sum_i \Pr\left(Cost(i,w_i) \ge \frac{\log^4 n}{n}\right)\right)^{1/2} \qquad (2.4) \\ &\le \frac{\log^4 n}{n} + (E(Y^2))^{1/2}\left(n \Pr\left(Cost(1,w_1) \ge \frac{\log^4 n}{n}\right)\right)^{1/2}. \end{aligned}$$

To bound $(E(Y^2))^{1/2}$, let $Z = \max\{C(i,j) : 1 \le i \le n, 1 \le j \le n\}$ and note that $(E(Y^2))^{1/2} \le E(Z^2)^{1/2}$. The variable Z has density $f(z) = n^2 e^{-z}(1-e^{-z})^{n^2-1}$, so

$$\begin{aligned} (E(Y^2))^{1/2} &\le \left(n^2 \int_0^\infty z^2 e^{-z}(1-e^{-z})^{n^2-1} dz\right)^{1/2} \\ &\le \left(n^2 \int_0^\infty z^2 e^{-z} dz\right)^{1/2} = \sqrt{2}n. \end{aligned}$$

Next, let $W = |\{j : C(1,j) \le \frac{\log^4 n}{n}\}|$ and let $\mathcal{U}$ denote the random set of vertices in $T_k(1)$ with in-degree equal to k. Since $W' := n - W \sim Bin(n, \exp(-\frac{\log^4 n}{n}))$, we have

$$\Pr(W \le \log^2 n) = P(W' \ge n - \log^2 n) \le \frac{e^{\log^2 n} E(e^{W'})}{e^n} < \frac{C}{e^{\log^3 n}} \qquad (2.5)$$

for some constant C which does not depend on n. Also, since at most $\frac{n}{k}$ vertices in $T_k(1)$ can have in-degree k, $|\mathcal{U}| \le \frac{n}{k}$ always. For $1 \le i \le n$, let $X_{(i)} = X_{(i)}(1)$ denote the vertex to which the i'th cheapest edge out of vertex 1 points, then $(X_{(1)}, X_{(2)}, ..., X_{(n)})$ is a uniformly distributed random permutation of the set $\{1, 2, ..., n\}$. The variables $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are measurable with respect to the σ-algebra generated by the variables $\{C(1,j) : 1 \le j \le n\}$, whereas the random tree $T_k(1)$, and hence the random set of unavailable vertices $\mathcal{U}$, is determined by the variables $\{C(i,j) : 2 \le i \le n, 1 \le j \le n\}$, so the variables $X_{(1)}, X_{(2)}, ..., X_{(n)}$ and the random set $\mathcal{U}$ are independent. Now *given* the event $W \ge b$ and the set $\mathcal{U}$, we have $Cost(1,w_1) \ge \frac{\log^4 n}{n}$ only if $X_{(1)}, X_{(2)}, ..., X_{(b)} \in \mathcal{U}$, i.e. we must reject at least the first b vertices that are examined and this happens only if all of those vertices are in $\mathcal{U}$. So

$$\begin{aligned} \Pr\left(Cost(1,w_1) \ge \frac{\log^4 n}{n} \Big| W \ge b, \mathcal{U} = U\right) &\le \Pr\left(X_{(1)}, X_{(2)}, ..., X_{(b)} \in U \Big| W \ge b, \mathcal{U} = U\right) \\ &= \left(\frac{|U|}{n}\right)\left(\frac{|U|-1}{n-1}\right)\cdots\left(\frac{|U|-b+1}{n-b+1}\right) \\ &\le \left(\frac{1}{k}\right)^b \left(1 - \frac{b}{n}\right)^{-b} \end{aligned}$$

since $|U| \leq \frac{n}{k}$. Using this bound, we obtain

$$\Pr\left(Cost(1, w_1)\right) \geq \frac{\log^4 n}{n})$$

$$\begin{aligned}
&\leq \sum_U \Pr\left(Cost(1, w_1) \geq \frac{\log^4 n}{n}\Big| W \geq b, \mathcal{U} = U\right) \Pr(W \geq b, \mathcal{U} = U) + \Pr(W \leq b) \\
&\leq \left(\frac{1}{k}\right)^b \left(1 - \frac{b}{n}\right)^{-b} \cdot \Pr(W \geq b) + \Pr(W \leq b)
\end{aligned} \tag{2.6}$$

where the sum is over all possible set values for the random set $\mathcal{U}$. It follows from (2.5) and (2.6) with $b = \log^2 n$, that

$$\Pr\left(Cost(1, w_1) \geq \frac{\log^4 n}{n}\right) \leq \frac{C'}{n^{(\log 2)(\log n)}}$$

where C' is a constant which may depend on k but which does not depend on n. Using this bound and the bound for $(E(Y^2))^{1/2}$ in (2.4) we obtain

$$E(Cost(r, w)) \leq E(Y) \leq \frac{2\log^4 n}{n}$$

for all large n.

Finally, since $Cost(T_k^*) \leq Cost(T_k^a)$ by the definition of T_k^*, and since $E(Cost(T_k^a)) = E(Cost(M_k^*)) + o(1)$ by Theorem 2.2, we have for all large n

$$E(Cost(M_k^*)) - \frac{2\log^4 n}{n} \leq E(Cost(T_k^*)) \leq E(Cost(T_k^a)) = E(Cost(M_k^*)) + o(1).$$

■

§3 Lower Bound

In this section we prove the following lower bound for the optimal tree's expected cost:

Theorem 3.1 *If $k \geq 2$, then*

$$\liminf_{n\to\infty} E(Cost(T_k^*)) \geq 1 + \frac{1}{e}\sum_{m>k}\frac{1}{m!}\sum_{\ell=1}^{m-k}\frac{m-k-\ell+1}{m-\ell+1}.$$

Proof. By Theorems 2.1 and 2.2, it suffices to prove that

$$\liminf_{n\to\infty} E(Cost(M_k^*)) \geq 1 + \frac{1}{e}\sum_{m>k}\frac{1}{m!}\sum_{\ell=1}^{m-k}\frac{m-k-\ell+1}{m-\ell+1}.$$

Since the optimal mapping M_k^* must have one edge out of each vertex, a crude lower bound is

$$Cost(M_k^*) \geq \sum_v c_{(1)}(v) = Cost(M^*)$$

where $c_{(i)}(v)$ denotes the cost of the i'th cheapest edge out of v, and $M^* = M_n^*$ is the cheapest map (with no restrictions on the in-degrees of the vertices). With high probability, a positive proportion of the vertices of M^* have in-degree larger than k. Hence a positive proportion of the edges in M^* cannot be used in M_k^*. Each time the cheapest edge is not used, we pay a penalty, and the idea is to estimate this penalty.

First some notation is needed. For any vertex v and $1 \leq i \leq n$, let $e_i(v)$ denote the i'th cheapest edge out of v, i.e. $Cost(e_i(v)) = c_{(i)}(v)$. Also, let $D(v) := c_{(2)}(v) - c_{(1)}(v) \sim \exp(n-1)$, and observe that

$$Cost(M_k^*) - Cost(M^*) \geq \sum_{v \text{ s.t. } e_1(v) \notin M_k^*} c_{(2)}(v) - c_{(1)}(v) = \sum_{v \text{ s.t. } e_1(v) \notin M_k^*} D(v).$$

We interpret $D(v)$ as the penalty which is paid for not using edge $e_1(v)$. For any map f from $\{1,2,...,n\}$ into $\{1,2,...,n\}$ and any vertex v, let $\rho(v,f)$ denote the in-degree of v in G_f, the directed graph representing the mapping f. For $0 \leq m \leq n$, let $\mathcal{V}_m(f) := \big\{v : \rho(v,f) = m\big\}$ and for any vertex y, let $\mathcal{B}(y,f) := \big\{v : f(v) = y\big\}$. In the special case where $f = M^*$, we write $\mathcal{V}_m$ for $\mathcal{V}_m(M^*)$ and $\mathcal{B}(y)$ for $\mathcal{B}(y, M^*)$.

Now given $y \in \mathcal{V}_m$ and $\mathcal{B}(y) = \big\{x_1, x_2, \ldots, x_m\big\}$, let $D_{(1)}(\mathcal{B}(y)), D_{(2)}(\mathcal{B}(y)), ...,$denote the order statistics of the variables $D(x_1), D(x_2), ..., D(x_m)$. For each vertex y, we define $X(y) = 0$ if $y \notin \mathcal{V}_m$ for some $m > k$ and $X(y) = \sum_{i=1}^{m-k} D_{(i)}(\mathcal{B}(y))$ if $y \in \mathcal{V}_m$ for some $m > k$. The variable $X(y)$ is a lower bound on the cost of redirecting edges that go into vertex y under M^*. In particular, if $y \notin \mathcal{V}_m$ for some $m > k$ then it is possible that none of the edges into y under M^* are redirected under the optimal solution M_k^*, whereas, if $y \in \mathcal{V}_m$ for some $m > k$, then at least $m - k$ of the edges into y (under M^*) *must* be redirected under M_k^* and the cost of redirecting edges will be at least $\sum_{i=1}^{m-k} D_{(i)}(\mathcal{B}(y))$. Using the variables $X(1), X(2), ..., X(n)$, we define a variable U_n that underestimates the cost of M_k^* :

$$Cost(M_k^*) \geq U_n := \text{Cost}(M^*) + \sum_{m>k} \sum_{y \in \mathcal{V}_m} \sum_{i=1}^{m-k} D_{(i)}(\mathcal{B}(y))) = Cost(M^*) + \sum_{y=1}^{n} X(y).$$

The variable U_n is the sum of the cost of the cheapest edge out of each vertex, plus an additional penalty for vertices y with in-degree $m > k$. The variables $X(1), X(2), ..., X(n)$ are identically distributed, so

$$E(U_n) = E(Cost(M^*)) + nE(X(1)) = 1 + nE(X(1))$$

since $E(Cost(M^*)) = nE(c_{(1)}(1)) = 1$.

Let $p_m(n) = \Pr(1 \in \mathcal{V}_m) = \ Pr(\rho(1, M^*) = m) = \binom{n}{m}(\frac{1}{n})^m(1-\frac{1}{n})^{n-m} \sim \frac{1}{em!}$, then we have

$$E(X(1)) = \sum_{m>k}^{n} E(X(1)\big|1 \in \mathcal{V}_m)p_m(n).$$

To compute $E(X(1)\big|1 \in \mathcal{V}_m)$ we write

$$E(X(1)\big|1 \in \mathcal{V}_m) = \sum_{\mathcal{A}} E(X(1)\big|\mathcal{B}(1) = \mathcal{A}) \Pr(\mathcal{B}(1) = \mathcal{A}\big|1 \in \mathcal{V}_m)$$

where the sum is over all subsets $\mathcal{A} \subseteq \{1, 2, ..., n\}$ such that $|\mathcal{A}| = m$. Recall that the σ-algebras $\sigma\{X_{(i)}(v) : 1 \leq i \leq n, 1 \leq v \leq n\}$ and $\sigma\{c_{(i)}(v) : 1 \leq i \leq n, 1 \leq v \leq n\}$ are independent. Now observe that given $\mathcal{B}(1) = \mathcal{A}$, then $X(1) = \sum_{i=1}^{m-k} D_{(i)}(\mathcal{A})$. For any subset $\mathcal{A}$, the event $\{\mathcal{B}(1) = \mathcal{A}\}$ is measurable with respect to $\sigma\{X_{(i)}(v) : 1 \leq i \leq n, 1 \leq v \leq n\}$ whereas the variable $X(1) = \sum_{i=1}^{m-k} D_{(i)}(\mathcal{A})$ is measurable with respect to $\sigma\{c_{(i)}(v) : 1 \leq i \leq n, 1 \leq v \leq n\}$. So by independence of the σ-algebras, we have

$$E(X(1)|\mathcal{B}(1) = \mathcal{A}) = E\Big(\sum_{i=1}^{m-k} D_{(i)}(\mathcal{A})\Big|\mathcal{B}(1) = \mathcal{A}\Big) = \sum_{i=1}^{m-k} E(D_{(i)}(\mathcal{A}))$$

for any subset $\mathcal{A} \in \{1, 2, ..., n\}$ such that $|\mathcal{A}| = m$. Also, for any subset $\mathcal{A} \in \{1, 2, ..., n\}$ we have

$$\sum_{i=1}^{m-k} E(D_{(i)}(\mathcal{A})) = \sum_{i=1}^{m-k} E(D_{(i)}(\mathcal{A}'))$$

where $\mathcal{A}' = \{1, 2, ..., m\}$, since the variables $\{D(i) : 1 \leq i \leq m\}$ and $\{D(x) : x \in \mathcal{A}\}$ have the same joint distribution. Since the variables $D(1), D(2), ..., D(m)$ are i.i.d. $\exp(n-1)$ random variables, it follows from Lemma 3.2 below that

$$\begin{aligned}
E(X(1))\big|1 \in \mathcal{V}_m) &= \sum_{\mathcal{A}} \sum_{i=1}^{m-k} E(D_{(i)}(\mathcal{A}')) \Pr(\mathcal{B}(1) = \mathcal{A}\big|1 \in \mathcal{V}_m) \\
&= \sum_{i=1}^{m-k} E(D_{(i)}(\mathcal{A}')) \\
&= \frac{1}{n-1} \sum_{l=1}^{m-k} \frac{m-k-l+1}{m-l+1}
\end{aligned}$$

where the sum is over all subsets $\mathcal{A} \subseteq \{1, 2, ..., n\}$ such that $|\mathcal{A}| = m$.

Finally, since $\lim_{n\to\infty} p_m(n) = \frac{1}{em!}$, and $p_m(n) \cdot \sum_{l=1}^{m-k} \frac{m-k-l+1}{m-l+1} \leq \frac{1}{m!} \cdot m$ for each

$m \geq 0$, we obtain

$$\begin{aligned}
\lim_{n\to\infty} E(U_n) &= 1 + \lim_{n\to\infty} nE(X(1)) \\
&= 1 + \lim_{n\to\infty} \frac{n}{n-1} \sum_{m>k} p_m(n) \sum_{l=1}^{m-k} \frac{m-k-l+1}{m-l+1} \\
&= 1 + \frac{1}{e} \sum_{m>k} \frac{1}{m!} \sum_{\ell=1}^{m-k} \frac{m-k-\ell+1}{m-\ell+1}
\end{aligned}$$

by dominated convergence, and the result follows. ■

The proof of Theorem 3.1 depends on a fact about the order statistics of the exponential distribution which we state as a lemma.

Lemma 3.2 *Let $X_1, X_2, \ldots, X_m$ be independent exponential random variables with parameter $\lambda > 0$. Let $S = \sum_{i=1}^{s} X_{(i)}$ be the sum of the s smallest of these m random variables. Then*

$$E(S) = \sum_{j=1}^{s} \frac{(s-j+1)}{\lambda(m-j+1)}.$$

Proof. The lemma follows from the fact that the expected value of the i'th order statistic from a sample of m exponential random variables with parameter λ is

$$E(X_{(i)}) = \sum_{j=1}^{i} \frac{1}{\lambda(m-j+1)},$$

and therefore

$$E(S) = \sum_{i=1}^{s} \sum_{j=1}^{i} \frac{1}{\lambda(m-j+1)} = \sum_{j=1}^{s} \sum_{i=j}^{s} \frac{1}{\lambda(m-j+1)} = \sum_{j=1}^{s} \frac{(s-j+1)}{\lambda(m-j+1)}$$

■

§4 Upper Bound

In this section prove

Theorem 4.1 *For $k \geq 2$*

$$\begin{aligned}\limsup_{n\to\infty} E(Cost(T_k^*)) \leq 1 &+ \sum_{m>k} \frac{1}{em!} \sum_{\ell=1}^{m-k} \frac{m-k-\ell+1}{m-\ell+1} \\ &+ \sum_{m=0}^{k} \frac{1}{em!} \left(\sum_{j=k-m+1}^{\infty} \frac{\lambda^j e^{-\lambda}}{j!} \sum_{l=1}^{j+m-k} \frac{j+m-k-l+1}{j+m-l+1} \right) \\ &+ \lambda \left(1 - \sum_{m=0}^{k} \frac{1}{em!} \right) \\ &+ \log(\frac{\alpha}{\alpha-\beta}) - \beta\end{aligned}$$

where

$$\begin{aligned}\lambda = \lambda(k) &= \sum_{m=0}^{k-1} \frac{(k-m)}{em!} - (k-1) \\ \alpha = \alpha(k) &= \sum_{m=0}^{k-1} \frac{1}{em!} \sum_{j=0}^{k-m-1} \frac{\lambda^j e^{-\lambda}}{j!} \\ \beta = \beta(k) &= \left(\sum_{m=0}^{k-1} \frac{1}{em!} \sum_{j=0}^{k-m-1} \frac{(k-m-j)\lambda^j e^{-\lambda}}{j!} \right) - (k-1).\end{aligned}$$

The proof of the theorem is based on the analysis of a greedy algorithm. The algorithm constructs a k-map M_k^a for which $E(Cost(M_k^a))$ can be bounded. We know that $Cost(M_k^*) \leq Cost(M_k^a)$, so Theorems 2.1 and 2.2 together imply that $E(Cost(T_k^*)) \leq E(Cost(M_k^a)) + o(1)$.

To construct the map M_k^a, we start with the optimal unrestricted random mapping, M^*, and make the modifications necessary to convert it to a k-map. This is carried out in three phases as follows. In Phase 1 we start with M^*, and at each vertex v with $\rho(v, M^*) > k$, cut $\rho(v, M^*) - k$ edges into v. These edges are selected so as to minimize the "redirection cost" at the beginning of Phase 2 when the cut edges are replaced by new, more expensive edges. The redirection process may result in overfull vertices. Hence there is another round of cutting at the end of Phase 2. The edges that are cut at the end of Phase 2 get replaced by more expensive edges in Phase 3 using a simple greedy procedure.

To describe the algorithm in more detail, we need some notation. Given a map f and a vertex v, recall that $\mathcal{B}(v, f) = \{x : f(x) = v\}$. For $x \in \mathcal{B}(v, f)$, let $D(x, f)$ be the difference between the cost of the current edge $(x, f(x))$ and the next cheapest edge, i.e. if $(x, f(x)) = c_i(x)$, then $D(x, f) = c_{(i+1)}(x) - c_{(i)}(x)$. If $\rho(v, f) = m$, let $D_{(1)}(v), D_{(2)}(v), ..., D_{(m)}(v)$ denote the order statistics of the variables $\{D(x, f) : x \in \mathcal{B}(v, f)\}$. In addition, if $\rho(v, f) = m > k$, then let $\mathcal{W}(v, f) = \{x \in \mathcal{B}(v, f) : D(x, v) = D_{(j)}(v, f)$ for some $1 \leq j \leq m - k\}$ be the $m-k$ vertices for which $D(x,f)$ is smallest. Finally, let $\mathcal{F}=\{v:\rho(v,M^*)>k\}$ and let $\mathcal{C}=\{v:\rho(v,M^*)\le k\}$. With this notation we can describe an algorithm constructs a heuristic mapping M_k^a.

```
PHASE 1
  • $f = M^*$;
    $\mathcal{R}(1) = \emptyset$;
    $\mathcal{R}(2) = \emptyset$;
  • For each edge $(x,v)$ such that $v \in \mathcal{F}$ and $x \in \mathcal{W}(v,f)$, delete $(x,v)$ from $f$ and add $x$
    to $\mathcal{R}(1)$.
PHASE 2
  • For all roots $x \in \mathcal{R}(1)$, add $e_2(x)$ to $f$.
  • For every $v \in \mathcal{F}$ that now has $\rho(v,f) > k$, delete all edges $(x,v)$ in $f$ such that
    $x \in \mathcal{R}(1)$, and add the vertex $x$ to $\mathcal{R}(2)$.
  • For every $v \in \mathcal{C}$ that now has $\rho(v,f) > k$, for every $x \in \mathcal{W}(v,f)$, delete $(x,v)$ from $f$
    and add $x$ to $\mathcal{R}(2)$.
PHASE 3
  • For $i = 1, \dots, |\mathcal{R}(2)|$,
        { Let $x_i$ be the vertex in $\mathcal{R}(2)$ with the $i$th smallest label. Add to $f$ the cheapest
        edge from $x_i$ to a vertex that is available to $x_i$ and which has not been rejected
        in Phases 1-2;
        }
  • Return $f$;
```

Our goal is to bound the expected cost of M_k^a which is the mapping returned by the algorithm. To this end, let

$$\mathbf{X}=\sum_{v=1}^{n} c_{(1)}(v) = Cost(M^*)$$

be the cost of f at the start of Phase 1, before any edges are deleted. If $e_1(x)$ is one of the edges removed in Phase 1, then $e_1(x)$ will not be used in M_k^a and at best the edge out of x will have cost $c_{(2)}(x)$ in M_k^a. For this reason we refer to $(c_{(2)}(x)-c_{(1)}(x))$ as the "penalty for deleting $e_1(x)$." Let $\mathbf{Y}$ be the total penalty for deleting edges in Phase 1:

$$\mathbf{Y}=\textstyle\sum_{x\in\mathcal{R}(1)}(c_{(2)}(x)-c_{(1)}(x)).$$

Similarly, let $\mathbf{W}$ be the penalty for rejecting edges in Phase 2 and let $\mathbf{Z}$ be the *additional* penalty for adding edges in Phase 3, i.e. the cost over and above that which is accounted for by $\mathbf{X}, \mathbf{Y}$, and $\mathbf{W}$. Then

$$Cost(M_k^a) = \mathbf{X} + \mathbf{Y} + \mathbf{W} + \mathbf{Z}. \tag{4.1}$$

Proof of Theorem 4.1. Fix $k \geq 2$. From the discussion above, it is enough to bound $E(Cost(M_k^a))$. We begin by noting that $\mathbf{X} + \mathbf{Y} = U_n$, so from the proof of Theorem 3.1 we have

$$E(\mathbf{X} + \mathbf{Y}) = 1 + \frac{1}{e} \sum_{m>k} \frac{1}{m!} \sum_{\ell=1}^{m-k} \frac{m-k-\ell+1}{m-\ell+1} + o(1). \tag{4.2}$$

Next, to bound $E(\mathbf{W})$, we write $\mathbf{W} = \sum_{v=1}^{n} \mathbf{Q}_v$, where $\mathbf{Q}_v$ is the penalty for rejecting edges into v during Phases 2. Let $\rho(v) = \rho(v, M^*)$, then we have

$$E(\mathbf{W}) = \sum_{v=1}^{n} E(\mathbf{Q}_v)$$

$$= \sum_{v=1}^{n} \left\{ \sum_{m=0}^{k} \Pr\big(\rho(v) = m\big) E\big(\mathbf{Q}_v | \rho(v) = m\big) + \Pr\big(\rho(v) > k\big) E\big(\mathbf{Q}_v | \rho(v) > k\big) \right\}$$

$$= n \sum_{m=0}^{k} \Pr\big(\rho = m\big) E\big(\mathbf{Q}_{v_1} | \rho = m\big) + n \Pr\big(\rho > k\big) E\big(\mathbf{Q}_{v_1} | \rho > k\big) \tag{4.3}$$

where v_1 is any fixed vertex and $\rho = \rho(v_1)$. Since $\Pr(\rho = m) \to \frac{1}{em!}$ as $n \to \infty$, it is enough to determine $\limsup_{n\to\infty} nE(\mathbf{Q}_{v_1} | \rho = m)$ and $\limsup_{n\to\infty} nE(\mathbf{Q}_{v_1} | \rho > k)$. These limits are obtained in Lemmas 4.3 and 4.4 below.

To prove the lemmas we first define variables $R = |\mathcal{R}(1)|$ and $V = |\{x \in \mathcal{R}(1) : e_2(x) = (x, v_1)\}|$ where v_1 is a fixed vertex. Then we have

Lemma 4.2 For $0 \leq m \leq k$ and $j \geq 0$

$$\lim_{n\to\infty} \Pr(V = j | \rho = m) = \frac{\lambda^j e^{-\lambda}}{j!}$$

where $\lambda = \lambda(k)$.

Proof. Observe that *given* $R = r$ and $\rho = m \leq k$, the distribution of V is $Bin(r, \frac{1}{n-1})$. So if $\lambda n - k^2 n^{3/4} \leq r \leq \lambda n + k^2 n^{3/4}$, it follows from the Poisson approximation for the Binomial distribution that

$$|\Pr(V = j | R = r, \rho = m) - \frac{e^{-\lambda}\lambda^j}{j!}| \leq \frac{C_j}{n^{1/4}}$$

where C_j is a constant which may depend on j but which does not depend on n. Martingale concentration results (see, for example, [15] and the appendix) establish that $\Pr(\lambda n - k^2 n^{3/4} \le R \le \lambda n + k^2 n^{3/4}) > 1 - \exp(-n^{1/4})$ for all large n. Furthermore, since $\Pr(\rho = m)$ is bounded away from zero as $n \to \infty$, we also have

$$\Pr\left(\lambda n - k^2 n^{3/4} \le R \le \lambda n + k^2 n^{3/4} \middle| \rho = m\right) > 1 - C_m \exp(-n^{1/4}),$$

where C_m is a constant which may depend on m but which does not depend on n. Using these bounds, we obtain

$$\begin{aligned}
&|\Pr(V = j|\rho = m) - \frac{e^{-\lambda}\lambda^j}{j!}| \\
&\le \sum_{\{r:|r-\lambda n|\le k^2 n^{3/4}\}} \left|\Pr(V = j|R = r, \rho = m) - \frac{e^{-\lambda}\lambda^j}{j!}\right| \Pr(R = r|\rho = m) \\
&+ \sum_{\{r:|r-\lambda n|> k^2 n^{3/4}\}} \left|\Pr(V = j|R = r, \rho = m) - \frac{e^{-\lambda}\lambda^j}{j!}\right| \Pr(R = r|\rho = m) \\
&\le \frac{C_j}{n^{1/4}} + 2C_m \exp(-n^{1/4})
\end{aligned}$$

and the lemma follows. ∎

Remark. We also note here that $\Pr(V = j|R = r, \rho = m) = \binom{r}{j}(\frac{1}{n-1})^j(1 - \frac{1}{n-1})^{r-j} \le \frac{1}{j!}$ for all possible values of r, so $\Pr(V = j|\rho = m) \le \frac{1}{j!}$ for all $j \ge 0$.

Lemma 4.3 *For* $0 \le m \le k$

$$\lim_{n\to\infty} nE(\mathbf{Q}_{v_1}|\rho = m) = \sum_{j=k-m+1}^{\infty} \frac{\lambda^j e^{-\lambda}}{j!} \sum_{l=1}^{j+m-k} \frac{j+m-k-l+1}{j+m-l+1}$$

where $\lambda = \lambda(k)$.

Proof. The first step is to write

$$\lim_{n\to\infty} nE(\mathbf{Q}_{v_1}|\rho = m) =$$

$$\lim_{n\to\infty} \sum_{j=k-m+1}^{n-m} nE(\mathbf{Q}_{v_1}|V = j, \rho = m)\Pr(V = j|\rho = m). \tag{4.4}$$

We compute $\lim_{n\to\infty} nE(\mathbf{Q}_{v_1}|V = j, \rho = m)$ by using a further conditioning argument. The calculation is divided into two cases.

Case 1: $m = 0$

Let $\mathcal{V} = \{x : x \in \mathcal{R}(1) \text{ and } e_2(x) = (x, v_1)\}$, then the event $\{\mathcal{V} = A, \rho = 0\}$, where $A \subseteq [n]$ and $|A| = j$, is measurable with respect to the σ-algebra generated by the variables $\{X_{(i)}(v) : i = 1, 2, \text{ and } 1 \leq v \leq n\} \cup \{c_{(2)}(v) - c_{(1)}(v) : 1 \leq v \leq n\}$ (as defined in Section 1). In particular, the event $\{\mathcal{V} = A, \rho = 0\}$ is independent of the variables $\{c_{(3)}(x) - c_{(2)}(x) : x \in A\}$. Now for any $A \subseteq [n]$ such that $|A| = j$, let $D_{(1)}(A), ..., D_{(j)}(A)$ denote the order statistics of the variables $\{c_{(3)}(x) - c_{(2)}(x) : x \in A\}$. It follows from Lemma 3.2 that

$$\begin{aligned} nE\left(\mathbf{Q}_{v_1}\Big|\mathcal{V} = A, |A| = j, \rho = 0\right) &= nE\left(\sum_{i=1}^{j-k} D_{(i)}(A)\Big|\mathcal{V} = A, |A| = j, \rho = 0\right) \\ &= nE\left(\sum_{i=1}^{j-k} D_{(i)}(A)\right) \\ &= \frac{n}{n-2}\sum_{\ell=1}^{j-k}\frac{j-k-\ell+1}{j-\ell+1} \end{aligned}$$

since the variables $\{c_{(3)}(x) - c_{(2)}(x) : x \in A\}$ are i.i.d. $\exp(n-2)$. It follows that

$$nE\left(\mathbf{Q}_{v_1}\big|V = j, \rho = 0\right) = \frac{n}{n-2}\sum_{\ell=1}^{j-k}\frac{j-k-\ell+1}{j-\ell+1} \leq 3j. \tag{4.5}$$

Since $\Pr(V = j|\rho = 0) \leq \frac{1}{j!}$ for $j \geq 0$, the result now follows for $m = 0$ by Lemma 4.2, (4.5) and dominated convergence applied to (4.4) with $m = 0$.

Case 2: $0 < m \leq k$.

Let $\mathcal{V}$ be defined as in Case 1 above, then the event $\{\mathcal{V} = A, (M^*)^{-1}(v_1) = B, \rho = m\}$, where $A, B \subseteq [n]$, $|A| = j$, $|B| = m$, and $B \cap A = \emptyset$, is measurable with respect to the σ-algebra generated by the variables $\{X_{(i)}(v) : i = 1, 2, \text{ and } 1 \leq v \leq n\} \cup \{c_{(2)}(v) - c_{(1)}(v) : 1 \leq v \leq n, v \notin B\}$. In particular, the event $\{\mathcal{V} = A, (M^*)^{-1}(v_1) = B, \rho = m\}$ is independent of the variables $\{c_{(2)}(x) - c_{(1)}(x) : x \in B\} \cup \{c_{(3)}(v) - c_{(2)}(v) : v \in A\}$. Let $A' = A \cup B$ and let $D_{(1)}(A'), ..., D_{(m+j)}(A')$ denote the order statistics of the variables $\{c_{(2)}(x) - c_{(1)}(x) : x \in B\} \cup \{c_{(3)}(v) - c_{(2)}(v) : v \in A\}$. Then

$$\begin{aligned} &nE\left(\mathbf{Q}_{v_1}\Big|\mathcal{V} = A, (M^*)^{-1}(v_1) = B, |A| = j, \rho = m\right) \\ &= nE\left(\sum_{i=1}^{j+m-k} D_{(i)}(A')\Big|\mathcal{V} = A, (M^*)^{-1}(v_1) = B, |A| = j, \rho = m\right) \\ &= nE\left(\sum_{i=1}^{j+m-k} D_{(i)}(A')\right). \end{aligned} \tag{4.6}$$

Now the calculation of $nE(\sum_{i=1}^{j+m-k} D_{(i)}(A'))$ requires some work since the variables $\{c_{(2)}(x) - c_{(1)}(x) : x \in B\} \cup \{c_{(3)}(v) - c_{(2)}(v) : v \in A\}$ are independent but not identically distributed. In particular, $c_{(2)}(x) - c_{(1)}(x) \sim \exp(n-1)$ for every $x \in B$, whereas $c_{(3)}(v) - c_{(2)}(v) \sim \exp(n-2)$ for every $v \in A$, so their joint density function is given by

$$f(u_1, ..., u_{j+m})$$

$$= (n-1)^m (n-2)^j \exp\left(-(n-1)\sum_{i=1}^{m} u_i - (n-2)\sum_{i=m+1}^{j+m} u_i\right)$$

on $(R^+)^{j+m}$. Define the function $g : (R^+)^{j+m} \to R^+$ by $g(u_1, ..., u_{j+m}) = \sum_{i=1}^{j+m-k} u_{(i)}$ where $u_{(i)}$ is the ith smallest of the coordinates $u_1, ..., u_{j+m}$. Then we have

$$nE\left(\sum_{i=1}^{j+m-k} D_{(i)}(A')\right) = n\int_{(R^+)^{j+m}} g(\vec{u}) \cdot f(\vec{u}) d\vec{u}$$

$$= n\int_{(R^+)^{j+m}} g(\vec{u}) \cdot \tilde{f}(\vec{u}) d\vec{u} + n\int_{(R^+)^{j+m}} g(\vec{u}) \cdot (f(\vec{u}) - \tilde{f}(\vec{u})) d\vec{u} \tag{4.7}$$

where $\tilde{f}(u_1, ..., u_{j+m}) = (n-2)^{j+m}\exp(-(n-2)\sum_{i=1}^{j+m} u_i)$ is the joint density of $j+m$ i.i.d. $\exp(n-2)$ random variables. By Lemma 3.2 we have

$$n\int_{(R^+)^{j+m}} g(u) \cdot \tilde{f}(u) du = \frac{n}{n-2}\sum_{\ell=1}^{j+m-k} \frac{j+m-k-\ell+1}{j+m-\ell+1} \tag{4.8}$$

To bound the second term on the RHS of (4.7), observe that

$$\begin{aligned} & n\Big|\int_{(R^+)^{j+m}} g(\vec{u}) \cdot (f(\vec{u}) - \tilde{f}(\vec{u})) d\vec{u}\Big| \\ & \leq n\int_{(R^+)^{j+m}} g(\vec{u}) \cdot \tilde{f}(\vec{u})\Big|1 - \frac{f(\vec{u})}{\tilde{f}(\vec{u})}\Big| d\vec{u} \\ & = n\int_{[0,\frac{1}{\sqrt{n}}]^{j+m}} g(\vec{u}) \cdot \tilde{f}(\vec{u})\Big|1 - \frac{f(\vec{u})}{\tilde{f}(\vec{u})}\Big| d\vec{u} \\ & + n\int_{([0,\frac{1}{\sqrt{n}}]^{j+m})^c} g(\vec{u}) \cdot \tilde{f}(\vec{u})\Big|1 - \frac{f(\vec{u})}{\tilde{f}(\vec{u})}\Big| d\vec{u}. \end{aligned} \tag{4.9}$$

For $\vec{u} \in [0, \frac{1}{\sqrt{n}}]^{j+m}$,

$$\Big|1 - \frac{f(\vec{u})}{\tilde{f}(\vec{u})}\Big| = \Big|1 - \left(\frac{n-1}{n-2}\right)^m e^{-\sum_{i=1}^m u_i}\Big| \leq \frac{C(k)}{\sqrt{n}} \tag{4.10}$$

where $C(k)$ is a constant which may depend on k but which does not depend on n. Thus

$$n\int_{[0,\frac{1}{\sqrt{n}}]^{j+m}} g(\vec{u}) \cdot \tilde{f}(\vec{u})\Big|1 - \frac{f(\vec{u})}{\tilde{f}(\vec{u})}\Big| d\vec{u} \leq \frac{C(k)}{\sqrt{n}} \cdot n\int_{(R^+)^{j+m}} g(u) \cdot \tilde{f}(u) du$$

$$= \frac{C(k)}{\sqrt{n}} \cdot \frac{n}{n-2}\sum_{\ell=1}^{j+m-k} \frac{j+m-k-\ell+1}{j+m-\ell+1} \leq \frac{3C(k)j}{\sqrt{n}}. \tag{4.11}$$

To bound the second term on the RHS of (4.9), observe that

$$
\begin{aligned}
n\int_{([0,\frac{1}{\sqrt{n}}]^{j+m})^c} g(\vec{u})\cdot \tilde{f}(\vec{u})\Big|1-\frac{f(\vec{u})}{\tilde{f}(\vec{u})}\Big|d\vec{u} &\le 2n\sum_{i=1}^{j+m}\int_{([0,\frac{1}{\sqrt{n}}]^{j+m})^c} u_i\tilde{f}(\vec{u})d\vec{u} \\
&= 2n(j+m)\int_{\frac{1}{\sqrt{n}}}^{\infty}(n-2)ue^{-(n-2)u}du \\
&\le 4n^2\cdot\left(\frac{2\exp(-(n-2)/\sqrt{n})}{\sqrt{n}}\right) \qquad (4.12)\\
&\le 8n^{3/2}\exp(-n^{1/3}).
\end{aligned}
$$

It follows from (4.6)-(4.12) that

$$
\Big|E\left(\sum_{i=1}^{j+m-k} D_{(i)}(A')\right)-\sum_{\ell=1}^{j+m-k}\frac{j+m-k-\ell+1}{j+m-\ell+1}\Big|\le\frac{C'(k)j}{\sqrt{n}}
$$

and hence

$$
\Big|nE(\mathbf{Q}_{v_1}|V=j,\rho=m)-\sum_{\ell=1}^{j+m-k}\frac{j+m-k-\ell+1}{j+m-\ell+1}\Big|\le\frac{C'(k)j}{\sqrt{n}} \qquad (4.13)
$$

where $C'(k)$ is a constant may depend on k but which does not depend on j or n. The result now follows from Lemma 4.2, (4.13) and dominated convergence applied to (4.4), since $\Pr(V=j|\rho=m)\le\frac{1}{j!}$ for all $j\ge 0$. ■

Lemma 4.4

$$
\limsup_{n\to\infty} nE(\mathbf{Q}_{v_1}|\rho>k)\Pr(\rho>k)\le\lambda(k)\cdot(1-\sum_{m=0}^{k}\frac{1}{em!}).
$$

Proof. If $\rho=m>k$, then all edges which are mapped to v_1 at the start of Phase 2 are rejected without making any comparisons. In particular, suppose that $x\in\mathcal{R}(1)$ and $e_2(x)=(x,v_1)$, then the edge $e_2(x)$ is rejected. The penalty paid for rejecting $e_2(x)$ is the minimum *additional* cost of finding a suitable edge out of x, i.e. the penalty for rejecting $e_2(x)$ is $c_{(3)}(x)-c_{(2)}(x)$. Arguments similar to those made in the proof of Lemma 4.3 establish that, that for any $A\subseteq[n]$,

$$
nE(\mathbf{Q}_{v_1}\Big|\mathcal{V}=A,\rho>k)=nE\left(\sum_{x\in A}c_{(3)}(x)-c_{(2)}(x)\right)=\frac{n|A|}{n-2}
$$

where $\mathcal{V}$ is as in the proof of Lemma 4.3. It follows that

$$\begin{aligned}
nE(\mathbf{Q}_{v_1}|\rho > k)\Pr(\rho > k) &= n\sum_{j=0}^{n} E(\mathbf{Q}_{v_1}\Big| V = j, \rho > k)\Pr(V = j|\rho > k)\Pr(\rho > k)\\
&= \frac{n}{n-2}\sum_{j=0}^{n} j\Pr(V = j|\rho > k)\Pr(\rho > k)\\
&= \frac{n}{n-2}E(V|\rho > k)\Pr(\rho > k)\\
&= \frac{n}{n-2}\sum_{m=k+1}^{n} E(V|\rho = m)\Pr(\rho = m)
\end{aligned} \tag{4.14}$$

where $V = |\mathcal{V}|$. Observe that *given* that $\rho = m > k$ and $R = r$ where $R = |\mathcal{R}(1)|$, $V \sim Bin(r - m + k, \frac{1}{n-1})$. So

$$E(V|\rho = m) = \frac{1}{n-1}E(R - m + k|\rho = m).$$

Since $\sum_{t=0}^{n} td_t = n = \sum_{t=0}^{n} d_t$, where $d_t = d_t(M^*)$ denotes the number of vertices in M^* with in-degree t, we have

$$R = \sum_{t>k}(t-k)d_t = \sum_{t=0}^{k-1}(k-t)d_t - (k-1)n < n,$$

Thus

$$\begin{aligned}
E(V|\rho = m) &= \frac{1}{n-1}\sum_{t=0}^{k-1}(k-t)E(d_t|\rho = m) - \frac{(k-1)n + (m-k)}{n-1}\\
&\le \frac{1}{n-1}\sum_{t=0}^{k-1}(k-t)E(d_t|\rho = m) - (k-1).
\end{aligned} \tag{4.15}$$

Now for $0 \le t \le k-1 < m$, we must have $E(d_t|\rho = m) \le n-1$, and hence $E(V|\rho = m) \le k^2/2$. To obtain a better bound, we note that for $0 \le t \le k-1$ and $k < m \le n-k$, we have

$$\begin{aligned}
E(d_t|\rho = m) &= (n-1)\Pr(\rho(v') = t|\rho(v_1) = m)\\
&= (n-1)\frac{(n-m)!}{t!(n-m-t)!}\left(\frac{1}{n}\right)^t \frac{(1-2/n)^{n-m-t}}{(1-1/n)^{n-m}}\\
&\le (n-1)\frac{e^{-1}e^{m/n}}{t!(1-2/n)^k}
\end{aligned}$$

where v' is any vertex other than v_1. Substitute this bound into (4.15) to obtain

$$E(V|\rho = m) \le (1-2/n)^{-k}\sum_{t=0}^{k-1}\frac{(k-t)e^{-1}e^{m/n}}{t!} - (k-1)$$

for $k < m < n-k$. Using this bound in (4.14), we obtain

$$\begin{aligned}
nE(\mathbf{Q}_v|\rho > k)\Pr(\rho > k) &\leq \left(\frac{n(1-2/n)^{-k}}{n-2}\right) e^{\frac{\log n}{n}} \left(\sum_{t=0}^{k-1} \frac{(k-t)e^{-1}}{t!}\right) \Pr(k < \rho < \log n) \\
&\quad + k^2 \Pr(\rho \geq \log n) - (k-1)\Pr(\rho > k) \\
&\leq \left(\sum_{t=0}^{k-1} \frac{(k-t)e^{-1}}{t!} - (k-1)\right) \Pr(\rho > k) + o(1) \\
&= \lambda(k)\left(1 - \sum_{m=0}^{k} \frac{1}{em!}\right) + o(1)
\end{aligned}$$

since $\Pr(\rho \geq \log n) \leq \frac{e}{(\log n)!}$ and $\Pr(\rho > k) = 1 - \sum_{m=0}^{k} \frac{1}{em!} + o(1)$. ∎

It follows from Lemmas 4.3 and 4.4 that

$$\begin{aligned}
\limsup_{n\to\infty} E(\mathbf{W}) \leq \sum_{m=0}^{k} \frac{1}{em!} &\left(\sum_{j=k-m+1}^{\infty} \frac{\lambda^j e^{-\lambda}}{j!} \sum_{l=1}^{j+m-k} \frac{j+m-k-l+1}{j+m-l+1}\right) \\
&+ \lambda\left(1 - \sum_{m=0}^{k} \frac{1}{em!}\right).
\end{aligned} \tag{4.16}$$

It only remains to bound $E(\mathbf{Z})$, where $\mathbf{Z}$ is the *additional cost* for adding edges in Phase 3, i.e. $\mathbf{Z}$ is the cost over and above that which is accounted for by $\mathbf{X}, \mathbf{Y}$, and $\mathbf{W}$. In order to execute Phase 3, we order the elements of $\mathcal{R}(2)$ according to the order of their labels. Suppose that x_i is the ith root in $\mathcal{R}(2)$ and let Γ_i denote the *additional cost* incurred by adding an edge out of x_i in Phase 3. So

$$E(\mathbf{Z}) = E(\sum_{i=1}^{R'} \Gamma_i)$$

where $R' = |\mathcal{R}(2)|$. Let $\mathcal{A}(2)$ denote the set of available vertices at the end of Phase 2 and let $A' = |\mathcal{A}(2)|$, then

$$E(\mathbf{Z}) = \sum_{1\leq r\leq a} \sum_{i=1}^{r} E(\Gamma_i | R' = r, A' = a)\Pr(R' = r, A' = a).$$

We claim that for $1 \leq i \leq r \leq a$

$$E(\Gamma_i | R' = r, A' = a) \leq \frac{n}{(n - 10\log n)(a - i + 1)} - \frac{1}{(n - 10\log n)} + \frac{1}{n^{3/2}}.$$

To prove the claim, fix $1 \leq i \leq r \leq a$. There are two cases to consider:

Case 1: $x_i \in \mathcal{R}(1) \bigcap \mathcal{R}(2)$.

In this case edges $e_1(x_i)$ and $e_2(x_i)$ have been rejected by the algorithm. In Phase 3 we examine edges $e_3(x_i), e_4(x_i), ...$ until an acceptable edge is found. Let τ_i denote the number of edges examined until an edge for x_i is found. Note that if A'_i is the number of available vertices for vertex x_i then τ_i corresponds to the number of draws without replacement until a black ball is drawn from an urn containing A'_i black balls and $n - A'_i - 2$ red balls.

Now if $\tau_i = 1$, then edge $e_3(x_i)$ is accepted, and in this case the additional cost is 0 since the incremental cost $c_{(3)}(x_i) - c_{(2)}(x_i)$ was included in the calculation of $\mathbf{W}$. If $\tau_i = m > 1$, then edge $e_{m+2}(x_i)$ is accepted and the *added* cost of accepting edge $e_{m+2}(x_i)$ is $c_{(m+2)}(x_i) - c_{(3)}(x_i)$. So for $n - 2 \geq m > 1$, we have

$$\begin{aligned}
&E(\Gamma_i \Big| x_i \in \mathcal{R}(1) \cap \mathcal{R}(2), \tau_i = m, R' = r, A' = a) \\
&= E(c_{m+2}(x_i) - c_3(x_i) \Big| x_i \in \mathcal{R}(1) \cap \mathcal{R}(2), \tau_i = m, R' = r, A' = a) \\
&= \frac{1}{n-3} + \frac{1}{n-4} + ... + \frac{1}{n-1-m} \\
&\leq \frac{m-1}{n - 10\log n} + 2\log n \cdot 1_{[m \geq 6\log n]}.
\end{aligned}$$

It follows from the proof of Lemma 2.9, that $|A'_i| \geq \frac{n}{4}$, so $\Pr(\tau_i \geq 6\log n) < (\frac{3}{4})^{6\log n} < \frac{1}{n^{3/2}}$ and thus

$$\begin{aligned}
&E(\Gamma_i | x_i \in \mathcal{R}(1) \cap \mathcal{R}(2), R' = r, A' = a) \\
&\leq \frac{E(\tau_i | x_i \in \mathcal{R}(1) \cap \mathcal{R}(2), R' = r, A' = a)}{n - 10\log n} - \frac{1}{(n - 10\log n)} + \frac{2\log n}{n^{3/2}}.
\end{aligned}$$

Now *given* $A'_i = a_i$, standard results for sampling without replacement yield

$$E(\tau_i | x_i \in \mathcal{R}(1) \cap \mathcal{R}(2), R' = r, A' = a, A'_i = a_i) \leq \frac{n}{a_i} \leq \frac{n}{a - i + 1}$$

always. The last inequality follows from a simple observation: if there are a available vertices at the start of Phase 3, then as each root finds a vertex the number of available vertices is reduced by *at most* 1, so $A'_i \geq A' - i + 1$ *always*. It follows that

$$E(\tau_i | x_i \in \mathcal{R}(1) \cap \mathcal{R}(2), R' = r, A' = a) \leq \frac{n}{a - i + 1}$$

and hence

$$E(\Gamma_i | x_i \in \mathcal{R}(1) \cap \mathcal{R}(2), R' = r, A' = a) \leq \frac{n}{(n - 10\log n)(a - i + 1)} - \frac{1}{n - 10\log n} + \frac{2\log n}{n^{3/2}}.$$

Case 2: $x_i \in \mathcal{R}(2) \setminus \mathcal{R}(1)$.

In this case edge $e_1(x_i)$ has been rejected by the algorithm during Phase 2. In Phase 3 edges $e_2(x_i), e_3(x_i), ...$ are examined until an acceptable edge is found. As in Case 1, $\Gamma_i = 0$ if $\tau_i = 1$, and for $n-1 \geq m > 1$

$$\begin{aligned}
&E(\Gamma_i | x_i \in \mathcal{R}(2) \setminus \mathcal{R}(1), \tau_i = m, R' = r, A' = a) \\
&= E(c_{m+1}(x_i) - c_2(x_i) | x_i \in \mathcal{R}(2) \setminus \mathcal{R}(1), \tau_i = m, R' = r, A' = a) \\
&= \frac{1}{n-2} + \frac{1}{n-3} + ... + \frac{1}{n-m} \\
&\leq \frac{m-1}{n - 10\log n} + 2\log n \cdot 1_{[k \geq 6\log n]}.
\end{aligned}$$

The same argument as given in Case 1 yields

$$E(\Gamma_i | x_i \in \mathcal{R}(2) \setminus \mathcal{R}(1), R' = r, A' = a) \leq \frac{n}{(n - 10\log n)(a - i + 1)} - \frac{1}{n - 10\log n} + \frac{2\log n}{n^{3/2}}.$$

So we in all cases, for $i \leq r$, we have

$$E(\Gamma_i | R' = r, A' = a) \leq \frac{n}{(n - 10\log n)(a - i + 1)} - \frac{1}{n - 10\log n} + \frac{2\log n}{n^{3/2}}.$$

Thus

$$\begin{aligned}
E(\mathbf{Z}) &\leq \sum_{1 \leq r \leq a} \sum_{i=1}^{r} E(\Gamma_i | R' = r, A' = a) \Pr(R' = r, A' = a) \\
&\leq \sum_{1 \leq r \leq a} \sum_{i=1}^{r} \left(\frac{n}{(n - 10\log n)(a - i + 1)} - \frac{1}{n - 10\log n} + \frac{2\log n}{n^{3/2}} \right) \Pr(R' = r, A' = a) \\
&\leq \frac{n}{n - 10\log n} \left(E\left(\log(\frac{A'}{A' - R'}) \right) - \frac{E(R')}{n} \right) + \frac{2E(R')\log n}{n^{3/2}}.
\end{aligned}$$

In the Appendix we show that there is a constant C_k, which may depend on k but which does not depend on n, such that for the event

$$\gamma = \{|A' - n\alpha(k)| \leq C_k n^{3/4}, |R' - n\beta(k)| \leq C_k n^{3/4}\}$$

we have

$$\Pr(\gamma) \geq 1 - 4\exp(-n^{1/4}).$$

It follows that for all large n

$$\begin{aligned}
E(\mathbf{Z}) &\leq \left(\log\left(\frac{\alpha(k)}{\alpha(k) - \beta(k)} \right) - \beta(k) + o(1) \right) \Pr(\gamma) + 2\log(n)\Pr(\gamma^c) \\
&= \log\left(\frac{\alpha(k)}{\alpha(k) - \beta(k)} \right) - \beta(k) + o(1)
\end{aligned} \tag{4.17}$$

and Theorem 4.1 now follows from (4.2), (4.16), and (4.17). ■

It is not difficult to describe various ways to improve the algorithm for constructing M_k^a. However, with each improvement of the algorithm, the analysis of the algorithm becomes more complicated.

§5 Conclusions

The table below summarizes our results for a few small values of k.

Asymptotic Bounds		
k	Lower	Upper
2	1.03892136	1.06806181
3	1.00656287	1.00755907
4	1.00097152	1.00100255
5	1.00012721	1.00012800
6	1.00001487	1.00001490
7	1.00000156	1.00000157

Limited simulation data for two to four hundred vertices suggest that, in the case $k = 2$, the lower bound is sharper than the upper bound. In this paper we have considered $exp(1)$ costs on the edges. It likely that, as in the $k = \infty$ case (Hansen[8]), the arguments can be carried out for other distributions as well, provided the order statistics of the chosen edge distribution are well behaved.

We note that for $exp(1)$ edge costs, Aldous[13] has shown that the expected cost of the optimal assignment converges to a constant as $n \to \infty$. This result also holds in the case $k = \infty$ for fairly general edge distributions. We speculate that a similar result might hold $2 \leq k < \infty$. The methods used in the case $k = \infty$ are insufficient to prove this, but it may be possible to adapt Aldous' "objective method"to obtain the result. In the case $k = \infty$, the limiting constant can be determined, but it is not known for the assignment problem with either $exp(1)$ or $U(0, 1)$ edge costs.

Acknowledgement. Our first version of this paper used uniform $U(0, 1)$ edge costs. We thank the anonymous referees for suggesting $exp(1)$, and also for catching an error in one of our proofs. The authors also thank Ljubomir Perkovic for use of his computer and optimization software.

References

[1] Goemans, M.X. and Kodialam, M.S., A lower bound on the expected cost of an optimal assignment, *Math. Oper. Res.* **18** (1993) 267-274.

[2] Birgitta Olin, Asymptotic properties of random assignment problems, Ph. D. thesis, Kungl Tekniska Hogskolan, (1992) Stockholm, Sweden.

[3] Don Coppersmith and Gregory Sorkin , Constructive bounds and exact expectations for the random assignment problem, IBM Research Report RC 21133(94490), (1998) 1 -35.

[4] A.M. Frieze, On the value of a random minimum spanning tree problem, *Discrete Applied Mathematics* **10** (1985) 47 – 56.

[5] A. Beveridge, A.M. Frieze and C. J. H. McDiarmid,Random minimum length spanning trees in regular graphs, preprint.

[6] A.M. Frieze and C. J. H. McDiarmid, On random minimum length spanning trees, *Combinatorica* **9** (4) (1989) 363 -374.

[7] Khuller, Raghavachari, and Young,Low-degree spanning trees of small weight, *SIAM J.Comput.* **25** (1996) 355–368.

[8] Jennie C. Hansen, Limit laws for the optimal directed tree with random costs, *Combinatorics, Probability and Computing* **6** (1997) 315-335 .

[9] Colin McDiarmid, On the greedy algorithm with random costs, *Mathematical Programming* **36** (1986) 245-255.

[10] Bazaraa, Jarvis, and Sherali, Linear programming and network flows, (1990), page 481.

[11] R.M. Karp and M. Steele, Probabilistic analysis of heuristics, in The traveling salesman problem, ed. Eugene L. Lawler et. al. , John Wiley & Sons Ltd.(1985) Great Britain, pages 181–205.

[12] P.Flajolet and M. Soria, General Combinatorial Schemas with Gausssian Limit Distributions and Exponential Tails, *Discrete Math.* **114** (1993) 159-180.

[13] D. Aldous, Asymptotics in the random assignment problem, *Probab. Theory Appl.* **93** (1992) 507-534.

[14] Garey and Johnson, Computers and Intractability, Freeman (1979),page 206.

[15] J. Michael Steele, Probability in Combinatorial Optimization, SIAM, (1997) Philadelphia, PA.

Appendix

Let Π_k be the following computational problem: given as input a cost matrix C, find a minimum cost k-tree. Then we have

Lemma A.1 *Π_k is NP-hard.*

Proof. Let U_k be the restriction of Π_k to *symmetric* matrices whose entries are all zeroes and ones. The instances of U_k correspond, in an obvious way, to instances of the NP-complete degree-k constrained spanning tree problem for undirected graphs (Garey and Johnson [14], page 206, comment to ND1). The correspondence is: include an edge $\{i, j\}$ in the undirected graph G if and only if $Cost((i, j)) = Cost((j, i)) = 0$. Then G has an undirected degree-k spanning tree iff the optimal (directed) k-tree for C has cost 0. Thus the existence of a polynomial time algorithm for Π_k would imply the existence of a polynomial time algorithm for an NP complete problem. ■

In the remainder of this appendix, we prove the martingale concentration results used in the proofs of the main theorems.

Lemma A.2 *Let M^* be a uniform random mapping on $\{1, 2, ..., n\}$, then for $0 \le k < n$ and $t > 0$,*

$$\Pr(|d_k - E(d_k)| \ge t) \le 2\exp(-t^2/2n)$$

where $d_k = d_k(M^*)$ denotes the number of vertices with in-degree k in M^*.

Proof. The proof of the lemma is an application of Azuma's inequality for martingale differences, and is based on the ideas in Steele [15]. Let $Z_1, Z_2, ..., Z_n$ be i.i.d. random variables such that Z_i corresponds to the 'choice' made by vertex i, i.e. $Z_i = j$ if and only if $M^*(i) = j$. We write $d_k = d_k(Z_1, Z_2, ..., Z_n)$ to emphasize the dependence of d_k on the variables $Z_1, Z_2, ..., Z_n$ which determine the mapping M^*. Next, let $\mathcal{F}_i = \sigma(Z_1, ..., Z_i)$ denote the σ-algebra generated by the variables $Z_1, Z_2, ..., Z_i$ and let $\mathcal{F}_0$ denote the trivial σ-algebra. The variables $E(d_k|\mathcal{F}_0), E(d_k|\mathcal{F}_1), ..., E(d_k|\mathcal{F}_n)$ form a martingale with $E(d_k|\mathcal{F}_0) = E(d_k)$ and $E(d_k|\mathcal{F}_n) = d_k$, and the variables $W_i = E(d_k|\mathcal{F}_i) - E(d_k|\mathcal{F}_{i-1}), i = 0, 1, 2, ..., n$, form a martingale difference sequence. Since $d_k - E(d_k) = \sum_{i=1}^n W_i$, Azuma's inequality for martingale differences tells us that

$$\Pr(|d_k - E(d_k)| \ge t) = \Pr(|\sum_{i=1}^{n} W_i| > t) \le 2\exp(-t^2/2\sum_{i=1}^{n} \|W_i\|_\infty^2).$$

So the lemma follows provided that $\|W_i\|_\infty^2 \le 1$ for all $1 \le i \le n$.

To bound on $\|W_i\|_\infty^2$ we use the following trick. Let $\hat{Z}_1, \hat{Z}_2, ..., \hat{Z}_n$ be a sequence of i.i.d. random variables which are uniform on $\{1, 2, ..., n\}$ and which are independent of the variables $Z_1, Z_2, ..., Z_n$. We claim that

$$E(d_k(Z_1, ..., Z_i, ..., Z_n)|\mathcal{F}_{i-1}) = E(d_k(Z_1, ..., Z_{i-1}, \hat{Z}_i, Z_{i+1}, ..., Z_n)|\mathcal{F}_i).$$

To see this, note that the conditional distribution of $d_k(Z_1, ..., Z_{i-1}, \hat{Z}_i, Z_{i+1}, ..., Z_n)$ given $\mathcal{F}_i$ is the same as the conditional distribution of $d_k(Z_1, ..., Z_{i-1}, Z_i, Z_{i+1}, ..., Z_n)$ given $\mathcal{F}_{i-1}$ since the σ-algebra $\mathcal{F}_i$ is independent of $\hat{Z}_i$ and only gives us information about $Z_1, Z_2, ..., Z_{i-1}$. Thus we can write

$$W_i = E(d_k(Z_1, ..., Z_i, ..., Z_n) - d_k(Z_1, ..., Z_{i-1}, \hat{Z}_i, Z_{i+1}, ..., Z_n)|\mathcal{F}_i)$$

i.e., W_i can be written as a single conditional expectation with respect to $\mathcal{F}_i$. Now it is easy to see that (without conditioning)

$$|d_k(Z_1, ..., Z_i, ..., Z_n) - d_k(Z_1, ..., Z_{i-1}, \hat{Z}_i, Z_{i+1}, ..., Z_n)| \leq 1$$

since all 'choices', except that made by vertex i, are the same and the choices made by Z_i and $\hat{Z}_i$ can only create a discrepency of at most 1 in the count of the number of vertices with in-degree k. Conditioning does not increase the L_∞ bound, so we get $\|W_i\|_\infty \leq 1$ too. It follows from Azuma's inequality that for any $\lambda > 0$,

$$\Pr(|d_k(f) - E(d_k(f)| \geq t) = \Pr(|\sum_{i=1}^{n} W_i| > t) \leq 2\exp(-t^2/2n).$$

■

As an application of Lemma A.2 we have

Corollary A.3 *For all large n and fixed $k \geq 2$,*

$$\Pr(|d_m - \frac{n}{em!}| < n^{3/4}, 0 \leq m \leq k) \geq 1 - \exp(-n^{1/4})$$

and

$$\Pr(|R - n\lambda(k)| \leq k^2 n^{3/4}) \geq 1 - \exp(-n^{1/4})$$

where $R = |\mathcal{R}(1)|$ and $\lambda(k) = \sum_{m=0}^{k-1} \frac{k-m}{em!} - (k-1)$.

Proof. Since $E(d_m) = n\binom{n}{m}(\frac{1}{n})^m(1-\frac{1}{n})^{n-m}$ for $0 \leq m \leq k$, there is a constant C_k which does not depend on n such that $|E(d_m) - \frac{n}{em!}| < C_k$ for $0 \leq m \leq k$. Applying Lemma A.2, we obtain

$$\Pr(|d_m - \frac{n}{em!}| > n^{3/4}) \leq \Pr(|d_m - E(d_m)| > \frac{n^{3/4}}{2}) \leq 2\exp(\frac{-n^{1/2}}{8})$$

for $0 \leq m \leq k$ and this establishes the first part of the result.

Now recall that

$$R = |\mathcal{R}(1)| = \sum_{m>k}(m-k)d_m = \sum_{m=0}^{k}(k-m)d_m - (k-1)n$$

since $\sum_{m=1}^{n} md_m = n = \sum_{m=0}^{n} d_m$, so the event $\{|d_m - \frac{n}{em!}| < n^{3/4}, 0 \leq m \leq k\}$ is contained in the event $\{|R - n\lambda(k)| \leq k^2 n^{3/4}\}$, and the result follows.

Next, recall that $\mathcal{A}(2)$ denotes the set of available vertices at the start of Phase 3 and $\mathcal{R}(2)$ denotes the roots at the start of Phase 3 in the algorithm from Section 4. Then adapting the proof of Lemma A.2, we obtain

Lemma A.4 *For all large n and fixed $k \geq 2$, there exists a constant C_k which does not depend on n, such that*

$$\Pr(|A' - \alpha(k)n| \leq C_k n^{3/4}, |R' - \beta(k)n| \leq C_k n^{3/4})$$

$$\geq 1 - 4\exp(-n^{1/4})$$

where $A' = |\mathcal{A}(2)|$, $R' = |\mathcal{R}(2)|$,

$$\alpha(k) = \sum_{m=0}^{k-1} \frac{1}{em!} \sum_{j=0}^{k-m-1} \frac{\lambda^j e^{-\lambda}}{j!},$$

$$\beta(k) = \left(\sum_{m=0}^{k-1} \frac{1}{em!} \sum_{j=0}^{k-m-1} \frac{(k-m-j)\lambda^j e^{-\lambda}}{j!} \right) - (k-1),$$

and $\lambda = \lambda(k)$ (as defined in Corollary A.3).

Proof. We define the random variable

$$g(d_0, d_1, ..., d_{k-1}) = \sum_{m=0}^{k-1} d_m \sum_{j=0}^{k-m-1} \binom{R}{j} \left(\frac{1}{n-1}\right)^j \left(\frac{n-2}{n-1}\right)^{R-j}$$

where $R = |\mathcal{R}(1)|$ is number of roots at the end of Phase 1. Standard calculations and approximations establish that if $|d_m - \frac{n}{em!}| < n^{3/4}$ for $0 \leq m \leq k$, then $|R - n\lambda(k)| \leq k^2 n^{3/4}$ and there is some constant C'_k which does not depend on n such that

$$|g(\vec{d}) - n\alpha(k)| \leq C'_k n^{3/4}.$$

We show that

$$\Pr\left(|A' - g(\vec{d})| < n^{3/4}, |d_m - \frac{n}{em!}| < n^{3/4}, 0 \leq m \leq k\right) \geq 1 - 2\exp(-n^{1/4}) \qquad (A.1)$$

and, since

$$\{|A' - g(\vec{d})| < n^{3/4}, |d_m - \frac{n}{em!}| < n^{3/4}, 0 \leq m \leq k\} \subseteq \{|A' - n\alpha(k)| \leq (C'_k + 1)n^{3/4}\},$$

we obtain

$$\Pr\left(|A' - n\alpha(k)| \leq (C'_k + 1)n^{3/4}\right) > 1 - 2\exp(-n^{1/4}). \qquad (A.2)$$

Let $\mathcal{A}(1)$ denote the set of available vertices at the end of Phase 1 and let ξ denote the event that $\mathcal{A}(1) = A, \mathcal{R}(1) = B$, and $M^* = f$, where f is a mapping such that $|d_m(f) - \frac{n}{em!}| < n^{3/4}$ for $0 \leq m \leq k$ and A and B are subsets of vertices such that $|B| \leq |A|$. We claim that $E(A'|\xi) = g(\vec{d}(f))$. To see this, suppose that $v \in A$ and $\rho(v, f) = m$. Then $v \in \mathcal{A}(2)$ only if the number of roots in B which are mapped to v in Phase 2 is less than $k - m$, so

$$\Pr(v \in \mathcal{A}(2)|\xi, \rho(v, f) = m) = \sum_{j=0}^{k-m-1} \binom{r}{j} \left(\frac{1}{n-1}\right)^j \left(\frac{n-2}{n-1}\right)^{r-j}$$

where $r = |B| = |\mathcal{R}(1)|$. It follows that

$$\begin{aligned} E(A'|\xi) &= \sum_{v \in A} \Pr(v \in \mathcal{A}(2)|\xi) \\ &= \sum_{m=0}^{k-1} d_m(f) \sum_{j=0}^{k-m-1} \binom{r}{j} \left(\frac{1}{n-1}\right)^j \left(\frac{n-2}{n-1}\right)^{r-j} \\ &= g(\vec{d}(f)). \end{aligned}$$

(Note that *given* the event ξ, the variable $g(\vec{d}(f)) = g(d_0(f), d_1(f), ..., d_k(f))$ is completely determined, since we have conditioned on $M^* = f$).

Now fix ξ , and let P_ξ denote the conditional probability measure $P(\cdot|\xi)$.
Let $x_1, x_2, ..., x_r$ denote the vertices in B and let Z_i denote the vertex that x_i is mapped to in Phase 2, i.e. $Z_i = v$ if $e_2(x_i) = (x_i, v)$. Now *given* the event ξ, the variables $Z_1, Z_2, ..., Z_r$ are independent and each variable Z_i is uniformly distributed over the set $\{1, 2, ..., n\} \setminus \{f(x_i)\}$. In this (conditional) probability space, let $\mathcal{F}_i = \sigma(Z_1, Z_2, ..., Z_i)$ denote the σ-algebra generated by the variables $Z_1, Z_2, ..., Z_i$ and let $\mathcal{F}_0$ denote the trivial σ- algebra. Let $W_k = E_\xi(A'|\mathcal{F}_i) - E_\xi(A'|\mathcal{F}_{i-1})$. So Azuma's inequality tells us that

$$P_\xi(|A' - g(\vec{d}(f))| > n^{3/4}) = P_\xi(|A' - E_\xi(A')| > n^{3/4}) \leq 2\exp(-n^{3/2}/2\sum_{i=1}^{r} \|W_i\|_\infty^2).$$

As in the proof of Lemma A.2, let $\hat{Z}_1, \hat{Z}_2, ..., \hat{Z}_r$ be a sequence of independent random variables such that $\hat{Z}_i \sim Z_i$ for $1 \leq i \leq r$ and which are also independent of the variables $Z_1, Z_2, ..., Z_r$. Then

$$W_i = E_\xi(A'(Z_1, ..., Z_i, ..., Z_r) - A'(Z_1, ..., Z_{i-1}, \hat{Z}_i, Z_{i+1}, ..., Z_r)|\mathcal{F}_i)$$

and $\|W_i\|_\infty \leq 1$ for $1 \leq i \leq r$ since given ξ,

$$|A'(Z_1, ..., Z_i, ..., Z_r) - A'(Z_1, ..., Z_{i-1}, \hat{Z}_i, Z_{i+1}, ..., Z_r)| \leq 1.$$

Thus

$$P_\xi(|A' - g(\vec{d}(f))| > n^{3/4}) \leq 2\exp(-n^{3/2}/2r) \leq \exp(-n^{1/4})$$

since $r = |B| \le n$. It follows from this and Lemma A.2, that

$$\Pr\left(|A' - g(\vec{g})| < n^{3/4}, |d_m - \frac{n}{em!}| < n^{3/4}, 0 \le m \le k\right) =$$

$$\begin{aligned}
&= \sum_{\xi} \Pr(|A' - g(\vec{d}(f))| < n^{3/4} | \xi) \Pr(\xi) \\
&\ge \sum_{\xi} (1 - \exp(-n^{1/4})) \Pr(\xi) \\
&= (1 - \exp(-n^{1/4}) \Pr\left(|d_m - \frac{n}{em!}| < n^{3/4}, 0 \le m \le k\right) \\
&\ge (1 - \exp(-n^{1/4})(1 - \exp(-n^{1/4}))
\end{aligned}$$

where the summation is over all events ξ such that $|d_m - \frac{n}{em!}| < n^{3/4}, 0 \le m \le k$. This establishes inequality (A.1) and hence (A.2).

Similar calculations yield a concentration result for R'. In this case, we define the function

$$\tilde{g}(d_0, d_1, ..., d_k) =$$

$$\begin{aligned}
&= R - \sum_{m=0}^{k-1} d_m \left(\sum_{j=0}^{k-m-1} (j - k + m) \binom{R}{j} \left(\frac{1}{n-1}\right)^j \left(\frac{n-2}{n-1}\right)^{R-j} + (k - m) \right) \\
&= \sum_{m=0}^{k-1} d_m \sum_{j=0}^{k-m-1} (k - m - j) \binom{R}{j} \left(\frac{1}{n-1}\right)^j \left(\frac{n-2}{n-1}\right)^{R-j} - (k-1).
\end{aligned}$$

(Note: we have used the identity $R = \sum_{m=0}^{k-1}(k-m)d_m - (k-1)$.) Again, routine approximations establish that if $|d_m - \frac{n}{em!}| < n^{3/4}$ for $0 \le m \le k$ then $|\tilde{g}(d_0, ..., d_k) - n\beta(k)| \le C'_k n^{3/4}$ for some constant C'_k which does not depend on n. So, as in the derivation of inequality $(A.2)$, it is enough to show that

$$\Pr\left(|R' - \tilde{g}(\vec{d})| < n^{3/4}, |d_m - \frac{n}{em!}| < n^{3/4}, 0 \le m \le k\right) \ge 1 - 2\exp(-n^{1/4}).$$

Again, let ξ denote the event $\mathcal{A}(1) = A, \mathcal{R}(1) = B$, and $M^* = f$, where f is a mapping such that $|d_m(f) - \frac{n}{em!}| < n^{3/4}$ for $0 \le m \le k$, and A and B are subsets of vertices with $|B| \le |A|$. We claim that

$$E_\xi(R') := E(R'|\xi) = \tilde{g}(\vec{d}(f)).$$

To see this, note that although it is not necessarily the case that $\mathcal{R}(2) \subseteq \mathcal{R}(1)$, it is *always* the case that $R' \le R$. In particular, the overall number of roots is only reduced by mapping vertices in B to available vertices in $\mathcal{A}(1) = A$ during Phase 2. Redirecting a vertex $x \in B$ to an *unavailable* vertex during Phase 2 may result in deleting x from the set of roots, but in this case *another* vertex is added to the set of roots and the *number* of roots is not decreased. For each $v \in A$, let Y_v denote the number of roots which are subtracted from

$|\mathcal{R}(1)| = |B| = r$ due to the roots in B which are mapped to v in Phase 2, then *given* the configuration ξ, we can write

$$R' = r - \sum_{v \in A} Y_v.$$

If $\rho(v,f) = m < k$ then Y_v is *at most* $k-m$ and for $0 \leq j \leq k-m-1$

$$\Pr(Y_v = j|\xi, \rho(v,f) = m) = \binom{r}{j}\left(\frac{1}{n-1}\right)^j \left(\frac{n-2}{n-1}\right)^{r-j},$$

whereas

$$\Pr(Y_v = k-m|\xi, \rho(v,f) = m) = 1 - \sum_{j=0}^{k-m-1} \binom{r}{j}\left(\frac{1}{n-1}\right)^j \left(\frac{n-2}{n-1}\right)^{r-j}.$$

It follows that

$$E(Y_v|\xi, \rho(v,f) = m) = \sum_{j=0}^{k-m-1} (j-k+m)\binom{r}{j}\left(\frac{1}{n-1}\right)^j \left(\frac{n-2}{n-1}\right)^{r-j} + (k-m)$$

and hence

$$E_\xi(R') = E\left(R - \sum_{v \in A} Y_v \middle| \xi\right) = \tilde{g}\left(\vec{d}(f)\right)$$

since $E(R|\xi) = r = \sum_{m=0}^{k-1}(k-m)d_m(f) - (k-1)$. Now let $x_1, x_2, ..., x_r$ denote the vertices in B. For $1 \leq i \leq r$, let Z_i denote the vertex that x_i is mapped to in Phase 2. Let $\mathcal{F}_i = \sigma(Z_1, Z_2, ..., Z_i)$ denote the σ-algebra generated by the variables $Z_1, Z_2, ..., Z_i$ and let $\mathcal{F}_0$ denote the trivial σ- algebra. Let $W_i' = E_\xi(R'|\mathcal{F}_i) - E_\xi(R'|\mathcal{F}_{i-1})$, then Azuma's inequality tells us that

$$P_\xi(|R' - \tilde{g}(\vec{d}(f))| > n^{3/4}) = P_\xi(|R' - E_\xi(R')| > n^{3/4}) \leq 2\exp(-n^{3/2}/2\sum_{i=1}^{r} \|W_i'\|_\infty^2).$$

It is straightforward to check that $\|W_i'\|_\infty \leq 1$ for $1 \leq i \leq r$, so

$$P_\xi(|R' - \tilde{g}(\vec{d}(f))| > n^{3/4}) \leq 2\exp(-n^{3/2}/2r) \leq \exp(-n^{1/4}).$$

Thus

$$\begin{aligned}
&\Pr\left(|R' - \tilde{g}(\vec{d})| < n^{3/4}, |d_m - \frac{n}{em!}| < n^{3/4}, 0 \leq m \leq k\right) = \\
&= \sum_\xi \Pr\left(|R' - \tilde{g}(\vec{d}(f))| < n^{3/4} \middle| \xi\right) \Pr(\xi) \\
&\geq \sum_\xi (1 - \exp(-n^{1/4}))\Pr(\xi) \\
&= (1 - \exp(-n^{1/4})\Pr\left(|d_m - \frac{n}{em!}| < n^{3/4}, 0 \leq m \leq k\right) \\
&\geq (1 - \exp(-n^{1/4})(1 - \exp(-n^{1/4}))
\end{aligned}$$

where the summation is over all events ξ such that $|d_m(f) - \frac{n}{em!}| < n^{3/4}$ for $0 \leq m \leq k$. Finally, since

$$\{|R' - \tilde{g}(\vec{d}(f))| < n^{3/4}, |d_m - \frac{n}{em!}| < n^{3/4}, 0 \leq m \leq k\}$$

$$\subseteq \{|R' - n\beta(k)| < (C'_k + 1)n^{3/4}\},$$

we obtain

$$\Pr(|R' - n\beta(k)| < (C'_k + 1)n^{3/4}) \geq 1 - 2\exp(-n^{1/4}). \qquad (A.3)$$

The lemma follows from inequalities (A.2) and (A.3). ∎